Visual Processing and Latent Representations in Biological and Artificial Neural Networks

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Christina Maria Funke
aus Illertissen

2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	13.11.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Matthias Bethge
2. Berichterstatter:	Prof. Dr. Fabian Sinz

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel

"Visual Processing and Latent Representations
in Biological and Artificial Neural Networks"

selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, den _____ _____

Datum Unterschrift

Acknowledgments

This dissertation would not have been possible without the help of many talented people. I would like to thank Matthias Bethge who was extremely supportive, both in terms of logistics and in shaping the scientific direction of my work. Many thanks to Thomas Wallis who was a very good advisor during my first years in the lab, and whose knowledge about human visual perception was almost unlimited. He gave me a great start in the world of science by having an outstanding attitude towards scientific quality. I thank Matthias Kümmerer who took over the advisory tasks very well when Tom got the call for new challenges. I am very grateful that he used his vast knowledge especially in the field of deep learning and mathematics to support me. Furthermore, I highly appreciate the valuable advice and ideas of Felix Wichmann and Jörg Stückler who complemented my thesis advisory committee. Thanks to the colleagues working with me on the individual projects and to all members of the Bethge laboratory. In particular, thanks to Heike König, Judith Lam and Melanie Ertle-Palm for their help with administrative tasks. Finally, I thank my parents, siblings and friends for their unconditional support all these years. Many thanks to my partner Simon for all his love and for always standing by me.

Summary

The human visual system performs the impressive task of converting light arriving at the retina into a useful representation that allows us to make sense of the visual environment. We can navigate easily in the three-dimensional world and recognize objects and their properties, even if they appear from different angles and under different lighting conditions. Artificial systems can also perform well on a variety of complex visual tasks. While they may not be as robust and versatile as their biological counterpart, they have surprising capabilities that are rapidly improving. Studying the two types of systems can help us understand what computations enable the transformation of low-level sensory data into an abstract representation. To this end, this dissertation follows three different pathways.

First, we *analyze* aspects of human perception. The focus is on the perception in the peripheral visual field and the relation to texture perception. Our work builds on a texture model that is based on the features of a deep neural network. We start by expanding the model to the temporal domain to capture dynamic textures such as flames or water. Next, we use psychophysical methods to investigate quantitatively whether humans can distinguish natural textures from samples that were generated by a texture model. Finally, we study images that cover the entire visual field and test whether matching the local summary statistics can produce metameric images independent of the image content.

Second, we *compare* the visual perception of humans and machines. We conduct three case studies that focus on the capabilities of artificial neural networks and the potential occurrence of biological phenomena in machine vision. We find that comparative studies are not always straightforward and propose a checklist on how to improve the robustness of the conclusions that we draw from such studies.

Third, we *address a fundamental discrepancy* between human and machine vision. One major strength of biological vision is its robustness to changes in the appearance of image content. For example, for unusual scenarios, such as a cow on a beach, the recognition performance of humans remains high. This ability is lacking in many artificial systems. We discuss on a conceptual level how to robustly disentangle attributes that are correlated during training, and test this on a number of datasets.

Zusammenfassung

Das menschliche Sehsystem ist in der Lage auf der Netzhaut eintreffendes Licht in eine abstrakte Repräsentation umzuwandeln, welche es erlaubt sich in einer dreidimensionalen Welt zurechtzufinden und Objekte sowie deren Eigenschaften unter verschiedensten Blickwinkeln und Lichtverhältnissen zu erkennen. Auch künstliche Systeme sind in der Lage, eine Vielzahl komplexer visueller Aufgaben zu bewältigen. Sie sind zwar nicht so robust und vielseitig wie das biologisches Gegenstück, verfügen aber über überraschende Fähigkeiten, die sich stetig verbessern. Die Verwendung beider Systemarten ermöglicht es, die Mechanismen zu verstehen, die es erlauben sensorische Daten in eine sinnvolle abstrakte Repräsentation zu verwandeln. Diese Dissertation verfolgt dazu drei Ansätze:

Zuerst analysieren wir Aspekte des menschlichen Sehens. Der Fokus liegt auf der Wahrnehmung im peripheren Gesichtsfeld und der Zusammenhang mit der Wahrnehmung von Texturen. Unsere Arbeit baut auf einem Texturmodell auf, welches die gelernten Muster eines neuronalen Netzes nutzt. Zuerst erweitern wir dieses Modell auf den zeitlichen Bereich, um dynamische Texturen wie Flammen oder Wasser zu erfassen. Anschließend untersuchen wir mit psychophysikalischen Methoden, ob Menschen natürliche und synthetisierte Texturen unterscheiden können. Schließlich untersuchen wir Bilder, die das gesamte Gesichtsfeld abdecken und testen ob Metamere entstehen, wenn die lokalen Statistiken mit denen des Originalbilds übereinstimmen.

Zweitens vergleichen wir die visuelle Wahrnehmung von Mensch und Maschine. Dazu führen wir drei Fallstudien durch, welche die Fähigkeiten von künstlichen neuronalen Netzwerken untersuchen, und testen, ob Phänomene die aus dem biologischen Sehen bekannt sind auch in künstlichen Systemen auftreten. Wir zeigen Probleme auf, die bei Vergleichsstudien auftreten können und erstellen eine Checkliste um robustere Schlussfolgerungen zu ziehen.

Drittens befassen wir uns mit einer grundlegenden Diskrepanz zwischen dem menschlichen und dem maschinellen Sehen. Das biologische Sehen ist weitestgehend robust gegenüber Veränderungen im Erscheinungsbild und erlaubt auch ungewöhnlichere Bilder zu verstehen, wie beispielsweise das Foto einer Kuh am Strand. Diese Fähigkeit fehlt in vielen künstlichen Systemen. Wir erörtern auf konzeptueller Ebene, wie man Attribute, die während des Trainings korreliert sind, robust entkoppeln kann und testen unseren Ansatz auf verschiedenen Datensätzen.

Contents

Introduction **13**

1 Background **15**
1.1 Texture Perception in Humans 15
1.2 Perception of Full-Field Scenes 17
1.3 Artificial Neural Networks as Models for Human Vision 19
1.4 Latent Representations and Disentanglement 22

2 Publications **27**
2.1 Synthesis of Dynamic Textures (**P1**) 27
2.2 Perceptual Quality of Texture Models (**P2**) 29
2.3 Summary Statistic Approach for Scene Metamers (**P3**) 31
2.4 On the Methodology of Comparison Studies (**P4**) 34
2.5 Disentanglement and Generalization under Correlation Shifts (**P5**) . . 40

3 General Discussion **45**
3.1 Methodology of Vision Studies 45
3.1.1 Levels for Analyzing Vision 45
3.1.2 Understanding Vision: Analyzing vs. Constructing 47
3.1.3 Types of Stimuli 48
3.2 Mechanisms Underlying Visual Processing 50
3.2.1 Local versus Global Processing 50
3.2.2 Importance of Recurrent Mechanisms 52
3.2.3 What Drives Generalization? 54
3.3 Conclusion 58

References **78**

Appendix **79**
P1: Synthesising dynamic textures using convolutional neural networks . . 79
P2: A parametric texture model based on deep convolutional features closely matches texture appearance for humans 89
P3: Image content is more important than Bouma's Law for scene metamers 119
P4: Five points to check when comparing visual perception in humans and machines 163
P5: Disentanglement and generalization under correlation shifts 187

Introduction

We humans rely heavily on our sense of sight. We trust that our brains are capable of capturing visual information and providing us with a rich picture of the environment. We use this inner picture to navigate in the physical environment, to recognize objects from great distance and to assess their properties. Even more, we can learn new objects and visual concepts from just a few observations and can identify them even in unfamiliar scenarios. But how does this powerful visual representation come about? Is it really as truthful and robust as we think? What mechanisms are involved in processing visual information and are they generally required for visual systems, be they biological or artificial?

These and similar questions are studied by several disciplines. On one hand, psychophysics studies behavioral properties by testing, for example, which stimuli can just be perceived. Neuroscience, on the other hand, examines neural data and tries to recognize patterns in this data. And finally, computer vision and machine learning research develop artificial systems that achieve good performance on a range of visual tasks. Combining observations and methodologies from the different fields is likely to be a fruitful way to learn about vision.

Here we follow three different pathways that comprise methods from the different fields. First, we analyze aspects of human perception. Second, we compare the visual perception of biological and artificial systems. And third, we investigate fundamental discrepancies between human and machine vision at a more conceptual level by studying how to achieve robust disentanglement. This set of approaches can help to learn about the mechanisms used by visual systems, and can shed light on whether certain components are necessary or just one of many possible solutions for visual processing.

This thesis is structured as follows. The first section introduces the relevant background and the research questions addressed in this dissertation. Section 2 describes the publications that resulted from my research and Section 3 discusses the findings at a more general level and previews future research directions.

1 Background

1.1 Texture Perception in Humans

Humans can perceive small details when focusing on specific parts in an image. On the other hand, we can build efficient representations of image content when it is not important to know all details. For example, think of a wall made of stones. Humans do not need to know the shape, color or exact position of each single stone to assess certain properties of the wall. Such properties may be the stability of the wall or the type of stone from which the wall is made, such as granite, limestone, sandstone, or fieldstone. Sometimes, we can even tell at a glance in what period and for what purpose a wall was built. In other words, we are able to perceive textural properties.

Visual textures form a substantial part of our visual input. The bark of a tree, the gravel in the yard, or the pattern of a dress are all examples of textures. Characteristic of textures is their homogeneity. From a mathematical point of view, textures can be described as statistically stationary (Petrou and Kamata, 2006). This means that the texture has the same local properties everywhere in the image. Or in other words, the statistical properties do not change when considering different parts of a texture image.

Understanding human perception of textures is of great interest to vision science (Balas, 2021; Rosenholtz, 2014). Humans are sensitive to small differences in visual textures, which allows to extract the properties of materials (Adelson, 2001). This can be very helpful, for example, when we need to judge whether a floor might be wet. Subtle differences in textures inform us about the orientation of surfaces, which is relevant for making inference about 3D shape (Li and Zaidi, 2000). Texture perception also allows us to locate the boundaries between objects. This is important for image segmentation and for distinguishing between figure and ground (Malik and Perona, 1990).

Being able to generate textures artificially and in a controlled manner is critical to the study of texture perception. Stimuli have evolved greatly over time. As nicely listed by Rosenholtz (2014), texture stimuli were initially created by hand: Researchers photographed wallpaper (Gibson, 1950), filled cells of tables to obtain random patters (Attneave, 1954), attached self-sticking black tape to white cardboard (Beck, 1967), or drew stimuli by hand with ink (Olson and Attneave, 1970). With access to computers, it became easier to generate stimuli in a more systematic and rapid manner (Julesz,

1962, 1965).

While early approaches focused on testing specific hypotheses by arranging line segments or designing patterns in controlled ways, later efforts focused on studying more natural textures. For example, textures were created by repeating sections of an original texture over larger areas (Efros and Freeman, 2001). Later, more sophisticated texture models were developed. The subset of *parametric* texture models is particularly interesting, because it allows systematic modification of specific aspects of a given texture as well as interpolation between different source textures. Two well-known texture models of this type are the Portilla and Simoncelli texture model (*PS-model*) (Portilla and Simoncelli, 2000) and a texture model based on the features of a deep convolutional neural network (*DNN texture model*) (Gatys et al., 2015). In both approaches, new samples of a given source texture are synthesized by iteratively changing random noise until their statistics match those of the original source image. The main differences between these models lies in the features that are matched. The PS-model relies on a steerable pyramid, while the features of the DNN texture model result from training a deep convolutional neural network for object recognition.

Textures can also extend to the temporal dimension. Examples for dynamic textures include the motion of flames or leaves moving in the wind. Now, the statistics are not only stationary in position, but also in time. In other words, for dynamic textures, the statistics are consistent for different time steps. The first research goal of this dissertation (**P1**) was to extend the DNN texture model to the temporal domain to generate dynamic textures.

Overall, the goal of texture models is to synthesize new samples of naturalistic textures that match human texture appearance. Visual inspection of the resulting textures can give a first impression of their quality. However, for a quantitative comparison of texture models, it is crucial to apply the precise methods of psychophysics. A number of such investigations was performed by Balas and colleagues on the PS-model (Balas, 2006, 2012). In their experiments they tested whether humans could tell the difference between original textures and the ones generated by a texture model. With this, they analyzed under which conditions the PS-model could match human texture appearance. A major finding of this work was that the model captured the appearance of texture better in the periphery than in the fovea. Our second research goal (**P2**) was to perform a similar analysis for the more recent DNN texture model (Gatys et al., 2015) and to test whether the more complex feature space increases the

perceptual quality of the synthesized textures.

Understanding texture perception in humans is an exciting goal in itself. Natural scenes, however, do not consist of only one texture. The next section deals with the perception of more complex scenes that fill the entire visual field.

1.2 Perception of Full-Field Scenes

Natural scenes often consist of inhomogenous image content. Different textures can occur in different areas of the image and, in addition, there is often image content that cannot be classified as texture. An example would be a beach scene with a water texture in the top half of the image, a sand texture in the bottom half of the image, and a deckchair in the middle. Besides the inhomogeneities in image content, the human visual system also exhibits inhomogeneities. While we can perceive a lot of details in the fovea, the perception in the periphery is much fuzzier. So the question is, how do these two types of inhomogeneity interact?

Essentially, there are two hypothesis that describe how full-frame scenes are processed. On the one hand, there could be a mechanism at play that segments the different parts of the image and processes them separately in a content-based manner. On the other hand, processing could be independent of image content and only depend on the peripheral eccentricity. Here, the hypothesis is that in the peripheral visual field, information is represented by texture-like statistics and that higher brain areas have access only to a summary statistical representation. With that, the relevance of texture representations would extend beyond texture-like image content. In the literature this idea is referred to as *compulsory texture perception* (Parkes et al., 2001; Lettvin et al., 1976; Rosenholtz, 2014).

Compulsory texture perception can explain human performance on a number of visual tasks. An example is the ability of humans to easily judge summary statistics of visual displays such as the average size or orientation of a group of elements (Ariely, 2001; Dakin and Watt, 1997). In another line of research, stimuli were texturized using the texture models described in the previous section. With that, researchers could determine by visual inspection which aspects of appearance cannot be recovered from a texture-like representation. Using this approach, it has, for example, been shown that the difficulty of visual search tasks can be largely explained by texture models of peripheral vision (Rosenholtz et al., 2012b).

Having access to only summary statistics is also interesting from the perspective of compression (Rosenholtz et al., 2012a). The visual system is thought to have at least one bottleneck (Nakayama, 1990), meaning that intermediate processing layers have limited capacity and cannot represent all the information presented to the eyes. The substantial loss of information that is associated with the computation of summary statistics is consistent with this bottleneck idea. In other words, the computation of summary statistics may originate naturally from the constraints posed by the architecture of the visual system (Balas et al., 2009). The idea of compressing information in intermediate representations also exists in machine learning. Autoencoders rely precisely on this idea, and there are approaches to use the information bottleneck principle for training DNNs (Amjad and Geiger, 2019; Tishby and Zaslavsky, 2015).

The phenomenon of visual crowding is also thought to be related to the hypothesis that one has only access to a summary statistic representation. As early as the 1920s, it was observed that a peripheral stimulus that would be recognizable if it stood alone cannot be perceived when similar visual content is nearby (Korte, 1923; Flom et al., 1963). Bouma's law (Bouma, 1970, 1973) describes this in a more quantitative way by specifying at which distance flankers can influence the perception of individual objects. Specifically, as retinal eccentricity increases, the size of the area where the flankers take effect increases. More recently, crowding has been associated with the hypothesis of compulsory texture perception. In particular, crowding could be explained by the computation of summary statistics within regions of a size approximately equal to Bouma's law (Rosenholtz et al., 2012a; Balas et al., 2009; Levi, 2008).

Freeman and Simoncelli (2011) connected these ideas with observations from physiology. The relevant property here is the *receptive field*, which denotes the spatial region of the input that can influence the activity of a given neuron. The diameter of the receptive fields increases linearly with eccentricity, and the *scale factor* describing this linear increase varies for the different brain areas. In particular, the scale factor is larger for regions located higher up in the ventral visual stream. The scale factor thus provides a signature that distinguishes different areas from each other. This is exciting because it could allow to allocate crowding to a specific brain area in the ventral visual stream. To this end, Freeman and Simoncelli (2011) have developed a model that computes summary statistics in local visual areas whose sizes increase with spatial eccentricity. Their model allows the synthesis of samples that locally match the summary statistics of original images. If the visual system represented the

periphery using summary statistics, these images would be indistinguishable when viewed centrally (they would be *metamers*). Freeman and Simoncelli found that this was indeed the case and that the scaling at which two synthesized images could not be told apart corresponded to the scale factor of visual area V2 and also matched the approximate value of Bouma's Law. This is an exciting finding, as it is thought to link receptive field scaling, crowding zones, and our perceptual experience (Cohen et al., 2016; Movshon and Simoncelli, 2014).

However, there are still two missing pieces to draw this conclusion. First, Freeman and Simoncelli compared only two *synthesized* samples with another. They never tested whether the synthesized images are metameric to the *original* images. Second, the scale factor at which two images can not be told appart must hold for all possible images. For these reasons, it is not yet clear whether summary statistic models can generate metamers for natural images. The third research question (**P3**) is directed towards this question.

The research on texture perception is an example of how artificial systems can be used to learn about human perception. The next part discusses the use of artificial systems as models for human vision more broadly.

1.3 Artificial Neural Networks as Models for Human Vision

Artificial systems have been considered as models for human vision for a long time. As early as 1943, McCulloch and Pitts (1943) created a basic model of a brain cell. Later, Rosenblatt (1958) extended it to a single-layer neural network called "perceptron". However, only linearly separably problems could be solved with this single layer (Minsky and Papert, 1969). A groundbreaking discovery was the existence of two major cell types in biological brains, namely simple and complex cells (Hubel and Wiesel, 1962). This finding was incorporated into the "neocognition" of Fukushima (1980), which consisted of multiple layers of simple and complex cells. This model already used location-invariant feature extractors. Another important step was the invention of backpropagation, which made it possible to optimize stacked layers (Rumelhart et al., 1986). Eventually, LeCun et al. (1989) introduced convolutional layers, similar to the ones that are nowadays used in many models that perform visual tasks. Other hierarchical models followed, one of them being the HMAX-model (Fukushima, 1980; Riesenhuber and Poggio, 1999). Building on these fundamentals, deep neural networks (DNNs) have become increasingly powerful and are now able

to perform well on a range of complex tasks.

One main advantage of DNNs is that they *learn* suitable features instead of relying on their manual design. Interestingly, by training on natural images, features are learned that simile the ones implemented in biological systems. For example, the features in the first layers resemble Gabor filters and color blobs (Krizhevsky et al., 2012). Furthermore, it was shown that the activity of real neurons could be predicted from the artificial units (Yamins et al., 2014) and that networks that performed better in object recognition also predicted neural activity better (Tacchetti et al., 2017). These findings raised the question of the extent to which DNNs resemble biological systems and could serve as models for biological vision (Lindsay, 2020; Majaj and Pelli, 2018).

Understanding where and how artificial and biological processing differ is an often pursued goal. Here, I will give a short, non-comprehensive overview over approaches used to judge the similarities of human and machine vision. In neuroscience, the representational similarity analysis (Kriegeskorte et al., 2008) and dissimilarity matrices (Khaligh-Razavi and Kriegeskorte, 2014) are commonly used. The Brain-Score (and the brain hierarchy score) evaluates how well DNN unit activation patterns can predict neuronal responses in primate visual areas (Schrimpf et al., 2020; Nonaka et al., 2021). Another set of approaches are visualization techniques that can reveal to some extent which image features are encoded by specific units or layers (Krizhevsky et al., 2012). On the behavioral level, the predictive performance on out-of-distribution data is used to measure the behavioral difference between human and machine vision (Geirhos et al., 2021). Other approaches study whether DNNs and humans match in how the similarity of images is perceived (King et al., 2019; Rosenfeld et al., 2018; Jozwik et al., 2017). The analysis of prediction errors can also provide valuable insights (Rajalingham et al., 2018). Golan et al. (2020) probe the discrepancies between models and human perception by studying images for which the predictions of machine models disagree. While some comparison studies measure the similarity on a very general level (as the one mentioned above), others are targeted to specific visual phenomena. As an example, psychological concepts such as the Gestalt principles are tested (Kim et al., 2019, 2021). One way to test perceptual phenomena known from vision is to rephrase them such that they can be measured in terms of the distance of two images in the latent space (Jacob et al., 2021). Other research addresses very specific phenomena such as the recognition gap, which studies whether there are "atoms" of vision (Ullman et al., 2016). The study of illusions is also of particular interest (Gomez-Villa et al., 2019). Other research questions target

the relevance of texture and shape for object recognition (Geirhos et al., 2019), crowding (Roig et al., 2018; Doerig et al., 2020), the perception of gloss (Storrs et al., 2021) or the ability to perform visual reasoning tasks (Kim et al., 2018; Yan and Zhou, 2017; Johnson et al., 2017).

Overall, comparing the visual processing of humans and machines is not straightforward. The basic architectural functionality of DNNs, namely the hierarchical stacking of neural connections, as well as their remarkable performance on a large range of visual tasks, support that DNNs could model biological neural networks. On the other hand, large differences in architectural and behavioral properties make it difficult to align the systems. This includes weight sharing and the learning mechanisms used in artificial models (such as backpropagation), as well as the discrepancy with the spiking properties of biological brains and Dale's law (ratio of inhibitory and excitatory weigths). On the behavioral side, DNNs cannot conquer with the versatility of the human system. They often lack generalization abilities and robustness (Geirhos et al., 2018), rely on spurious information (Geirhos et al., 2020a), are susceptible to adversarial noise (Szegedy et al., 2014), and have difficulty learning abstract concepts (Stabinger et al., 2016; Kim et al., 2018; Yan and Zhou, 2017; Johnson et al., 2017).

Given the differences on the behavioral and functional level, researchers have thought about how to compare DNNs to the human visual system. Suggestions include the usage of short stimulus presentation times as correspondence of a forward pass in a DNN (Tang et al., 2018), and the use of challenging set-ups (Wichmann et al., 2017). The fourth research project of this dissertation (**P4**) contributes to discussions centering around the question "how" to compare human and machine perception. We conduct three case studies and present a checklist that can help avoid common pitfalls.

When studying visual perception of humans and machines, it can be particularly insightful to focus on the resulting internal representations. The analysis of visual representations and their properties is a main aspect of the summary statistics idea and the scene metamers described in Section 1.2. The comparison of internal representations is also common in neuroscience (Kriegeskorte et al., 2008; Khaligh-Razavi and Kriegeskorte, 2014) and is an essential part of many comparative studies between biological and artificial systems (Golan et al., 2020; Jacob et al., 2021). In the final part, I will delve further into latent representations, focusing on a property called *disentanglement*. An example of disentanglement is the separation of content and style as

done by Gatys et al. (2016). Here, texture statistics of one image (see Section 1.1) are combined with the content of another image allowing for transferring the style of one image to another.

1.4 Latent Representations and Disentanglement

As nicely described by DiCarlo and Cox (2007), there are two different ways to look at visual perception. On one hand, the computations can be viewed as "complex decision functions [...] that operate on the retinal image representation". On the other hand, the process of visual perception can be divided into two steps, so that it can be recast as "finding operations that progressively transform this retinal representation into a new form of representation, followed by the application of relatively simple decision functions". The latter view allows for treating visual tasks as a problem of finding useful data representations also denoted as *internal* or *latent representations*. Due to the progressive transformation of information, the resulting intermediate representations form a hierarchical structure. More specifically, in the human visual system, the retinal image passes through different visual areas such as V1 and V2 before arriving in the IT cortex. Similarly, DNNs consist of multiple layers resulting in representations at increasingly abstract levels (LeCun et al., 2015). For both biological and artificial networks, the information encoded in the upper and lower layers differs significantly. In particular, along the depth of DNNs, features transition from general to task specific (Yosinski et al., 2014; Long et al., 2015).

The field of representation learning aims at finding "useful" representations of the input data. A natural question is what constitutes a good representation and how to determine whether one representation is better than another (Bengio et al., 2013). Most research agrees that a useful representation enables good performance on downstream tasks, that it leads to robustness to changes that are not relevant for a chosen task, that it "untangles" attribute manifolds (DiCarlo and Cox, 2007), and that it is more compact with respect to the input data. Contrastive learning (Zimmermann et al., 2021; Oord et al., 2018; Islam et al., 2021) is one approach to find meaningful representations and does so by encouraging a property that is intuitively important: Similar samples (positive pairs) should have similar representations, i.e., be close to each other in the embedding space, while distinct samples (negative pairs) should be far away. Another useful feature of representations could be to disentangle certain attributes: It may be beneficial to encode different attributes, such as the identity of an object and its appearance, in separate parts of the latent representation.

Disentanglement is often motivated by the *vision as inverse graphics* paradigm (Mansinghka et al., 2013; Kulkarni et al., 2015). In computer graphics, images are rendered based on a compact description of the scene that defines properties such as the location, shape, pose and, texture of the objects, as well as the lighting conditions and camera parameters. This resembles the physical process that is responsible for generating images in the real world. Vision is sometimes described as approximate *inverse graphics*, meaning that it tries to reverse-engineer this generative process (Loper and Black, 2014) to find the *generative factors*. This is closely related to blind source separation (Cardoso, 1989; Jutten and Herault, 1991), where the goal is to recover source signals that have been mixed with an unknown mixing function. The question of identifiability is relevant in this context as it addresses whether it is theoretically possible to recover the source variables for a given mixing function and model.

Another way to motivate disentanglement is to consider the transformations that can be applied to the input data without changing the identity of an object. Examples of such transformations include changing the illumination, angle of view or color of an object. The attributes associated with these transformations are often referred to as *nuisance factors*. A good property of a representation is to be *invariant* to such transformations, meaning that the transformations should not affect the property of interest such as the identity of an object. One way to handle this is to drop the nuisance factors during visual processing so that they are no longer present in the more compact latent representation. This was the case with the metamer idea discussed earlier. The more sophisticated approach is to explicitly encode the nuisance variables. This is called *equivariance*. Here, the nuisance variables are still present in the latent representation, but are encoded in separate parts. While for both versions the application of the transformation to the input data does not affect the identity of the object, in the equivariance approach the knowledge about the nuisance variables is preserved. This corresponds to the fact that people are able to recognize the color of an object, although the color has no influence on their ability to recognize the object. Deriving disentanglement from the invariance against transformations is similar to the study of symmetry in physics (Cohen and Welling, 2014; Higgins et al., 2018). In physics, symmetry transformations led to the discovery of new concepts, as transformations that leave certain properties invariant yield exploitable structure. Similarly, disentangled representations should capture the "symmetry transformations of the world state" (Higgins et al., 2018).

Disentangled representations could have a number of practical benefits. Humans can

do combinatorial generalization (Fodor and Pylyshyn, 1988), meaning that they can still identify the color of a red elephant even if they have never seen this combination of attributes before. Disentangled representations could facilitate robustness to new attribute values or combinations of attribute values not seen in the training data (Montero et al., 2021; Madan et al., 2020; Schott et al., 2021). In addition, disentanglement can improve interpretability, as one can better predict how the source image would change if the representation were varied and vice versa. In a disentangled setting, applying a transformation in the input space (for example by rotating an object), should produce the same result as applying the transformation in the latent space (for example by modifying the subspace encoding the rotation). Furthermore, disentanglement is useful in the light of fairness, could improve the ability to perform well on downstream tasks and could improve the robustness against changes in the test domain.

Most works on disentanglement consider the independent and identically distributed (i.i.d.) setting. However, correlations are common in the real world. As an example, the foreground and the background of images is highly correlated. The "cow on the beach" is a well known example: Since most pictures of cows are taken in grassy environments, networks have difficulty recognizing the cow when it is pictured in an unusual setting such as a beach or next to a boat. This is an example of how training on a correlated dataset can lead to problems, as missclassifications are likely for out-of distribution data (Beery et al., 2018; Arjovsky et al., 2019). Another example is that it is likely that there are multiple objects in an image (Beyer et al., 2020; Tsipras et al., 2020). In natural images, their occurrence is not random, and certain objects tend to appear together. For example, keyboards and monitors are often in the same image. Importantly, correlated data and correlation shifts in the test data can occur in areas that affect people's lives, including healthcare and fairness applications.

Such questions are typically addressed in *domain adaptation* and *domain generalization* (Zhao et al., 2019). Here one has access to multiple labeled source domains (and in the adaptation case, also to images of the target domain). The goal is to generalize to the target domain, which is attempted by finding an intermediate representation that is invariant between the source and target domains. Another approach to improve out-of-distribution (ood) robustness is to train on counterfactual data (Sauer and Geiger, 2020).

However, the study of correlated attributes is also interesting in terms of disentan-

glement. More recently, it has been pointed out that disentangling correlated factors can be problematic. In the experiment of Träuble et al. (2021), two attributes were correlated in the training data. After training a variational auto encoder on this data, the resulting latent subspace encoded a mixture of these attributes. Moreover, they found that common disentanglement metrics could not reveal such pairwise correlations. To counter this, they introduced partial supervision as a solution to cope with correlations. The fifth research topic (**P5**) follows up on this research. We show that minimizing mutual information between latent subspaces can fail even with full supervision. Therefore, we discuss what should be the correct objective for disentanglement in the face of correlations.

2 Publications

This dissertation follows three pathways to learn about visual perception and the latent representations in biological and artificial neural networks. It starts from analyzing texture perception in human vision (**P1**-**P3**). Then it compares the visual perception of humans and machines (**P4**). And finally it studies fundamental discrepancies between human and machine vision on a principled level by discussing the requirements for robust disentanglement (**P5**). The five publications that resulted from this work will be summarized in the following. Each study will be motivated, the main results described and the implications discussed. The full publications and the description of the authors' contributions can be found in the Appendix.

2.1 Synthesis of Dynamic Textures (P1)

This section summarizes:

Christina M. Funke*, Leon A. Gatys*, Alexander S. Ecker, Matthias Bethge (2017). Synthesising dynamic textures using convolutional neural networks. *arXiv:1702.07006*.

The full publication and author contributions can be found in the appendix on page 79.

* joint first authors

Motivation

The work of Gatys et al. (2015) has significantly advanced the generation of visual textures by employing the feature space of a DNN for texture synthesis. The resulting model can create new samples of a given texture that have a high visual quality. Here, we extend the DNN texture model to the temporal domain. While the original model captures static textures, we now aim at generating samples for dynamic textures, also denoted as *texturized motion* (Wang and Zhu, 2002). Examples for such dynamic textures are leaves that are moving in the wind or the characteristic motion of water and flames.

Previous works used a range of different approaches to synthesize dynamic textures. One class of approaches are physics-based alogrithms which simulate the dynamic behaviour by building a physical model (Ebert et al., 1994). Another class of approaches are image-based approaches. While some works combine existing image patches in a smart way such that they form a dynamic texture (Kwatra et al., 2003; Schödl et al., 2000), other approaches attempt to model the statistical proper-

ties (Szummer and Picard, 1996; Wei and Levoy, 2000; Wang and Zhu, 2002; Doretto et al., 2003). Most recently, Xie et al. (2017) used a generative ConvNet to synthesize dynamic textures. Our approach falls into the category of image-based approaches that capture statistical properties.

Results

We developed a model that captures second-order dependencies between spatial features. Our model was similar to the static DNN texture model (Gatys et al., 2015), but computed the statistics over multiple consecutive frames. These spatio-temporal summary statistics were then used for the generation of new textures. Similar to the static approach, we synthesized new samples via gradient descent to match the features extracted from the source texture. We tested our model on a range of source textures and found that in most cases the results already looked decent when only two frames of the original texture were used to synthesize new samples. In an additional experiment, we varied the size of the temporal window over which the summary statistics were computed and found that this parameter had little effect on the quality of the resulting videos.

Discussion

We developed a simple model that captures second order dependencies between spatial features and showed that we can reach good synthetic results using as little as two adjacent frames of the original texture. Unlike previous work (Xie et al., 2017), our model did not require the network to be re-trained for each new source texture. Our model had difficulty capturing more complex dynamic textures where the spatial and/or temporal dependencies extend over larger areas. Subsequent studies improved the generation of dynamic textures by factorizing spatial appearance and dynamics using optical flow prediction (Tesfaldet et al., 2018) and by accounting for long range dependencies (Zhang et al., 2021).

2.2 Perceptual Quality of Texture Models (P2)

This section summarizes:

Thomas S.A. Wallis, Christina M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, Matthias Bethge (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision,17(12), 5–5*.

The full publication and author contributions can be found in the appendix on page 89.

Motivation

The goal of texture models is to synthesize new samples of naturalistic textures. While a visual inspection of the synthesized texture images can be a first test bed, only a thorough experimental investigation can provide information on whether a model really matches human texture appearance.

A number of such investigations were performed by Balas and colleagues (Balas, 2006, 2012) on the PS-model (Portilla and Simoncelli, 2000). In these experiments a set of texture images was shown to human observers, from which they had to select the image that differed from the others (the "oddball"). The oddball could be either a crop of the original image or a crop of the synthesized image. This experimental design ensured that all stimuli were physically different. So instead of measuring whether a human could detect differences at the pixel level, the subjective appearance was tested. By performing this kind of experiments under foveal and peripheral presentation, Balas and colleagues could draw conclusions about the ability of the PS-model to model human texture appearance. In our study we perform a similar analysis for the more recent DNN texture model (Gatys et al., 2015). In particular, we are interested in whether the more complex feature space increases the perceptual quality of the synthesized textures.

Results

We selected twelve images showing natural textures such as gravel, crumpled paper, or roof tiles. For each source texture, we synthesized new images using the PS-texture model and the DNN texture model. In a three-alternative oddity paradigm procedure, we tested whether humans could distinguish the model-generated textures from the original textures. Importantly, all three stimuli were physically different: Two of them were crops of the original image and one was a crop of the synthetic

image (or vice versa). This procedure allowed us to test whether the appearance of the textures matches, rather than testing whether individual samples of the same texture could be distinguished. We assessed the discriminablity under two presentation conditions, namely a parafoveal condition (short presentation time) and an inspection condition (long presentation time).

We found that both texture models were good in matching appearance in the parafoveal condition. In the inspection condition, the DNN texture model performed better than the PS-model on average. At the level of individual source textures, we found that the DNN-model was successful for nine original textures, whereas the PS-model was successful on five textures. For one texture, the PS-model performed better than the DNN-model and for two textures, both models failed.

A total of three textures were difficult for the DNN model. A closer inspection of these textures revealed that they could be considered "quasiperiodic", i.e., they contained a regular structure extending across the entire image (e.g., the roof tiles). In a follow-up experiment, we found that for two of these textures, the appearance could be improved by using a DNN model that additionally captured the power spectrum (Liu et al., 2016). In a control analysis, we ensured that the good results were not due to the models simply learning to copy the source images.

Discussion

Our results showed that the PS texture model (Portilla and Simoncelli, 2000) was good when the stimuli where shown only briefly in the parafovea, but was less successful under inspection. The DNN texture model (Gatys et al., 2015) was a good model of texture appearance for most textures, even when viewed foveally. Finally, we found that adding a power spectrum constraint improved the results for source textures with long-range structures.

2.3 Summary Statistic Approach for Scene Metamers (P3)

This section summarizes:

Thomas S.A. Wallis*, Christina M. Funke*, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, Matthias Bethge (2019). Image content is more important than Bouma's Law for scene metamers. *ELife,8, e42512*.

The full publication and author contributions can be found in the appendix on page 119.

* joint first authors

Motivation

Freeman and Simoncelli (2011) tested the hypothesis that the downstream visual system in the periphery has access only to a summary statistic representation. For this goal, they synthesized images that match the summary statistics of an original images in pooling regions of a certain size. They found that two synthesized samples could not be told apart for pooling regions that corresponded to the scaling of visual area V2, i.e., a scale factor of 0.5. This work led to much excitement in the community, as it was perceived as linking crowding zones, receptive field scaling, and perceptual experience. However, this conclusion relies on the assumption that the introduced summary statistic model produces metamers for natural scenes at a scale factor of 0.5. We question this for two reasons.

First, we argue that it is necessary to compare the synthesized samples to the original image, rather than only comparing two synthesized samples to another. While it may be a necessary condition that synthesized samples are metameric, this is not a sufficient condition for the summary statistic hypothesis. If the system had only access to summary statistics, the synthesized samples should also be indistinguishable from the original image.

Second, we argue that the critical scale must be considered across all possible images. The critical scale is the maximal scale factor that still allows for metamerism. For each individual image, the critical scale may vary. However, if the visual system computes a summary statistic representation and has access only to this compressed representation, there must be a scale factor (the *system critical scale*) that holds for all images — be they texture-like or scene-like.

Results

In total, we estimated the critical scale for 20 images. These images were selected based on a pilot experiment such that half of them could be described as "texture-like" and the other half as "scene-like". For each original image, we used the FS-model to synthesize multiple versions with different scale factors. In a three alternative oddity paradigm, we measured human discrimination performance and used these results to estimate the critical scale for each image.

We conducted this study for two comparison conditions. A model synthesis was compared either to another model synthesis (which had a different initialization) or to the corresponding original image. When comparing two model syntheses with each other, we found only a small difference in the critical scale factor between texture-like and scene-like images, which is consistent with the study of Freeman and Simoncelli (2011). However, when comparing to the original image, the average critical scale was lower for scene-like images than for texture-like images. In other words, it was easier to distinguish the synthesis from the original when the image had scene-like image content. For two of the scene-like images, human performance was nearly perfect even at a scale factor of only 0.25. Importantly, in both image categories the critical scale was below 0.5.

Our selection of the scene-like and texture-like images, however, is questionable because our selection was based on a pilot experiment. Moreover, this distinction makes only limited sense for whole images, since most images contain both types of image content in different regions of the images. Therefore, in a second experiment, we distorted only small patches of the images, as these were easier to assign to one category or the other. The distortion was obtained by averaging the summary statistics of the DNN texture model (Gatys et al., 2015) only for these small image regions. We found that their visibility was strongly dependent on the type of the image content. Even small distortions were highly visible in scene-like regions, while large distortions could go unnoticed in texture-like regions.

Discussion

In summary, we found that the critical scale factor is smaller than previously thought and that this factor depends strongly on the image content. Similar results were found by contemporary work by Deza et al. (2017). In particular, our estimate for the critical scale was smaller than the V2 scaling and rather corresponded to the scaling of

V1 or less. This is in contrast to the popular idea that appearance matching can be described by summary statistics computed on a scale similar to the size of V2's receptive field. Also, the strong dependence on image content contradicts the notion that the human system has access only to summary statistic, and thereby discards non-textural information. Based on these findings, we hypothesized that the sole computation of summary statistic may not be the right approach to capture human visual appearance.

We identified three caveats to this hypothesis. First, it might be possible that metamerism can be achieved for smaller scale factors. However, we argued that the induced compression for such small pooling regions is very limited (the syntheses hardly differ from the original), which in turn calls into question the relevance of the summary statistic approach. Second, the summary statistics computed by Freeman and Simoncelli (2011) may not be the ones used by the human visual system. This is a legitimate concern, since matching the "wrong" features can obviously result in images that are not metamers for humans. This limitation is difficult to overcome because one cannot prove the non-existence of a correct set of features. However, the analysis of two additional summary statistics models that rely on the features of a DNN (namely the NeuroFovea model of Deza et al. (2017) and an additional DNN model that we introduce in our paper) can remove some doubts: None of these models was able to capture the appearance of scenes and they also depended on the image content. Third, all models we considered use an optimization procedure to find syntheses that match the statistics of the original image. This entails that in many cases the final losses are different from zero, which implies that the features may not have been accurately matched. In our study, we addressed this issue by performing a control analysis, which showed that the observed difference between texture-like and scene-like images could not be explained by the final loss value alone. Nevertheless, further studies on the effects of the optimization procedure would help to shed further light on this issue.

However, if our hypothesis above is correct despite these concerns, it would imply that one component is missing to capture the appearance of scenes. One candidate is mechanisms responsible for perceptual organization, such as segmentation and grouping mechanisms. This is consistent with other research (Herzog et al., 2015; Clarke et al., 2014; Manassi et al., 2013).

2.4 On the Methodology of Comparison Studies (P4)

This section summarizes:

Christina M. Funke*, Judy Borowski*, Karolina Stosio, Wieland Brendel†, Thomas S.A. Wallis†, Matthias Bethge† (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision, 21(3), 16–16*.

The full publication and author contributions can be found in the appendix on page 163.

* joint first authors, † joint senior authors

Motivation

Comparative psychology and psychophysics have a long history of performing cross-species comparisons. In these fields, researchers have thought a lot about what can be learned about one system by studying the other (Romanes, 1883; Köhler, 1925; Koehler, 1943; Haun et al., 2011; Boesch, 2007; Tomasello and Call, 2008). With the wave of excitement about DNNs as a new model of the human visual system, again it is important to discuss how to compare the fundamentally different systems (Majaj and Pelli, 2018; Barrett et al., 2019; Cichy and Kaiser, 2019; Buckner, 2019).

In our work, we discuss how to draw robust conclusions from such studies. To this end, we conducted three case studies comparing humans and machines on specific tasks. The first case study concerns the perception of closed contours in humans and machines. While at first glance there appear to be similarities, a closer look reveals that the two systems are not so similar at all. The other two case studies concern experiments that seem to reveal discrepancies between human and machine vision. We show, however, that the discrepancies can be largely eliminated by changing the experimental design.

Results

- *Closed Contour Detection.* In our first case study, we tested how well humans and DNNs can distinguish closed from open contours. For this purpose, we created a custom dataset consisting of two classes of images that differed only in whether they contained a closed contour. To ensure that the task could not be solved based on differences in the number of dark pixels, the image in both cases contained a main contour consisting of 3 to 9 straight line segments. Additionally, lines with one or two segments were added to make the task more

difficult. In psychophysical experiments, we confirmed that humans can distinguish the two classes. When testing a standard DNN on this task, we found that it could achieve similar high performance as humans. However, simply comparing accuracy does not tell us whether the model uses the same decision making process as humans. To better understand this, we tested the model without further training on variations of the dataset. If our DNN solved the task similar to humans, it should also be able to perform well on these generalization test that are still clear to humans. In fact, we found that the model generalized well to variations of the dataset that we considered to be challenging. For example, the model still made the correct predictions when the main contour was no longer a polygon but was generated by a radial frequency function. On the other hand, testing the model on further, seemingly easy, variations of the dataset revealed that the decision making process must be different from the one of humans: Varying the thickness or color of the lines caused a severe drop in performance. In a second experiment, we analyzed the importance of local and global features. Intuitively, detecting whether a contour line is closing requires access to large portions of an image, since the line could have a gap at any position. This is in contrast to the type of information preferred by DNNs, as they often rely on local image content (Geirhos et al., 2019). For this reason, we investigated the importance of global processing by using a constrained model that only has access to small image patches only (Brendel and Bethge, 2019). We found that this model still performed well on our task, suggesting that, contrary to our original assumption, global features were not required to perform well on the task we designed.

- *Synthetic Visual Reasoning Test.* Humans are very good at solving tasks that require abstract visual reasoning. An interesting research question is how well machines can perform such tasks. One dataset to study this is the Synthetic Visual Reasoning Test (SVRT) (Fleuret et al., 2011). While humans can easily solve all 23 problems in this dataset, DNNs were found to have difficulty with a subset of them. Specifically, lower performance was obtained when the task required judging whether two shapes were the same (*same-different tasks*), as opposed to judging the spatial arrangement of shapes (*spatial tasks*) (Stabinger et al., 2016; Kim et al., 2018). Using a parameterized version of the SVRT dataset, Kim et al. (2018) analyzed this further and found a difference in the learning curves of feedforward models: Same-different tasks were more difficult to learn than spatial tasks. In further experiments, they showed that an attentive version

of the model was able to learn both types of tasks. Based on these findings, the authors concluded that feedback mechanisms, as present in biological systems, are likely important for efficiently performing same-different tasks. However, we argued that despite the large number of experiments, one still cannot conclude that DNNs without feedback mechanisms are not capable of performing well on same-different tasks. First, the observed difference in learning complexity is no evidence that one type of task is harder than another: Humans could also have different learning curves, but these differences would not show up when testing adults who have been exposed to visual input throughout their lives. Second, the differences in the learning complexity could also be explainable by other differences between biological and artificial systems or by a poor choice of network architecture or training procedure. Third, the finding that a model with an attentive mechanism can perform well on same-different tasks should not be taken as evidence that feedback is required. For these reasons, we argued that it is not yet clear whether feedback mechanisms are necessary for same-different tasks. In fact, we showed that DNNs can indeed perform well on all SVRT tasks if an appropriate architecture and training scheme are chosen.

- *Recognition Gap.* When an image is reduced in size, the content becomes unrecognizable after a certain point. Interestingly, this drop in recognition performance is very sharp. In particular, there is a certain subimage for which further reduction of the image renders it unrecognisable to a large percentage of observers (Ullman et al., 2016). This sharp drop in performance is referred to as a "recognition gap". Ullman et al. (2016) investigated whether a similar performance drop exists for DNNs. To this end, they tested a DNN on the image pairs for which humans experienced the recognition gap. They found no recognition gap and concluded "that the human visual system uses features and processes that are not used by current models and that are critical for recognition". While we agreed with the conclusion that DNNs may not use the same features as humans, we disagreed with the statement that these features are "critical for recognition". To gain further insights, we modified the experiment to allow a fairer comparison with humans. Instead of evaluating the model on image pairs that were selected by humans, the model could select its own patches in an iterative process similar to that used when testing humans. In this setup, we found a recognition gap similar to that in humans. This showed that the machines were also very sensitive to even small reductions in image size; only the precise minimal features differed from those used by humans.

Discussion

This project started when we studied the closed contour detection task to compare the inductive biases and decision making processes between humans and machines. In this process we noticed that the conclusions drawn from such comparisons can depend severely on small details in the design, analysis, and interpretation. This observation was also true for our two additional case studies on SVRT and the recognition gap. While our findings on each case study are certainly insightful for the respective research questions, the main contribution of this work is our systematic assessment of common pitfalls that can arise in comparative studies. In particular, we identified aspects that may influence the outcome of comparison studies and structured them in the form of a checklist.

As outlined in Section 1.3, there is a long history in comparative psychology of studying model systems such as monkeys or mice and discussing how to draw conclusions from such comparisons. Our checklist builds on these results, but is tailored to the peculiarities of artificial systems. A such, it complements discussions of the methodology of human-machine comparisons (Majaj and Pelli, 2018; Barrett et al., 2019; Cichy and Kaiser, 2019; Buckner, 2019). Addressing the individual points of our checklist can sometimes be very difficult and certainly there are more aspects than those we mention. Nevertheless, taking these points in account when planning, analyzing and interpreting experiments can improve the robustness of conclusions drawn from comparative studies between humans and machines. In the following, I will describe the individual points of our checklist using the three case studies as well as other examples from the literature.

1. **Isolating implementational or functional properties.** By their very nature, the human and machine systems exhibit a wide variety of differences. Therefore, it is difficult to isolate a particular component responsible for a given observation. For example, as pointed out in the SVRT case study, the difference in the learning behavior between humans and machines may hinder the usefulness of the learning complexity as a tool for comparing the difficulty of task types. A promising approach for isolating specific aspects can be to constrain the network such that it differs only in one dimension, as was done when comparing ResNets to BagNets on ImageNet (Brendel and Bethge, 2019) and also on our closed contour study.
2. **Aligning experimental conditions for both systems.** A fair comparison requires a good alignment of experimental conditions. Accounting for the plethora

of differences between the human and machine visual system is important to ensure equivalent settings. The case study on the recognition gap is an example of how the alignment of experimental conditions can change the outcome of a study. The critical point here was the selection procedure of the patches used in the machine experiment. Instead of using the patches found in the experiments on humans, we used the performance of the machines for patch selection. Other challenges in aligning the experimental conditions between humans and machines include the mismatch of stimuli presentation times (DiCarlo et al., 2012; Serre et al., 2007; VanRullen, 2007) and differences in the hardware such as memory or capacity of the systems (Firestone, 2020).

3. **Differentiating between necessary and sufficient mechanisms.** In many cases, there are several ways to solve a task. Therefore, one must be careful not to be biased to assume that one way is better than another. The studies on SVRT and closed contour detection showed that multiple mechanisms can enable good performance. In particular, our results on SVRT showed that recurrent processes are not necessary to perform well on same-different tasks. Similarly, we showed that global processing was not the only way to solve our closed contour detection task. Instead, contrary to our intuitions, local features could also lead to high performance. Similarly, for object classification, both texture and shape features can allow for high performance (Kubilius et al., 2016; Geirhos et al., 2019).
4. **Testing generalization of mechanisms.** The decisions of an artificial system can be very sensitive to small changes in the dataset or task. Thus, it is important to specify for which settings a result is intended to hold. Assessing whether a system has learned the underlying concept is particularly difficult. Generalization tests, such as those conducted for the closed contour task, can provide some insights into whether the features used are similar to those that humans use to make decisions (Yan and Zhou, 2017). One could also ask for the SVRT study whether our model understood the concept. We intentionally did not test this in our study because we were only interested in whether a good performance on the tasks could be reached, which was not clear at the time. Nevertheless, it is an interesting question whether the concept of sameness was learned. This was addressed by Puebla and Bowers (2021), who found that our model did not perform well on most generalization tests, suggesting that it did not learn the underlying concept.
5. **Resisting human bias.** Our human reference point can influence the design of

studies and the interpretation of results. The closed contour study illustrates the difficulty of overcoming our human bias. Although it would be tempting to conclude from initial results that a model has learned a human-like concept, several additional experiments may be required to find alternative solutions as it was done in the closed contour study. The study on the recognition gap showed how a bias in the experimental design, namely the selection of stimuli, can affect the outcome of a study. The selection of stimuli and labels has also been shown to be important when studying adversarial examples (Dujmović et al., 2020). Other examples for the influence of our human reference point are Braitenberg vehicles (Braitenberg, 1984), the effect of anthromorphizing (Buckner, 2019), and the experimenter outcome-bias (Rosenthal and Fode, 1961).

2.5 Disentanglement and Generalization under Correlation Shifts (P5)

This section summarizes:

Christina M. Funke*, Paul Vicol*, Kuan-Chieh Wang, Matthias Kümmerer†, Richard Zemel†, Matthias Bethge† (2022). Disentanglement and generalization under correlation shifts. *Oral presentation at Conference on Lifelong Learning Agents, 2022.*

The full publication and author contributions can be found in the appendix on page 187.

* joint first authors, † joint senior authors

Motivation

One remarkable property of human vision is its robustness. We can easily recognize objects or animals in new environments. Artificial algorithms, on the other hand, rely heavily on the structure present in the training data. As a result, they often fail for unseen or rare situations, such as a cow on the beach (Beery et al., 2018). Understanding the concept of background and foreground and the ability to disentangle them could be important to achieve robustness to changes in the environment. The goal of our work was to identify the objective that allows for learning disentangled representations in settings with correlated training data.

Disentanglement methods, such as variational auto encoders, typically minimize the mutual information between latent subspaces. However, Träuble et al. (2021) found that for these methods, training on correlated data results in subspaces that encode a mixture of the correlated attributes. Furthermore, they found that common disentanglement metrics cannot detect pairwise correlations. As a solution, they propose partial supervision. Here we show that even with full supervision, the objective of minimizing the mutual information between latent subspaces may fail. We discuss how the objective needs to be adapted to support learning disentangled representations for correlated training data.

More formally, the problem we want to target can be described as follows [1]. Suppose we observe noisy data $\mathbf{x} \in \mathbb{R}^m$ obtained from an (unknown) generative process $\mathbf{x} = g(\mathbf{s})$ where $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ are the *underlying factors of variation*, also called source variables or attributes, which may be correlated with each other. We wish to find a transformation $f : \mathbb{R}^m \to \mathbb{R}^n$ to a latent space $f(\mathbf{x}) = \mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K)$ such that

[1] The problem statement was taken almost verbatim from our corresponding paper.

each of the original attributes s_k can be recovered from the corresponding subspace $\mathbf{z}_k$ by a linear mapping $\mathbf{R}_k$, e.g., $\hat{s}_k = \mathbf{R}_k \mathbf{z}_k$. We consider three different objectives for learning the latent subspaces: 1) minimizing a supervised loss L (e.g., mean squared error or cross-entropy), $\sum_{k=1}^{K} L(\hat{s}_k, s_k)$; 2) minimizing the *unconditional mutual information between subspaces* in addition to the supervised loss, $\sum_k L(\hat{s}_k, s_k) + I(\mathbf{z}_1, \ldots, \mathbf{z}_K)$; and 3) minimizing the *conditional mutual information between subspaces conditioned on observed attributes*, in addition to the supervised loss, $\sum_k L(\hat{s}_k, s_k) + I(\mathbf{z}_k; \mathbf{z}_{-k} \mid s_k)$. We denote by $\mathbf{z}_{-k}$ the set of subspaces $\{\mathbf{z}_1, \ldots, \mathbf{z}_{k-1}, \mathbf{z}_{k+1}, \ldots, \mathbf{z}_K\}$. After optimizing the models for these three objectives, we evaluated them on data where the underlying attributes were differently correlated.

Results

First, we demonstrated that the problem of disentangling correlated attributes arises even in simple problems such as linear regression of Gaussian data. For this, we assumed the following setup: The generation process is given by $\mathbf{x} = g(\mathbf{s}) = \mathbf{A}\mathbf{s} + \mathbf{n}$, where $\mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s})$ are the underlying source variables with covariance matrix $\mathbf{C_s}$ and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$ is i.i.d. Gaussian noise. $\mathbf{A}$ is a linear transformation (in the simplest case $\mathbf{A} = \mathbf{I}$). The latent representation is given by $\mathbf{z} = f(\mathbf{x}) = \mathbf{W}\mathbf{x}$ and the predictions of the model can obtained linearly by $\hat{s}_k = R_k z_k$. In this example, the optimal $\mathbf{W}$ and $\mathbf{R}$ for the three objectives as well us the corresponding performances under correlation shift can be computed analytically. In short, we found that

- minimizing only the regression loss does not lead to disentanglement. More precisely, we found that optimal linear regression does not equal the inverse $\mathbf{A^{-1}}$ of the generative model but depends on $\mathbf{C_s}$ and $\mathbf{C_n}$. In other words, the correlation structure $\mathbf{C_s}$ is exploited to overcome the noise, leading to wrong predictions when the correlation structure changes.
- minimizing the MI between latent subspaces in addition to the regression loss does not help. The reason is that this constraint forces the model to remove all correlations between the latent subspaces. However, since the source variables are correlated, this leads to poor predictive performance — even on the training data.
- minimizing the CMI in addition to the regression loss leads to robustness under correlation shift. In contrast to unconditional independence, conditional independence retains the shared information necessary to account for the correlation between the source variables. In particular, we found that for optimal regression under the constraint of conditional independence, the underlying system

matrix is recovered ($\mathbf{W} = \mathbf{A}^{-1}$).

In addition to the regression task, we compared the three different objectives on a toy *classification* task. For both tasks, we examined the dependence on the noise level and the strength of the correlation present in the training data. We found that for the baseline model, the drop in performance under correlation shift was largest for strong correlations and intermediate noise levels. When enforcing conditional independence, accuracy remained high under correlation shift for all noise and correlation levels.

Next, we presented a method for minimizing CMI for more complex cases. In short, this method selects samples that match a particular attribute value (e.g., $a_k = 0$) and batch-wise shuffles their latent spaces such that only the subspace corresponding to the attribute remains unchanged. A discriminator is trained to distinguish the original samples $p(z_1, \ldots, z_K \mid a_k)$ from the shuffled ones $p(z_k \mid a_k)p(z_{-k} \mid a_k)$. By adversarial training, we teach the encoder to fool the discriminator. This enforces $p(z_1, \ldots, z_K \mid a_k) = p(z_k \mid a_k)p(z_{-k} \mid a_k)$ which corresponds to the CMI being minimized. This optimization is done for all attributes and all their possible values simultaneously.

We applied this method to two classification tasks with unknown generation processes. In one of them, we placed two handwritten digits (LeCun et al., 1998) side by side so that the identity of the digits was correlated during training. Noise was induced by partial occlusion of the images. For the second task, we used a subset of the CelebA dataset (Liu et al., 2015) containing faces for which the attributes `Male|Female` and `Smiling|Not Smiling` were correlated during training. For both datasets, we found that conditional independence leads to improved robustness under correlation shift compared to the baseline objectives. In addition, we evaluated common disentanglement metrics (Locatello et al., 2019) on the uncorrelated test data to confirm our findings and showed that our results also hold for weakly supervised settings.

Discussion

The main contribution of this work was to establish *conditional* mutual information (CMI) as correct objective for disentanglement in the presence of correlations. In particular, we showed that the usual approach of minimizing mutual information is conceptually incorrect in this setting. This became apparent when evaluating the predic-

tive performance under correlation shift. While the CMI objective showed its benefit most clearly for highly correlated attributes, it is still relevant in terms of fairness for weaker correlations. We saw that the baseline models made use of the structure present in the training data to improve their performance. Or in other words, datapoints were systematically treated incorrectly based on biases present in historical data. This can be very problematic when the correlations involve sensitive attributes such as gender or ethnicity. The CMI objective, on the other hand, prohibits the exploitation of such undesirable correlations in the training data.

In summary, we have shown that minimizing CMI is the correct objective in the face of correlations. However, our method for minimizing CMI has a few limitations. First, it requires labels for the attributes. Second, it is only applicable to categorical attributes. Third, it increases the computational cost and can be challenging to train. An advantage of our method is that it operates on the latent space and is therefore agnostic to the encoder architecture. Future work could consider combining our approach with other methods such as Invariant Risk Minimization (Arjovsky et al., 2019) or ideas regarding the information bottleneck (Sauer and Geiger, 2020).

3 General Discussion

In this dissertation, three paths were taken to improve our understanding of visual processing and the resulting internal representations. In the first part of the thesis, we analyzed aspects of human vision. For this, we conducted psychophysical experiments with stimuli generated by artificial models to learn about human texture perception and peripheral vision. The second line of research focused on the comparison between human and machine perception. We discussed at a methodological level how to draw robust conclusions from comparative studies and conducted three case studies, on closed contour detection, visual reasoning tasks, and a phenomenon called recognition gap. In the third approach, we addressed a fundamental discrepancy between humans and machine vision and made it computationally explicit. In particular, we targeted the robustness to correlation shifts and established conditional independence as the correct objective for disentanglement in correlated settings.

The general discussion is divided into two parts. The first part focuses on the methodological level and addresses *how* we can gain knowledge about visual perception (Section 3.1). The second part discusses *what* we have learned about visual processing and the mechanisms that might be crucial (Section 3.2).

3.1 Methodology of Vision Studies

One major question is *how* to learn about vision. Our checklist for comparative studies (**P4**) contributes to this ongoing discussion. Here, I will discuss the methodology of understanding vision on a broader level. I selected three major questions and will discuss my research in this context. The three questions are at which levels can vision be analyzed (Section 3.1.1), what approaches can a researcher take to learn about vision (Section 3.1.2), and what types of stimuli can be used (Section 3.1.3).

3.1.1 Levels for Analyzing Vision

David Marr's seminal work in the 1970s identified three levels for analyzing vision (Marr and Poggio, 1976; Poggio, 1981). In this multi-level description, the principal problem that is attempted to be solved is termed as *computational level*. The strategies to realize these computations form the *algorithmic level*. Finally, the building blocks used to implement these algorithms are denoted as the *implementational level*. Although the distinction between these levels is not always straightforward, it is important to consider carefully which level of description is suitable for the research

goal at hand. When modeling the human system with artificial systems, there is often no intention to adopt all aspects of the implementation. For example, it is neither feasible nor desirable to mimic the chemical processes of biological vision. Instead, performing the analysis on the algorithmic or even the computational level is more fruitful.

My research often targets the intersection of the computational and algorithmic level and makes use of having access to two fundamentally different systems that can perform visual tasks, namely biological and artificial systems. In the study on peripheral vision (**P3**) the goal of compressing the image in the periphery could be considered as the computational level, while the strategy of computing specific summary statistics would correspond to the algorithmic level. Whether this is done with synapses or computer code would then be the implementational level. The ideas discussed in (**P4**) mainly concern the algorithmic level. One point of our checklist highlights that multiple solutions can allow for good performance on a given task. Here, we emphasize the importance of distinguishing between necessary and sufficient mechanisms and pointed out that multiple solutions can allow for good performance. In particular, the solution used by the human visual system need not be the only one or even the best one. We illustrated this with our closed contour detection task, where we saw that multiple approaches, namely global and local processing, can lead to good performance. Similarly, when discussing the challenge of making same-different judgments, we discussed the importance of one possible mechanism, namely recurrency. And finally, in the work on disentanglement (**P5**), we targeted the computational level. In particular, we envisioned robustness to correlation shift and proposed to implement it by minimizing the conditional MI objective, which is likely to be very different from the way robustness is achieved in biological systems.

A common objection raised with respect to studies comparing humans and machines is that such comparisons do not make sense given the many differences at the implementational level. These claims neglect that comparisons are also possible at the algorithmic and computational level, which can be largely independent of the exact implementation. When trying to identify the *fundamental components* that are important for successful visual processing, it may actually be an advantage that the systems differ at the implementational (or even the algorithmic) level. Access to multiple systems that perform visual tasks in different ways can help distinguish the critical components from mere design choices or solutions that evolution has produced (Nilsson, 2021).

3.1.2 Understanding Vision: Analyzing vs. Constructing

In vision research, the various fields take different approaches to learning about vision. Psychophysics and neuroscience analyze existing systems to derive knowledge. Typically behavioral or neural data is collected, and researchers attempt to find models that reproduce biological phenomena. While this is an exciting goal, it underestimates the challenge of modeling the *functional properties* of the brain. Computer vision, on the other hand, aims more at achieving a high performance on specific tasks. Here, biological phenomena are of little interest, although they may provide some inspiration. Often the starting point is to define benchmarks for a task of interest and to compare a range of models on these benchmarks. This can lead to an understanding of what characterizes good and bad models, providing the researcher with a good starting point for further model improvement. This idea of understanding by constructing follows the famous dictum of the physicist Richard Feynman "What I cannot create, I do not understand". For further discussion see also Anderson and Kreiman (2011).

This dissertation took three approaches to learning about vision that incorporate methods of different fields of vision research. Specifically, we *analyzed* aspects of biological vision, we *compared* biological and artificial neural networks and we *addressed fundamental discrepancies* between these systems. All three approaches incorporated ideas from both the analyzing and constructing concepts described above. While the studies on texture perception and peripheral vision (**P1**-**P3**) focused on the analysis of human processing, the development of texture models was a major part. Here, the generation of textures that appear natural to humans is closely related to the aspect of constructing. The comparison of visual systems (**P4**), on the other hand, involved extensive analysis of the individual systems. Finally, we addressed a fundamental discrepancy between biological and artificial systems (**P5**). For this, we examined the problem on a conceptual level, constructed a solution that incorporated the resulting ideas and analyzed whether this closed the gap in performance.

Overall, it is important to have a wide set of methods available to obtain a comprehensive picture of visual processes. From this point of view, the different fields of vision research complement each other and it is worth combining their approaches.

3.1.3 Types of Stimuli

In traditional psychophysics, vision is being studied with artificial stimuli such as bars, sinusoidal gratings, or ensembles of simple stimuli. With the upcoming of more computational power it became feasible to experiment with more natural stimuli. Such studies concerned, for example, the usability of image statistics for animal recognition (Torralba and Oliva, 2003; Wichmann et al., 2010) or studied the orientation of contours in natural images (Coppola et al., 1998). The use of natural and ecologically relevant stimuli is promising, as this is thought to provide a more meaningful test of neural models (Felsen and Dan, 2005; Olshausen and Field, 2005). From an intuitive point of view, it makes sense to use natural images, as the biological visual system has also evolved to process natural images. Furthermore, the design of artificial stimuli always involves assumptions about the stimulus parameters that might be relevant to the research question. Therefore, the designed stimuli lack the richness of natural stimuli, which can prevent us from filling gaps in our knowledge about vision.

There are also voices arguing for the benefits of synthetic stimuli (Rust and Movshon, 2005; Martinez-Garcia et al., 2019). For one, their properties can be better controlled, which allows for testing specific hypotheses. This can shed light on why a model might have failed on natural images. In addition, natural image datasets can also be biased and even misrepresent basic visual phenomena (Torralba and Efros, 2011). Since the photographer already decides which aspect of reality to capture in a picture, such datasets might be less natural than one might expect (Adams, 1980; Wichmann et al., 2010). Similar concerns apply to the "natural" datasets that are commonly used for training DNNs, such as ImageNet (Deng et al., 2009). The naturalness of the images, the choice of classes and the labeling of the images is highly debatable. However, there are some recent attempts to improve the ecological relevance of such datasets (Mehrer et al., 2021). In summary, for all kinds of stimuli, the selection is likely biased by our human perspective. The closed contour detection and detection gap studies (**P4**) are examples of how stimulus selection has a large impact on the result.

There are use cases for a wide range of stimuli. Sometimes, even mathematical examples like Gaussian distributions are well suited. This was the case in our study on disentanglement (**P5**), in which we targeted conceptual questions. Specially designed stimuli on the other hand may be suitable to test whether DNNs learned abstract

concepts (**P4**) or whether they are susceptible to visual illusions. Parametric models (such as the texture models used in **P1**-**P3**) can help to generate stimuli that share important properties with natural stimuli, but at the same time can be synthesized in a well-controlled manner. Another promising avenue is the generation of synthetic images using 3D CAD models (Aubry et al., 2014). Natural images can also be useful for a variety of studies. An example for the superiority of natural images is provided by Borowski et al. (2020). A typical way to visualize the features of DNNs is to synthesize images that maximally activate specific layers or neurons (Erhan et al., 2009). However, Borowski et al. (2020) have shown that the resulting images are actually less informative than natural images. Overall, choosing the right kind of stimuli is a difficult, yet very crucial task.

One approach used in this dissertation are *metamers*. This kind of stimulus is particularly interesting, and thus it is worth reviewing the underlying idea and its implications in some more detail. For non-bijective models, multiple input images can result in the same representation. This effect is called metamerism. A typical example for metamerism is human color perception (Wyszecki and Stiles, 1982; Cohen and Kappauf, 1982). Here, certain color stimuli are perceived as identical, although their spectral power distributions differ. Systematic study of these color metamers has revealed the three dimensions of human color vision, which is implemented by three types of cones, each responding to a range of wavelengths. While color metamers arise from retinal properties, it is likely that metamers also exist further down the ventral stream. Their study, in a similar way as for the color metamers, could shed light on how visual information is processed. One candidate is the metamers of peripheral vision, for which it is assumed that information is discarded by computing summary statistics (Freeman and Simoncelli, 2011, **P3**). Metamerism can also occur in other domains. For example, ocean sound recorded at different times may be indistinguishable. Studying which stimuli produce the same responses at some stage of a network's representation can be very insightful and this approach has gained interest in recent years. In particular, Feather et al. (2019) have promoted to use metamers of neural networks to show deviations from the human system. The idea is that metamers can reveal to which image features an encoder is invariant. From this point of view, adversarial examples have been described as complementary. As phrased by Feather et al. (2019), "adversarial examples are metameric (perceived similarly) for humans but are not metameric to the network they are derived for, demonstrating that the network lacks some invariances present in humans". This is explored in detail by Jacobsen et al. (2019).

3.2 Mechanisms Underlying Visual Processing

Another major question in vision research concerns the computations that produce a visual representation that enables the robust and efficient performance of visual tasks. The role of specific mechanisms is not yet fully understood and the topic in the vision community. Among other topics, such discussions revolve around the question whether information is processed locally or globally (Section 3.2.1), the potential benefits of recurrent mechanisms (Section 3.2.2) and the factors that are important for generalization and robustness (Section 3.2.3).

3.2.1 Local versus Global Processing

The question of whether visual perception is based on local or global features has a long history in psychology. Wilhelm Wundt and Edward Titchener established the structural psychology, proposing that perception could be explained by identifying the individual components and putting them together to describe complex processes (Titchener, 1929; Biederman, 1987). This view changed with the emerge of the Gestalt theory (Koffka, 1922) following the premise "the whole is greater than the sum of its parts" (Aristotle). One main implications is that the arrangement of items can influence the perception of the individual parts. This finding is summed up by the famous statement that we see the "forest before trees" (Navon, 1977; Kimchi, 1994).

The importance of global structure for humans is evident from a number of studies. It is, for example, well known that context has a strong influence on our perception. One and the same local patch can be perceived very differently depending on the surrounding image content. Even more, in low-resolution images where only the gist of an object is visible, people hallucinate objects that fit into the scene. For example, people assume that the blackish area below a computer monitor must be a keyboard, even though there are no local features to support this hypothesis (Torralba et al., 2010; Murphy et al., 2006). The study of two-tone images revealed the importance of context in the form of prior knowledge. Here, the local structure of images was removed by limiting the images to black and white pixels (Hayes, 1988). The recognizability of these two-tone images was found to be strongly influenced by high-level prior knowledge (Teufel et al., 2018). The importance of global structure for human perception is also evident in children's learning behavior. When learning words for new objects, the shape of the objects is weighted more heavily than their size or texture (Landau et al., 1988). Furthermore, objects that have the texture characteristics of one category and the shape of another (so-called cue-conflict images), revealed that

humans are biased towards shape when classifying objects (Geirhos et al., 2019). Finally, it has been shown that people ignore features that are highly predictable and clearly visible, relying instead on less predictable global features (Malhotra et al., 2021).

With the advent of artificial systems it is debated whether machine vision has a similar preference for global information. While some studies report a shape-bias similar to humans (Ritter et al., 2017; Kubilius et al., 2016), most work has found that local cues are more important for DNNs. First, the success of DNN features in predicting neural responses was found to be due to the textural properties rather than contour properties (Laskar et al., 2018; Long and Konkle, 2018). Second, while context is important for human perception, machines have been shown to lack awareness for context. In search tasks, humans were found to often miss targets when their size is inconsistent with the rest of the scene (Eckstein et al., 2017). This was not the case for the DNNs tested, which was explained by their reliance on local features. Third, it was found that standard DNNs trained for object classification on ImageNet have a texture-bias for cue-conflict images (Geirhos et al., 2019). However, it should be noted that the training data has a strong influence on this preference (Geirhos et al., 2019; Hermann et al., 2020; Hosseini et al., 2018). Finally, DNNs restricted to local information were still able to perform surprisingly well in image classification (Gatys et al., 2015; Brendel and Bethge, 2019). Similarly, our research (**P4**) has shown that a locally constrained network can perform well on a closed contour detection task where one would have intuitively assumed that the task would not be solvable without access to global features.

The question of local and global processing is also prevalent in discussions about peripheral vision. The loss of visual information in the periphery and effects such as crowding are often explained by the pooling of visual features in local areas. The hypothesis is that higher layers have only access to these summary statistics (Rosenholtz, 2014; Freeman and Simoncelli, 2011). Our work, however, has shown that this local summary statistic approach does not capture appearance. In particular, we found that long-range spatial dependencies pose a problem. This became already evident in our study of the perceptual quality of texture models (**P2**), where we found that textures with long-range dependencies, such as roof tiles, were problematic. Similarly, for dynamic textures, we found that the global structure was not captured by the purely local representation (**P1**). Most directly, however, the problem of long-range structures showed in our third study (**P3**). We found that the sum-

mary statistics approach could not produce full field metamers when the synthesized images were compared to the original image. Instead, we found that the ability to generate metamers depended strongly on the image content. In particular, we found that features that span over large areas of the image are indeed relevant. This is consistent with previous findings showing that flankers far away from the target can in fact affect how the target is perceived. These phenomena are known as *uncrowding* (Manassi et al., 2013, 2016) and *supercrowding* (Vickery et al., 2009). As summarized by Herzog et al. (2015), "the spatial configuration across the entire visual field determines crowding".

Later work aimed to improve the foveated models to better capture the appearance of scenes across the field of view. While some researchers focused on increasing the speed and usability of the models (Walton et al., 2021; Long et al., 2018), others discussed how to include additional methods to incorporate global aspects (Herrera-Esposito et al., 2021; Doerig et al., 2019). A number of recent papers made use of the resulting stimuli to investigate the relation between local versus global processing, foveation operations, and the properties of resulting representations (Harrington and Deza, 2021; Ziemba and Simoncelli, 2021; Deza and Konkle, 2020). The work of Jagadeesh and Gardner (2022) is particularly interesting, as it offers a different perspective on whether human perception of objects is based on texture or shape. More specifically, they find that the human visual cortex also represents texture-like information, but unlike CNNs, is sensitive to the spatial arrangement of the features.

Overall, there is widespread agreement that many human visual processes can not be explained by combining local features alone. Mechanisms that support grouping and segmentation such as attention and recurrency, are promising additions (Doerig et al., 2020; Linsley et al., 2018a; Spoerer et al., 2019). A relatively recent development in this vein are transformer architectures (Vaswani et al., 2017), which allow flexible allocation of attention over larger spatial scales. These architectures have been successfully used for visual tasks (Parmar et al., 2018; Dosovitskiy et al., 2020) and can overcome the texture-bias on cue-conflict images (Tuli et al., 2021).

3.2.2 Importance of Recurrent Mechanisms

The human brain is known to have connections not only from lower to higher visual areas. Rather, the higher visual areas can influence the activation of neurons in lower visual areas. Also, there exist connections between neurons of the same visual area

(Lamme and Roelfsema, 2000). Often, feedforward-only processing is associated with pre-attentive vision, whereas recurrent processing corresponds to attentive vision. In this sense, a feedforward sweep is thought to provide an initial coarse visual representation sufficient for rapid categorization tasks (Serre et al., 2007; VanRullen, 2007). The feedback mechanisms and lateral connections, on the other hand, are thought to be important for perceptual organization and segmentation.

Unlike biological brains, many standard DNNs have only feedforward connections. However, the inclusion of recurrent mechanisms may be important for visual processing (O'Reilly et al., 2013; Lindsay, 2020; Nayebi et al., 2018) and can help overcome observed differences between humans and machines (Kar et al., 2019; Kietzmann et al., 2019; Kim et al., 2018; Linsley et al., 2018b). For example, Spoerer et al. (2017) has found that object recognition in the presence of clutter and occlusions can be improved by lateral and feedback connections. Recurrent mechanisms and attention can also help for tasks with long-range dependencies (Linsley et al., 2019; Kim et al., 2020). This becomes evident in transformer architectures, which, in a targeted manner, take larger spatial regions into account (see previous section).

In terms of Marr's level of description, the question arises whether recurrency is an implementational choice or a functional property. While the literature listed in the previous section seems to favor the latter, there is also evidence for the former hypothesis. For one, any finite-time recurrent network can be unrolled into a feedforward network with weight sharing (Liao and Poggio, 2016; van Bergen and Kriegeskorte, 2020). For another, the error patterns of recurrent and feedforward models are remarkably similar (Geirhos et al., 2020b). Furthermore, our study on the SVRT dataset (**P4**) has shown that, contrary to previous assumptions, feedback mechanisms are not necessary to perform well on same-different tasks: While previous studies found that recurrent connections could be crucial for efficient performance of these tasks (Kim et al., 2018), we argued that they are not required to reach a good performance and showed that a standard DNN, namely ResNet-50, could perform well on same-different tasks. This supports the hypothesis that recurrency is an implementational choice rather than a functional property.

Since we published our work, the importance of recurrency has been further explored and researchers found that recurrent networks can have some advantages over feedforward architectures in practice. Recurrent architectures are thought to be more computationally efficient (e.g., if the hardware is limited), more flexible (e.g., if the

task or dataset changes), and allow exploitation of temporal dependencies in the data (van Bergen and Kriegeskorte, 2020; Kreiman and Serre, 2020).

In particular, it has been discussed whether attentive mechanisms might be beneficial when considering performance on data outside of the training distribution. For this, a range of architectures with attentive mechanisms were tested on the SVRT same-different tasks, namely ResNets with attention modules (Vaishnav et al., 2021), Recurrent Vision Transformers (Messina et al., 2021), Relation Networks (Santoro et al., 2017; Kim et al., 2018) and Siamese architectures (Kim et al., 2018). Vaishnav et al. (2021) analyzed the dependence on the number of training examples more systematically, created a taxonomy of the SVRT tasks and discussed their computational demands. Importantly, they distinguished between feature learning and rule learning. Feature learning tests whether a high accuracy can be reached on the test set, whereas rule learning tests whether the underlying concept has been learned (similar to **P4**, where we tested whether the concept of closedness has been learned). The researchers found that attention processes contribute to rule learning. This may indicate that while feedforward processing works well in feature learning, recurrent mechanisms may facilitate learning the abstract rule. This is consistent with the finding of Puebla and Bowers (2021) who showed that the pure ResNet model that we used in our experiments does not perform well on a range of generalization tests.

While these findings seem to suggest that recurrency is a functional property for learning the rule of sameness, there are still debatable points. For one, Puebla and Bowers (2021) showed that the Relation Network suffers from the same limitation as ResNet, even though it was designed for relational reasoning problems. Moreover, as will be elaborated in the next section, it is questionable what can be inferred from generalization tests for rule learning abilities.

3.2.3 What Drives Generalization?

Often DNNs achieve good accuracy on the task for which they were trained, but have problems on test data that has other properties. Such data is denoted as *out-of-distribution* (ood) data and the change of the distribution is often denoted as *domain shift*. One can vaguely distinguish between two types of shifts in the test distribution. First, the appearance of the scene or image can be affected, which means that the low-level image statistics change. This pixel-wise structure is typically assumed to be encoded in the lower layers of neural networks. In the real world, this corresponds

to variations in lighting conditions or camera settings (DiCarlo and Cox, 2007; Cox, 2014). Such effects can be simulated by adding noise, by blurring the image or by changing its contrast (Geirhos et al., 2018; Ghodrati et al., 2014). Another example for a change in the appearance is the modification in color or thickness of the lines in the closed contour case study (**P4**). Second, the shifts may be more at the level of objects. Loosely speaking, this would refer to features that are sometimes thought to be encoded in higher layers of neural networks. Returning to the closed contour example, this corresponds to generalization tests that modify the shape of the contours. Another example could be a change of the probability for co-occurrence of objects in the same image (Beery et al., 2018) — or more generally the correlation structure between attributes (**P5**).

Testing humans and machines on a range of different test datasets can be exciting, as the differences in the generalization performance can teach us something about the similarities and differences in the decision making process. We (**P4**) and other researchers (Zhang et al., 2018, 2019; Puebla and Bowers, 2021) have used this idea to discuss whether machines have learned the same concept that humans use to make their decision. Such concepts include closedness, sameness or counting abilities. The idea here is that if machines fail on generalization tests that humans can do easily, it is unlikely that they have "understood" the concept.

However, drawing conclusions about what a model has or has not understood is not easy. Suppose a model performs well on all variants of a dataset designed by a researcher. Would this be sufficient evidence that the model understood the abstract relationship? In our closed contour example, and in the aforementioned study of Puebla and Bowers (2021), all variants had in common that they consisted of simple lines and basic shapes. Wouldn't a model that understood the concept also have to work for images that contain photographs of real objects? Would failure on these test sets automatically mean that the model did not understand the abstract relation? In a sense, no one would be surprised if a model trained on line segments had difficultiy with such complex images, because one could argue that the domain shift was too large and that the model simply could not handle that kind of input. In other words, there seems to be a threshold for how large the differences between training and testing domain may be in order to still allow for testing whether a concept has been learned without obscuring the results by other factors. In this sense, in our study on closed contour detection, as in the study on same-different tasks, it could be the case that the model understood the rule, even though it failed on some generalization

tests. This discussion is closely related to the ideas of Firestone (2020) who address the distinction between performance and competence.

With all these concerns in mind, one major question remains, namely, what factors drive generalization. It is well known that the biases of a network determine generalization abilities (Mitchell, 1980), but it is not yet clear what "good" biases are and how they can be obtained. Overall, there are three aspects that have a large impact on biases and generalization performance: 1) the data used for training, 2) the architecture of the model, and 3) the training objective.

First, the data that a visual system faces has a large impact. For example, if the training data allows for shortcuts, it is not surprising that a model will use of these easier solutions and have difficulty if the shortcuts are not available in the test environment (Geirhos et al., 2020a; Malhotra et al., 2021). A striking example is the closed contour task (**P4**), where we have found that there are distinctive features that are not obvious to human observers. We saw that exploiting these features allowed for a reasonable performance on within-distribution data, but led to errors in ood data. One can take advantage of the strong influence of data on the generalization performance to improve robustness. A common method to do this is data augmentation.

Second, the architecture of the visual system affects the generalization performance. A large body of research is directed towards finding good model architectures (Li et al., 2017; Ding and Fu, 2017). The results of Geirhos et al. (2020b), on the other hand, indicate that the decision boundaries are hardly affected by the choice of architecture, which implies that the generalization performance depends less on the architecture as one may have expected.

Third, the functional goal of the visual system is of great importance for generalization abilities. The research on shortcut learning (Geirhos et al., 2020a) has shown that generalization performance decreases if a manipulation of the dataset affects a property that was used by a system to make its decision. If only features or correlation structures that were not used for decision making are changed, performance is not affected. This was also the case for the many successful generalization tests in the closed contour study (**P4**). Overall, the demands on a visual system can strongly influence on which features a model relies on. For a machine learning model that has the single goal of classifying objects on the ImageNet dataset, there is no need to become robust against changes in illumination, viewpoint or background. For humans,

on the other hand, this robustness is highly beneficial (DiCarlo and Cox, 2007; Cox, 2014). Conversely, DNNs are hardly affected by changes where texture statistics remain intact, such as the shuffling of small patches of the image (Brendel and Bethge, 2019). Since such a shuffling of the image content does not occur in real life scenarios, it is not surprising that humans have not developed robustness to such modifications. Changing the demands on the artificial systems can go a long way toward improving robustness (Carlucci et al., 2019). In a similar way, we have seen (**P5**) that choosing the correct objective function can counteract the exploitation of unwanted correlations and thereby improve generalization under correlation shift.

3.3 Conclusion

The processing of visual information has evolved over many million years, leading to a multitude of biological visual systems. Together with the diversity of artificial systems, researchers now have access to a variety of neural networks that perform well on visual tasks. By comparing their similarities and differences we are closer than ever to understanding the computations that enable the transformation of low-level information into meaningful abstract representations.

This dissertation follows three pathways to learn about visual processing and thereby combines methods from psychophysics, computer vision and machine learning. The first set of studies analyzes human texture perception. The next study addresses comparisons between human and machine perception. And finally, a fundamental discrepancy between human and machine vision is explored at a principled level by discussing the requirements for robust disentanglement.

This series of studies made it possible to gain insights into the importance of certain mechanisms. The focus was on three mechanisms, namely local versus global processing (Section 3.2.1), recurrent mechanisms (Section 3.2.2) and factors that drive generalization (Section 3.2.3). In addition, the dissertation addresses the methodological side of learning about these mechanisms. It puts a particular emphasis on three aspects, namely the levels at which vision can be analyzed (Section 3.1.1), possible approaches to learn about vision (Section 3.1.2) and the types of stimuli that can be used (Section 3.1.3).

Despite the great advance in computer vision and machine learning, biological vision is in many ways superior to artificial systems. Human representation of visual information is usually both robust and efficient. We can grasp concepts, understand underlying structures, and learn new tasks from very little training data. A key difference is that machine systems often only have access to limited training data, which might prevent them from learning a more universal feature space. Machine models could benefit from the prerequisites that humans have during their visual development, namely the ability to navigate and explore a three-dimensional environment. It will be exciting to see if recent ideas on continual and active learning can bridge the gap between biological and artificial systems.

References

Adams, A. (1980). *The camera. New Ansel Adams Photography Series*. New York Graphic Society.

Adelson, E. H. (2001). On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, volume 4299, pages 1–12. International Society for Optics and Photonics.

Amjad, R. A. and Geiger, B. C. (2019). Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9):2225–2239.

Anderson, W. S. and Kreiman, G. (2011). Neuroscience: What we cannot model, we do not understand. *Current Biology*, 21(3):R123–R125.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological science*, 12(2):157–162.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, 61(3):183.

Aubry, M., Maturana, D., Efros, A. A., Russell, B. C., and Sivic, J. (2014). Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3762–3769.

Balas, B. (2012). Contrast negation and texture synthesis differentially disrupt natural texture appearance. *Frontiers in psychology*, 3:515.

Balas, B. (2021). Texture perception. In *Oxford Research Encyclopedia of Psychology*.

Balas, B., Nakano, L., and Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of vision*, 9(12):13–13.

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to

study texture representations in the human visual system. *Vision research*, 46(3):299–309.

Barrett, D. G., Morcos, A. S., and Macke, J. H. (2019). Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64.

Beck, J. (1967). Perceptual grouping produced by line figures. *Perception & Psychophysics*, 2(11):491–495.

Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. (2020). Are we done with imagenet? *arXiv preprint arXiv:2006.07159*.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115.

Boesch, C. (2007). What makes us human (homo sapiens)? the challenge of cognitive cross-species comparison. *Journal of Comparative Psychology*, 121(3):227.

Borowski, J., Zimmermann, R. S., Schepers, J., Geirhos, R., Wallis, T. S., Bethge, M., and Brendel, W. (2020). Exemplary natural images explain cnn activations better than state-of-the-art feature visualization. *International Conference on Learning Representations*.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, 226(5241):177–178.

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision research*, 13(4):767–782.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. The MIT press, Cambridge, Massachusetts.

Brendel, W. and Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *International Conference on Learning Represen- tations*.

Buckner, C. (2019). The comparative psychology of artificial intelligences. *Philsci. Archive: 16128*.

Cardoso, J.-F. (1989). Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 2109–2112. IEEE.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238.

Cichy, R. M. and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317.

Clarke, A. M., Herzog, M. H., and Francis, G. (2014). Visual crowding illustrates the inadequacy of local vs. global and feedforward vs. feedback distinctions in modeling visual perception. *Frontiers in psychology*, 5:1193.

Cohen, J. B. and Kappauf, W. E. (1982). Metameric color stimuli, fundamental metamers, and wyszecki's metameric blacks. *The American journal of psychology*, pages 537–564.

Cohen, M. A., Dennett, D. C., and Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in cognitive sciences*, 20(5):324–335.

Cohen, T. and Welling, M. (2014). Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, pages 1755–1763. PMLR.

Coppola, D. M., Purves, H. R., McCoy, A. N., and Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences*, 95(7):4002–4006.

Cox, D. D. (2014). Do we understand high-level vision? *Current opinion in neurobiology*, 25:187–193.

Dakin, S. C. and Watt, R. (1997). The computation of orientation statistics from visual texture. *Vision research*, 37(22):3181–3192.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Deza, A., Jonnalagadda, A., and Eckstein, M. (2017). Towards metamerism via foveated style transfer. *International Conference on Learning Representations*.

Deza, A. and Konkle, T. (2020). Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Ding, Z. and Fu, Y. (2017). Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 27(1):304–313.

Doerig, A., Bornet, A., Choung, O.-H., and Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision research*, 167:39–45.

Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., and Herzog, M. H. (2019). Beyond bouma's window: How to explain global aspects of crowding? *PLoS computational biology*, 15(5):e1006580.

Doretto, G., Chiuso, A., Wu, Y. N., and Soatto, S. (2003). Dynamic textures. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference On*, volume 2, pages 439–446. IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dujmović, M., Malhotra, G., and Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife*, 9:e55978.

Ebert, D. S., Carlson, W. E., and Parent, R. E. (1994). Solid spaces and inverse particle systems for controlling the animation of gases and fluids. *The Visual Computer*, 10(4):179–190.

Eckstein, M. P., Koehler, K., Welbourne, L. E., and Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, 27(18):2827–2832.

Efros, A. A. and Freeman, W. T. (2001). Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

Feather, J., Durango, A., Gonzalez, R., and McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. In *NeurIPS*, pages 10078–10089.

Felsen, G. and Dan, Y. (2005). A natural approach to studying vision. *Nature neuroscience*, 8(12):1643–1646.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., and Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625.

Flom, M. C., Heath, G. G., and Takahashi, E. (1963). Contour interaction and visual resolution: Contralateral effects. *Science*, 142(3594):979–980.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature*

neuroscience, 14(9):1195–1201.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202.

Gatys, L., Ecker, A. S., and Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28:262–270.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020a). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Geirhos, R., Meding, K., and Wichmann, F. A. (2020b). Beyond accuracy: Quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems 33*.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *arXiv preprint arXiv:2106.07411*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems 31*.

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in computational neuroscience*, 8:74.

Gibson, J. J. (1950). The perception of visual surfaces. *The American journal of psychology*, 63(3):367–384.

Golan, T., Raju, P. C., and Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337.

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., and Bertalmío, M. (2019). Convolutional neural networks can be deceived by visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12317.

Harrington, A. and Deza, A. (2021). Finding biological plausibility for adversarially robust features via metameric tasks. In *SVRHM 2021 Workshop@ NeurIPS*.

Haun, D. B., Jordan, F. M., Vallortigara, G., and Clayton, N. S. (2011). Origins of spatial, temporal, and numerical cognition: Insights from comparative psychology. In *Space, Time and Number in the Brain*, pages 191–206. Elsevier.

Hayes, A. (1988). Identification of two-tone images; some implications for high-and low-spatial-frequency processes in human vision. *Perception*, 17(4):429–436.

Hermann, K., Chen, T., and Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015.

Herrera-Esposito, D., Coen-Cagli, R., and Gomez-Sena, L. (2021). Flexible contextual modulation of naturalistic texture perception in peripheral vision. *Journal of vision*, 21(1):1–1.

Herzog, M. H., Sayim, B., Chicherov, V., and Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of vision*, 15(6):5–5.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*.

Hosseini, H., Xiao, B., Jaiswal, M., and Poovendran, R. (2018). Assessing shape bias property of convolutional neural networks. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition Workshops, pages 1923–1931.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154.

Islam, A., Chen, C.-F. R., Panda, R., Karlinsky, L., Radke, R., and Feris, R. (2021). A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8845–8855.

Jacob, G., Pramod, R., Katti, H., and Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, 12(1):1–14.

Jacobsen, J.-H., Behrmann, J., Zemel, R., and Bethge, M. (2019). Excessive invariance causes adversarial vulnerability. *International Conference on Learning Representations*.

Jagadeesh, A. V. and Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *bioRxiv*.

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., and Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, 8:1726.

Julesz, B. (1962). Visual pattern discrimination. *IRE transactions on Information Theory*, 8(2):84–92.

Julesz, B. (1965). Texture and visual perception. *Scientific American*, 212(2):38–49.

Jutten, C. and Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recog-

nition behavior. *Nature neuroscience*, 22(6):974–983.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863.

Kim, B., Reif, E., Wattenberg, M., and Bengio, S. (2019). Do neural networks show gestalt phenomena? an exploration of the law of closure. *arXiv preprint arXiv:1903.01069*, 2(8).

Kim, B., Reif, E., Wattenberg, M., Bengio, S., and Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, pages 1–13.

Kim, J., Linsley, D., Thakkar, K., and Serre, T. (2020). Disentangling neural mechanisms for perceptual grouping. *International Conference on Learning Representations*.

Kim, J., Ricci, M., and Serre, T. (2018). Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4):20180011.

Kimchi, R. (1994). The role of wholistic/configural properties versus global properties in visual form perception. *Perception*, 23(5):489–504.

King, M. L., Groen, I. I., Steel, A., Kravitz, D. J., and Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage*, 197:368–382.

Koehler, O. (1943). Zaehl-versuche an einem kolkraben und vergleichsversuche an menschen. *Zeitschrift für Tierpsychologie*, 5(3):575–712.

Koffka, K. (1922). Perception: an introduction to the gestalt-theorie. *Psychological bulletin*, 19(10):531.

Köhler, W. (1925). The mentality of apes. *New York: Kegan Paul, Trench, Trubner & Co*.

Korte, W. (1923). Über die gestaltauffassung im indirekten sehen. *Zeitschrift für Psychologie*, 93:17–82.

Kreiman, G. and Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464(1):222–241.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. (2015). Deep convolutional inverse graphics network. *Advances in neural information processing systems*, 28.

Kwatra, V., Schödl, A., Essa, I., Turk, G., and Bobick, A. (2003). Graphcut textures: image and video synthesis using graph cuts. In *ACM Transactions on Graphics (ToG)*, volume 22, pages 277–286. ACM.

Lamme, V. A. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, 23(11):571–579.

Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321.

Laskar, M. N. U., Giraldo, L. G. S., and Schwartz, O. (2018). Correspondence of deep neural networks and the brain for visual textures. *arXiv preprint arXiv:1806.02888*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and

Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lettvin, J. Y. et al. (1976). On seeing sidelong. *The Sciences*, 16(4):10–20.

Levi, D. M. (2008). Crowding—an essential bottleneck for object recognition: A mini-review. *Vision research*, 48(5):635–654.

Li, A. and Zaidi, Q. (2000). Perception of three-dimensional shape from texture is based on patterns of oriented energy. *Vision research*, 40(2):217–242.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017). Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550.

Liao, Q. and Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*.

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of cognitive neuroscience*, pages 1–15.

Linsley, D., Kim, J., Berson, D., and Serre, T. (2018a). Robust neural circuit reconstruction from serial electron microscopy with convolutional recurrent networks. *arXiv preprint arXiv:1811.11356*.

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., and Serre, T. (2018b). Learning long-range spatial dependencies with horizontal gated recurrent units. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 152–164.

Linsley, D., Shiebler, D., Eberhardt, S., and Serre, T. (2019). Learning what and where to attend. *International Conference on Learning Representations*.

Liu, G., Gousseau, Y., and Xia, G.-S. (2016). Texture synthesis through convolutional neural networks and spectrum constraints. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3234–3239. IEEE.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124.

Long, B. and Konkle, T. (2018). The role of textural statistics vs. outer contours in deep cnn and neural responses to objects. In *Conference on Computational Cognitive Neuroscience*, page 4.

Long, B., Yu, C.-P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR.

Loper, M. M. and Black, M. J. (2014). Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision*, pages 154–169. Springer.

Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., Durand, F., Pfister, H., and Boix, X. (2020). On the capability of neural networks to generalize to unseen category-pose combinations. Technical report, Center for Brains, Minds and Machines (CBMM).

Majaj, N. J. and Pelli, D. G. (2018). Deep learning—using machine learning to study biological vision. *Journal of vision*, 18(13):2–2.

Malhotra, G., Dujmovic, M., and Bowers, J. S. (2021). Feature blindness: a challenge for understanding and modelling visual object recognition. *bioRxiv*.

Malik, J. and Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *JOSA A*, 7(5):923–932.

Manassi, M., Lonchampt, S., Clarke, A., and Herzog, M. H. (2016). What crowding

can tell us about object representations. *Journal of Vision*, 16(3):35–35.

Manassi, M., Sayim, B., and Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of vision*, 13(13):10–10.

Mansinghka, V. K., Kulkarni, T. D., Perov, Y. N., and Tenenbaum, J. (2013). Approximate bayesian image interpretation using generative probabilistic graphics programs. *Advances in Neural Information Processing Systems*, 26:1520–1528.

Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. *AI Memos AIM-357, MIT*.

Martinez-Garcia, M., Bertalmío, M., and Malo, J. (2019). In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in neuroscience*, 13:8.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8).

Messina, N., Amato, G., Carrara, F., Gennaro, C., and Falchi, F. (2021). Recurrent vision transformer for solving visual reasoning problems. *arXiv preprint arXiv:2111.14576*.

Minsky, M. L. and Papert, S. A. (1969). Perceptrons.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, Rutgers University.

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., and Bowers, J. S. (2021). The role of disentanglement in generalisation. *International Conference on Learning Representations (ICLR)*.

Movshon, J. A. and Simoncelli, E. P. (2014). Representation of naturalistic image structure in the primate visual cortex. In *Cold Spring Harbor symposia on quantitative biology*, volume 79, pages 115–122. Cold Spring Harbor Laboratory Press.

Murphy, K., Torralba, A., Eaton, D., and Freeman, W. (2006). Object detection and localization using local and global features. In *Toward category-level object recognition*, pages 382–400. Springer.

Nakayama, K. (1990). The iconic bottleneck and the tenuous link between early visual processing and perception. *Vision: Coding and efficiency*, 411422.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, 31.

Nilsson, D.-E. (2021). The diversity of eyes and vision. *Annual Review of Vision Science*, 7.

Nonaka, S., Majima, K., Aoki, S. C., and Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, page 103013.

Olshausen, B. A. and Field, D. J. (2005). How close are we to understanding v1? *Neural computation*, 17(8):1665–1699.

Olson, R. K. and Attneave, F. (1970). What variables produce similarity grouping? *The American Journal of Psychology*, pages 1–21.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., and Jilk, D. J. (2013). Recurrent processing during object recognition. *Frontiers in psychology*, 4:124.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., and Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature neuroscience*, 4(7):739–744.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., and Tran, D. (2018). Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR.

Petrou, M. M. and Kamata, S.-i. (2006). *Image processing: dealing with texture*. John Wiley & Sons.

Poggio, T. (1981). Marr's computational approach to vision. *Trends in neurosciences*, 4:258–262.

Portilla, J. and Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40(1):49–70.

Puebla, G. and Bowers, J. (2021). Can deep convolutional neural networks learn same-different relations? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025.

Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pages 2940–2949. PMLR.

Roig, G., Volokitin, A., and Poggio, T. (2018). Do deep neural networks suffer from crowding? *Journal of Vision*, 18(10):902–902.

Romanes, G. J. (1883). *Animal intelligence*. D. Appleton.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rosenfeld, A., Solbach, M. D., and Tsotsos, J. K. (2018). Totally looks like-how humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1961–1964.

Rosenholtz, R. (2014). Texture perception. *Oxford handbook of perceptual organization*,

167:186.

Rosenholtz, R., Huang, J., and Ehinger, K. A. (2012a). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in psychology*, 3:13.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., and Ilie, L. (2012b). A summary statistic representation in peripheral vision explains visual search. *Journal of vision*, 12(4):14–14.

Rosenthal, R. and Fode, K. L. (1961). The problem of experimenter outcome-bias. *Series research in social psychology. Washington, DC: National Institute of Social and Behavioral Science*.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Rust, N. C. and Movshon, J. A. (2005). In praise of artifice. *Nature neuroscience*, 8(12):1647–1650.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Sauer, A. and Geiger, A. (2020). Counterfactual generative networks. *International Conference on Learning Representations*.

Schott, L., von Kügelgen, J., Träuble, F., Gehler, P., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., and Brendel, W. (2021). Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2020). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

Schödl, A., Szeliski, R., Salesin, D. H., and Essa, I. (2000). Video Textures. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 489–498, New York, NY, USA. ACM Press/Addison-Wesley

Publishing Co.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429.

Spoerer, C. J., Kietzmann, T. C., and Kriegeskorte, N. (2019). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. *BioRxiv*, page 677237.

Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551.

Stabinger, S., Rodríguez-Sánchez, A., and Piater, J. (2016). 25 years of cnns: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks*, pages 380–387, Cham. Springer, Springer International Publishing.

Storrs, K. R., Anderson, B. L., and Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, pages 1–16.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*.

Szummer, M. and Picard, R. W. (1996). Temporal texture modeling. In *Image Processing, 1996. Proceedings., International Conference on*, volume 3, pages 823–826. IEEE.

Tacchetti, A., Isik, L., and Poggio, T. (2017). Invariant recognition drives neural representations of action sequences. *PLoS computational biology*, 13(12):e1005859.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840.

Tesfaldet, M., Brubaker, M. A., and Derpanis, K. G. (2018). Two-stream convolutional

networks for dynamic texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6703–6712.

Teufel, C., Dakin, S. C., and Fletcher, P. C. (2018). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific reports*, 8(1):1–12.

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE.

Titchener, E. B. (1929). Systematic psychology: Prolegomena.

Tomasello, M. and Call, J. (2008). Assessing the validity of ape-human comparisons: A reply to boesch (2007).

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2010). Using the forest to see the trees: exploiting context for visual object detection and localization. *Communications of the ACM*, 53(3):107–114.

Torralba, A. and Oliva, A. (2003). Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., and Bauer, S. (2021). On disentangled representations learned from correlated data. In *International Conference on Machine Learning*, pages 10401–10412. PMLR.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. (2020). From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR.

Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. L. (2021). Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*.

Ullman, S., Assif, L., Fetaya, E., and Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10):2744–2749.

Vaishnav, M., Cadene, R., Alamia, A., Linsley, D., Vanrullen, R., and Serre, T. (2021). Understanding the computational demands underlying visual reasoning. *arXiv preprint arXiv:2108.03603*.

van Bergen, R. S. and Kriegeskorte, N. (2020). Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193.

VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology*, 3(1-2):167.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vickery, T. J., Shim, W. M., Chakravarthi, R., Jiang, Y. V., and Luedeman, R. (2009). Supercrowding: Weakly masking a target expands the range of crowding. *Journal of Vision*, 9(2):12–12.

Walton, D. R., Anjos, R. K. D., Friston, S., Swapp, D., Akşit, K., Steed, A., and Ritschel, T. (2021). Beyond blur: real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–14.

Wang, Y. and Zhu, S.-C. (2002). A generative method for textured motion: Analysis and synthesis. In *European Conference on Computer Vision*, pages 583–598. Springer.

Wei, L.-Y. and Levoy, M. (2000). Fast texture synthesis using tree-structured vector quantization. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 479–488. ACM Press/Addison-Wesley Publishing Co.

Wichmann, F. A., Drewes, J., Rosas, P., and Gegenfurtner, K. R. (2010). Animal detection in natural scenes: critical features revisited. *Journal of Vision*, 10(4):6–6.

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., and Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, 2017(14):36–45.

Wyszecki, G. and Stiles, W. S. (1982). *Color science*, volume 8. Wiley New York.

Xie, J., Zhu, S.-C., and Nian Wu, Y. (2017). Synthesizing dynamic patterns by spatial-

temporal generative convnet. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 7093–7101.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624.

Yan, Z. and Zhou, X. S. (2017). How intelligent are convolutional neural networks? *arXiv preprint arXiv:1709.06126*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Zhang, K., Wang, B., Chen, H.-S., Lei, X., Wang, Y., and Kuo, C.-C. J. (2021). Dynamic texture synthesis by incorporating long-range spatial and temporal correlations. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE.

Zhang, X., Watkins, Y., and Kenyon, G. T. (2018). Can deep learning learn the principle of closed contour detection? In Bebis, G., Boyle, R., Parvin, B., Koracin, D., Turek, M., Ramalingam, S., Xu, K., Lin, S., Alsallakh, B., Yang, J., Cuervo, E., and Ventura, J., editors, *Advances in Visual Computing*, pages 455–460, Cham. Springer International Publishing.

Zhang, X., Wu, X., and Du, J. (2019). Challenge of spatial cognition for deep learning. *arXiv preprint arXiv:1908.04396*.

Zhao, H., Des Combes, R. T., Zhang, K., and Gordon, G. (2019). On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR.

Ziemba, C. M. and Simoncelli, E. P. (2021). Opposing effects of selectivity and invariance in peripheral vision. *Nature communications*, 12(1):1–11.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR.

Appendix

The appendix contains the publications included in this cumulative dissertation. For each of these collaborative projects, the authors' contributions are listed. Whenever available, the contribution statements are reprinted from the corresponding publication.

P1: Synthesising dynamic textures using convolutional neural networks

Christina M. Funke*, Leon A. Gatys*, Alexander S. Ecker, Matthias Bethge

* joint first authors

Technical report on *arXiv:1702.07006*.

Contributions CMF, LAG, ASE, and MB designed the study. CMF implemented the dynamic texture model and performed the experiments with help of LAG. The manuscript was jointly written by CMF and LAG with input from ASE and MB.

Synthesising Dynamic Textures using Convolutional Neural Networks

Christina M. Funke*[†, 1, 2, 3], Leon A. Gatys[†, 1, 2, 4], Alexander S. Ecker[1, 2, 5] and Matthias Bethge[1, 2, 3, 6]

[†]contributed equally
[1]Werner Reichardt Center for Integrative Neuroscience, University of Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, Tübingen
[3]Max Planck Institute for Biological Cybernetics, Tübingen
[4]Graduate School of Neural Information Processing, University of Tübingen, Germany
[5]Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
[6]Institute for Theoretical Physics, University of Tübingen, Germany

Abstract

Here we present a parametric model for dynamic textures. The model is based on spatiotemporal summary statistics computed from the feature representations of a Convolutional Neural Network (CNN) trained on object recognition. We demonstrate how the model can be used to synthesise new samples of dynamic textures and to predict motion in simple movies.

1 Introduction

Dynamic or video textures are movies that are stationary both in space and time. Common examples are movies of flame patterns in a fire or waves in the ocean. There exists a long history in synthesising dynamic textures (e.g. [1, 2, 3, 4, 5, 6, 7]) and recently spatio-temporal Convolutional Neural Networks (CNNs) were proposed to generate samples of dynamic textures [8]. In this note we introduce a much simpler approach based on feature spaces of a CNN trained on object recognition [9]. We demonstrate that our model leads to comparable synthesis results without the need to train a separate network for every input texture.

*Corresponding Author: christina.funke@bethgelab.org

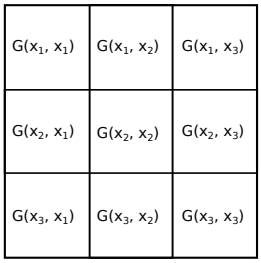


Figure 1: Illustration of the components of the Gram matrix for Δt=3. On the diagonal blocks are the Gram matrix of the frames, which are identical to the ones of the static texture model from [10]. The other blocks contain the correlations between the adjacent frames.

2 Dynamic texture model

Our model directly extends the static CNN texture model of Gatys et al. [10]. In order to model a dynamic texture, we compute a set of spatio-temporal summary statistics from a given example movie of that texture. While the static texture model from [10] only captures spatial summary statistics of a single image, our model additionally includes temporal correlations over several video frames.

We start with a given example video texture $\mathbf{X}$ consisting of T frames $\mathbf{x}_t$, for $t \in \{1, 2, ...T\}$. For each frame we compute the feature maps $\mathbf{F}_\ell(\mathbf{x}_t)$ in layer ℓ of a pre-trained CNN. Each column of $\mathbf{F}_\ell(\mathbf{x}_t)$ is a vectorised feature map and thus $\mathbf{F}_\ell(\mathbf{x}_t) \in \mathcal{R}^{M_\ell(\mathbf{x}_t) \times N_\ell}$ where N_ℓ is the number of feature maps in layer ℓ and $M_\ell(\mathbf{x}_t) = H_\ell(\mathbf{x}_t) \times W_\ell(\mathbf{x}_t)$ is the product of height and width of each feature map.

In the static texture model from [10], a texture is described by a set of Gram Matrices computed from the feature responses of the layers included in the texture model. A Gram Matrix from the feature maps in layer ℓ in response to image $\mathbf{x}$ is defined as $\mathbf{G}_\ell(\mathbf{x}) = \frac{1}{M_\ell(\mathbf{x})} \mathbf{F}_\ell(\mathbf{x})^\top \mathbf{F}_\ell(\mathbf{x})$.

To include temporal dependencies in our dynamic texture model we combine the feature maps of Δt consecutive frames and compute one large Gram Matrix from them (Fig.1). We first concatenate the feature maps from the Δt frames along the second axis: $\mathbf{F}_{\ell,\Delta t}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{\Delta t}) = [\mathbf{F}_\ell(\mathbf{x}_1), \mathbf{F}_\ell(\mathbf{x}_2), ..., \mathbf{F}_\ell(\mathbf{x}_{\Delta t})]$ such that $\mathbf{F}_{\ell,\Delta t} \in \mathcal{R}^{M_\ell \times \Delta t N_\ell}$. Then we use this large feature matrix to compute a Gram Matrix $\mathbf{G}_{\ell,\Delta t} = \frac{1}{M_\ell} \mathbf{F}_{\ell,\Delta t}^\top \mathbf{F}_{\ell,\Delta t}$ that now also captures temporal dependencies of the order Δt (Fig.1). Finally this Gram Matrix is averaged over all time windows Δt_i for $i \in [1, T-(\Delta t-1)]$. Thus our model describes a dynamic texture by the spatio-temporal summary statistics

$$\mathbf{G}_{\ell,\Delta t}(\mathbf{X}) = \frac{1}{M_\ell} \sum_{i=1}^{T-(\Delta t-1)} \mathbf{F}_{\ell,\Delta t_i}^\top \mathbf{F}_{\ell,\Delta t_i} \tag{1}$$

computed at all layers ℓ included in the model. Compared to the static texture model [10] this increases the number of parameters by a factor of Δt^2.

3 Texture generation

After extracting the spatio-temporal summary statistics from an example movie they can be used to generate new samples of the video texture. To that end we sequentially generate frames that match the extracted summary statistics. Each frame is generated by a gradient based pre-image search that starts from a white noise image to find an image that matches the texture statistics of the original video.

Thus, to synthesise a frame $\hat{\mathbf{x}}_t$ given the previous frames $[\hat{\mathbf{x}}_{t-\Delta t+1}, ..., \hat{\mathbf{x}}_{t-1}]$ we minimise the following loss function with respect to $\hat{\mathbf{x}}_t$:

$$\mathcal{L} = \sum_{\ell} w_\ell E_\ell(\hat{\mathbf{x}}_t) \tag{2}$$

$$E_\ell(\hat{\mathbf{x}}_t) = \frac{1}{4N_\ell^2} \sum_{ij} \left(\mathbf{G}_{\ell,\Delta t+1}(\hat{\mathbf{x}}_{t-\Delta t}, ..., \hat{\mathbf{x}}_t) - \mathbf{G}_{\ell,\Delta t}(\mathbf{X})\right)^2_{ij} \tag{3}$$

For all results presented here we included the layers 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1' and 'conv5_1' of the VGG-19 network [9] in the texture model and weighted them equally ($w_l = w$).

The initial Δt -1 frames can be taken from the example movie, which allows the direct extrapolation of an existing video. Alternatively they can be generated jointly by starting with Δt randomly initialised frames and minimising $\mathcal{L}$ jointly with respect to $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, ..., \hat{\mathbf{x}}_{\Delta t}$.

In general this procedure can generate movies of arbitrary length because the extracted spatio-temporal summary statistics naturally do not depend on the length of the source video.

4 Experiments and Results

Here we present dynamic textures generated by our model. We used example video textures from the DynTex database [11] and from the Internet. Each frame was generated by minimising the loss function for 500 iterations of the L-BFGS algorighm [12]. All source textures and generated results can be found at `https://bethgelab.org/media/uploads/dynamic_textures/`.

First we show the results for $\Delta t = 2$ and random initialisation of the initial frames (Fig. 2). We extracted the texture parameters from either $T = 42$ frames of the source movie or just from a pair of frames $T = 2$. Surprisingly we find that extracting the texture parameters from only two frames is often sufficient to generate diverse dynamic textures of arbitrary length (Fig. 2, bottom rows). However, the entropy of the generated frames is clearly higher for $T = 42$ and

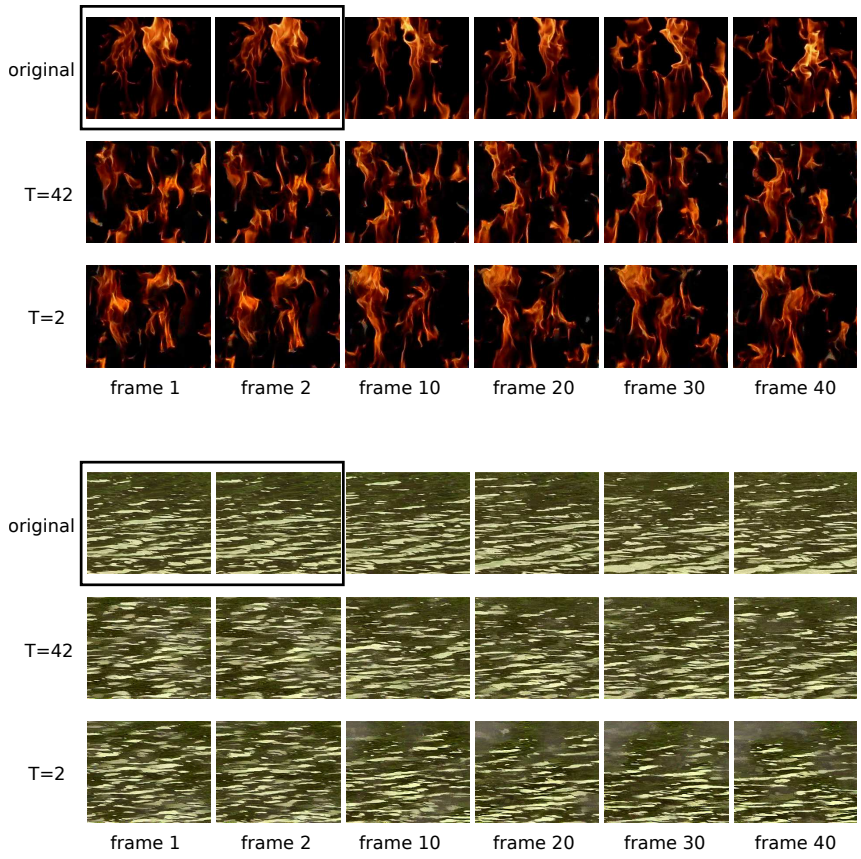


Figure 2: Examples of generated video textures for $\Delta t = 2$ and two example textures. In the top rows frames of the original video are shown. For the frames in the middle rows, 42 original frames were used. For the frames in the bottom rows two original frames were used (the ones in the black box). The full videos can be found at `https://bethgelab.org/media/uploads/dynamic_textures/figure2/`.

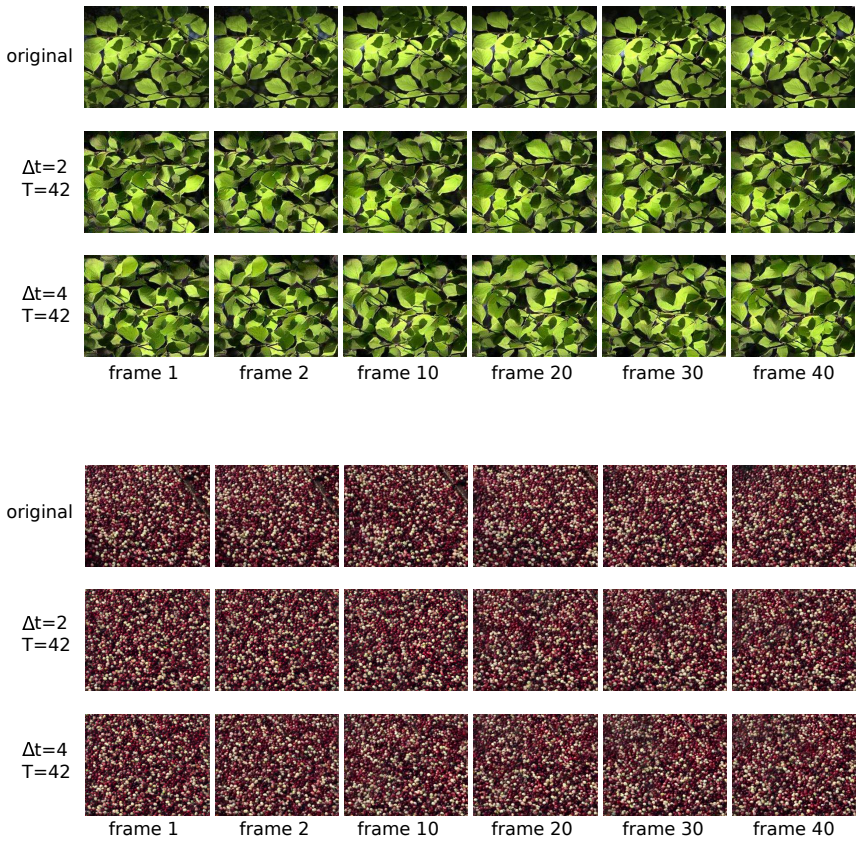


Figure 3: Examples of generated videos for $\Delta t = 2$ (middle rows) and $\Delta t = 4$ (bottom rows). In the top rows frames of the original video are shown. 42 original frames were used. The global structure of the motion is not preserved. The full videos can be found at `https://bethgelab.org/media/uploads/dynamic_textures/figure3/`

for some videos (example: water) greyish regions appear in the generated texture if only two original frames are used.

Next we explored the effect of increasing the size of the time window Δt . Here we show results for $\Delta t = 2$ and $\Delta t = 4$. In general we noted that for most video textures varying the size of the time window Δt has little effect. We observed differences, however, in cases where the motion is more structured. For example, given a movie of a branch of a tree moving in the wind (Fig. 3, top row), the leaves are only moving slightly up and down for $\Delta t = 2$ (Fig. 3, middle row), whereas for $\Delta t = 4$ the motion extends over a larger range (Fig. 3, bottom row).

Still, even for $\Delta t = 4$, the generated video fails to capture the motion of the original texture. In particular, it fails to reproduce the global coherence of the motion in the source video. While in the source video, all leaves move together with the branch up or down, in the synthesised one some leaves move up while some move down at the same time. The disability to capture the global structure of the motion is even more apparent in the second example in Fig. 3 and illustrates a limitation of our model.

Finally, instead of generating a video texture from a random initialisation, we can also initialise with $\Delta t - 1$ frames from the example movie. In that way the spatial arrangement is kept and we are predicting the next frames of the movie based on the initial motion. We use three frames of the original video were to define the texture statistics ($\Delta t = 3$, $T = 3$) (Fig. 4). The first two frames of the new movie are taken from the example and the following frames were sequentially generated as described in section 3. In the resulting video the different elements keep moving in the same direction: The squirrel continues flying to the top left, while the plants move upwards. If an element disappears from the image, it reappears somewhere else in the image. The generated movie can be arbitrarily long. In this case we used only the initial 3 frames to generate over 600 frames of a squirrel flying through the image and did not observe a decrease in image quality.

5 Discussion

We introduced a parametric model for dynamic textures based on the feature representations of a CNN trained on object recognition [9]. In contrast to the CNN-based dynamic texture model by Xie et al. [7], our model can capture a large range of dynamic textures without the need to re-train the network for every given input texture.

Surprisingly we find that even when the temporal dependencies are extracted from as little as two adjacent frames our model still produces diverse looking dynamic textures (Fig. 2). This is also true for non-texture movies with simple motion. We see that in this case we can generate a theoretically infinite movie repeating the same motion (Fig. 4.).

However, our model fails to capture structured motion with more complex

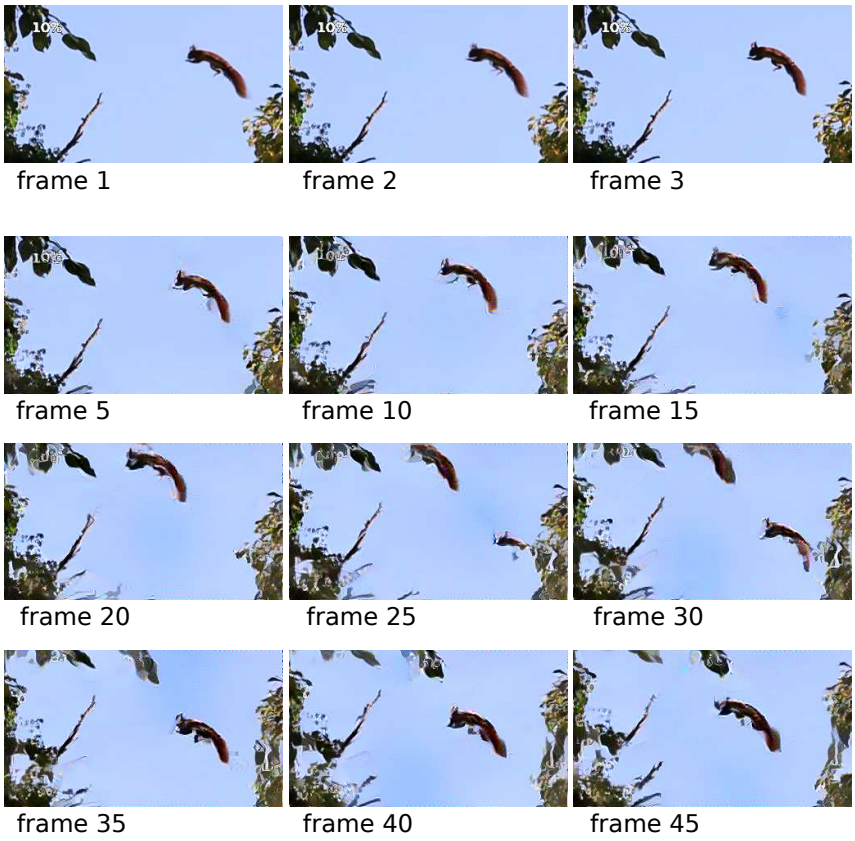


Figure 4: Initialisation of the new video with the original frames. The first three frames shown are the original frames, the others are generated by our model. The full video can be found at `https://bethgelab.org/media/uploads/dynamic_textures/figure4/`.

temporal dependencies (Fig. 3). Possibly spatio-temporal CNN features or the inclusion of optical flow measures [13] might help to model temporal dependencies of that kind.

In general though we find that for many dynamic textures the temporal statistics can be captured by second order dependencies between complex spatial features leading to a simple yet powerful parametric model for dynamic textures.

References

[1] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference On*, vol. 2, pp. 439–446, IEEE, 2003.

[2] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," in *ACM Transactions on Graphics (ToG)*, vol. 22, pp. 277–286, ACM, 2003.

[3] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa, "Video Textures," in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, (New York, NY, USA), pp. 489–498, ACM Press/Addison-Wesley Publishing Co., 2000.

[4] M. Szummer and R. W. Picard, "Temporal texture modeling," in *Image Processing, 1996. Proceedings., International Conference on*, vol. 3, pp. 823–826, IEEE, 1996.

[5] Y. Wang and S.-C. Zhu, "A generative method for textured motion: Analysis and synthesis," in *European Conference on Computer Vision*, pp. 583–598, Springer, 2002.

[6] L.-Y. Wei and M. Levoy, "Fast texture synthesis using tree-structured vector quantization," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 479–488, ACM Press/Addison-Wesley Publishing Co., 2000.

[7] J. Xie, S.-C. Zhu, and Y. N. Wu, "Synthesizing Dynamic Textures and Sounds by Spatial-Temporal Generative ConvNet," *arXiv preprint arXiv:1606.00972*, 2016.

[8] Z. Zhu, X. You, S. Yu, J. Zou, and H. Zhao, "Dynamic texture modeling and synthesis using multi-kernel Gaussian process dynamic model," *Signal Processing*, vol. 124, pp. 63–71, July 2016.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Sept. 2014. arXiv: 1409.1556.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," *Advances in Neural Information Processing Systems 28*, 2015.

[11] R. Péteri, S. Fazekas, and M. J. Huiskes, "DynTex: A comprehensive database of dynamic textures," *Pattern Recognition Letters*, vol. 31, pp. 1627–1632, Sept. 2010.

[12] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.

[13] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *German Conference on Pattern Recognition*, pp. 26–36, Springer, 2016.

P2: A parametric texture model based on deep convolutional features closely matches texture appearance for humans

Thomas S.A. Wallis, Christina M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, Matthias Bethge

Published in *Journal of Vision,17(12), 5–5*.

Contributions Designed the experiments: TSAW, ASE, CMF, LAG, FAW, MB. Programmed the experiments: TSAW. Collected the data: CMF, TSAW. Analyzed the data: TSAW, ASE, CMF. Wrote the paper: TSAW. Revised the paper: CMF, ASE, LAG, FAW, MB.

A parametric texture model based on deep convolutional features closely matches texture appearance for humans

Thomas S. A. Wallis
Werner Reichardt Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen, and the Bernstein Center for Computational Neuroscience, Tübingen, Germany

Christina M. Funke
Werner Reichardt Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen, and the Bernstein Center for Computational Neuroscience, Tübingen, Germany

Alexander S. Ecker
Werner Reichardt Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen, and Bernstein Center for Computational Neuroscience, Tübingen, Germany, and Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA

Leon A. Gatys
Werner Reichardt Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen and the Bernstein Center for Computational Neuroscience, Tübingen, Germany

Felix A. Wichmann
Neural Information Processing Group, Faculty of Science, Eberhard Karls Universität Tübingen, Bernstein Center for Computational Neuroscience, and the Max Planck Institute for Intelligent Systems, Empirical Inference Department, Tübingen, Germany

Matthias Bethge
Werner Reichardt Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen, Bernstein Center for Computational Neuroscience, Institute for Theoretical Physics, Eberhard Karls Universität Tübingen, and the Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Our visual environment is full of texture—"stuff" like cloth, bark, or gravel as distinct from "things" like dresses, trees, or paths—and humans are adept at perceiving subtle variations in material properties. To investigate image features important for texture perception, we psychophysically compare a recent parametric model of texture appearance (convolutional neural network [CNN] model) that uses the features encoded by a deep CNN (VGG-19) with two other models: the venerable Portilla and Simoncelli model and an extension of the CNN model in which the power spectrum is additionally matched. Observers discriminated model-generated textures from original natural textures in a spatial three-alternative oddity paradigm under two viewing conditions: when test patches were briefly presented to the near-periphery

Citation: Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, *17*(12):5, 1–29, doi: 10.1167/17.12.5.

doi: 10.1167/17.12.5 Received March 24, 2017; published October 5, 2017 ISSN 1534-7362

("parafoveal") and when observers were able to make eye movements to all three patches ("inspection"). Under parafoveal viewing, observers were unable to discriminate 10 of 12 original images from CNN model images, and remarkably, the simpler Portilla and Simoncelli model performed slightly better than the CNN model (11 textures). Under foveal inspection, matching CNN features captured appearance substantially better than the Portilla and Simoncelli model (nine compared to four textures), and including the power spectrum improved appearance matching for two of the three remaining textures. None of the models we test here could produce indiscriminable images for one of the 12 textures under the inspection condition. While deep CNN (VGG-19) features can often be used to synthesize textures that humans cannot discriminate from natural textures, there is currently no uniformly best model for all textures and viewing conditions.

Introduction

Textures are characterized by the repetition of smaller elements, sometimes with variation, to make up a pattern. Significant portions of the visual environment can be thought of as textures ("stuff" as distinct from "things"; Adelson & Bergen, 1991): your neighbor's pink floral wallpaper, the internal structure of dark German bread, the weave of a wicker basket, the gnarled bark of an old tree trunk, a bowl full of prawns ready for the barbie. Texture is an important material property whose perception is of adaptive value (Adelson, 2001; Fleming, 2014). For example, we can readily discriminate wet from dry stones (e.g., Ho, Landy, & Maloney, 2008), separating the underlying spatial texture from potentially temporary characteristics like glossiness. Where surfaces of different textures form occlusion boundaries, texture can provide a powerful segmentation cue; conversely, occlusion borders of similarly textured surfaces can camouflage the occlusion (hiding a tiger among the leaves). Given the importance and ubiquity of visual textures, it is little wonder that they have received much scientific attention, not only from within vision science but also in computer vision, graphics, and art (see Dakin, 2014; Landy, 2013; Pappas, 2013; Rosenholtz, 2014, for comprehensive recent reviews of this field).

Studying texture perception with parametric texture models

Seminal early work on visual texture perception includes that by Gibson (Beck & Gibson, 1955; Gibson, 1950) and by Julesz (Julesz, 1962, 1981; Julesz, Gilbert, & Victor, 1978). Julesz's thinking remains an important influence on approaches to texture perception, in particular the idea that there exists some set of statistics (parameters in a parametric model) that are both necessary and sufficient for matching the appearance of textures (see also Portilla & Simoncelli, 2000). For computer vision applications, where a goal might be to match the appearance of some region of texture to facilitate image compression, the most effective approaches can be nonparametric—for example, by *quilting* repetitions of a base level crop over the area of the texture (e.g., Efros & Freeman, 2001). However, nonparametric approaches have little to teach us about the human visual system because they make no explicit hypotheses about what features are represented. In this paper we will therefore focus on parametric texture models.

Parametric models that aim to match the appearance of natural textures are typically assessed by examining artificial textures synthesized by the model (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000; Safranek & Johnston, 1989; Safranek, Johnston, & Rosenholtz, 1990; Zhu, Wu, & Mumford, 1998). The statistics of a model are first computed on a target image, then a new image is synthesized to approximately match the statistics of the target image (often via gradient descent). This approach carries forward Julesz's "necessary and sufficient statistics" idea by assuming that texture appearance can be captured by the coefficients of some specified set of image statistics. Note that this focus on naturalistic appearance is distinct from a complementary approach which starts from local analysis of luminance distributions to posit an "alphabet" of independent microtexture dimensions (Victor, Thengone, & Conte, 2013), but does not seek to match the appearance of natural textures.

A number of parametric texture models operate by assuming a plausible image representation for the early primate visual system, decomposing the target image into some number of frequency and orientation bands (Cano & Minh, 1988; Heeger & Bergen, 1995; Malik & Perona, 1990; Porat & Zeevi, 1989; Portilla & Simoncelli, 2000; Simoncelli & Portilla, 1998; Zhu et al., 1998). The spatially averaged responses in some combination of these bands form the parameters of the model, whose values are then matched by the synthesis procedure. The parametric texture model of Portilla and Simoncelli (Portilla & Simoncelli, 2000; Simoncelli & Portilla, 1998) extended this approach by additionally matching the correlations between channels and other statistics, producing more realistic appearance matches to textures. This model has since had broad impact on the field of human perception and neuroscience: the texture statistic representation may provide a fruitful way to understand the processing in mid-ventral visual areas (Freeman & Simoncelli, 2011; Freeman, Ziemba, Heeger, Simoncelli, & Movshon,

2013; Movshon & Simoncelli, 2014; Okazawa, Tajima, & Komatsu, 2015; Ziemba, Freeman, Movshon, & Simoncelli, 2016), and it has been argued to provide a good approximation of the type of information encoded in the periphery, and thus a model for tasks such as crowding and visual search (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Keshvari & Rosenholtz, 2016; Rosenholtz, 2011; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012)—though other evidence questions the more general adequacy of this representation for explaining crowding and peripheral appearance (Agaoglu & Chung, 2016; Clarke, Herzog, & Francis, 2014; Herzog, Sayim, Chicherov, & Manassi, 2015; Wallis, Bethge, & Wichmann, 2016).

How, though, does it perform as a model of texture appearance in humans? Balas and colleagues (Balas, 2006, 2008, 2012; Balas & Conlin, 2015) have reported a number of psychophysical investigations using the Portilla and Simoncelli (hereafter, PS) texture model that are relevant to this question. Balas (2006) quantified the relative importance of subsets of the PS statistics compared to the full set for matching the appearance of different classes of texture (periodic, structured, or asymmetric). He used a task in which human observers chose the "oddball" image from a set of three (a three-alternative oddity task) that were presented briefly to the near-periphery. Two of the images were drawn from original textures whereas the oddball was drawn from a model synthesis matched to the original texture (or vice versa; the oddball could be either original or synthetic). Importantly, all three images were physically different from each other (consisting of subcrops of larger images). The oddity judgment therefore concerns the subjective dissimilarity of the images—which image is "produced by a different process"—rather than exact appearance matching. In this study, the importance of different parameter subsets depended on the class of texture, and including the full set of statistics brought average discrimination performance quite close to chance (around 40% correct on average), showing that the PS statistics do a reasonably good job in capturing texture appearance under brief peripheral viewing conditions.

Balas (2012) used a four-alternative oddity task to investigate the discriminability of real and synthetic textures. Observers were allowed to view each stimulus array for unlimited time and to foveate the images. Under these viewing conditions, observers could easily discriminate original natural textures and PS-synthesized images from each other, whether the oddball was real or synthetic (average performance 85%–90%). However, when the original images were sourced from abstract artworks rather than photographs of fruits and vegetables, performance for discriminating real from PS-synthesized images was worse and depended on whether the oddball was real or synthesized (with performance around 55% for the former and 65% for the latter). Together with the results of Balas (2006), these results suggest that the PS model better captures texture appearance in the periphery than in the fovea, and that the perceptual fidelity of the matching depends on the image or texture type.

Finally, Balas and Conlin (2015) assessed whether the influence of illumination change on human texture perception could be captured by PS synthesis. Observers performed a match-to-sample task, in which they decided which of two match images depicted the same texture as a previously presented sample. Performance was quite high (above 90%) when the illumination between the sample and correct match image was constant (in this case, the match image was physically identical to the sample), whether the images were real or synthesized. When the correct match image was presented with different illumination to the sample, performance declined to around 70% correct for synthetic images but remained high for real images. That is, observers could easily ignore illumination changes when matching real textures, but their judgments were impaired by illumination change when discriminating synthesized images. Note that the foil images (the nontarget match image) were selected to be "approximately visually matched" by the experimenters; it is likely that the results (but perhaps not conclusions) will depend on this choice. Similar results were obtained after equalizing the luminance and power spectra of the images, and when match and sample images were physically different (cropped from different areas of the same texture). These results show that the PS feature space does not perfectly preserve the necessary statistics to match texture appearance across changes in illumination.

Together, the experiments show that while aspects of human texture perception are not captured by or fall outside the scope of the PS feature space, it does succeed in capturing key aspects of texture appearance for many classes of natural texture. The PS feature space is based on the idea—amply supported by psychophysical and neurophysiological evidence—that the human visual system decomposes an image into a number of spatial and orientation subbands. To what extent will a more complex feature space improve on the PS model?

A new parametric texture model based on deep features

Gatys, Ecker, and Bethge (2015) recently introduced a new parametric texture model that produces subjectively high-quality matches to texture appearance, and whose features can be used to separate the "style" of an

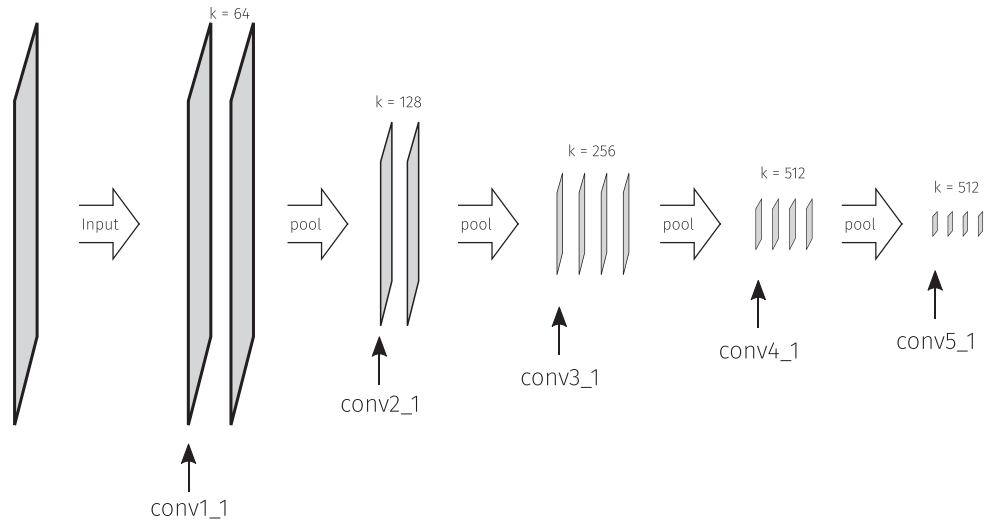


Figure 1. The architecture of the VGG-19 convolutional neural network (Simonyan & Zisserman, 2015), whose pretrained features are used by the Gatys, Ecker, and Bethge (2015) texture model. The network consists of stacks of convolutional stages followed by max pooling. In higher network layers, the feature map sizes decrease (depicted as the increasingly small panels), the corresponding "receptive field" sizes of the units increase, and the number of feature maps (k) increase. In this article we synthesize textures using the first convolutional layer from each stack after the max pooling.

image from its content (Gatys, Ecker, & Bethge, 2016). This texture synthesis procedure (see "CNN texture model" section) is based on the pretrained features of a deep convolutional neural network (the VGG-19; Simonyan & Zisserman, 2015; Figure 1) that achieves near state-of-the-art performance on the Imagenet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015): basically, returning labels for the likely objects present in an image. Due to their success on benchmarks like the Imagenet Large Scale Visual Recognition Challenge, CNNs have become the dominant approach to many visual inference problems in the field of computer vision, with some networks showing impressive transfer learning performance (doing well on new tasks with only minimal changes to the network; e.g., Donahue, Jia, & Vinyals, 2013).

Briefly, a single-layer convolutional neural network (CNN) learns (via supervised training) the weights of filters that are convolved with input images, creating a spatial feature map of activations, similar to a traditional bank of Gabor filters familiar to vision scientists. Using convolutional filters allows the detection of spatial patterns at any position in the image (translation equivariance), and also facilitates learning through weight sharing—the intuition here is that features useful to know about at one spatial location are likely to be useful for all spatial locations. All convolutional layer activations are then passed through a pointwise nonlinearity, typically a rectified linear ("relu") function $f(x) = \max(0, x)$. These feature maps can then be pooled (in VGG by taking the maximum of activations in a small area), creating local spatial invariance, and combined with downsampling to reduce the spatial dimensions of the feature maps (see Figure 1). Stacking such operations repeatedly (passing the outputs of one convolutional or max-pool layer as the input to another, creating a "deep" CNN with at least one hidden layer) has several effects. The spatial area of the input image to which features respond are larger for higher layers (analogous to the increase in receptive field size from V1 to IT cortex), and the features to which higher convolutional layers respond becoming increasingly nonlinear functions of the input pixels (analogous to the feature selectivity from V1 to IT cortex). It is this accumulating nonlinear behavior that allows complex properties such as object identity (and many other properties; Hong, Yamins, Majaj, & DiCarlo, 2016) to be linearly decoded from the higher network layers. For more comprehensive recent reviews, see Kietzmann, McClure, and Kriegeskorte (2017); LeCun, Bengio, and Hinton (2015); and Yamins and DiCarlo (2016).

CNNs are interesting for the study of human vision first and foremost because they perform interesting tasks. Until recently, there was only one known class of system ("biological brains") that could detect and recognize objects in photographic images with high accuracy; now there are two. The second reason that human vision researchers might be curious about CNNs is that there is growing evidence that the *way* in which CNNs perform these tasks has intriguing similarities to some biological visual systems. For example, there is now quantitative evidence that performance-optimized CNN features predict ventral

stream brain signals in monkeys and humans using the stimulus input better than existing models built explicitly for that purpose (Cadieu et al., 2014; Cichy, Khosla, Pantazis, & Oliva, 2016; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Guclu & van Gerven, 2015; Hong et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins, Hong, Cadieu, & DiCarlo, 2013; Yamins et al., 2014). CNN models also show similarities to human psychophysical object recognition performance under brief presentation conditions (Hong et al., 2016; Yamins et al., 2014). A recent paper reported that CNNs trained on ImageNet (natural photos) can still partially recognize objects from silhouette information only, and show other human-similar shape biases (Kubilius, Bracci, & Op de Beeck, 2016). There are of course important ways that current CNNs are *un*like primate visual systems. For example, a subtle modification of an image that is nearly imperceptible to a human can cause a deep network to misclassify an object with high confidence (Szegedy et al., 2013; see Yamins & DiCarlo, 2016, for additional discussion). Furthermore, human object recognition remains remarkably robust in images degraded by white noise, whereas the original VGG network is strongly impaired (Geirhos et al., 2017). Bearing these caveats in mind, an exciting possibility is that the study of CNNs may help to elucidate some fundamental mechanisms of human perception.

In this article we pursue a less lofty goal: to measure how well humans can discriminate textures synthesised by the Gatys et al. (2015) model from natural textures. How well do CNN texture features match the appearance of the original textures? To address this question we compare the model of Gatys et al. (2015) to the PS model (Portilla & Simoncelli, 2000) and to a recent modification of the Gatys model (Liu, Goussau, & Xia, 2016). Experimentally, we closely follow the approach of Balas (2006), described above.[1] Using images that are all physically different measures the extent to which model syntheses are *categorically* or *structurally* lossless (in that they could both be considered samples from original images; Pappas, 2013), as opposed to being *perceptually* lossless (unable to be told apart) compared either to each other (Freeman & Simoncelli, 2011) or the original source images (Wallis et al., 2016). Perceptual losslessness could be important for understanding visual encoding in general but categorical losslessness is arguably more useful for understanding the perceptual representation of texture.

In addition to assessing the discriminability of brief, peripherally presented textures (as in Balas, 2006), we are also interested in how this changes when longer foveal comparison is possible (as in Balas, 2012). We therefore include two presentation conditions: a parafoveal condition and an inspection condition.[2] Note that depending on the spatial scale of the most informative differences, sensitivity to some aspects of texture can be *better* in the parafovea than in the fovea under some conditions (Gurnsey, Pearson, & Day, 1996; Kehrer, 1987, 1989). Therefore, differences in psychophysical performance between these conditions are informative about the extent to which the texture models under consideration capture, or fail to capture, features that are important for both foveal and near peripheral texture perception.

General methods

All stimuli, data, and code to reproduce the figures and statistics reported in this article are provided online (raw data and code at http://doi.org/10.5281/zenodo.836726, stimuli at http://doi.org/10.5281/zenodo.438031). This document was prepared using the knitr package (Xie, 2013, 2015) in the R statistical environment (Arnold, 2016; Auguie, 2016; R Core Development Team, 2016; Wickham, 2009, 2011; Wickham & Francois, 2016) to improve its reproducibility.

Apparatus

Stimuli were displayed on a VIEWPixx 3D LCD (VPIXX Technologies, Saint-Bruno-de-Montarville, Quebec, Canada; spatial resolution 1920 × 1080 px, temporal resolution 120 Hz, operating with the scanning backlight turned off in high-bitdepth grayscale mode). Outside the stimulus image the monitor was set to mean gray. Observers viewed the display from 60 cm (maintained via a chinrest) in a darkened chamber. At this distance, pixels subtended approximately 0.024° on average (41 px per degree of visual angle [dva]). The monitor was linearized (maximum luminance 260 cd/m^2) using a Konica-Minolta LS-100 photometer (Konica-Minolta Inc., Tokyo, Japan). Stimulus presentation and data collection was controlled via a desktop computer (Intel Core i5-4460 CPU, AMD Radeon R9 380 GPU) running Ubuntu Linux (16.04 LTS), using the Psychtoolbox Library (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997, version 3.0.12) and our internal iShow library (http://dx.doi.org/10.5281/zenodo.34217) under MATLAB (R2015b; The Mathworks, Inc., Natick, MA).

Source images

Twelve unique texture images[3] (see Figure 2) were selected to provide a variety of texture-like structure (including some with obvious periodicity and others

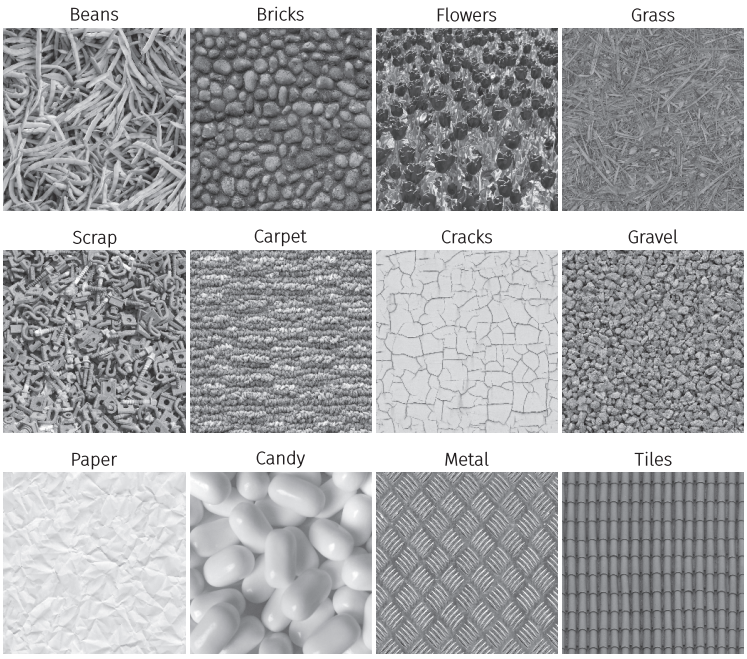


Figure 2. The 12 original texture images used in the experiments. Arranged to correspond to Figure 7. These images are copyrighted by www.textures.com (used with permission).

that were asymmetric) but were also chosen to exhibit some nontexture naturalistic structure (such as the size gradient visible in the flowerbed image). Images were converted to grayscale using scikit-image's io.imread function (van der Walt et al., 2014), then cropped to the largest possible square from the center of the image. The original images all had at least one dimension of 1024 px. We then downsampled all images to 256×256 px using the cubic interpolation of skimage.transform.resize. To preserve the naturalistic appearance of the images we did not standardize the mean or variance of intensities. Since all texture models considered here also match the low-level image statistics, this will not impact our results. For each image model (conv1–conv5 and PS for Experiment 1, conv5, PS and powerspec in Experiment 2; see below) we generated 10 unique synthesised images of size 256 from each original image, resulting in a final stimulus set of 732 images for Experiment 1 and 372 images for Experiment 2. All images were stored as 16-bit .png files.

CNN texture model

The CNN texture model (Gatys et al., 2015) uses the pretrained features of the VGG-19 network (Simonyan & Zisserman, 2015), which shows near state-of-the-art performance on the object recognition ImageNet challenge (Russakovsky et al., 2015). While there are now CNN models that outperform the VGG network on object recognition, the VGG network remains appealing because of its relatively simple architecture (Figure 1), and because it produces more introspectively appealing textures and style transfer than those networks that currently perform better on ImageNet. It consists of two operations, stacked many times: convolutions with k 3×3 filters (where k is the number of input feature maps) followed by a 2×2 max-pooling in nonoverlapping regions. The model uses five pooling and 16 convolutional layers (plus three fully connected layers which we do not use here). The layers are typically labeled with the stack (e.g., "conv1" or "pool1") with an underscore denoting the sublayer. For example, "conv1_1" refers to the first convolutional layer of the network, whereas "conv3_2" would be the second convolutional layer of the third stack (Figure 1). We use a subset of these feature maps for texture synthesis (see below). The code was implemented in Theano using the Lasagne framework, and may be downloaded from https://github.com/leongatys/DeepTextures. The weights of the VGG-19 network are scaled such that the mean activation of each filter over images and positions is equal to 1.

The first step of the texture synthesis algorithm is to pass the original image through the network, generating responses in all network layers. For the feature responses of a subset of layers (described below) the Gram matrices are computed (the Gram matrix is the dot product of the vectorized feature maps; each entry in the resulting matrix is the correlation between two

features in response to a particular input image). The basic idea of the texture synthesis is to create an image with the same Gram matrix representation via gradient descent (the same synthesis principle as in Portilla and Simoncelli (2000) using different features). We start with a white noise image and minimize the mean-squared distance between the entries of the Gram matrices of the original image and the Gram matrix of the image being generated. For the optimization we use the L-BFGS method from the the SciPy package (Jones, Oliphant, & Peterson, 2001) using 1,000 iterations, which was sufficient to bring the loss to an acceptable (but usually nonzero) value. Note that this procedure (using a unique random initialization and converging on nonzero loss) can therefore generate an effectively infinite number of physically unique synthesised images. We discuss the gradient descent further in the Appendix. After gradient descent, the intensity histogram of the resulting image was matched to the intensity histogram of the original image (ensuring that the images have the same global luminance, contrast, skew, and kurtosis).

The network was trained on RGB images and expects three-channel input. We duplicated the grayscale original images into three channels, and to ensure that the outputs of the synthesis remained grayscale, we averaged the gradients of each color channel during optimization. The layers conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 were used for texture synthesis by taking the activations after rectification. For simplicity, we label the texture models used below with the name of the highest convolutional stack used. We match all the Gram matrices cumulatively up to the named layer (e.g., the model we label "conv3" below matches Gram matrices for layers conv1_1, conv2_1 and conv3_1). For each layer l with n_l feature maps, $n_l(n_l + 1) / 2$ parameters are matched (division by two is because the Gram matrices are symmetrical). The approximate number of parameters in each CNN texture model are shown in Figure 5. Outputs were saved as 16-bit .png images. Example syntheses can be seen in Figure 3.

CNN plus power spectrum model

To capture long-range correlations (such as contours that extend over large sections of the image) the model can be extended by additionally matching the power spectrum of the original image when performing the gradient descent to find texture syntheses (Liu et al., 2016). The new loss function is $L = L_{\mathrm{CNN}} + \beta L_{\mathrm{spe}}$ and the new gradient is $\Delta = \Delta_{\mathrm{CNN}} + \beta\Delta_{\mathrm{spe}}$, where L_{CNN} is the loss function and Δ_{CNN} is the gradient from the pure CNN texture model, L_{spe} and Δ_{spe} are related to the distance between the current image and the target Fourier spectrum, and $\beta = 10^5$. That is, the additional constraints are simply added into the loss function and gradient (see Liu et al., 2016, for further details).

To synthesize these stimuli we used code provided by Liu et al. (2016). There are a number of differences between the implementation of the power spectrum model and the base CNN model described above. First, the code is written using Matconvnet instead of Lasagne. The network and the images are normalized to [0, 1] and the stopping criterion of the optimization process is different. In the power spectrum model we used up to 2,000 iterations (as distinct to 1,000 iterations for the base model). The power spectrum model matches different layers of the VGG compared to our CNN model: Conv1_1, Pooling1, Pooling2, Pooling3, and Pooling4. The power spectrum constraint adds 32,768 parameters (half the size of the image because phase is discarded), yielding a total of 209,408 parameters (Figure 5). While we have not run extensive experiments, we argue that the most consequential change between the models for the results we report is the inclusion of the power spectrum matching constraint rather than other implementation differences.

PS texture model

Portilla and Simoncelli (2000) texture images were generated using the publically available MATLAB toolbox (http://www.cns.nyu.edu/lcv/texture/). The texture synth representation we used consisted of four spatial scales and orientations, and a spatial neighborhood of 11 px (these are the most common settings used in the literature where reported (e.g., Balas et al., 2009; Freeman & Simoncelli, 2011). The gradient descent procedure was based on 50 iterations. The PS model matches approximately 1,000 parameters (Figure 5). Outputs were saved as 16-bit .png images.

Procedure and design

On each trial observers were presented with three physically different image patches. Two were from the original image and one from a model synthesis image matched to that original image (or vice versa—two patches could come from the same model synthesis and one patch from the original image). That is, the oddball image could be either original or synthesized with equal probability, so a "pick the natural-looking image" strategy would not succeed. The three image patches (size 128×128 px) were cropped from a larger image (size 256). To obtain two nonoverlapping crops from the same physical image (for the nontarget intervals) one could simply use the image quadrants. To increase the physical variation in the images across trials we

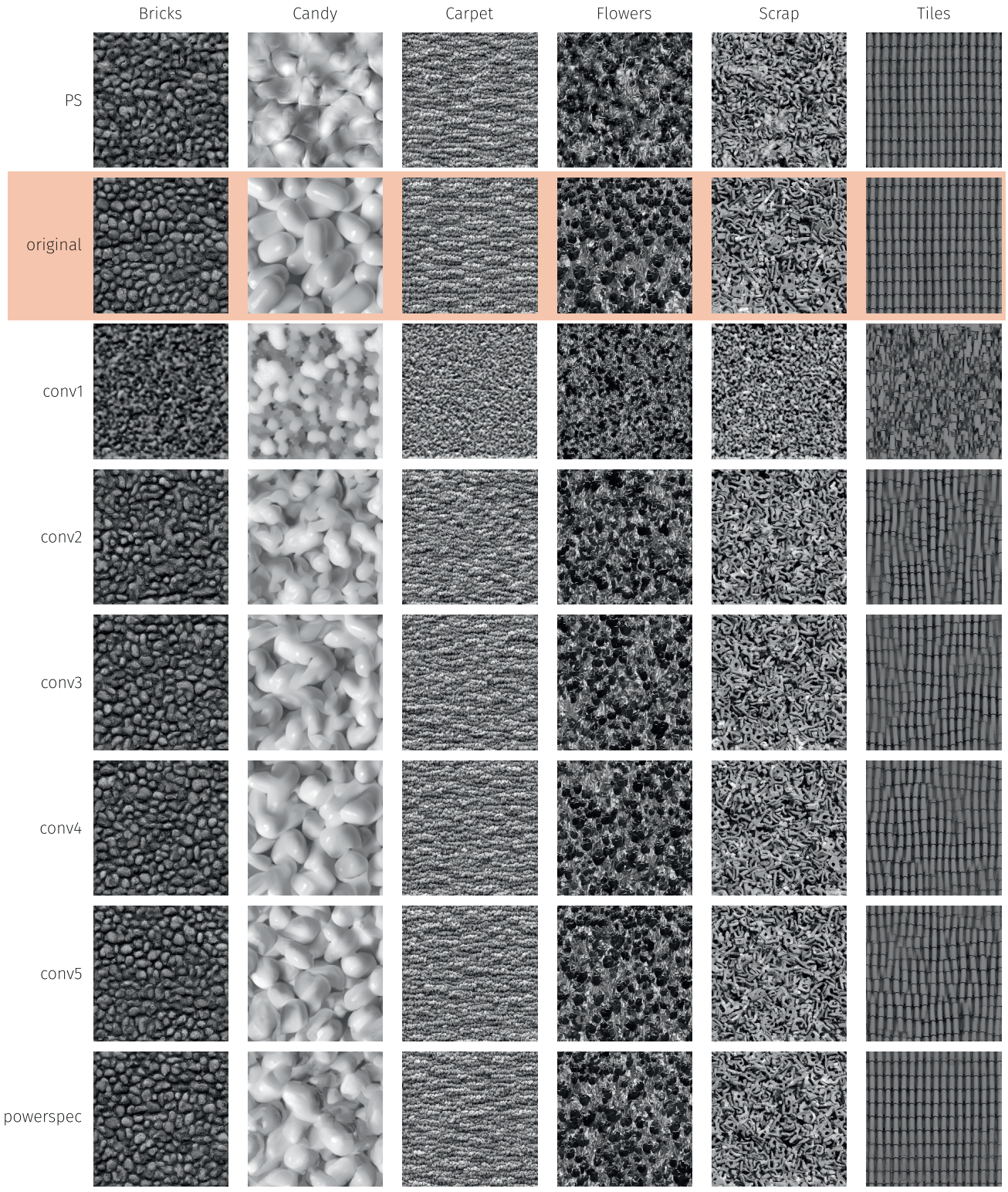


Figure 3. Example experimental stimuli used in Experiment 1 (PS, conv1–conv5) and Experiment 2 (PS, conv5 and powerspec).

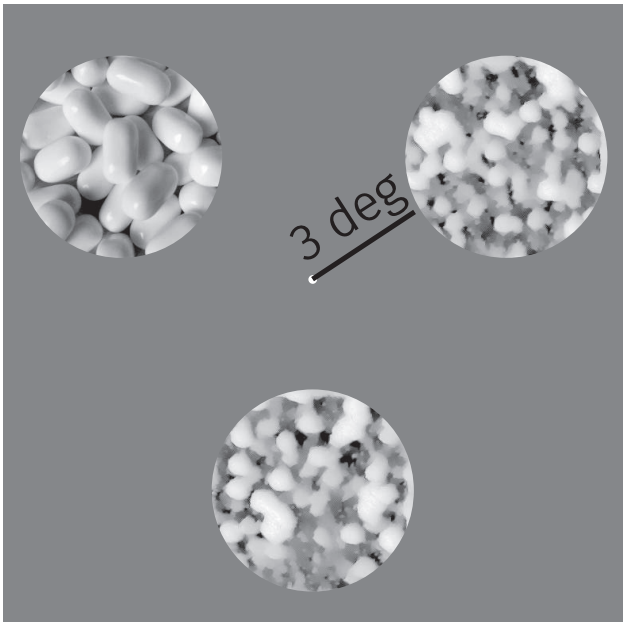


Figure 4. Example experimental display (not to scale). Distance bar not shown in actual experiment.

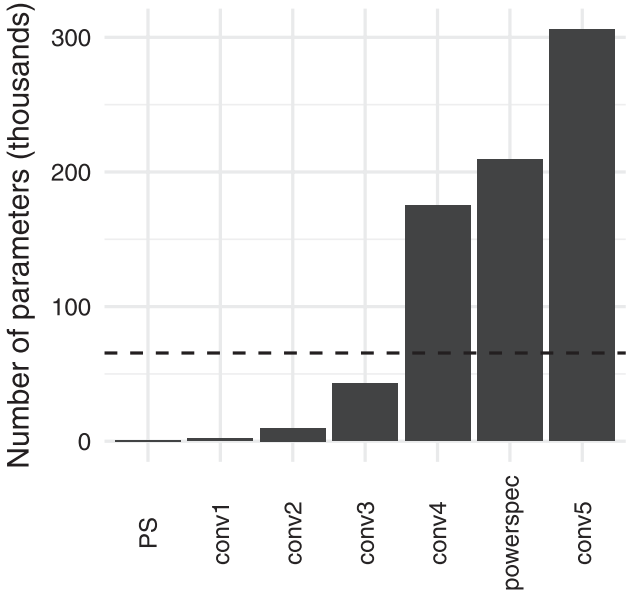


Figure 5. Approximate number of parameters matched by each texture model assessed in the present paper. The dashed line shows the number of pixels in a 256 px^2 image. Models above the line are overcomplete.

instead chose two adjacent crops drawn from non-overlapping but otherwise jittered image sections. On half of the trials the crops were from adjacent "rows" with the vertical dimension randomly sampled, whereas on the other half the crops were from adjacent columns with horizontal dimension randomized. This strategy eliminated the possibility that observers could match specific features of the images within a trial (as in Balas, 2006).

The oddball image could appear at any one of three locations with equal probability (see Figure 4). The observers' task was to report which of three simultaneously presented images was different to the other two, in that it was "generated by a different process" (rather than being physically the same). Observers fixated a spot (best for steady fixation from Thaler, Schütz, Goodale, & Gegenfurtner, 2013) in the center of the screen, and the images were arranged around the fixation in a downward-pointing equilateral triangle configuration. The images were windowed by a circular cosine, ramping the contrast to zero in the space of 6 px. The distance between the fixation point and the nearest edge of the image was 3 dva, and the image patches subtended 3.1 dva.

The stimulus display was presented for either 200 ms, with observers instructed to maintain fixation (the parafoveal condition) or for 2000 ms with observers allowed to make eye movements freely (the inspection condition). Observers then had 1200 ms to respond (responses could also be made while the stimulus remained on the screen). The intertrial interval was 400 ms. To reduce the possibility that observers could learn specific strategies for different images based on familiarity, no trial-to-trial feedback was provided. Instead, a break screen was presented every 72 trials telling the observer their mean performance on the previous trials.

Within a block of trials observers saw five repetitions of the 72 combinations of image model (six levels) and source image (12 levels), for a total of 360 trials per block. Trials were pseudorandomly interleaved throughout a block, with the constraint that trials using the same source image were required to be separated by at least two intervening trials. Presentation condition was blocked to allow observers to anticipate the trial timing and adjust their strategy accordingly.

At the beginning of the experiment, naive observers performed 30 trials with a 2-s presentation time to allow them to become familiar with the task. All observers then performed a practice session of 30 trials at the relevant presentation time for the upcoming block.

Data analysis

We analyzed the data using a logistic generalized linear mixed model, estimated using Bayesian inference. Experimentally manipulated fixed effects of presentation condition and image model were estimated along with random effects for observer and image. The model

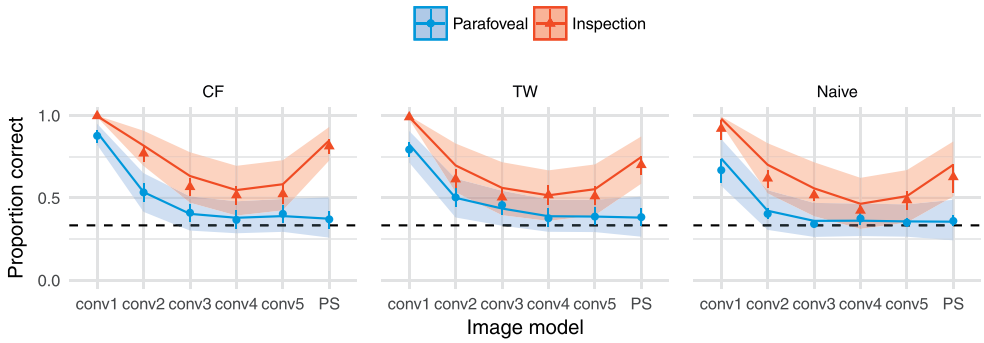


Figure 6. Performance as a function of image model in Experiment 1, averaging over images. For the authors (CF and TW), points show the mean proportion correct and error bars show 95% bootstrapped confidence intervals. Each data point represents 300 trials. Solid lines show mixed-effects model predictions for this observer (mean of posterior regression line), ignoring random effects of image. For naive observers (right panel, N = 10), points show grand mean and 95% bootstrapped confidence intervals based on individual observer means; lines represent mixed-effects model predictions and uncertainty for the population fixed effects, ignoring random effects of observer and image. The dashed horizontal lines in all panels show chance performance. Shaded regions in all panels show 95% credible intervals for the given model. Note these are independent, and so overestimate the uncertainty for making any pairwise comparison between conditions (see Appendix for details).

parameters were given conservative, weakly informative prior distributions such that we assumed no effects of our experimental manipulations (by using priors for regression parameters centered on zero) but with high uncertainty. This biases the model against finding spuriously large effects. Bayesian model estimation offers two practical advantages here: first, posterior credible intervals over model parameters have an intuitively appealing meaning (they represent our belief that the "true" parameter lies within some interval with a given probability, conditioned on the priors, model and data). Second, the priors act to sensibly regularize the model estimates to ensure all parameters are identifiable. More details and analysis are provided in the Appendix.

Experiment 1: Original texture model

This experiment compares textures produced by the CNN texture model to the PS model under two observation conditions. This experiment was conducted on two groups of observers. The first (Experiment 1a) consisted of two of the authors, who were familiar with the stimuli, experienced with the psychophysical task, and optically corrected as appropriate. The authors completed five experiment sessions (each consisting of one parafoveal and one inspection block), for a total of 3,600 trials each. The order of presentation conditions was pseudorandomly determined for each author in each experiment session. The dataset consisted of 7,200 trials.

The second group (Experiment 1b) consisted of ten naive observers[4] (median age 25 years, min = 21, max = 36), who completed only one experimental session each (i.e., one block of each presentation time).[5] They were paid 10 EUR for the 1-hr session. Half the observers saw the parafoveal condition first, whereas the other half performed the inspection condition first. All protocols conformed to Standard 8 of the American Psychological Associations Ethical Principles of Psychologists and Code of Conduct (2010) and to the Declaration of Helsinki (with the exception of Article 35 concerning preregistration in a public database). The final dataset consisted of 7,200 trials.

Results

Performance as a function of image model and presentation time, averaging over images, is shown in Figure 6. More complex CNN models (matching more parameters) tend to produce poorer psychophysical performance (i.e., better matches to natural appearance), and the performance in the parafoveal condition is poorer than the inspection condition. The PS model produces better psychophysical performance (i.e., is not as good at matching appearance) than the higher layer CNN models under the inspection condition but not under the parafoveal condition. The average pattern of results for the 10 naive observers is qualitatively similar to the data shown by the two authors, with the exception that performance is slightly lower. The figure additionally demonstrates what might be believed about the "population of texture images" from our results. Estimates and credible intervals from the mixed-effects model are shown as lines and shaded

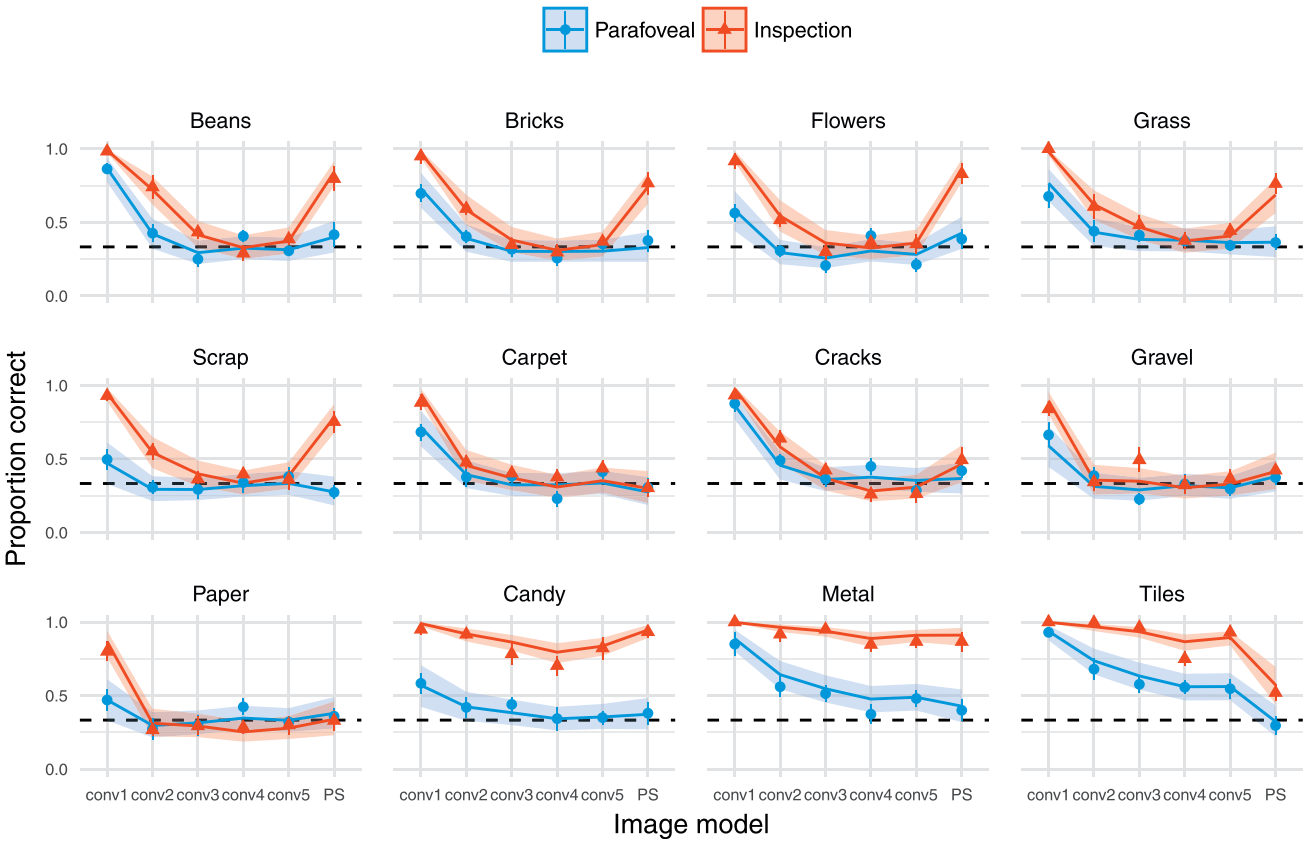


Figure 7. Performance for each image in Experiment 1. Points show the grand mean of all observer means (based on 25 trials for authors and 5 trials for naives). Error bars on points show ±1 SEM. Lines show mixed-effects model estimates (posterior mean, including random effects of image but excluding random effects of subject) and shaded regions show 95% credible intervals. That is, the model predicts mean performance to lie in the shaded area with 95% probability, if the image was shown to an average, unknown subject. Images have been arranged according to consistent patterns of results (reading left-to-right). The original images can be seen in Figure 2.

areas in Figure 6 (further details and quantification are provided in the Appendix).

We observe distinctly different effects of image model and presentation time at the level of individual images (Figure 7). Five images (beans, bricks, flowers, grass, and scrap) show a similar pattern of results as in the average data. Unlike the first five images, the PS model also succeeds in matching appearance for carpet, cracks, gravel, and paper under the inspection condition. In addition, for these images there is less evidence of a difference between the parafoveal and inspection conditions after the conv1 model. These results suggest these four images are easier for all models to synthesize than the first five images. Conversely, all models fail to match the appearance of metal and candy under the inspection condition (psychophysical performance well above chance), whereas the parafoveal condition has a marked effect such that performance drops nearly to chance for the higher convolutional and PS models. Finally, the Tiles image is interesting because here the PS model produces better matches to appearance than the CNN models (the syntheses are more difficult to discriminate).

Experiment 2: Power spectrum constraint

In Experiment 1, the CNN texture model failed to match textures that could be considered "quasiperiodic," in that they contain global regularities spaced across the whole texture image (for example, the roof tiles or the metal floor textures). Liu et al. (2016) recently showed that such textures can be more closely modeled by adding a power spectrum constraint to the synthesis procedure in CNN texture models. That is, the gradient descent procedure now aims to match both the CNN features and the global Fourier power

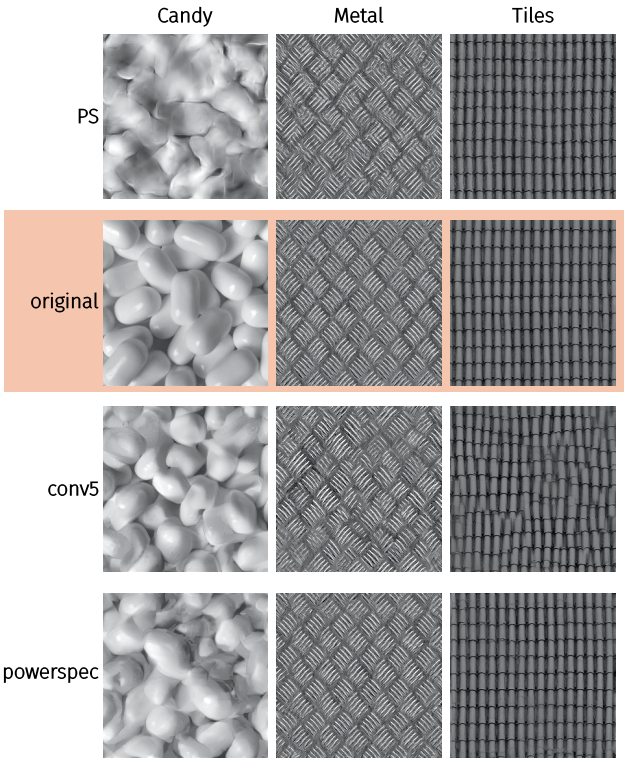


Figure 8. Example experimental stimuli, Experiment 2. Original texture images are copyrighted by www.textures.com (used with permission).

spectrum of the original image. In an image like the tiles, the periodic regularity shows up as a strong orientation-and-frequency component in the power spectrum. Matching this improves the perceptual quality of such textures (see Figure 8). In this experiment we seek to quantify this improvement with respect to the unconstrained conv5 model and the PS model for our 12 texture images, using the same procedure as in Experiment 1.

Five observers participated in this experiment, consisting of two authors (CF and TW) and three naive observers, one of whom had participated in the first experiment. All observers completed two experiment sessions (each consisting of one parafoveal and one inspection block) for a total of 1,440 trials, with the exception of S1, who did not return for a second session of testing and so completed only 720 trials.

Results

For average performance over images (Figure 9) and at the individual image level (Figure 10), the results of Experiment 2 are similar to those of Experiment 1 for the conv5 and PS models. The powerspec model produces similar performance to the conv5 model for

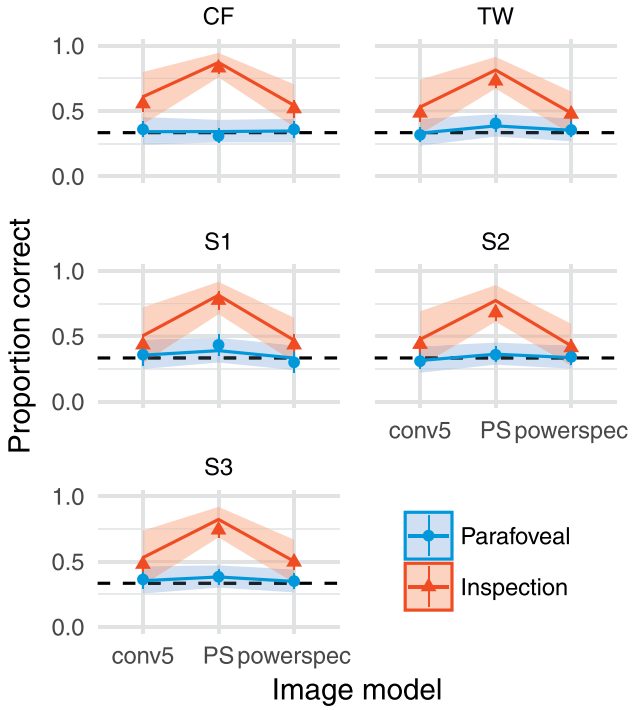


Figure 9. Performance as a function of image model in Experiment 2, averaging over images. Points show mean and 95% confidence intervals on performance (each based on 240 trials for all observers except S2). Lines show mixed-effects model predictions for each observer (mean of posterior regression line) and shaded regions show model 95% credible intervals, ignoring random effects of image.

most images, with the possible exceptions of beans, bricks, flowers, and grass, in which human performance is slightly higher than for conv5 (i.e. the powerspec model is less effective at matching appearance than conv5). For images with significant long-range regularities (metal and tiles) whose appearance failed to be matched by conv5, the powerspec model drastically reduced psychophysical performance. That is, the model syntheses are now approximately matched to the visual appearance of these original images even under foveal inspection (see Appendix). Note, however, that one observer (author TW) still achieved high accuracy for the powerspec model of metal, showing that the model fails to capture some important features that at least one observer can see. Finally, all models fail to capture the appearance of the candy image under inspection.

Control analysis: Cross-correlation of image crops

The experiments reported above show that the CNN texture model (specifically the power spectrum match-

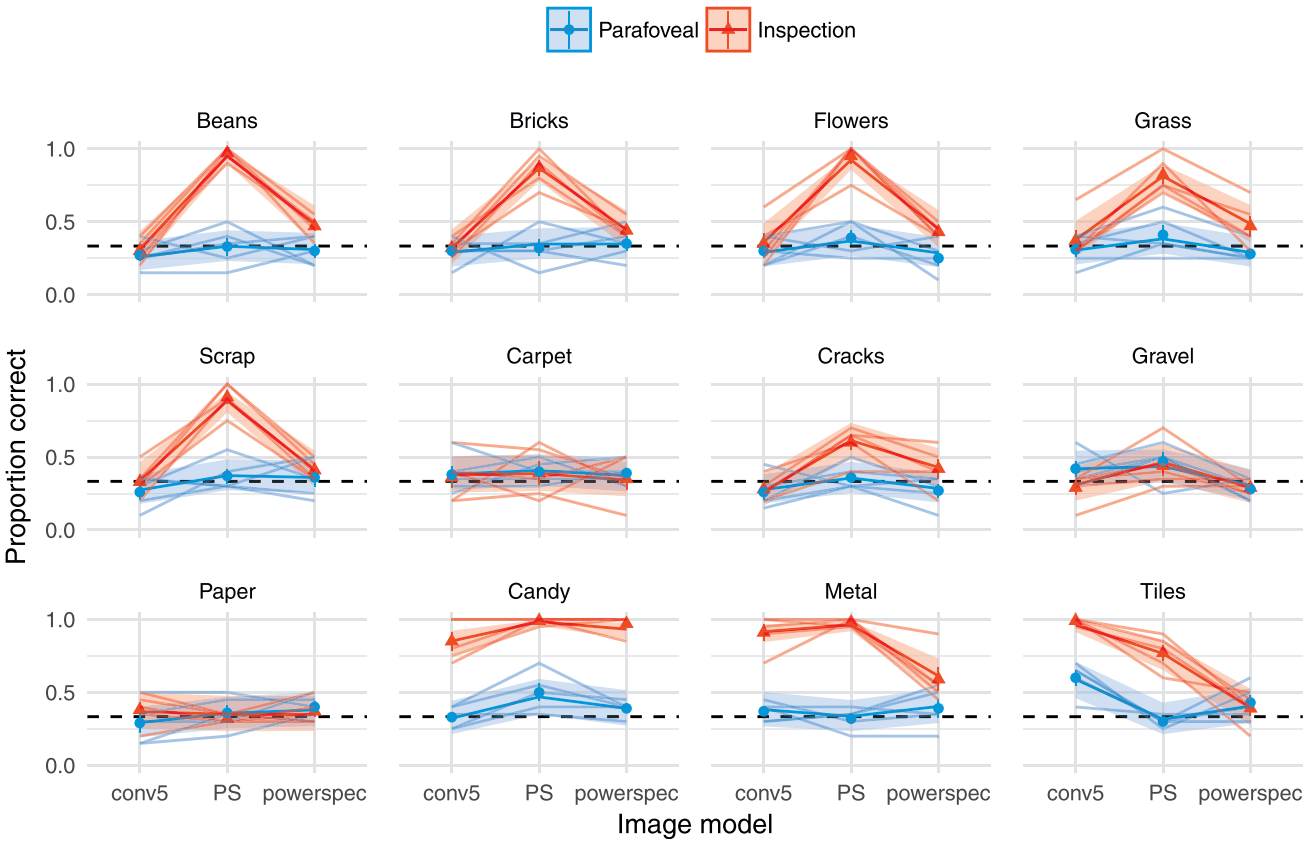


Figure 10. Performance for each image in Experiment 2. Points show means and ±1*S EM* over observer means. Faint lines link individual observer mean performance (based on 20 trials for all observers except S2). Solid lines show mixed-effects model estimates (posterior mean) and shaded regions show 95% credible intervals.

ing variant) can match the appearance of a range of textures even under foveal viewing. One concern with this result is that the model may be overfitting on the target texture image. Consider a "copy machine" model that would exactly copy the image up to a phase shift. Samples generated by this model would likely be indistinguishable from the original image, because our experimental design (taking nonoverlapping crops) enforces the samples to be physically different. Consequently, if a model was acting like a copy machine, this could not show up in our existing results. If this were the case, one could argue that the model has not learned anything about the structure of textures per se but rather how to copy pixels.

To investigate this issue, we computed the normalized maximum cross-correlation between different texture samples and the corresponding original texture. If the algorithm simply copies and phase-shifts the image, the maximum cross-correlation with the original will be one. Specifically, for each of the 10 unique texture samples of size 256×256 synthesized by each model in Experiment 2, we took one $N \times N$ crop of the center plus 10 additional random crops of edge N px, for each of $N = \{32, 64, 128\}$. Each crop is then normalized to have zero mean and unit variance, before computing the cross-correlation function between crop and original and taking the maximum. Finally, we take the average of this maximum across the 11 crops.

For certain textures however, it may be the case that a synthesis algorithm *needs* to act like a copy machine (up to a spatial shift) to match the appearance of the texture. For example, textures with strong periodicities and little variation between individual texture elements (e.g., metal or tiles) might require copying for appearance to be matched, whereas the appearance of less regular structure (grass or beans) might be sufficiently captured by far less. To account for this image-specific variation, we additionally computed the maximum cross-correlation between an $N \times N$ center crop from the original texture, and the full 256×256 px image itself (after excluding shifts of ±16 px around the center, which would trivially return one). This value can be seen as a measure of self-similarity.

The maximum cross-correlation values for the images used in this paper are shown in Figure 11. This result shows that crops of synthesized textures are not

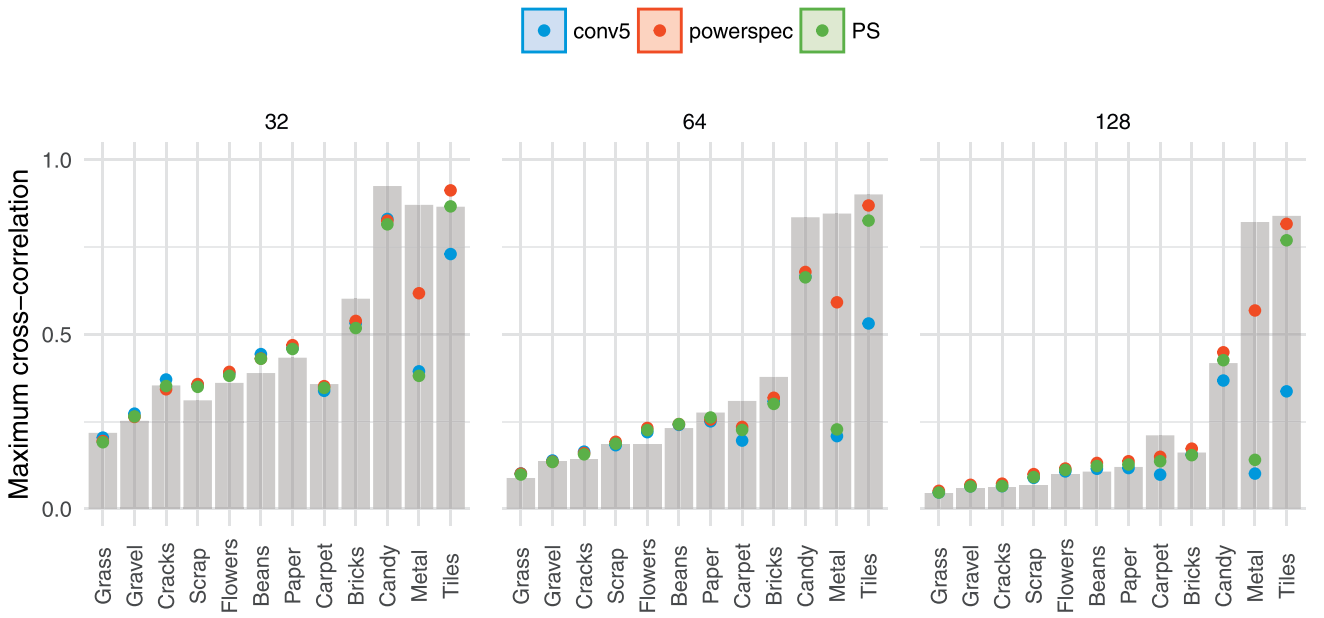


Figure 11. Controlling for texture model overfit. Points show the average maximum cross-correlations between crops of model syntheses (colors) and the original images, for three different crop sizes (32, 64, and 128 px; panels). If the model simply copied and phase-shifted the original, these values would be approximately one. Bars show the baseline of a crop from the original image correlated to itself. Some images are more self-similar, and thus require some degree of copying to match appearance.

more similar to the best matching crop in the corresponding original image than are any two crops taken from the original image. Thus, none of the models are simply copying the original images at any of the spatial scales we tested. The metal and tiles images are the most self-similar (gray bars) at all scales, and these were also the images for which adding the power spectrum constraint to the CNN texture model helped most (compare conv5 and powerspec cross-correlation values).

General discussion

We have shown that the CNN texture model of Gatys et al. (2015) can produce artificial images that are indiscriminable from a range of source textures even under foveal viewing. That is, images synthesised from the Gatys model could pass as natural materials, at least for nine of the 12 images we test here and for similar viewing conditions. A model that matches both a selection of deep CNN features and the power spectrum of the original image (Liu et al., 2016) greatly improves the perceptual fidelity of two of the remaining three images not captured by the Gatys model (Experiment 2). These results were not attributable to simply copying the target images (Figure 11). The most popular existing parametric texture model (PS; Portilla & Simoncelli, 2000) can capture texture appearance for many images briefly presented to the parafovea, but is less successful under foveal inspection (matching appearance for four of the images—see Figure 7). These results regarding the PS model corroborate the findings of Balas (2006) and Balas (2012) respectively. Taken together, our results show that the natural image statistics represented by the CNN model (and the power spectrum variant) can capture important aspects of material perception in humans, but are not sufficient to capture the appearance of all textures.

The patterns of performance in Figures 7 and 10 suggest that for the purposes of assessing parametric texture models, natural textures may be parsed into at least four clusters.[6] First, one cluster of images (beans, bricks, flowers, grass, and scrap) can be matched by the CNN texture model's higher layers even for foveal inspection, but only for parafoveal viewing by the PS model. These images feature readily discernable texture elements that do not follow a regular periodic arrangement. The second cluster (carpet, cracks, gravel, and paper) can be matched by all but the simplest CNN texture model under both parafoveal and inspection conditions. For these images, it is possible that individual textons (single texture elements; Julesz, 1981) were difficult to resolve even foveally, allowing models that failed to capture individual textons to nevertheless sufficiently match appearance. Third, the metal and tiles images include regular structure that can only be effectively matched by the CNN+powerspectrum model. These are both strongly periodic textures

with easily resolvable textons. Finally, the candy image cannot be matched by any of the models tested here for foveal inspection. It contains large textons with interesting material properties (glossiness)[7] as well as occlusions and shading suggesting depth. These clusters may provide useful test cases for parametric texture models in the future. In particular, a single image from each class may be sufficient to provide a generalizable test of a texture model. More generally, psychophysics may offer an approach to find equivalence classes of textures that are useful for discriminating between texture models.[8] The failure of all models we test here to capture the candy image shows that the CNN features we test here are still not sufficient to capture the appearance of all textures.

An additional noteworthy feature of the data is that for many images, the conv5 model is slightly worse at matching appearance (psychophysical performance is better) than the conv4 model (e.g., Figure 12). This is particularly evident for example for the candy and tiles images under inspection (Figure 7, though note these data points are also affected by large oddball type biases—see Figure 16). Assuming this effect is robust, it could be related to the observation that the conv5 model results in higher final total loss values after optimization than the conv4 model (Figure 18).

Model complexity versus feature complexity

Why do the features used in the CNN texture model often succeed in capturing texture appearance? One possibility is that training to recognize objects in images causes deep CNNs to abstract a set of statistics from images that support quite general visual inferences (transfer learning; Donahue et al., 2013). An alternative possibility is suggested by Ustyuzhaninov, Brendel, Gatys, and Bethge (2016), who found that single-layer CNNs using many filters with random weights could produce textures of surprisingly good perceptual quality (assessed via introspection). That is, high-quality texture synthesis from CNNs may require neither a hierarchical (deep) representation nor filters learned on any particular task—many random filters could instead be sufficient (the random-multiscale model from that paper uses about 2 million random parameters, which is significantly more than all models in this paper—Figure 5). If the latter is the case, this would suggest that the improved appearance matching as more convolutional layers are included is because there are simply more features, not that they are "better."

However, we do not believe the improved appearance matching is only due to the number of parameters matched. Gatys et al. (2015) showed that the number of parameters in the CNN model could be reduced by computing Gram matrices on only the first k principle components of each network layer. Textures synthesized using approximately 10,000 parameters from VGG layers conv1_1, conv2_1, conv3_1, conv4_1, and conv5_1 produced (introspectively) much higher quality textures than only using all parameters from conv1_1 and conv2_1 (about 12,000). A second piece of evidence that speaks to this point is that having more parameters—even having more parameters than pixels (i.e., being overcomplete)—does not necessarily result in introspectively high-quality textures (Ustyuzhaninov et al., 2016). Thus, features from the higher network layers seem to improve texture synthesis because they are "better" features, not simply because they add more parameters.

Why are higher layer network features (with the possible exception of conv5_1; see above) better? Recall that deep convolutional networks stack nonlinearities (Figure 1), allowing increasingly complex functions of the input image to be explicitly (linearly) decoded. Higher layers might therefore be better for texture synthesis because they learn to represent complex information. Alternatively, it could just be that higher layers have larger receptive fields than lower layers, and a model that includes both high and low layer information improves because of its multiscale structure. Ustyuzhaninov et al. (2016) showed that having features at multiple scales improves texture synthesis. On one hand, the fact that trained features produce (introspectively) better textures than the random multiscale network using fewer parameters implies that our texture models including higher VGG layers are not better exclusively because they model information at more spatial scales. Another possibility is that it is easier to optimize trained features than random features, which leads to better texture synthesis but does not mean deep features are "better" for parametric texture modelling in general.

Ultimately we think the models in this article perform well due to a mixture of both more numerous and more complex features, and that this is not simply a function of including information at multiple scales. Future psychophysical comparisons could be used to add quantitative rigor to this discussion. For example, comparing the perceptual quality of the random-filter and trained CNN model textures (with and without compression) would quantify the importance of learned features. Similarly, comparing hierarchical (cumulative) and nonhierarchical models could be used to quantify the importance of scale information.

Finally, we would like to emphasize that for those textures the CNN model can mimic, the model features likely represent a superset of the necessary statistics. One important challenge now is to compress these representations into a minimal set of features, in order to develop a parsimonious and intuitive description of

the critical aspects of the feature space. As noted above, Gatys et al. (2015) showed that qualitatively reasonable results could be obtained for a principle component analysis-reduced feature space with 10,000 parameters, compared to the 175,000 of the conv4 or 306,000 of the conv5 models used here. Of course, the PS model matches substantially fewer (about 1,000) parameters than even this, and so its performance for parafoveal images is impressive. The difference between the two models, more substantively quantified, could yield insights into the differences in foveal and peripheral encoding of texture.

Categorical losslessness

Our experiments show that humans cannot tell which of three physically different images were "generated by a different process" (for all but one of the images we test). This condition could be termed "categorical" or "structural" losslessness (Pappas, 2013): Under our experimental conditions, the model syntheses can pass as natural textures (they are perceived as the same category). Images that are perceptually equivalent along some dimension can also be called "eidolons" (Koenderink, Valsecchi, van Doorn, Wagemans, & Gegenfurtner, 2017). Achieving categorical losslessness in an image-computable model is an important step toward understanding human material perception, because the model encodes sufficient statistics for capturing the appearance of these textures. Categorical losslessness must, however, be distinguished from perceptual losslessness: humans are likely able to tell that the three images in our experiments are different from each other (and thus we avoid using the term *metamer* here, which refers to physically different images that cannot be told apart). The latter criterion may be important for understanding information loss in the visual system more generally (Freeman & Simoncelli, 2011; Koenderink & van Doorn, 1996; Wallis et al., 2016; Wandell, 1995).

Caveats

Three caveats should be borne in mind when interpreting our results. First, we have considered only one relationship between input image size and CNN feature scaling (specifically, we used input images of 256×256 px, which is close to the 224×224 px images on which the VGG features were learned). Because the network layers have a fixed receptive field size (the pixels of the original image associated with a given unit), downsampling or upsampling the input images will cause the same network layers to respond to different image structure. For example, it is possible that there is a relationship between the degree to which texture appearance is successfully captured by the model and the size of the texture elements in the image. One possible reason that the candy image (Figure 2) fails to be matched for foveal viewing is that the textons (individual candies) and their overlap are too large to be captured by single filters at some critical layer within the network, even though features in the highest layers are large enough to cover groups of candies. We have tried rescaling the images but this did not seem to improve the syntheses, indicating that this relationship is perhaps not trivial.

A second caveat is that the fidelity of the resulting textures could depend on the number of iterations of the gradient descent used to minimize the loss between the original and the new image (see Appendix, Figure 18). Because this loss is never exactly zero for the more complex models, more iterations could only improve synthesis fidelity—though in our experience, the coarse structure of the images is largely fixed within 200 iterations, and further iterations mostly reduce high-spatial frequency noise. In theory, as long as all features are perfectly matched (i.e., if the loss is exactly zero), more features can only lead to more similar patterns. However, given that the optimization of texture synthesis algorithms typically yields a residual loss, more features do not necessarily improve perceptual quality, and the design of good features is not straightforward and may depend on various factors including the type of textures to be synthesized. As it stands, different models are ideal for different purposes. For peripheral texture perception the PS model achieves best performance with relatively small number of parameters, for random fields with pairwise interactions the scattering network provides a very compact representation for texture synthesis (Joan Bruna, personal communication) and for foveal inspection of textures the VGG features seem particularly useful.

Finally, in our experiments we closely followed the oddity method used by Balas (2006). We believe this paradigm has many desirable properties as a measure of categorical losslessness, but our results also point to a caveat. By cropping from inhomogeneous images (e.g., the flowers image, which contains a size gradient) we introduce greater perceptual variability in the stimuli shown to subjects. Depending on the relative (in)homogeneity of original and synthesized images, this may lead to differences in performance depending on the class of the oddball and potentially to below-chance performance (e.g., in the flowers image). We discuss these issues and present further analysis in the Appendix. While we believe this property will have little effect on our overall conclusions, it is nevertheless useful to consider for future studies.

Conclusion

We have shown that the texture model of Gatys et al. (2015), which uses the features learned by a convolutional neural network trained to recognize objects in images, provides a high-fidelity model of texture appearance for many textures even in the fovea. Overall, however, our results do not identify a uniformly best parametric model for matching texture appearance. Instead, different models may be appropriate for different use cases. The PS model is the best (and the most simple) model to use if textures are intended to be viewed briefly in the parafovea. For textures intended to be foveated, incorporating the power spectrum constraint will be critical for textures with strong periodicities (Liu et al., 2016), whereas the CNN model (conv4) performs best for most other textures we test here. It would obviously be desirable to identify a uniformly best model in future work, and the single failure case we identify here (the candy image) may provide a useful benchmark for testing such models.

Keywords: spatial vision, natural scenes, texture perception, peripheral vision

Acknowledgments

Designed the experiments: TSAW, ASE, CMF, LAG, FAW, MB. Programmed the experiments: TSAW. Collected the data: CMF, TSAW. Analyzed the data: TSAW, AE, CMF. Wrote the paper: TSAW. Revised the paper: CMF, ASE, LAG, FAW, MB. Thanks to Paul-Christian Bürkner for his assistance with fitting the generalized linear mixed-effects model in brms, Heiko Schütt for helpful comments on presentation, and www.textures.com for permission to use the images. Funded, in part, by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the German Science Foundation (DFG; priority program 1527, BE 3848/2-1 and Collaborative Research Centre 1233).

Commercial relationships: none.
Corresponding author: Thomas S. A. Wallis.
Email: thomas.wallis@uni-tuebingen.de.
Address: AG Bethge, Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen, Tübingen, Germany.

Footnotes

[1] As Balas (2006) writes, "The 3AFC [three-alternative forced-choice procedure] task presented here represents a modest contribution towards the formulation of texture discrimination tasks that make explicit the importance of local texture analysis in the human visual system." We agree.

[2] These are analogous to Balas' preattentive and attentive conditions, but we consider these terms somewhat of a historical misnomer: Because there is no spatial or temporal uncertainty, observers can presumably accurately deploy spatial attention to the stimuli in both cases.

[3] These images are copyrighted by www.textures.com (used with permission). Copies of the texture images used in the experiments are available with the online materials of this article (redistributed with permission).

[4] Ten was chosen a priori based on pilot testing.

[5] Observer S9 completed 144 trials of the inspection condition before this data was lost due to computer malfunction. The observer repeated the full testing session; thus this observer had more practice and exposure to the images than the other observers.

[6] Since we have only used 12 texture images in the present study, it is likely that a number of additional clusters exist that were not represented in the set of images we used.

[7] While the structure of the candy image is never successfully captured by the CNN model, one intriguing feature of the syntheses is that they appear glossy as for the original image (compare for example the conv3 and conv4 syntheses in Figure 3). This glossy appearance is not captured by the PS model.

[8] Balas (2006) subjectively delineated three texture categories: *pseudoperiodic* (containing strongly periodic structure), *structured* (repeated structural elements with no periodicity), and *asymmetric* (containing asymmetric lighting giving the impression of depth). Our cluster containing metal and tiles is equivalent to Balas' pseudoperiodic textures, but our other three data-determined clusters do not trivially map onto Balas' other categories (e.g., bricks and grass are structured, whereas flowers, beans, and scrap contain asymmetric lighting and other depth cues).

[9] A three-alternative forced-choice procedure as we use here has a chance performance rate of 1/3. If we were interested in estimating some "threshold" of a psychometric function, the standard logistic link function might be considered inappropriate for these data: It could predict that performance falls below 0.33, which if it occurs in observed data can only be due to measurement error or to observers incorrectly switching responses (and is therefore not a desirable prediction to make in general; though see our third

caveat in the General discussion). However, we are not estimating thresholds here, but rather we wish to quantify performance differences between discrete levels and also the extent to which performance is different to chance performance. The standard logistic link function is therefore more desirable.

References

Adelson, E. H. (2001). On seeing stuff: The perception of materials by humans and machines. In *Photonics West 2001-Electronic imaging* (pp. 1–12). Bellingham, AW: International Society for Optics and Photonics.

Adelson, E. H., & Bergen, J. R. (1991). The plenoptic function and the elements of early vision. *Computational Models of Visual Processing*, *1*, 3–20.

Agaoglu, M. N., & Chung, S. T. L. (2016, December). Can (should) theories of crowding be unified? *Journal of Vision*, *16*(15):10, 1–22, doi:10.1167/16.15.10. [PubMed] [Article]

Arnold, J. B. (2017). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 3.4.0. https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2016). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.2.1. https://CRAN.R-project.org/package=gridExtra

Baayen, R., Davidson, D., & Bates, D. (2008, November). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412, doi:10.1016/j.jml.2007.12.005.

Balas, B. J. (2006, February). Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Research*, *46*(3), 299–309, doi:10.1016/j.visres.2005.04.013.

Balas, B. J. (2008). Attentive texture similarity as a categorization task: Comparing texture synthesis models. *Pattern Recognition*, *41*(3), 972–982.

Balas, B. J. (2012, January). Contrast negation and texture synthesis differentially disrupt natural texture appearance. *Frontiers in Psychology*, *3*, 515, doi:10.3389/fpsyg.2012.00515.

Balas, B. J., & Conlin, C. (2015, October). Invariant texture perception is harder with synthetic textures: Implications for models of texture processing. *Vision Research*, *115*, 271–279, doi:10.1016/j.visres.2015.01.022.

Balas, B. J., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12):13, 1–18, doi:10.1167/9.12.13. [PubMed] [Article]

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48, doi:10.18637/jss.v067.i01.

Beck, J., & Gibson, J. J. (1955). The relation of apparent shape to apparent slant in the perception of objects. *Journal of Experimental Psychology*, *50*(2), 125–133, doi:10.1037/h0045219.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Bürkner, P.-C. (in press). Brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014, December). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*(12), e1003963, doi:10.1371/journal.pcbi.1003963.

Cano, D., & Minh, T. (1988, September). Texture synthesis using hierarchical linear transforms. *Signal Processing*, *15*(2), 131–148, doi:10.1016/0165-1684(88) 90066-7.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., ...Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cheung, S.-H., Kallie, C. S., Legge, G. E., & Cheong, A. M. Y. (2008). Nonlinear mixed-effects modeling of MNREAD data. *Investigative Ophthalmology & Visual Science*, *49*(2), 828–835. [PubMed] [Article]

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2016, April). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 146–158, doi:10.1016/j.neuroimage.2016.03.063.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016, June). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755, doi:10.1038/srep27755.

Clarke, A. M., Herzog, M. H., & Francis, G. (2014, October). Visual crowding illustrates the inadequacy of local vs. global and feedforward vs. feedback distinctions in modeling visual perception. *Frontiers in Psychology*, *5*, 1193, doi:10.3389/fpsyg.2014.01193.

Dakin, S. (2014). Seeing statistical regularities: Texture and pattern perception. In *Oxford handbook of perceptual organization* (pp. 1–19). Oxford, UK: Oxford University Press.

Donahue, J., Jia, Y., & Vinyals, O. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv*, 1–10, arXiv:1310.1531.

Efros, A. A., & Freeman, W. T. (2001). Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 341–346). New York: ACM.

Fleming, R. (2014). Visual perception of materials and their properties. *Vision Research*, *94*(C), 62–75.

Freeman, J., & Simoncelli, E. P. (2011). Metameres of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201.

Freeman, J., Ziemba, C., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981.

Freeman, J., Ziemba, C. M., Simoncelli, E. P., & Movshon, J. A. (2013, November). Functionally partitioning the ventral stream with controlled natural stimuli. In *Annual Meeting, Neuroscience*. San Diego, CA: Society for Neuroscience.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015, May). Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems 28*. La Jolla, CA: Neural Information Processing Systems Foundation.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016, June). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE.

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *arXiv:1706.06969*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gibson, J. J. (1950, July). The perception of visual surfaces. *The American Journal of Psychology*, *63*(3), 367, doi:10.2307/1418003.

Guclu, U., & van Gerven, M. A. J. (2015, July). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014, doi:10.1523/JNEUROSCI.5023-14.2015.

Gurnsey, R., Pearson, P., & Day, D. (1996). Texture segmentation along the horizontal meridian: Nonmonotonic changes in performance with eccentricity. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 738.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques* (pp. 229–238). New York: ACM.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision*, *15*(6):5, 1–18, doi:10.1167/15.6.5. [PubMed] [Article]

Ho, Y.-X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*, *19*(2), 196–204.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(Apr), 1593–1623.

Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016, February). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622, doi:10.1038/nn.4247.

Jones, E., Oliphant, E., Peterson, P., et al. SciPy: Open Source Scientific Tools for Python, 2001-. http://www.scipy.org/

Julesz, B. (1962, February). Visual pattern discrimination. *IEEE Transactions on Information Theory*, *8*(2), 84–92, doi:10.1109/TIT.1962.1057698.

Julesz, B. (1981, March). Textons, the elements of texture perception, and their interactions. *Nature*, *290*(5802), 91–97, doi:10.1038/290091a0.

Julesz, B., Gilbert, E. N., & Victor, J. D. (1978). Visual discrimination of textures with identical third-order statistics. *Biological Cybernetics*, *31*(3), 137–140.

Kehrer, L. (1987). Perceptual segregation and retinal position. *Spatial Vision*, *2*(4), 247–261.

Kehrer, L. (1989). Central performance drop on perceptual segregation tasks. *Spatial Vision*, *4*(1), 45–62.

Keshvari, S., & Rosenholtz, R. (2016, February).

Pooling of continuous features provides a unifying account of crowding. *Journal of Vision*, *16*(3):39, 1–15, doi:10.1167/16.3.39. [PubMed] [Article]

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, November). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), e1003915, doi:10.1371/journal.pcbi.1003915.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *bioRxiv*, 133504.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception, 36*. (ECVP Abstract Supplement).

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York: Springer.

Koenderink, J., Valsecchi, M., van Doorn, A., Wagemans, J., & Gegenfurtner, K. (2017, March). Eidolons: Novel stimuli for vision research. *Journal of Vision*, *17*(2):7, 1–36, doi:10.1167/17.2.7. [PubMed] [Article]

Koenderink, J., & van Doorn, A. J. (1996). Metamerism in complete sets of image operators. In K. Bowyer & N. Ahuja (Eds.), *Advances in image understanding: A Festschrift for Azriel Rosenfeld*. Los Alamitos, CA: IEEE Computer Society Press.

Kruschke, J. K. (2011). *Doing Bayesian data analysis*. Burlington, MA: Elsevier.

Kruschke, J. K., & Liddell, T. M. (2017, February). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, E-pub ahead of print, doi:10.3758/s13423-016-1221-4.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), e1004896.

Landy, M. S. (2013). Texture analysis and perception. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neurosciences* (pp. 639–652). Cambridge, MA: MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444, doi:10.1038/nature14539.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Liu, G., Goussau, Y., & Xia, G.-S. (2016). Texture synthesis through convolutional neural networks and spectrum constraints. *arXiv preprint arXiv: 1605.01141*.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A*, *7*(5), 923–932.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (No. 122). Boca Raton, LA: CRC Press.

Miller, J., & Ulrich, R. (2015, November). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic Bulletin & Review*, *23*(1), 124–130, doi:10.3758/s13423-015-0859-7.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015, October). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123, doi:10.3758/s13423-015-0947-8.

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E.-J. (2015, November). Continued misinterpretation of confidence intervals: Response to Miller and Ulrich. *Psychonomic Bulletin & Review*, *23*(1), 131–140, doi:10.3758/s13423-015-0955-8.

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision*, *12*(11):26, 1–17, doi:10.1167/12.11.26. [PubMed] [Article]

Movshon, J. A., & Simoncelli, E. P. (2014). Representation of naturalistic image structure in the primate visual cortex. *Cold Spring Harbor Symposia on Quantitative Biology*, *79*, 115–122, doi:10.1101/sqb.2014.79.024844.

Okazawa, G., Tajima, S., & Komatsu, H. (2015, January). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences*, *112*(4), E351–E360, doi:10.1073/pnas.1415146112.

Pappas, T. N. (2013). The rough side of texture: Texture analysis through the lens of HVEI. In *IS&T/SPIE Electronic Imaging*(pp. 86510P–86510P). Bellingham, WA: International Society for Optics and Photonics.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Porat, M., & Zeevi, Y. (1989, January). Localized

texture processing in vision: Analysis and synthesis in the Gaborian space. *IEEE Transactions on Biomedical Engineering*, *36*(1), 115–129, doi:10.1109/10.16457.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

R Core Development Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenholtz, R. (2011). What your visual system sees where you are not looking. In B. E. Rogowitz & T. N. Pappas (Eds.), *Proceedings of SPIE: Human vision and electronic imaging*, *XVI* (pp. 786510–786514). Bellingham, WA: SPIE.

Rosenholtz, R. (2014). Texture perception. In *Oxford handbook of perceptual organization* (pp. 1–24). Oxford, UK: Oxford University Press.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: Effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology*, *3*, 13.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, doi:10.1167/12.4.14. [PubMed] [Article]

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252, doi:10.1007/s11263-015-0816-y.

Safranek, R. J., & Johnston, J. (1989). A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Acoustics, speech, and signal processing, 1989. ICASSP-89., 1989 international conference on* (pp. 1945–1948). Glasgow, UK: IEEE, doi:10.1109/ICASSP.1989.266837.

Safranek, R. J., Johnston, J. D., & Rosenholtz, R. E. (1990, October). Perceptually tuned sub-band image coder. *IEEE Transactions on Circuits and Systems for Video Technology*, *5*(6), 284–293, doi:10.1117/12.19678.

Simoncelli, E., & Portilla, J. (1998). Texture characterization via joint statistics of wavelet coefficient magnitudes. In *Proceedings of fifth international conference on image processing* (Vol. 1, pp. 62–66). Chicago, IL: IEEE, doi: 10.1109/ICIP.1998.723417.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. Presented at the International Conference on Learning Representations, May 7–9, 2015, San Diego, CA.

Stan Development Team. (2017). Stan Modeling Language Users Guide and Reference Manual, Version 2.15.1. http://mc-stan.org

Stan Development Team. (2017). The Stan Core Library, Version 2.15.1. http://mc-stan.org

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint*, arXiv:1312.6199.

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42.

Ustyuzhaninov, I., Brendel, W., Gatys, L. A., & Bethge, M. (2016, June). Texture synthesis using shallow convolutional networks with random filters, *arXiv*, arXiv:1606.00021.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... the scikit-image contributors (2014, June). Scikit-image: Image processing in Python. *PeerJ*, *2*, e453, doi:10.7717/peerj.453.

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC$_*$. *arXiv preprint*, arXiv:1507.04544.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Victor, J. D., Thengone, D. J., & Conte, M. M. (2013, March). Perception of second- and third-order orientation signals and their interactions. *Journal of Vision*, *13*(4):21, 1–21, doi:10.1167/13.4.21. [PubMed] [Article]

Wallis, T. S. A., Bethge, M., & Wichmann, F. A. (2016, March). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision*, *16*(2):4, 1–30, doi:10.1167/16.2.4. [PubMed] [Article]

Wandell, B. A. (1995). *Foundations of vision*. Sunderland, MA: Sinauer Associates.

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wickham, H., Romain, F., Henry, L., & Müller, K. (2017). dplyr: A grammar of data manipulation. R

package version 0.7.2. https://CRAN.R-project.org/package=dplyr

Xie, Y. (2013). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing Reproducible Computational Research*. Boca Raton, FL: CRC Press.

Xie, Y. (2015). *Dynamic documents with R and Knitr* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Yamins, D. L. K., & DiCarlo, J. J. (2016, February). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365, doi:10.1038/nn.4244.

Yamins, D. L. K., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, June). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624, doi:10.1073/pnas.1403112111.

Zhu, S. C., Wu, Y., & Mumford, D. (1998). Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, *27*(2), 107–126.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016, May). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149, doi:10.1073/pnas.1510847113.

Appendix

Bayesian multilevel modelling

To analyze the data, we first made the (standard) assumptions that the observers' responses on each trial (correct/incorrect) reflected a Bernoulli process, and that the response on a given trial was not dependent on previous responses. We estimate the success probability of this Bernoulli process using a generalized linear mixed-effects model with a logistic link function whose parameters were estimated in a Bayesian framework.[9] A mixed-effects model (a type of hierarchical or multilevel model) includes some number of "fixed" effect parameters that quantify how the response depends on the predictor variables at a population level, and some "random" (also called group-level) effects that allow the fixed effect coefficients to vary over discrete levels that are assumed to be nonexhaustive samples from a larger population. Our model contains two fixed-effect factors: the image model (with six levels, entered into the model design matrix using successive difference coding using contr.sdif from the MASS package for R; Venables & Ripley, 2002) and the presentation condition (with two levels, parafoveal and inspection, coded with sum contrasts [1, −1]). We included the interaction terms between these factors such that the model consisted of 12 fixed effect coefficients. The variation caused by observers and images are modeled as random effects, which are coded as offsets added to the fixed effect coefficients whose variance is estimated. Note that we make an additional simplifying assumption by ignoring other sources of variance, such as the synthesized image used on a trial and the random crop location (see Methods). We assume that each fixed effect coefficient can vary by observer or by image, and that the variance could be correlated. The specification of the model in R formula syntax (lme4 / brms) was

$$\begin{aligned}&\text{model\_formula} < -\text{correct} \sim \\ &\quad \text{image\_model} * \text{presentation\_cond} + \\ &\quad (\text{image\_model} * \text{presentation\_cond}|\text{subj}) + \\ &\quad (\text{image\_model} * \text{presentation\_cond}|\text{image\_code})\end{aligned}$$

We used conservative, weakly informative prior distributions in the sense that they bias estimates towards the middle of the range of possible values and away from indicating large effects. Consider that the model coefficients are defined on the linear predictor scale, whose effective range runs from approximately −5 (returning an expected success probability of 0.007) to 5 (returning 0.993; a linear predictor value of zero gives 0.5). We therefore expect that no standardized fixed-effect coefficient to be larger than ±5 (i.e., the difference between two factor levels runs from the lowest to the highest observable success probabilities, other effects being equal), and they will very likely be smaller than this. We therefore place Gaussian priors over all fixed-effect coefficients for factors with mean zero (i.e., our a priori expectation is for no effect), standard deviation of 2 (indicating a weak implausibility of large coefficients). These are therefore weak, but not flat (uniform) prior distributions. We also place priors over the variation in random effects; following the logic for effective range of the linear predictor we expect that the effect sizes of our fixed effects are unlikely to vary by more than two on average (i.e., the standard deviation is very unlikely to be larger than 2). We use half-Cauchy priors (i.e., with a lower bound of

zero, as recommended by Gelman & Hill, 2007) over the standard deviation parameters for each random effect, with a mode of zero (i.e., our maximum a priori assumption is that subjects and images are no different) and a standard deviation of 1, reflecting large uncertainty. Finally, we set a prior over the correlation matrix for observer and image-level offsets in the fixed effects that assumes that smaller correlations are slightly more likely than larger ones (an "lkj[2]" prior, see Lewandowski, Kurowicka, & Joe, 2009; Stan Development Team, 2015, for details). While the priors we use here are informed by the scale of the model and by common practice for Bayesian regression models (see for example Gelman, 2006; Gelman & Hill, 2007; Kruschke, 2011), the specific choices we make are somewhat arbitrary. As we see above, the model provides a good fit to the data, but the reader should bear in mind that as always, our inferences depend on the model we assume.

We estimate the posterior distribution over model parameters using a Markov Chain Monte Carlo procedure implemented in the Stan language (version 2.15.1; Carpenter et al., 2017; Hoffman & Gelman, 2014; Stan Development Team, 2017), using the model wrapper package brms (version 1.7.0; Bürkner, in press) in the R statistical environment. The brms package allows the specification of flexible mixed-effects Stan models using formula syntax similar to the popular lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Samples were drawn using the NUTS sampling algorithm (Hoffman & Gelman, 2014) with six independent chains, each sampled with 30,000 samples of which 10,000 were used to adaptively tune the sampler (warmup). To reduce the final file size we saved every sixth sample. This procedure resulted in a final total of 20,000 postwarmup samples. Chain convergence was assessed using the $\widehat{R}$ statistic (Gelman & Rubin, 1992) and visual inspection of trace plots. Readers are encouraged to consult the online code for further details.

The resulting posterior distribution is summarized as Bayesian credible intervals on marginal parameter values and predictions. Unlike frequentist confidence intervals in general, credible intervals have the desirable property that they represent a coherent statement of belief about the parameters' likely values, given the model, priors and data. A 95% credible interval means that the "true" parameter value (conditioned on model, prior, and data) has a 95% probability of lying within the interval (see Miller & Ulrich, 2015; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015; Morey, Hoekstra, Rouder, & Wagenmakers, 2015, for recent discussion on this issue), which many readers will find intuitively appealing. We report 95% credible intervals (rather than 67% or 82% or any other interval) merely as convention. The model's belief about the data is represented by the full posterior distribution, which can be summarized into arbitrary intervals (see McElreath, 2016, p. 58 for related discussion). Readers should avoid mental hypothesis testing (rejecting null values that lie outside the interval). Using Bayesian credible intervals to reject null values in this way suffers two of the same problems as null hypothesis significance testing using p values: It can only reject but never accept a null value, and if used with optional stopping of data collection it will always reject null values even if they are true (Kruschke & Liddell, 2017). Instead, the credible intervals serve to give information about the magnitude and precision of likely effects.

Another advantage of a Bayesian approach in this context is that the weakly informed priors we use act as a regularizer for the model, ensuring that parameters are identifiable (indeed, in our hands the lme4 package had troubles fitting this model). Using zero-centered prior distributions on regression parameters biases the parameters against finding spuriously large effects. One caveat is that credible intervals in general, unlike confidence intervals, are not guaranteed to result in a prespecified error rate for binary inferences (e.g., effect/no effect) in the long run. Given that some decisions about our analyses were made after seeing the data (making this *exploratory* research), frequentist p values would not have their nominal false-alarm rates in any case. For these reasons we report a Bayesian analysis here; readers wishing to apply other analyses are encouraged to do so using the raw data provided online.

Where it makes sense to compare discrete models, we do so using an approximation to the out-of-sample (leave-one-out) prediction error provided by the R package loo (v 1.1.0; Vehtari, Gelman, & Gabry, 2016). Loosely, this value estimates the ability of the model to predict new data (smaller values are better). We report differences between models and their standard errors on the deviance scale (–2 times the expected log pointwise predictive density estimated by the loo package, called LOOIC).

Experiment 1

Figure 6 shows model predictions for both individual observers (authors CF and TW) and for the average of the naive observers. For the individual observer model estimates (CF and TW) we show the model prediction conditioned on observer. The observer's mean performance is 95% likely to lie within the shaded area for an average, unknown image (Baayen, Davidson, & Bates, 2008). The "naive" panel shows the average performance for the naive observers. The model predictions here exclude both observer and image random effects: Mean performance has a 95% probability to lie within the shaded area for an average, unknown image and an

average, unknown subject. Note that the model uncertainties shown in Figure 6 depict the expected spread of population averages across images, but are not appropriate for comparing between presentation conditions because they do not take into account the paired nature of these data (the design was within-subjects and within-images).

To quantify the differences between conditions more appropriately we examine the mixed-effects model coefficients. First, we quantify the performance difference between the inspection and parafoveal conditions, marginalizing over image models and all random effects variance. The posterior median of the difference between these conditions on the linear predictor scale is 1.23. Considering the exponent of this value as log odds, this means that correct trials are $\exp(1.23) = 3.41$ times more likely under the inspection condition than the parafoveal condition, if all other effects are held at zero. In other words, for every 10 correct responses in the parafoveal condition we expect about 34 correct responses in the inspection condition, on average. The 95% credible interval tells us to believe that the difference has a 95% probability (conditioned on the data, model and prior) of lying between 0.65 and 1.82. To indicate the likely sign of an effect we report the posterior probability that the coefficient is negative (if this value is small, the coefficient is likely positive; if the value is 0.5 then the coefficient is equally likely to be positive or negative). The inspection condition is very likely to elicit higher performance than the parafoveal condition, because the coefficient coding their difference has only a small probability of being negative, $p(\beta < 0) = 9.998\text{e-}05$. To make future quantifications more concise, for the remainder of this section we report them as ($\beta = 1.23$, 95% CI = [0.65, 1.82], $p[\beta < 0] < 0.001$).

Next, we examine whether the differences between image models depended on the presentation condition. An interaction is clearly evident in Figure 6. This subjective impression was supported by a model comparison between a linear and an interaction model using a measure of each model's ability to generalize to new data (the LOOIC; the interaction model had a lower LOOIC by 294 [$SE = 33$]). We therefore further consider the differences between image models conditioned on the presentation condition.

For the parafoveal condition, image models above conv2 and also the PS model produced performance at approximately chance level (see below). Our model quantifies the sequential differences between the models, with the coefficients coding the difference between two models on the linear predictor scale. Performance in conv2 was worse than convl, $\beta = -1.35$, 95% CI = [−1.94, −0.78], $p(\beta < 0) > 0.999$, and conv3 was worse than conv2, $\beta = -0.26$, 95% CI = [−0.58, 0.06], $p(\beta < 0) = 0.947$. However, because performance was now approximately at chance, there was no evidence that conv4 was different to conv3, $\beta = 0$, 95% CI = [−0.26, 0.29], $p(\beta < 0) = 0.487$, or that conv5 was different to conv4, $\beta = -0.02$, 95% CI = [−0.25, 0.21], $p(\beta < 0) = 0.561$. Similarly, the PS model was also not different to conv5, $\beta = -0.01$, 95% CI = [−0.51, 0.52], $p(\beta < 0) = 0.508$.

The inspection condition showed similar results as the parafoveal condition with two exceptions: First, performance remained approximately above chance, and psychophysical performance was better for the PS model than the conv5 model (i.e., synthesized and natural textures were easier to discriminate). The conv2 model produced worse performance than conv1, $\beta = -3.08$, 95% CI = [−3.87, −2.32], $p(\beta < 0) > 0.999$, and conv3 produced worse performance than conv2, $\beta = -0.62$, 95% CI = [−0.98, −0.26], $p(\beta < 0) = 0.999$. Conv4 produced worse performance than conv3 in that the coefficient coding their difference was likely to be negative, $\beta = -0.38$, 95% CI = [−0.68, −0.09], $p(\beta < 0) = 0.995$. Performance for the conv5 model was approximately equal to conv4, $\beta = 0.19$, 95% CI = [−0.05, 0.43], $p(\beta < 0) = 0.056$. Finally, there was weak evidence that PS model produced better psychophysical performance than the conv5 model when observers could inspect the images, $\beta = 0.82$, 95% CI = [0.09, 1.54], $p(\beta < 0) = 0.014$.

To summarize, the two most important characteristics of these data are first, that psychophysical performance is effectively at chance for the parafoveal condition for the conv4, conv5, and PS models. Second, under inspection the PS model produces poorer matches to appearance (better psychophysical performance) than the conv5 and conv4 CNN texture models. Taken together, the data show that the PS model features are sufficient to capture the appearance of natural textures under brief, parafoveal viewing conditions, but that the increased complexity of the CNN model features improves appearance-matching performance under inspection.

The attentive reader may wonder why the model's uncertainty estimates in Figure 6 are so large relative to the confidence intervals on the data (particularly in the author plots, which are quite precisely measured). We believe this highlights a particular strength of mixed modeling for psychophysical data (Cheung, Kallie, Legge, & Cheong, 2008; Knoblauch & Maloney, 2012; Moscatelli, Mezzetti, & Lacquaniti, 2012): Multiple sources of variability can be accounted for and incorporated into predictions at various levels (e.g., the observer and image level, or the subject level ignoring images, or the population level). In this case, averaging over the images and displaying credible intervals that ignore the pairwise experimental design (as in Figure 6) disguises the fact that different images show distinctly different effects of image model and presentation time.

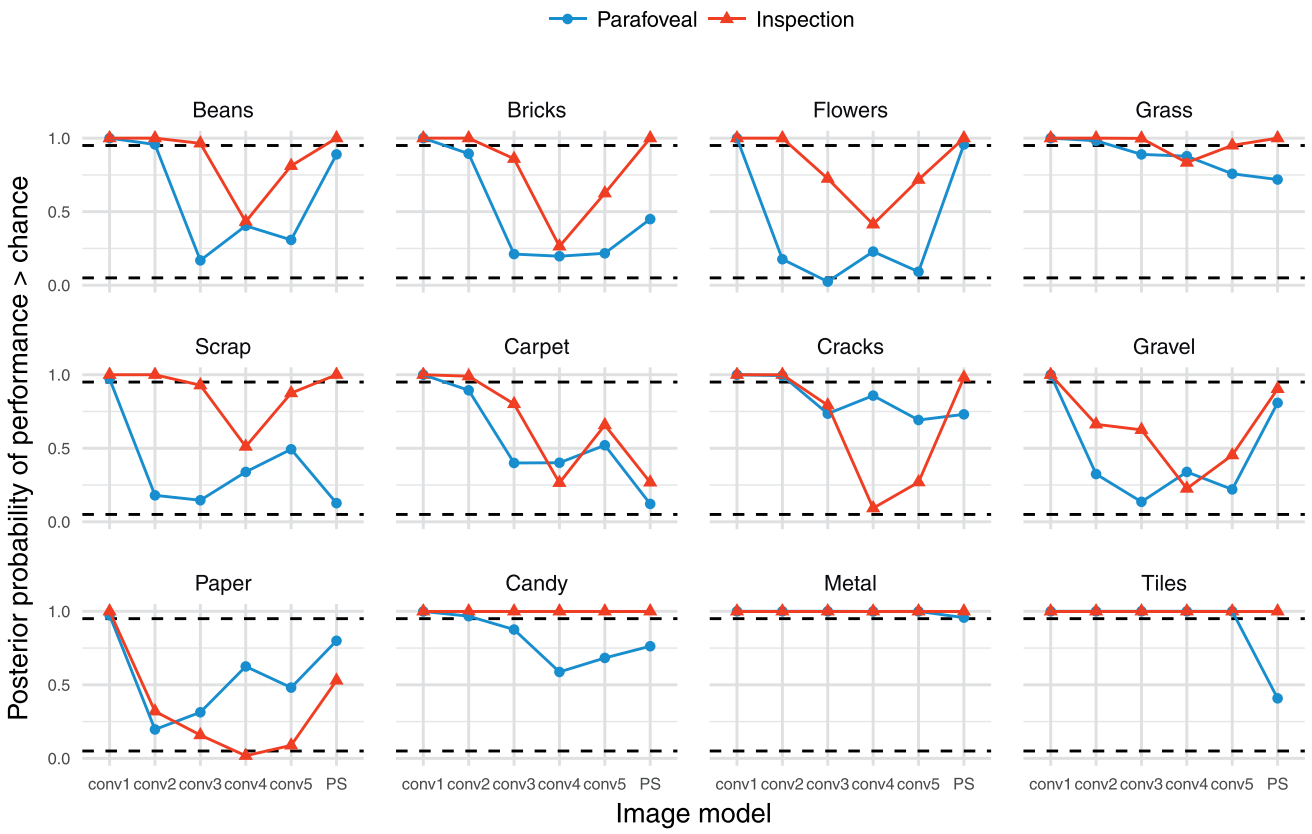


Figure 12. Posterior probability that performance for discriminating each image and image model in Experiment 1 lies above chance. Conditions falling above the dashed horizontal line at 0.95 have a greater than 95% probability of being discriminable, conditions falling below the dashed horizontal line at 0.05 are more than 95% likely to be below-chance. Conditions for which the model predicts exactly chance performance would fall at 0.5.

For example, for each fixed-effect coefficient we can ask whether more variance in the data is caused by variation over observers or images. On average, the variance associated with images is 2.1 times greater than that associated with observers. The linear predictor difference between PS and conv5 averaged over presentation condition is associated with about 3.3 times more variance from images than from observers. That is, this difference tends to depend strongly on the image (Figure 7). The model uncertainties in Figure 6 are large because the "average" or population-level behavior is uncertain in light of this; indeed, it may make little sense to talk about a "population level" over images from these data. In contrast, Figure 7 shows model estimates that are far more constrained relative to Figure 6, because the uncertainty in the estimates now reflects between-subject variability rather than between-image variability.

Chance performance in the oddity task indicates the original and synthesized images are not discriminable from each other. To what degree do our data suggest observers perform above chance for each image and viewing condition? One way to quantify this is to compute the proportion of posterior probability density lying above chance performance. This estimates, for every condition, the probability of observers being sensitive to the difference between original and synthesized textures. Conditions that lie above the dashed horizontal line are those for which we can be more than 95% certain (conditional on model and priors) that observers are sensitive to the difference between original and synthesized images. These dashed lines are provided as a guide rather than to encourage dichotomous decision making about "different or not." The posterior probabilities confirm, in general, our qualitative statements made in the manuscript (Figure 12).

Experiment 2

The results of Experiment 2 for the conv5 and PS models replicate the results of Experiment 1. When stimuli are presented briefly to the parafovea, observers are effectively at chance to discriminate both conv5 and

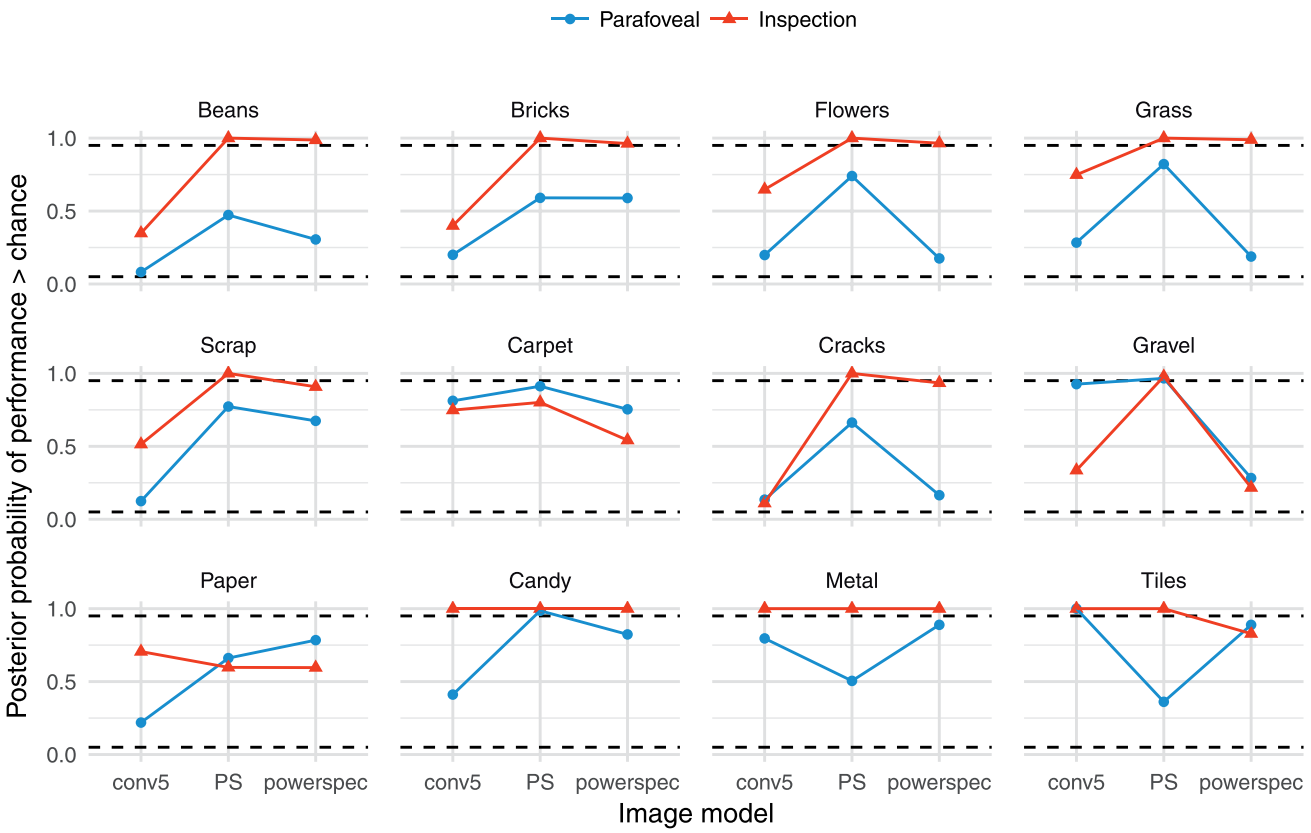


Figure 13. Posterior probability that performance for each image in Experiment 2 lies above chance. Plot elements as in Figure 12.

PS from the original textures, and there was evidence that the models did not differ, $\beta = 0.15$, 95% CI = [−0.4, 0.71], $p(\beta < 0) = 0.275$, whereas under inspection the PS model was easier to discriminate from the original images than the conv5 model, $\beta = 1.45$, 95% CI = [0.48, 2.41], $p(\beta < 0) = 0.003$. Additionally matching the power spectrum ("powerspec" model) produced similarly indistinguishable performance from the PS model in the parafovea, $\beta = -0.13$, 95% CI = [−0.63, 0.36], $p(\beta < 0) = 0.712$, but better performance than the PS model under inspection, $\beta = -1.64$, 95% CI = [−2.35, −0.95], $p(\beta < 0) > 0.999$.

Posterior probabilities that performance lies above chance for each image and viewing condition are shown in Figure 13. As for Experiment 1, these values generally support our qualitative statements made in the manuscript.

Performance as a function of oddball type

Consider that some data points appear to be reliably *below* chance performance (see for example the conv3 model in the flowers image). Below-chance performance in a forced-choice task generally only occurs in observed data due to measurement error or to observers incorrectly switching responses. However, in our experiments, it is also possible that below-chance performance could be caused in part by cropping from inhomogeneous images. For example, the original flowers image (Figure 2) contains a size gradient such that flowers on the bottom are larger and sparser than flowers on the top of the image, and this size gradient may result in greater inhomogeneity in the synthesized textures. More generally it may be the case that performance will depend on the relative (in)homogeneity of the original or synthesized images.

To investigate this further we computed performance for trials where the oddball image was an original compared to a model synthesis. When averaging over observers and images (Figure 14), performance is generally slightly higher if the oddball image is a model synthesis rather than an original image. The size of this effect depends on the particular image. For example, in the parafoveal viewing condition (Figure 15) the advantage for synthetic oddballs is quite strong for metal and tiles. Similarly, under inspection (Figure 16) observers remain highly sensitive to oddball candy and tile syntheses, whereas their performance is relatively poor when the oddball is an original image. This seems

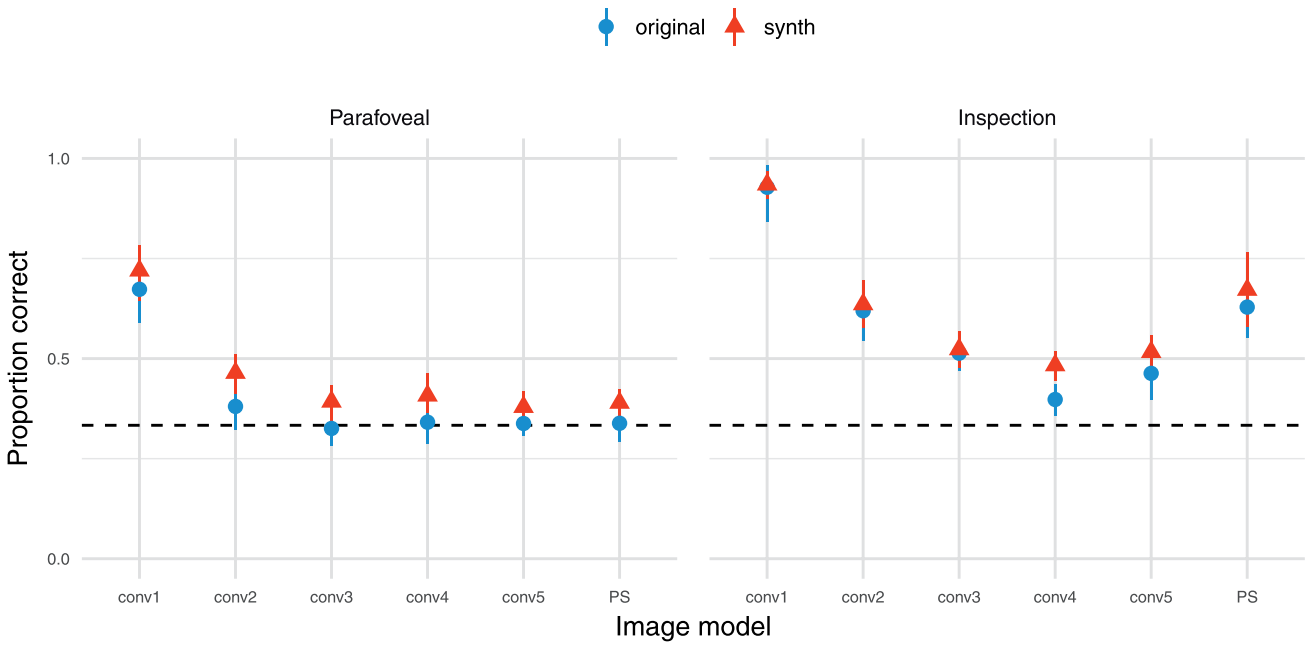


Figure 14. Performance in Experiment 1 according to whether the oddball image was an original or a model synthesis ("synth"), averaging over images. Points show grand mean across observer means, error bars show SEM.

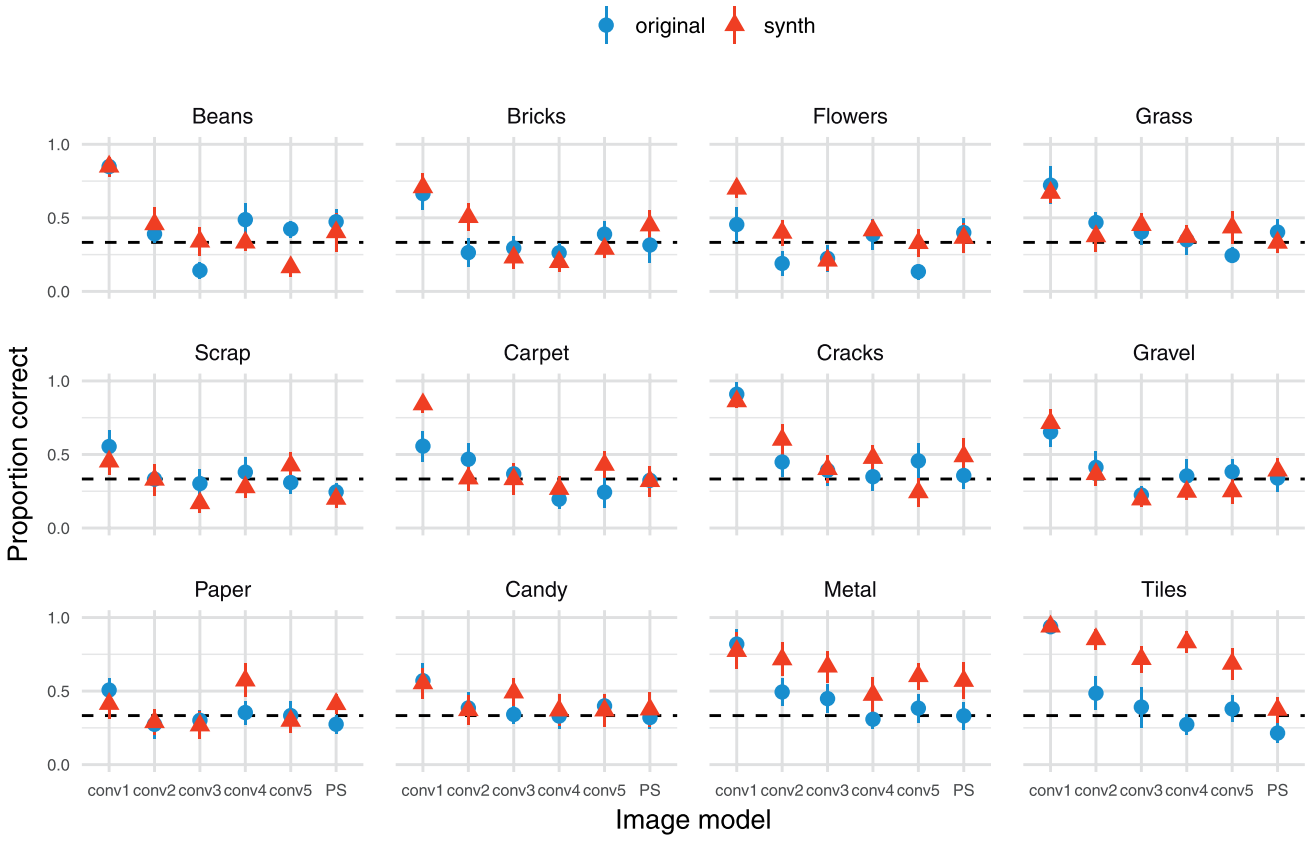


Figure 15. Parafoveal performance in Experiment 1 according to whether the oddball image was an original or a model synthesis ("synth"), for each image. Points show grand mean across observer means, error bars show SEM.

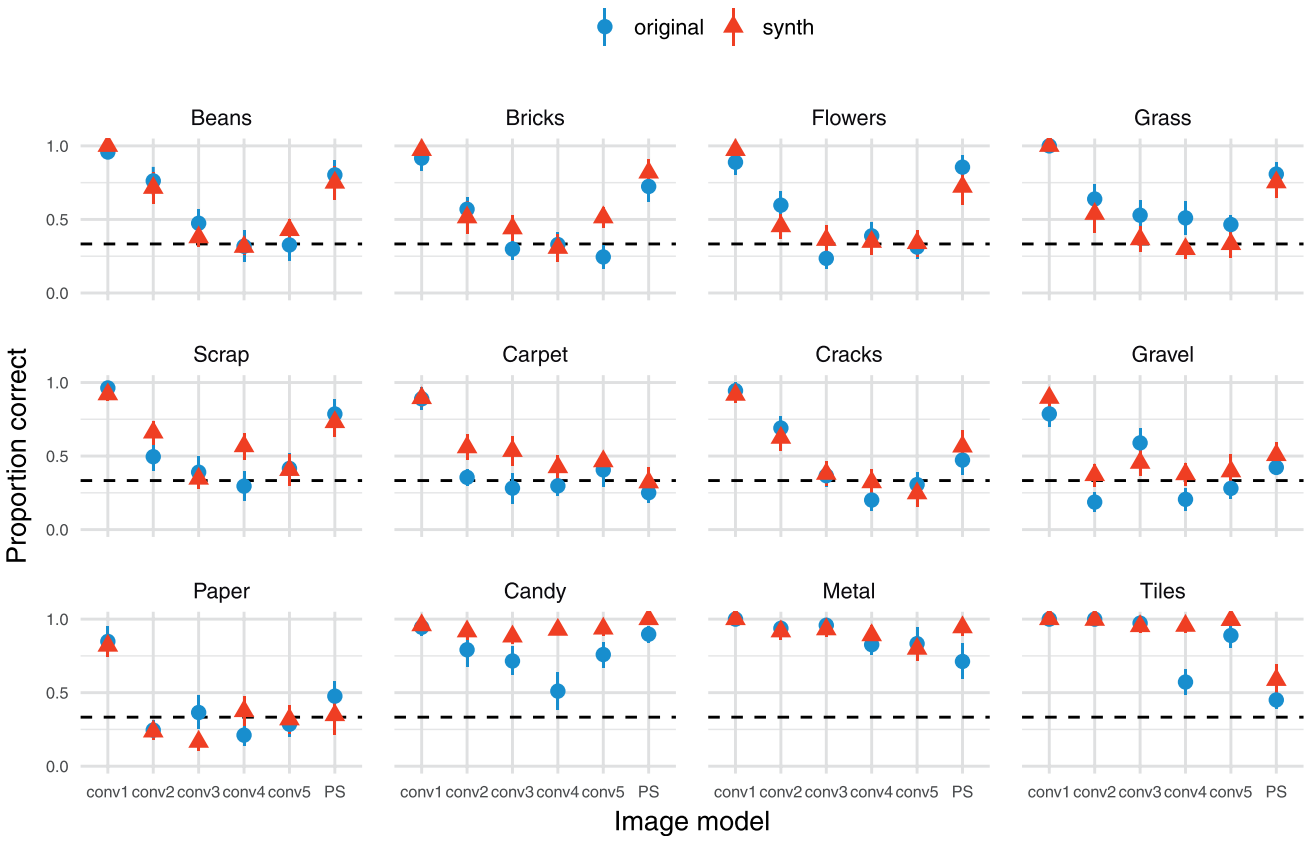


Figure 16. Inspection performance in Experiment 1 according to whether the oddball image was an original or a model synthesis ("synth"), for each image. Points show grand mean across observer means, error bars show SEM.

particularly strong for the conv4 model, explaining the lower average performance under this model condition.

These differences according to oddball type are generally consistent with the perceptual variability account above. If crops from the synthesized images appear different to each other and to the original, but crops from the original are quite self-similar, then on trials with an original oddball each of the three images looks different to the others. One of the synthesized images may appear "most different" (Figure 17a), and the observer incorrectly chooses that. Conversely, on trials where the synthesized image is the oddball, the two intervals containing the original images look similar to each other but different to the synthesized image (Figure 17b), making the task easier. This perceptual variability explanation is particularly appealing for images where the model fails to match appearance, such as for candy, metal, and tiles, and is also consistent with the larger self-similarity of those images (Figure 11). Other, not mutually exclusive possibilities include that observers are influenced by nonperceptual factors, such as the use of a suboptimal decision strategy ("pick the unnatural-looking image") on some trials, or of exogenous orienting of spatial attention to unnatural images. Whatever the cause(s) of the oddball differences we observe, note that traditional observer models for the oddity paradigm assume both unbiased responding and that the stimulus classes have equal variance (Macmillan & Creelman, 2005, p. 235); thus, computing *d'* from our data with the intention of comparing sensitivity to other paradigms should be performed cautiously or with a model explicitly including bias/variance terms for each trial type.

Loss

For the stimuli used in this study, the CNN texture models conv4, conv5, and powerspec are overcomplete (have more parameters than pixels in the image). Thus the loss of the gradient descent for those models does not converge to zero, but ends in a local minimum. Figure 18A shows a typical convergence function, where the gradient descent for conv1 terminates early (after reaching convergence within tolerance) but for more complex models (conv3–conv5) loss appears to find a local minimum, remaining relatively stable after 750 iterations. The final loss after 1,000 iterations is

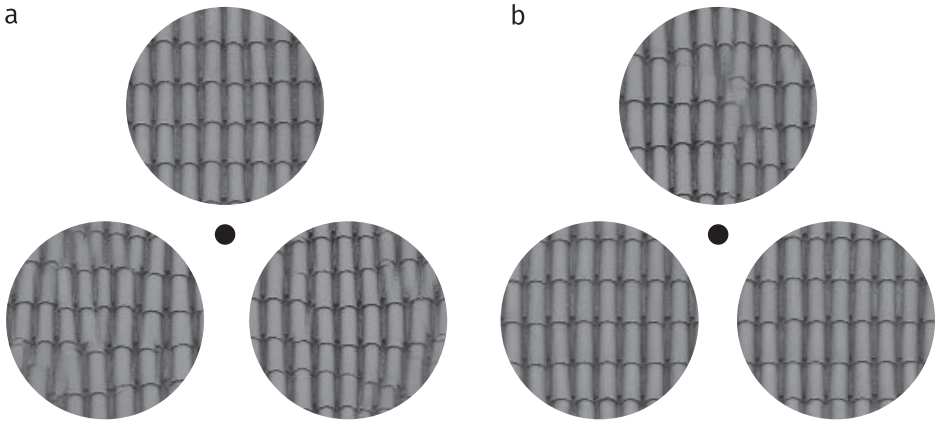


Figure 17. A depiction of an oddball "original" trial (a) and an oddball "synth" trial (b). In both cases the oddball is the top image. All images are physically different. When model syntheses look different to the original and each other, and the original images are very self-similar, then the perceptual variability of all stimulus intervals is larger on oddball original than oddball synth trials.

superlinear (Figure 18B): for example, conv5 has a little less than double the number of parameters as conv4, but about 23 times higher final loss.

Given that we interleaved 10 unique syntheses for each original image within our experiment, it would be interesting to assess whether a correlation exists between the final loss of each synthesis and psychophysical performance. A positive correlation between loss and performance would mean that images that show greater difference to the original under the model would also be easier for humans to discriminate. Unfortunately however, we did not save the final loss of the images after gradient descent but prior to histogram matching. Because histogram matching substantially alters the loss values under the model, including changing the order of syntheses, we are unable to assess a correlation between performance and final loss in this dataset.

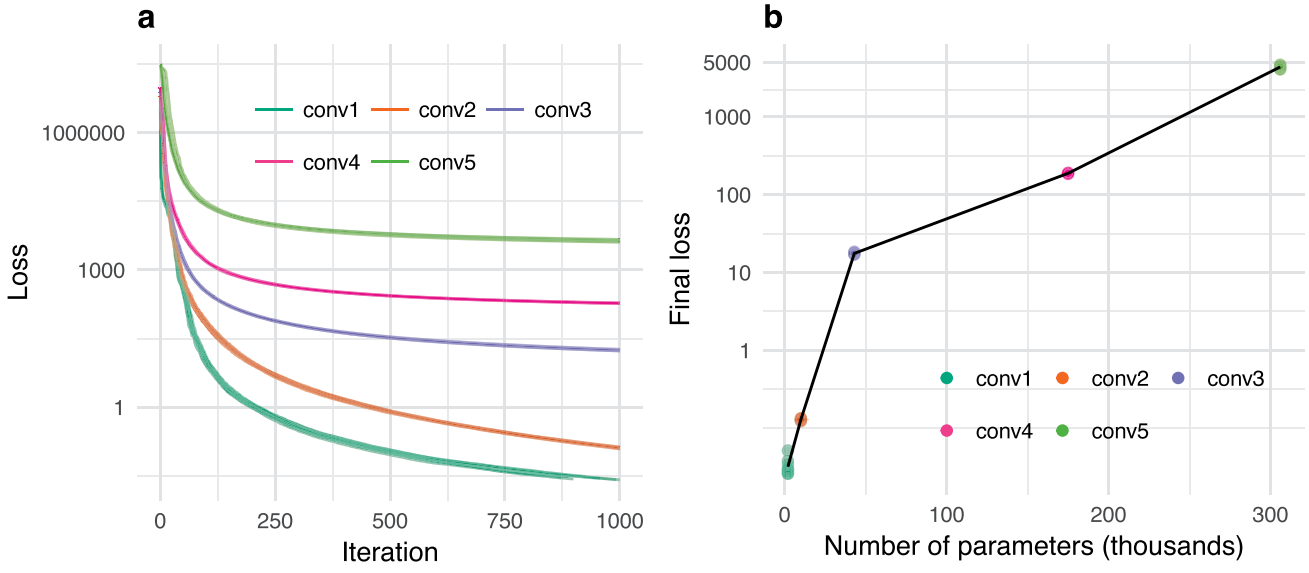


Figure 18. (a) Decrease of the loss over iterations on a logarithmic scale for ten syntheses (lines) of one example image (Bricks). Loss for simple models (e.g., conv1) converges to zero whereas for more complex models (conv3, conv4 and conv5) it stabilizes in a local minimum. (b) Final loss (logarithmic scale) for the synthesized images in (a) as a function of number of parameters in the model. Points show individual syntheses, lines link means within a model. Final loss is superlinear in the number of parameters.

P3: Image content is more important than Bouma's Law for scene metamers

Thomas S.A. Wallis*, Christina M. Funke*, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, Matthias Bethge

* joint first authors

Published in *ELife,8, e42512*.

Contributions Thomas S.A. Wallis, Conceptualization, Resources, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing — original draft, Project administration, Writing — review and editing; Christina M. Funke, Conceptualization, Resources, Software, Formal analysis, Validation, Investigation, Methodology, Writing — original draft, Writing—review and editing; Alexander S. Ecker, Conceptualization, Supervision, Methodology, Project administration, Writing—review and editing; Leon A. Gatys, Resources, Software, Methodology, Writing — review and editing; Felix A. Wichmann, Conceptualization, Supervision, Funding acquisition, Methodology, Project administration, Writing — review and editing; Matthias Bethge, Conceptualization, Supervision, Funding acquisition, Project administration, Writing — review and editing.

Parts of this publication were subject of my master thesis. In particular, this was the case for the foveated DNN texture model presented in the appendix, the method to fill in texturized local patches as well as some of the preliminary experiments. However, my contributions to the ideas and experiments that make up the bulk of this publication can be attributed to my doctoral studies.

eLIFE elifesciences.org

SHORT REPORT

Image content is more important than Bouma's Law for scene metamers

Thomas SA Wallis[1,2†]*, Christina M Funke[1,2†], Alexander S Ecker[1,2,3,4], Leon A Gatys[1‡], Felix A Wichmann[5], Matthias Bethge[3,4,6]

[1]Werner Reichardt Center for Integrative Neuroscience, Eberhard Karls Universität Tübingen, Tübingen, Germany; [2]Bernstein Center for Computational Neuroscience, Berlin, Germany; [3]Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, United States; [4]Institute for Theoretical Physics, Eberhard Karls Universität Tübingen, Tübingen, Germany; [5]Neural Information Processing Group, Faculty of Science, Eberhard Karls Universität Tübingen, Tübingen, Germany; [6]Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract We subjectively perceive our visual field with high fidelity, yet peripheral distortions can go unnoticed and peripheral objects can be difficult to identify (crowding). Prior work showed that humans could not discriminate images synthesised to match the responses of a mid-level ventral visual stream model when information was averaged in receptive fields with a scaling of about half their retinal eccentricity. This result implicated ventral visual area V2, approximated 'Bouma's Law' of crowding, and has subsequently been interpreted as a link between crowding zones, receptive field scaling, and our perceptual experience. However, this experiment never assessed natural images. We find that humans can easily discriminate real and model-generated images at V2 scaling, requiring scales at least as small as V1 receptive fields to generate metamers. We speculate that explaining why scenes look as they do may require incorporating segmentation and global organisational constraints in addition to local pooling.

DOI: https://doi.org/10.7554/eLife.42512.001

***For correspondence:** thomas.wallis@uni-tuebingen.de

[†]These authors contributed equally to this work

Present address: [‡]Apple Inc, Cupertino, United States

Competing interest: See page 16

Funding: See page 16

Received: 03 October 2018
Accepted: 09 March 2019
Published: 30 April 2019

Reviewing editor: Michael Herzog, EPFL, Switzerland

Introduction

Vision science seeks to understand why things look as they do (***Koffka, 1935***). Typically, our entire visual field looks subjectively crisp and clear. Yet our perception of the scene falling onto the peripheral retina is actually limited by at least three distinct sources: the optics of the eye, retinal sampling, and the mechanism(s) giving rise to crowding, in which our ability to identify and discriminate objects in the periphery is limited by the presence of nearby items (***Bouma, 1970***; ***Pelli and Tillman, 2008***). Many other phenomena also demonstrate striking 'failures' of visual perception, for example change blindness (***Rensink et al., 1997***; ***O'Regan et al., 1999***) and inattentional blindness (***Mack and Rock, 1998***), though there is some discussion as to what extent these are distinct from crowding (***Rosenholtz, 2016***). Whatever the case, it is clear that we can be insensitive to significant changes in the world despite our rich subjective experience.

Visual crowding has been characterised as compulsory texture perception (***Parkes et al., 2001***; ***Lettvin, 1976***) and compression (***Balas et al., 2009***; ***Rosenholtz et al., 2012a***). This idea entails that we cannot perceive the precise structure of the visual world in the periphery. Rather, we are aware only of some set of summary statistics or ensemble properties of visual displays, such as the average size or orientation of a group of elements (***Ariely, 2001***; ***Dakin and Watt, 1997***). One of the appeals of the summary statistic idea is that it can be directly motivated from the perspective of efficient coding as a form of compression. Image-computable texture summary statistics have been shown to be correlated with human performance in various tasks requiring the judgment of peripheral

eLife digest As you read this digest, your eyes move to follow the lines of text. But now try to hold your eyes in one position, while reading the text on either side and below: it soon becomes clear that peripheral vision is not as good as we tend to assume. It is not possible to read text far away from the center of your line of vision, but you can see 'something' out of the corner of your eye. You can see that there is text there, even if you cannot read it, and you can see where your screen or page ends. So how does the brain generate peripheral vision, and why does it differ from what you see when you look straight ahead?

One idea is that the visual system averages information over areas of the peripheral visual field. This gives rise to texture-like patterns, as opposed to images made up of fine details. Imagine looking at an expanse of foliage, gravel or fur, for example. Your eyes cannot make out the individual leaves, pebbles or hairs. Instead, you perceive an overall pattern in the form of a texture. Our peripheral vision may also consist of such textures, created when the brain averages information over areas of space.

Wallis, Funke et al. have now tested this idea using an existing computer model that averages visual input in this way. By giving the model a series of photographs to process, Wallis, Funke et al. obtained images that should in theory simulate peripheral vision. If the model mimics the mechanisms that generate peripheral vision, then healthy volunteers should be unable to distinguish the processed images from the original photographs. But in fact, the participants could easily discriminate the two sets of images. This suggests that the visual system does not solely use textures to represent information in the peripheral visual field. Wallis, Funke et al. propose that other factors, such as how the visual system separates and groups objects, may instead determine what we see in our peripheral vision.

This knowledge could ultimately benefit patients with eye diseases such as macular degeneration, a condition that causes loss of vision in the center of the visual field and forces patients to rely on their peripheral vision.

DOI: https://doi.org/10.7554/eLife.42512.002

information, such as crowding and visual search (***Rosenholtz et al., 2012a***; ***Balas et al., 2009***; ***Freeman and Simoncelli, 2011***; ***Rosenholtz, 2016***; ***Ehinger and Rosenholtz, 2016***). Recently, it has even been suggested that summary statistics underlie our rich phenomenal experience itself—in the absence of focussed attention, we perceive only a texture-like visual world (***Cohen et al., 2016***).

Across many tasks, summary statistic representations seem to capture aspects of peripheral vision when the scaling of their pooling regions corresponds to 'Bouma's Law' (***Rosenholtz et al., 2012a***; ***Balas et al., 2009***; ***Freeman and Simoncelli, 2011***; ***Wallis and Bex, 2012***; ***Ehinger and Rosenholtz, 2016***). Bouma's Law states that objects will crowd (correspondingly, statistics will be pooled) over spatial regions corresponding to approximately half the retinal eccentricity (***Bouma, 1970***; ***Pelli and Tillman, 2008***; though see ***Rosen et al., 2014***). While the precise value of Bouma's law can vary substantially even over different visual quadrants within an individual (see e.g. ***Petrov and Meleshkevich, 2011***), we refer here to the broader notion that summary statistics are pooled over an area that increases linearly with eccentricity, rather than the exact factor of this increase (the exact factor becomes important in the paragraph below). If the visual system does indeed represent the periphery using summary statistics, then Bouma's scaling implies that as retinal eccentricity increases, increasingly large regions of space are texturised by the visual system. If a model captured these statistics and their pooling, and the model was amenable to being run in a generative mode, then images could be created that are indistinguishable from the original despite being physically different (metamers). These images would be equivalent to the model and to the human visual system (***Freeman and Simoncelli, 2011***; ***Wallis et al., 2016***; ***Portilla and Simoncelli, 2000***; ***Koenderink et al., 2017***).

Freeman and Simoncelli (2011) developed a model (hereafter, FS-model) in which texture-like summary statistics are pooled over spatial regions inspired by the receptive fields in primate visual cortex. The size of neural receptive fields in ventral visual stream areas increases as a function of retinal eccentricity, and as one moves downstream from V1 to V2 and V4 at a given eccentricity. Each

visual area therefore has a signature scale factor, defined as the ratio of the receptive field diameter to retinal eccentricity (**Freeman and Simoncelli, 2011**). Similarly, the pooling regions of the FS-model also increase with retinal eccentricity with a definable scale factor. New images can be synthesised that match the summary statistics of original images at this scale factor. As scale factor increases, texture statistics are pooled over increasingly large regions of space, resulting in more distorted synthesised images relative to the original (that is, more information is discarded).

The maximum scale factor for which the images remain indistinguishable (the critical scale) characterises perceptually-relevant compression in the visual system's representation. If the scale factor of the model corresponded to the scaling of the visual system in the responsible visual area, and information in upstream areas was irretrievably lost, then the images synthesised by the model should be indistinguishable while discarding as much information as possible. That is, we seek the maximum compression that is perceptually lossless:

$$s_{\mathrm{crit}}(I) = \max_{s:\, \mathrm{d}(\hat{I}_s, I)=0} s,$$

where $s_{\mathrm{crit}}(I)$ is the critical scale for an image I, $\hat{I}_s$ is a synthesised image at scale s and d is a perceptual distance. Larger scale factors discard more information than the relevant visual area and therefore the images should look different. Smaller scale factors preserve information that could be discarded without any perceptual effect.

Crucially, it is the *minimum* critical scale over images that is important for the scaling theory. If the visual system computes summary statistics over fixed (image-independent) pooling regions in the same way as the model, then the model must be able to produce metamers for all images. While images may vary in their individual critical scales, the image with the smallest critical scale determines the maximum compression for appearance to be matched by the visual system in general, assuming an image-independent representation:

$$s_{\mathrm{system}} = \min_{I} s_{\mathrm{crit}}(I)$$

Freeman and Simoncelli showed that the largest scale factor for which two synthesised images could not be told apart was approximately 0.5, or pooling regions of about half the eccentricity. This scaling matched the signature of area V2, and also matched the approximate value of Bouma's Law. Subsequently, this result has been interpreted as demonstrating a link between receptive field scaling, crowding, and our rich phenomenal experience (e.g. **Block, 2013**; **Cohen et al., 2016**, **Landy, 2013**, **Movshon and Simoncelli, 2014**, **Seth, 2014**). These interpretations imply that the FS-model creates metamers for natural scenes. However, observers in Freeman and Simoncelli's experiment never saw the original scenes, but only compared synthesised images to each other. Showing that two model samples are indiscriminable from each other could yield trivial results. For example, two white noise samples matched to the mean and contrast of a natural scene would be easy to discriminate from the scene but hard to discriminate from each other. Furthemore, since synthesised images represent a specific subset of images, and the system critical scale s_{system} is the minimum over all possible images, the s_{system} estimated in **Freeman and Simoncelli (2011)** is likely to be an overestimate.

No previous paper has estimated s_{system} for the FS-model using natural images. **Wallis et al., 2016** tested the related **Portilla and Simoncelli (2000)** model textures, and found that observers could easily discriminate these textures from original images in the periphery. However, the Portilla and Simoncelli model makes no explicit connection to neural receptive field scaling. In addition, relative to the textures tested by **Wallis et al., 2016**, the pooling region overlap used in the FS-model provides a strong constraint on the resulting syntheses, making the images much more similar to the originals. It is therefore still possible that the FS-model produces metamers for natural scenes for scale factors of 0.5.

Results

Measuring critical scale in the FS-model

We tested whether the FS-model can produce metamers using an oddity design in which the observer had to pick the odd image out of three successively shown images (***Figure 1E***). In a three-alternative oddity paradigm, performance for metamerism would lie at 1/3 (dashed horizontal line, ***Figure 1F***). We used two comparison conditions: either observers compared two model syntheses to each other (synth vs synth; as in ***Freeman and Simoncelli, 2011***) or the original image to a model synthesis (orig vs synth). As in the original paper (***Freeman and Simoncelli, 2011***) we measured the performance of human observers for images synthesised with different scale factors (using Freeman and Simoncelli's code, see Materials and methods). To quantify the critical scale factor we fit the same nonlinear model as Freeman and Simoncelli, which parameterises sensitivity as a function of critical scale and gain, but using a mixed-effects model with random effects of participant and image (see Materials and methods).

We used 20 images to test the FS model. These images are split into two classes of ten images each, which we labelled 'scene-like' and 'texture-like'. The distinction of these two classes is based on the results of a pilot experiment with a model we developed, which is inspired by the FS model but based on a different set of image features (those extracted by a convolutional neural network; see Materials and methods and ***Appendix 2—figure 1***). In this pilot experiment, we found that some images are easier to discriminate than others (***Appendix 2—figure 7—figure 9***). Easily-discriminable images tended to contain larger areas of inhomogenous structure, long edges, borders between different surfaces or objects, and angled edges providing perspective cues ('scene-like'). Difficult images tended to contain more visual textures: homogenous structure, patterned content, or materials ('texture-like'"). For example, images from the first class tended to contain more structure such as faces, text, skylines, buildings, and clearly segmented objects or people, whereas images from the second class tended to contain larger areas of visual texture such as grass, leaves, gravel, or fur. A similar distinction could also be made along the lines of 'human-made' versus 'natural' image structure, but we suspect the visual structure itself rather than its origin is of causal importance and so used that level of description.

While our labelling of images in this way is debatable (for example, 'texture-like' regions contain some 'scene-like' content and vice versa) and to some degree based on subjective judgment, we hypothesised that this classification distinguishes the types of image content that are critical. If the visual system indeed created a texture-like summary in the periphery and the FS-model was a sufficient approximation of that process, then we should observe no difference in the average critical scale factor of images in each group (because image content would be irrelevant to the distribution of $s_{\mathrm{crit}}(I)$).

We start by considering the condition where participants compared synthesised images to each other—as in ***Freeman and Simoncelli (2011)***. Under this condition, there was little evidence that the critical scale depended on the image content (see curves in ***Figure 1F***, synth vs synth). The critical scale (posterior mean with 95% credible interval quantiles) for scene-like images was 0.28, 95% CI [0.21, 0.36] and the critical scale for texture-like images was 0.37, 95% CI [0.27, 0.5] (***Figure 1G***). Though these critical scales are lower than those reported by ***Freeman and Simoncelli (2011)***, they are within the range of other reported critical scale factors (***Freeman and Simoncelli, 2013***). There was weak evidence for a difference in critical scale between texture-like and scene-like images, with the posterior distribution of scale differences being 0.09, 95% CI [−0.03, 0.24], $\mathrm{p}(\beta<0)=0.078$ (where $\mathrm{p}(\beta<0)$ is the posterior probability of the difference being negative; symmetrical posterior distributions centered on zero would have $\mathrm{p}(\beta<0)=0.5$). However, this evidence should be interpreted cautiously: because asymptotic performance never reaches high values, critical scale estimates are more uncertain than in the orig vs synth condition below (***Figure 1G***). This poor asymptotic performance may be because we used more images in our experiment than Freeman and Simoncelli, so participants were less familiar with the distortions that could appear. To make sure this difference did not arise due to different experimental paradigms (oddity vs. ABX), we repeated the experiment using the same ABX task as in Freeman and Simoncelli (***Appendix 1—figure 4***). This experiment again showed poor asymptotic performance, and furthermore demonstrated no evidence for a critical scale difference between the scene- and texture-like images. Taken together, our synth vs synth results are somewhat consistent with Freeman and Simoncelli, who

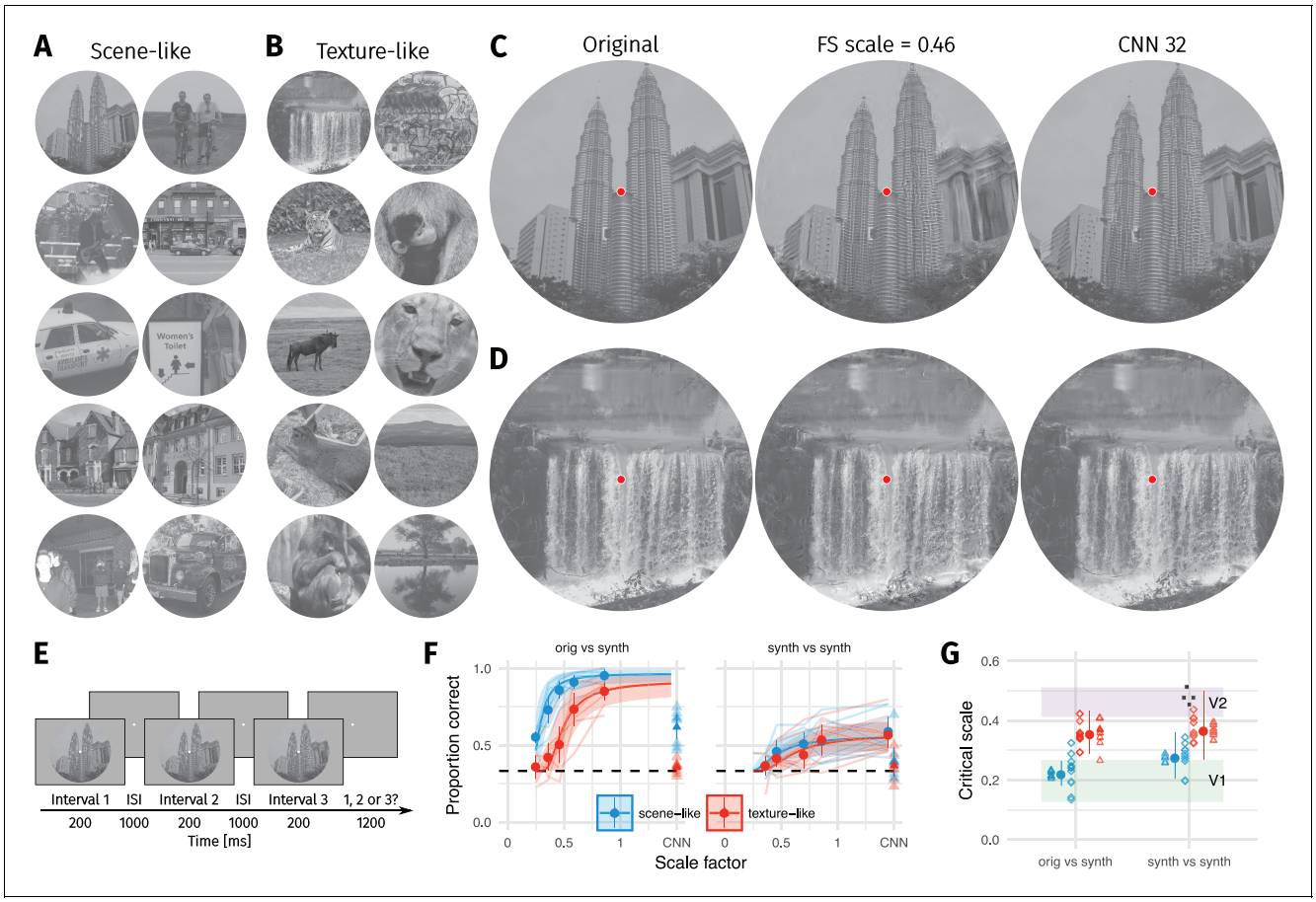


Figure 1. Two texture pooling models fail to match arbitrary scene appearance. We selected ten scene-like (**A**) and ten texture-like (**B**) images from the MIT 1003 dataset (***Judd et al., 2009***, https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) and synthesised images to match them using the Freeman and Simoncelli model (FS scale 0.46 shown) or a model using CNN texture features (CNN 32; example scene and texture-like stimuli shown in (**C**) and (**D**) respectively). Images reproduced under a CC-BY license (https://creativecommons.org/licenses/by/3.0/) with changes as described in the Methods. (**E**): The oddity paradigm. Three images were presented in sequence, with two being physically-identical and one being the oddball. Participants indicated which image was the oddball (1, 2 or 3). On 'orig vs synth' trials participants compared real and synthesised images, whereas on 'synth vs synth' trials participants compared two images synthesised from the same model. (**F**): Performance as a function of scale factor (pooling region diameter divided by eccentricity) in the Freeman-Simoncelli model (circles) and for the CNN 32 model (triangles; arbitrary x-axis location). Points show grand mean ±2 SE over participants; faint lines link individual participant performance levels (FS-model) and faint triangles show individual CNN 32 performance. Solid curves and shaded regions show the fit of a nonlinear mixed-effects model estimating the critical scale and gain. Participants are still above chance for scene-like images in the original vs synth condition for the lowest scale factor of the FS-model we could generate, and for the CNN 32 model, indicating that neither model succeeds in producing metamers. (**G**): When comparing original and synthesised images, estimated critical scales (scale at which performance rises above chance) are lower for scene-like than for texture-like images. Points with error bars show population mean and 95% credible intervals. Triangles show posterior means for participants; diamonds show posterior means for images. Black squares show critical scale estimates of the four participants from ***Freeman and Simoncelli (2011)*** (x-position jittered to reduce overplotting); shaded regions denote the receptive field scaling of V1 and V2 estimated by ***Freeman and Simoncelli (2011)***. Data reproduced from ***Freeman and Simoncelli (2011)*** using WebPlotDigitizer v. 4.0.0 (Rohatgi, A., software under the GNU Affero General Public License v3, https://www.gnu.org/licenses/agpl-3.0.en.html).

DOI: https://doi.org/10.7554/eLife.42512.003

The following figure supplement is available for figure 1:

Figure supplement 1. The ten scene-like and ten texture-like images used in our main experiments, along with example syntheses from the FS-0.46 and CNN 32 models (best viewed with zoom).

DOI: https://doi.org/10.7554/eLife.42512.004

reported no dependency of $s_{\mathrm{crit}(I)}$ on image. It seems likely that this is because comparing synthesised images to each other means that the model has removed higher-order structure that might allow discrimination. All images appear distorted, and the task becomes one of identifying a specific distortion pattern.

Comparing the original image to model syntheses yielded a different pattern of results. First, participants were able to discriminate the original images from their FS-model syntheses at scale factors of 0.5 (***Figure 1F***). Performance lay well above chance for all participants. This result held for both scene-like and texture-like images. Furthermore, there was evidence that critical scale depended on the image type. Model syntheses matched the texture-like images on average with scale factors of 0.36, 95% CI [0.29, 0.43]. In contrast, the scene-like images were quite discriminable from their model syntheses even at the smallest scale we could generate (0.25). The critical scale estimated for scene-like images was 0.22, 95% CI [0.18, 0.27]. Texture-like images had higher critical scales than scene-like images on average (scale difference = 0.13, 95% CI [0.06, 0.22], $\mathrm{p}(\beta<0)=0.001$).

This difference in critical scale was not attributable to differences in the success of the synthesis procedure between scene-like and texture-like images. Scene-like images had higher final loss (distance between the original and synthesised images in model space) than texture-like images on average (see Materials and methods). This is a corollary of the importance of image content: since a texture summary model is a poor description of scene-like content, the model's optimisation procedure is also more likely to find local minima with relatively high loss. We checked that our main result was not explained by this difference by performing a control analysis in which we refit the model after equating the average loss in the two groups by excluding images with highest final loss until the groups were matched (resulting in four scene-like images being excluded; see Materials and methods). The remaining scene-like images had a critical scale of 0.24, 95% CI [0.2, 0.28] in the orig vs synth condition, texture-like images again showed a critical scale of 0.36, 95% CI [0.3, 0.42] and the difference distribution had a mean of 0.12, 95% CI [0.06, 0.19], $\mathrm{p}(\beta<0)<0.001$. Thus, differences in synthesis loss do not explain our findings.

As noted above, the image with the minimum critical scale determines the largest compression that can be applied for the scaling model to hold (s_{system}). For two images (***Figure 2A and E***) the nonlinear mixed-effects model estimated critical scales of approximately 0.14 (see ***Figure 1G***, diamonds; the minimum critical scale after excluding high-loss images in the control analysis reported above was 0.19). However, examining the individual data for these images (***Figure 2D and H***) reveals that these critical scale estimates are largely determined by the hierarchical nature of the mixed-effects model, not the data itself. Both images were easy to discriminate from the original even for the lowest scale factor we could generate. This suggests that the true scale factor required to generate metamers may be even lower than estimated by the mixed-effects model.

Our results show that smaller pooling regions are required to make metamers for scene-like images than for texture-like images. Human observers can reliably detect relatively small distortions produced by the FS-model at scale factors of 0.25 in scene-like image content (compare ***Figure 2B and F*** at scale 0.25 and ***C*** and ***G*** at scale 0.46 to images ***A*** and ***B***). Thus, syntheses at these scales are not metamers for natural scenes.

Local image structure determines the visibility of texture-like distortions

In our first experiment we found that scene-like images yielded lower critical scales than texture-like images. However, this categorisation is crude: 'texture-ness' in photographs of natural scenes is a property of local regions of the image rather than the image as a whole. In addition, the classification of images above was based in part on the difficulty of these images in a pilot experiment.

We therefore ran a second experiment to test the importance of local image structure more directly (***Bex, 2010***; ***Koenderink et al., 2017***; ***Valsecchi et al., 2018***; ***Wallis and Bex, 2012***), using a set of images whose selection was not based on pilot discrimination results. Participants detected a localised texture-like distortion (generated by the texture model of ***Gatys et al., 2015***) blended into either a scene-like or texture-like region (***Figure 3A–C***). These image regions were classified by author CF (non-authors showed high agreement with this classification—see Materials and methods). The patches were always centered at an eccentricity of six degrees, and we varied the radius of the circular patch (***Figure 3D***). This is loosely analogous to creating summary statistics in a single pooling

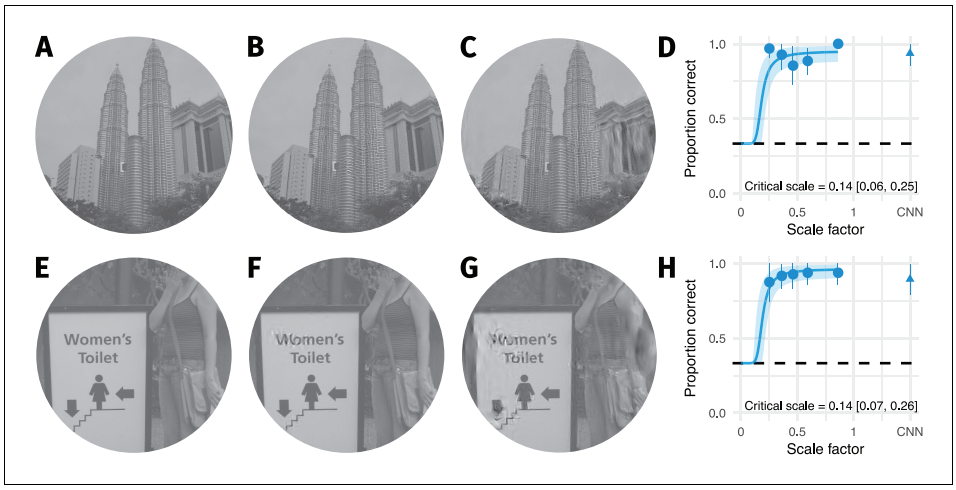


Figure 2. The two images with smallest critical scale estimates are highly discriminable even for the lowest scale factor we could generate. (**A**) The original image. (**B**) An example FS synthesis at scale factor 0.25. (**C**) An example FS synthesis at scale factor 0.46. Images in B and C reproduced from the MIT 1003 Database (***Judd et al., 2009***), https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) under a CC-BY license (https://creativecommons.org/licenses/by/3.0/) with changes as described in the Methods. (**D**) The average data for this image. Points and error bars show grand mean and ±2 SE over participants, solid curve and shaded area show posterior mean and 95% credible intervals from the mixed-effects model. Embedded text shows posterior mean and 95% credible interval on the critical scale estimate for this image. (**E–H**) Same as A–D for the image with the second-lowest critical scale. Note that in both cases the model is likely to overestimate critical scale.

DOI: https://doi.org/10.7554/eLife.42512.005

The following figure supplement is available for figure 2:

Figure supplement 1. Images with the highest and lowest critical scale estimates within the scene-like and texture-like categories for the orig vs synth comparison.

DOI: https://doi.org/10.7554/eLife.42512.006

region (***Wallis et al., 2016***). Participants discriminated between the original image and an image containing a local distortion in a 2IFC paradigm (***Figure 3E***).

The results showed that the visibility of texture-like distortions depended strongly on the underlying image content. Participants were quite insensitive to even large texture-like distortions occurring in texture-like image regions (***Figure 3F***). Performance for distortions of nearly five degrees radius (i.e. nearly entering the foveal fixation point) was still close to chance. Conversely, distorting scene-like regions is readily detectable for the three largest distortion patch sizes.

Discussion

It is a popular idea that the appearance of scenes in the periphery is described by summary statistic textures captured at the scaling of V2 neural populations. In contrast, here we show that humans are very sensitive to the difference between original and model-matched images at this scale (***Figure 1***). A recent preprint (***Deza et al., 2017***) finds a similar result in a set of 50 images, and our results are also consistent with the speculations made by Wallis et al. based on their experiments with Portilla and Simoncelli textures (***Wallis et al., 2016***). Together, these results show that the pooling of texture-like features in the FS-model at the scaling of V2 receptive fields does not explain the appearance of natural images.

One exciting aspect of ***Freeman and Simoncelli (2011)*** was the promise of inferring a critical brain region via a receptive field size prediction derived from psychophysics. Indeed, aspects of this promise have since received empirical support: the presence of texture-like features can discriminate V2 neurons from V1 neurons (***Freeman et al., 2013***; ***Ziemba et al., 2016***; see also ***Okazawa et al., 2015***). Discarding all higher-order structure not captured by the candidate model by comparing syntheses to each other, thereby isolating only features that change, may be a useful way to distinguish the feedforward component of sequential processing stages in neurons.

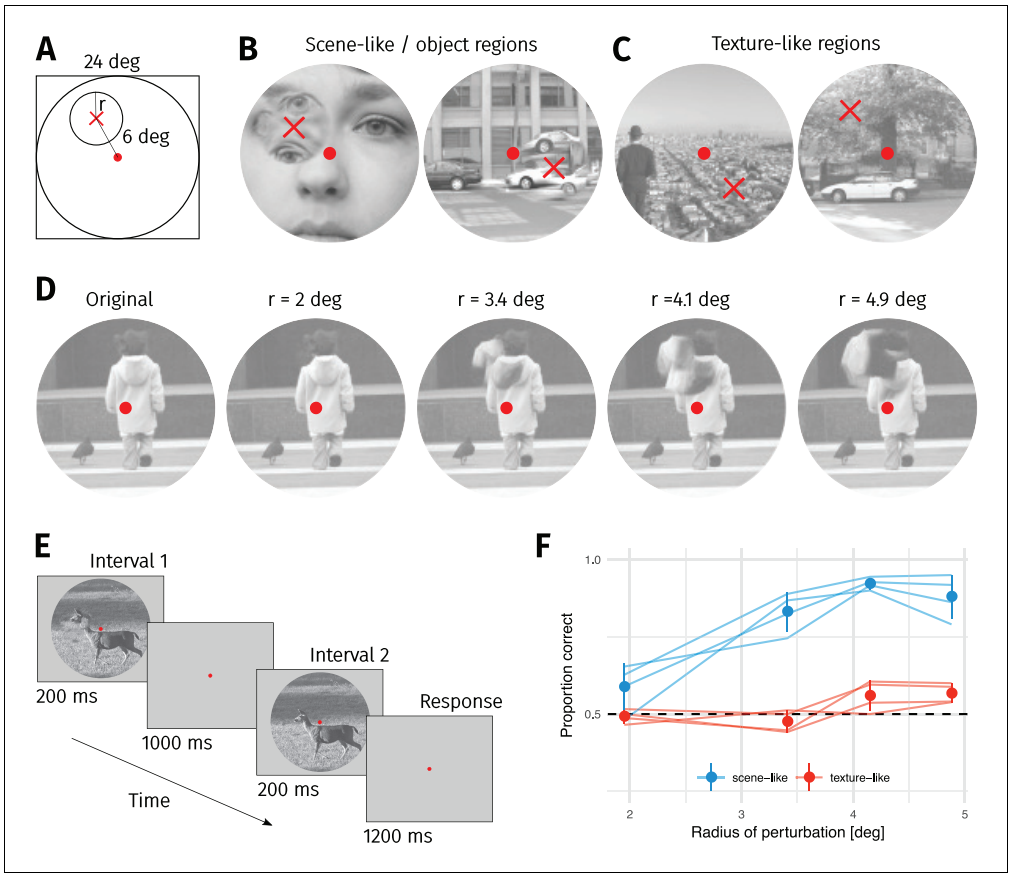


Figure 3. Sensitivity to local texture distortions depends on image content. (**A**) A circular patch of an image was replaced with a texture-like distortion. In different experimental conditions the radius of the patch was varied. (**B**) Two example images in which a 'scene-like' or inhomogenous region is distorted (red cross). (**C**) Two example images in which a 'texture-like' or homogenous region is distorted (red cross). (**D**) Examples of an original image and the four distortion sizes used in the experiment. Images in B–D reproduced from the MIT 1003 Database (***Judd et al., 2009***), https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) under a CC-BY license (https://creativecommons.org/licenses/by/3.0/) with changes as described in the Methods. (**E**) Depiction of the 2IFC task, in which the observer reported whether the first or second image contained the distortion. (**F**) Proportion correct as a function of distortion radius in scene-like (blue) and texture-like (red) image regions. Lines link the performance of each observer (each point based on a median of 51.5 trials; min 31, max 62). Points show mean of observer means, error bars show ±2 SEM.
DOI: https://doi.org/10.7554/eLife.42512.007

While texture-like representations may therefore be important for understanding neural encoding (***Movshon and Simoncelli, 2014***), our results call into question the link between receptive field scaling and scene appearance. If the peripheral appearance of visual scenes is explained by image-independent pooling of texture-like features, then the pooling regions must be small. Consider that participants in our experiment could easily discriminate the images in ***Figure 2B and F*** from those in ***Figure 2A and E*** respectively. Therefore, images synthesised at a truly metameric scaling must remain extremely close to the original: s_{system} must be at least as small as V1 neurons, and perhaps even lower (***Figure 2***). This may even be consistent with scaling in precortical visual areas. For example, the scaling of retinal ganglion cell receptive fields at the average eccentricity of our stimuli (six degrees) is approximately 0.08 for the surround (***Croner and Kaplan, 1995***) and 0.009 for the centre (***Dacey and Petersen, 1992***). It becomes questionable how much is learned about compression in the ventral pathway using such an approach, beyond the aforementioned, relatively well-studied limits of optics and retinal sampling (e.g. ***Wandell, 1995***; ***Watson, 2014***).

A second main finding from our paper is that the ability of the FS-model to synthesise visual metamers at a given scale factor depends on image content. Images containing predominantly

Figure 4. The visibility of texture-like distortions depends on image content. (**A**) 'Geotemporal Anomaly' by Pete Birkinshaw (2010: https://www.flickr.com/photos/binaryape/5203086981, re-used under a CC-BY 2.0 license: https://creativecommons.org/licenses/by/2.0/uk/). The image has been resized and a circular bullseye has been added to the centre. (**B**) Two texture-like distortions have been introduced into circular regions of the scene in A (see ***Figure 4—figure supplement 1*** for higher resolution). The distortion in the upper-left is quite visible, even with central fixation on the bullseye, because it breaks up the high-contrast contours of the text. The second distortion occurs on the brickwork centered on the bullseye, and is more difficult to see (you may not have noticed it until reading this caption). The visibility of texture-like distortions can depend more on image content than on retinal eccentricity (see also ***Figure 3***). (**C**) Results synthesised from the FS-model at scale 0.46 for comparison. Pooling regions depicted for one angular meridian as overlapping red circles; real pooling regions are smooth functions tiling the whole image. Pooling in this fashion reduces large distortions compared to B, but our results show that this is insufficient to match appearance.

DOI: https://doi.org/10.7554/eLife.42512.008

The following figure supplement is available for figure 4:

Figure supplement 1. Higher-resolution versions of the images from ***Figure 4***.

DOI: https://doi.org/10.7554/eLife.42512.009

'scene-like' content tended to be more difficult to match (requiring lower scale factors in the case of the FS-model) than images containing 'texture-like' content (***Figure 1F and G***). In a second experiment measuring the visibility of local texture distortions, we found that people can be quite insensitive to even large texture-like distortions so long as these fall on texture-like regions of the input image (***Figure 3***). This confirms the importance of the distinction between 'things' (scene-like content) and 'stuff' (texture-like content; ***Adelson, 2001***) for peripheral scene appearance.

This result can be experienced via simple demonstration. The 'China Lane' sign in ***Figure 4A*** has been distorted in ***Figure 4B*** (using local texture distortions as in ***Figure 3***), and is readily visible in the periphery (with central fixation on the circular bullseye). The same type of distortion in a texture-like region of the image is far less visible (the brickwork in the image centre; FS-model result ***Figure 4C***), despite appearing in the parafovea. It is the image content, not retinal eccentricity, that is the primary determinant of the visibility of at least some summary statistic distortions. Requiring information to be preserved at V1 or smaller scaling would therefore be inefficient from the standpoint of compression: small scale factors will preserve texture-like structure that could be compressed without affecting appearance.

It may seem trivial that a texture statistic model better captures the appearance of textures than non-textures. However, if the human visual system represents the periphery as a texture-like summary, and these models are sufficient approximations of this representation, then image content should not matter—because scene-like retinal inputs in the periphery are transformed into textures by the visual system.

Perhaps the V2 scaling theory holds but the FS-model texture features are insufficient to capture natural scene appearance. To test whether improved texture features (***Gatys et al., 2015***) could help in matching appearance for scenes, we developed a new model (CNN-model; see Materials

and methods and *Appendix 2—figures 1–4*) that was inspired by the FS-model but uses the texture features of a convolutional neural network (VGG-19, *Simonyan and Zisserman, 2015*) that have previously been shown to better capture the appearance of some textures than the Portilla and Simoncelli texture features (*Wallis et al., 2017*). As for the FS-model, discrimination performance becomes poorer as pooling region sizes become smaller (*Appendix 2—figure 3*). The CNN 32 model shows very similar behaviour to the FS-model such that human performance for scene-like images is higher than for texture-like images (triangles in *Figure 1D* and *Figure 2*). Thus, the syntheses from both models are not metamers for natural scenes. Nevertheless, our results cannot rule out that a hereto unknown summary statistic model exists that will create metamers for all images at V2 scales or higher. However, that two additional summary statistic models (the CNN-model and the NeuroFovea model of *Deza et al., 2017*) also fail to capture scene appearance and show dependence on image content adds some generality to our claim that these models are insufficient descriptions of peripheral visual scene appearance.

If this claim was correct, this begs the question: what is the missing ingredient that could capture appearance while compressing as much information as possible? Through the Gestalt tradition, it has long been known that the appearance of local image elements can crucially depend on the context in which they are placed and their interpretation in the scene (for overviews of recent work, see *Jäkel et al., 2016*; *Wagemans et al., 2012a*; *Wagemans et al., 2012b*). We speculate that mechanisms of perceptual organisation (such as segmentation and grouping) need to be considered if one wants to capture appearance in general—yet current models that texturise local regions do not explicitly include these mechanisms (*Herzog et al., 2015*; *Clarke et al., 2014*). If segmentation and grouping processes are critical for efficiently matching scene appearance, then uniformly computing summary statistics without including these processes will require preserving much of the original image structure by making pooling regions very small. A parsimonious model capable of compressing as much information as possible might need to adapt either the size and arrangement of pooling regions or the feature representations to the image content.

Local vs global mechanisms

These segmentation and grouping mechanisms could be mediated by local interactions between nearby image features, global properties of the scene, or both. The present results do not allow us to distinguish these alternatives.

In favour of the importance of local interactions, studies of contour integration in Gabor fields show that the arrangement of local orientation structure can influence the discrimination of contour shape (*Dakin and Baruch, 2009*) and contour localisation (*Robol et al., 2012*), and that these effects are consistent with crowding (*Robol et al., 2012*). In these stimuli, crowding between nearby contour elements is the primary determinant of global contour judgments (see also *Dakin et al., 2009*). Specifically, contours consisting of parallel Gabor elements ('snakes') were more easily perceived when adjacent Gabor elements were oriented perpendicularly to the main contour. A related study (*Van der Burg et al., 2017*) used an evolutionary algorithm to select dense line element displays that maximally alleviated crowding in an orientation discrimination task. Displays evolved using human responses showed that a substantial reduction of crowding was obtained by orienting the two line segments nearest the target (separated by only 0.75° at 6° eccentricity) to be perpendicular to the target's mean orientation (forming 'T' and/or 'I' junctions). In contrast, simulations based on Bouma's Law predicted that much larger areas of the display (relative to the human data) would need to be adjusted. These results are consistent with our finding that humans can be far more sensitive to image structure in the periphery than predicted by Bouma-like scaling.

The studies above suggest the possibility that T-junctions may be critical local cues to segmentation in the periphery. The potential importance of different junction types in segmentation and grouping has long been noted (*Biederman, 1987*). In real scenes, T-junctions usually signal occlusion edges between rigid surfaces, whereas Y-, L- and arrow-junctions are created by projecting the corners of 3D objects into 2D. Histograms of junction distributions are diagnostic of scene category (*Walther and Shen, 2014*), with human-made scenes such as city streets and offices tending to contain more T-junctions than more natural environments like beaches and mountains. A recent study also highlights the importance of local contour symmetry for scene categorisation (*Wilder et al., 2019*). Finally, *Loschky et al. (2010)* found that participants were extremely poor at classifying scene category from *Portilla and Simoncelli (2000)* global textures of scene images. These results suggest

that the Portilla and Simoncelli texture statistics (used in the FS-model) do not adequately preserve junction information.

Taken together, these studies give rise to the following hypothesis: images with more junctions (particularly T-junctions; ***Van der Burg et al., 2017***) will require smaller pooling regions to match and thus will show lower critical scale estimates in the FS-model. We applied the junction detection algorithm of ***Xia et al. (2014)*** to each of the 20 original images used in our first experiment. Consistent with the (post-hoc) hypothesis above, lower critical scales were associated with more frequent junctions, particularly if 'less meaningful' junctions (defined by the algorithm) were excluded (T-junction correlation $r = -0.54$; L-junctions $r = -0.63$; ***Appendix 1—figure 3***). If confirmed by a targeted experiment (and dissociated from general edge density), this relationship would suggest a clear avenue for future improvement of scene appearance models: they must successfully capture junction information in images.

Other evidence supports the role of global information (the arrangement and organisation of objects over large retinal areas) in segmentation and grouping. In crowding, ***Manassi et al. (2013)*** found that configurations of stimuli well outside the region of Bouma's law could modulate the crowding effectiveness of the same flankers (see also ***Manassi et al., 2012***; ***Saarela et al., 2009***; ***Vickery et al., 2009***; ***Levi and Carney, 2009***). ***Neri (2017)*** reported evidence from a variety of experiments in support of a fast segmentation process, operating over large regions of space, that can strongly modulate the perceptual interpretation of—and sensitivity to—local edge elements in a scene according to the figure-ground organisation of the scene (see also ***Teufel et al., 2018***). Our findings could be explained by the fact that the texture summary statistic models we examine here do not include any such global segmentation processes. The importance of these mechanisms could be examined in future studies, and potentially dissociated from the local information discussed above, by using image manipulations thought to disrupt the activity of global grouping mechanisms such as polarity inversion or image two-toning (***Neri, 2017***; ***Balas, 2012***; ***Teufel et al., 2018***).

Summary statistics, performance and phenomenology

Our results do not undermine the considerable empirical support for the periphery-as-summary-statistic theory as a description of visual performance. Humans can judge summary statistics of visual displays (***Ariely, 2001***; ***Dakin and Watt, 1997***), summary statistics can influence judgments where other information is lost (***Fischer and Whitney, 2011***; ***Faivre et al., 2012***), and the information preserved by summary statistic stimuli may offer an explanation for performance in various visual tasks (***Rosenholtz et al., 2012b***; ***Balas et al., 2009***; ***Rosenholtz et al., 2012a***; ***Keshvari and Rosenholtz, 2016***; ***Chang and Rosenholtz, 2016***; ***Zhang et al., 2015***; ***Whitney et al., 2014***; ***Long et al., 2016***; though see ***Agaoglu and Chung, 2016***; ***Herzog et al., 2015***; ***Francis et al., 2017***). Texture-like statistics may even provide the primitives from which form is constructed (***Lettvin, 1976***)—after appropriate segmentation, grouping and organisation. However, one additional point merits further discussion. The studies by Rosenholtz and colleagues primarily test summary statistic representations by showing that performance with summary statistic stimuli viewed foveally is correlated with peripheral performance with real stimuli. This means that the summary statistics preserve sufficient information to explain the performance of tasks in the periphery. Our results show that these summary statistics are insufficient to match scene appearance, at least under the pooling scheme used in the Freeman and Simoncelli model at computationally feasible scales. This shows the usefulness of scene appearance matching as a test: a parsimonious model that matches scene appearance would be expected to also preserve enough information to show correlations with peripheral task performance; the converse does not hold.

While it may be useful to consider summary statistic pooling in accounts of visual performance, to say that summary statistics can account for phenomenological experience of the visual periphery (***Cohen et al., 2016***; see also ***Block, 2013***; ***Seth, 2014***) seems premature in light of our results (see also ***Haun et al., 2017***). ***Cohen et al. (2016)*** additionally posit that focussed spatial attention can in some cases overcome the limitations imposed by a summary statistic representation. We instead find little evidence that participants' ability to discriminate real from synthesised images is improved by cueing spatial attention, at least in our experimental paradigm and for our CNN-model (***Appendix 2—figure 6***).

Conclusion

Our results show that the appearance of scenes in the periphery cannot be captured by the ***Freeman and Simoncelli (2011)*** summary statistic model at receptive field scalings similar to V2. We suggest that peripheral appearance models emphasising pooling processes that depend on retinal eccentricity will instead need to explore input-dependent grouping and segmentation. We speculate that mechanisms of perceptual organisation (either local or global) are critical to explaining visual appearance and efficient peripheral encoding. Models of the visual system that assume image content is processed in feedforward, fixed pooling regions—including current convolutional neural networks—lack these mechanisms.

Materials and methods

All stimuli, data and code to reproduce the figures and statistics reported in this paper are available at http://dx.doi.org/10.5281/zenodo.1475111. This document was prepared using the knitr package (***Xie, 2013***; ***Xie, 2016***) in the R statistical environment (***R Core Team, 2017***; ***Wickham and Francois, 2016***; ***Wickham, 2009***, ***Wickham, 2011***; ***Auguie, 2016***; ***Arnold, 2016***) to improve its reproducibility.

Participants

Eight observers participated in the first experiment (***Figure 1***): authors CF and TW, a research assistant unfamiliar with the experimental hypotheses, and five naïve participants recruited from an online advertisement pool who were paid 10 Euro per hr for two one-hour sessions. An additional naïve participant was recruited but showed insufficient eyetracking accuracy (see below). Four observers participated in the second experiment (***Figure 3***); authors CF and TW plus two naïve observers paid 10 Euro per hour. All participants signed a consent form prior to participating. Participants reported normal or corrected-to-normal visual acuity. All procedures conformed to Standard 8 of the American Psychological Association's 'Ethical Principles of Psychologists and Code of Conduct' (2010).

Stimuli

Images were taken from the MIT 1003 scene dataset (***Judd et al., 2012***; ***Judd et al., 2009***). A square was cropped from the center of the original image and downsampled to 512 $\times$ 512 px. The images were converted to grayscale and standardized to have a mean gray value of 0.5 (scaled [0,1]) and an RMS contrast (σ/μ) of 0.3. For the first experiment, images were selected as described in the Results and ***Appendix 2—figure 7—figure 9***.

Freeman and Simoncelli syntheses

We synthesised images using the FS-model (***Freeman and Simoncelli, 2011***, code available from https://github.com/freeman-lab/metamers). Four unique syntheses were created for each source image at each of eight scale factors (0.25, 0.36, 0.46, 0.59, 0.7, 0.86, 1.09, 1.45), using 50 gradient steps as in Freeman and Simoncelli's main experiment. Pilot experiments with stimuli generated with 100 gradient steps produced similar results. ***Freeman and Simoncelli (2011)*** computed the final loss between original and synthesised images as 'mean squared error, normalized by the parameter variance'. We take this to mean the following: for a matrix of model parameters from an original image $\mathrm{X}_{\mathrm{orig}}$ (rows are parameters and columns are pooling regions) and the corresponding parameters for the synthesised image $\mathrm{X}_{\mathrm{synth}}$, we compute the normalised MSE as $\mathrm{MSE} = \mathrm{mean}((\mathrm{X}_{\mathrm{orig}} - \mathrm{X}_{\mathrm{synth}})^2)/\mathrm{Var}(\mathrm{X}_{\mathrm{orig}})$. Freeman and Simoncelli report that this metric was 0.01 ± 0.015 (mean ± s.d.) across all images and scales in their experiment. For our experiment, the same metric across all images and scales was 0.06 ± 0.2. These higher final loss values were driven by the scene-like images, which had a mean loss of 0.11 ± 0.27 compared to the texture-like images (0.01 ± 0.05). Excluding the four highest-loss images (all scene-like) reduced the average loss of the scene-like category to 0.01 ± 0.02, which is similar to the range of the syntheses used by ***Freeman and Simoncelli (2011)*** and to the texture-like images. A control analysis showed the difference in critical scale between the image categories remained after matching the average loss (Results).

To successfully synthesise images at scale factors of 0.25 and 0.36 it was necessary to increase the central region of the image in which the original pixels were perfectly preserved (pooling regions near the fovea become too small to compute correlation matrices). Scales of 0.25 used a central radius of 32 px (0.8 dva in our viewing conditions) and scales 0.36 used 16 px (0.4 dva). This change should, if anything, make syntheses even harder to discriminate from the original image. All other parameters of the model were as in Freeman and Simoncelli. Synthesising an image with scale factor 0.25 took approximately 35 hr, making a larger set of syntheses or source images infeasible. It was not possible to reliably generate images with scale factors lower than 0.25 using the code above.

CNN model syntheses

The CNN pooling model (triangles in ***Figure 1***) was inspired by the model of Freeman and Simoncelli, with two primary differences: first, we replaced the ***Portilla and Simoncelli (2000)*** texture features with the texture features derived from a convolutional neural network (***Gatys et al., 2015***), and second, we simplified the 'foveated' pooling scheme for computational reasons. Specifically, for the CNN 32 model presented above, the image was divided up into 32 angular regions and 28 radial regions, spanning the outer border of the image and an inner radius of 64 px. Within each of these regions we computed the mean activation of the feature maps from a subset of the VGG-19 network layers (conv1_1, conv2_1, conv3_1, conv4_1, conv5_1). To better capture long-range correlations in image structure, we computed these radial and angular regions over three spatial scales, by computing three networks over input sizes 128, 256 and 512 px. Using this multiscale radial and angular pooling representation of an image, we synthesised new images to match the representation of the original image via iterative gradient descent (***Gatys et al., 2015***). Specifically, we minimised the mean-squared distance between the original and a target image, starting from Gaussian noise outside the central 64 px region, using the L-BFGS optimiser as implemented in scipy (***Jones et al., 2001***) for 1000 gradient steps, which we found in pilot experiments was sufficient to produce small (but not zero) loss. Further details, including tests of other variants of this model, are provided in Appendix 2.

Local distortion experiment

We identified local regions that were scene-like or texture-like, whose centre-of-mass was approximately 128 px (±5 px; approximately 6 degrees) from the centre of the image. Because we are not aware of any algorithmic method to distinguish these types of image structure, these were chosen based on our definition of scene-like and texture-like image content (see Results) by author CF. Specifically, a Python script was used to display the 1003 images of the MIT database with a circle of radius 128 px superimposed. CF clicked on a point on the circle that lay in a texture- or scene-like region; if no such region was identified this image was discarded. The coordinates of this point as well as its classification were stored. This procedure resulted in 389 unique images, of which 229 contained a 'scene-like' region and 160 contained a 'texture-like' region.

Non-authors generally agreed with this classification. We conducted a pilot experiment to measure agreement in five participants. Participants were shown each of the 389 images above with a circle (of radius 100 px) superimposed over the region defined by CF. They were instructed to classify the circled region as 'scene-like' (defined as 'tend to contain larger areas of inhomogenous structure, long edges, borders between different surfaces or objects, and angled edges providing perspective cues') or 'texture-like' (defined as 'homogenous structure, patterned content, or materials') in a single-interval binary response task. We found a mean agreement of 88.6% with CF's classification (individual accuracies of 74.8, 90.2, 92.5, 92.8, 92.8%, mean d' = 2.81, with a mean bias to respond 'scene-like', $\log\beta = -1.39$). In this experiment (conducted approximately two years after the initial classification), CF showed a retest agreement of 97.4%.

For each image we perturbed a circular patch in the center of the texture/object region using the texture model of ***Gatys et al. (2015)***. Note that this is the texture model not the CNN-model using radial and angular pooling regions. For each original image, we generated new images containing distortions of different sizes (radii of 40, 70, 85 and 100 px, corresponding to approximately 2, 3.4, 4.1 and 4.9 dva). The local texture features were computed as the (square) Gram matrices in the same VGG-19 layers as used in the CNN-model over an area equal to the radius plus 24 px (square side length $2(r + 24)$). Texture synthesis was then performed via gradient descent as in the CNN-

model, with the exception that the loss function included a circular cosine spatial windowing function which ramped between the synthesised and original pixels over a region of 12 px, in order to smoothly blend the texture distortion with the surrounding image structure. Some example images are shown in ***Figure 3***. In total we therefore used 389 unique images and 389*4 synthesised images as stimuli in this experiment.

Equipment

Stimuli were displayed on a VIEWPixx 3D LCD (VPIXX Technologies Inc, Saint-Bruno-de-Montarville, Canada; spatial resolution 1920 $\times$ 1080 pixels, temporal resolution 120 Hz, operating with the scanning backlight turned off in normal colour mode). Outside the stimulus image the monitor was set to mean grey. Participants viewed the display from 57 cm (maintained via a chinrest) in a darkened chamber. At this distance, pixels subtended approximately 0.025 degrees on average (approximately 40 pixels per degree of visual angle). The monitor was linearised (maximum luminance 260 $\mathrm{cd/m^2}$) using a Konica-Minolta LS-100 (Konica-Minolta Inc, Tokyo, Japan). Stimulus presentation and data collection was controlled via a desktop computer (Intel Core i5-4460 CPU, AMD Radeon R9 380 GPU) running Ubuntu Linux (16.04 LTS), using the Psychtoolbox Library (version 3.0.12, ***Brainard, 1997***; ***Kleiner et al., 2007***; ***Pelli, 1997***), the Eyelink toolbox (***Cornelissen et al., 2002***) and our internal iShow library (http://dx.doi.org/10.5281/zenodo.34217) under MATLAB (The Mathworks Inc, Natick MA, USA; R2015b). Participants' gaze position was monitored by an Eyelink 1000 (SR Research) video-based eyetracker.

Procedure

In the first experiment, participants were shown three images in succession on each trial. Two images were identical, one image was different (the 'oddball', which could occur first, second or third with equal probability). The oddball could be either a synthesised or a natural image (in the orig vs synth condition; counterbalanced), whereas the other two images were physically the same as each other and from the opposite class as the oddball. In the synth vs synth condition (as used in Freeman and Simoncelli), both oddball and foil images were (physically different) model synths. The participant identified the temporal position of the oddball image via button press. Participants were told to fixate on a central point (***Thaler et al., 2013***) presented in the center of the screen. The images were centred around this spot and displayed with a radius of 512 pixels (i.e. images were upsampled by a factor of two for display), subtending $\approx$12.8˚ at the eye. Images were windowed by a circular cosine, ramping the contrast to zero in the space of 52 pixels. The stimuli were presented for 200 ms, with an inter-stimulus interval of 1000 ms (making it unlikely participants could use motion cues to detect changes), followed by a 1200 ms response window. Feedback was provided by a 100 ms change in fixation cross brightness. Gaze position was recorded during the trial. If the participant moved the eye more than 1.5 degrees away from the fixation spot, the trial immediately ended and no response was recorded; participants saw a feedback signal (sad face image) indicating a fixation break. Prior to the next trial, the state of the participant's eye position was monitored for 50 ms; if the eye position was reported as more than 1.5 degrees away from the fixation spot a recalibration was triggered. The inter-trial interval was 400 ms.

Scene-like and texture-like images were compared under two comparison conditions (orig vs synth and synth vs synth; see main text). Image types and scale factors were randomly interleaved within a block of trials (with a minimum of one trial from another image in between) whereas comparison condition was blocked. Participants first practiced the task and fixation control in the orig vs synth comparison condition (scales 0.7, 0.86 and 1.45); the same images used in the experiment were also used in practice to familiarise participants with the images. Participants performed at least 60 practice trials, and were required to achieve at least 50% correct responses and fewer than 20% fixation breaks before proceeding (as noted above, one participant failed). Following successful practice, participants performed one block of orig vs synth trials, which consisted of five FS-model scale factors (0.25, 0.36, 0.46, 0.59, 0.86) plus the CNN 32 model, repeated once for each image to give a total of 120 trials. The participant then practiced the synth vs synth condition for at least one block (30 trials), before continuing to a normal synth vs synth block (120 trials; scale factors of 0.36, 0.46, 0.7, 0.86, 1.45). Over two one-hour sessions, naïve participants completed a total of four

blocks of each comparison condition in alternating order (except for one participant who ran out of time to complete the final block). Authors performed more blocks (total 11).

In the second experiment, observers discriminated which image contained the distortion in a 2IFC paradigm. Each image was presented for 200 ms with a 1000 ms inter-stimulus interval, after which the observer had 1200 ms to respond. The original, unmodified image could appear either first or second; the other image was the same but contained the circular distortion. Observers fixated a spot (**Thaler et al., 2013**) in the centre of the screen. Feedback was provided, and eyetracking was not used. All observers performed 389 trials. To avoid effects of familiarity with the distortion region, each observer saw each original image only once (that is, each original image was randomly assigned to one of the four distortion scales for each observer). While authors were familiar with the images, naïve observers were not. The consistency of effects between authors and naïves suggests that familiarity does not play a major role in this experiment.

Data analysis

In the first experiment, we discarded trials for which participants made no response (N = 66) and broke fixation (N = 239), leaving a total of 7555 trials for further analysis. The median number of responses for each image at each scale for each subject in each condition was 4 trials (min 1, max 7). The individual observer data for the FS-model averaged over images (faint lines in **Figure 1F**) were based on a median of 39 trials (min 20, max 70) for each scale in each condition. The individual observer performance as a function of condition (each psychometric function of FS-scale) was based on a median of 192.5 responses (min 136, max 290).

In the second experiment we discarded trials with no response (N = 8), and did not record eye movements, leaving 1548 trials for further analysis.

To quantify the critical scale as a function of the scale factor s, we used the same 2-parameter function for discriminability d' fitted by Freeman and Simoncelli:

$$d'(s) = \begin{cases} \alpha\left(1 - \frac{s_c^2}{s^2}\right), & s > s_c \\ 0, & s \leq s_c \end{cases}$$

consisting of the critical scale s_c (below which the participant cannot discriminate the stimuli) and a gain parameter α (asymptotic performance level in units of d'). This d' value was transformed to proportion correct using a Weibull function as in **Wallis et al., 2016**:

$$\mathrm{p(correct)} = \frac{1}{m} + (1 - \frac{1}{m})(1 - \exp(-(d'/\lambda)^k))$$

with m set to three (the number of alternatives), and scale λ and shape k parameters chosen by minimising the squared difference between the Weibull and simulated results for oddity as in **Craven (1992)**. The posterior distribution over model parameters (s_c and α) was estimated in a nonlinear mixed-effects model with fixed effects for the experimental conditions (comparison and image type) and random effects for participant (crossed with comparison and image type) and image (crossed with comparison, nested within image type), assuming binomial variability. Note that s_c here is shorthand for a population-level critical scale and group-level offsets estimated for each participant and image; $s_{\mathrm{crit}}(I)$ is the image-specific s_c estimate. Estimates were obtained by a Markov Chain Monte Carlo (MCMC) procedure implemented in the Stan language (version 2.16.2, **Stan Development Team, 2017**; **Hoffman and Gelman, 2014**), with the model wrapper package brms (version 1.10.2, **Bürkner, 2017**; **Bürkner, 2018**) in the R statistical environment. MCMC sampling was conducted with four chains, each with 20,000 iterations (10,000 warmup), resulting in 40,000 post-warmup samples in total. Convergence was assessed using the $\hat{R}$ statistic (**Brooks and Gelman, 1998**) and by examining traceplots. The model parameters were given weakly-informative prior distributions, which provide information about the plausible scale of parameters but do not bias the direction of inference. Specifically, both critical scale and gain were estimated on the natural logarithmic scale; the mean log critical scale (intercept) was given a Gaussian distribution prior with mean -0.69 (corresponding to a critical scale of approximately 0.5—that is centred on the result from Freeman and Simoncelli) and sd 1, other fixed-effect coefficients were given Gaussian priors with mean 0 and sd 0.5, and the group-level standard deviation parameters were given positive-truncated Cauchy priors with mean 0 and sd 0.1. Priors for the log gain parameter were the same,

except the intercept prior had mean 1 (linear gain estimate of 2.72 in d' units) and sd 1. The posterior distribution represents the model's beliefs about the parameters given the priors and data. This distribution is summarised above as posterior mean, 95% credible intervals and posterior probabilities for the fixed-effects parameters to be negative (the latter computed via the empirical cumulative distribution of the relevant MCMC samples).

Acknowledgments

Funded by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the Deutsche Forschungsgemeinschaft (DFG; priority program 1527, BE 3848/2–1 and Projektnummer 276693517 – SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP03). We acknowledge support by the Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of University of Tübingen. We thank Wiebke Ringels for assistance with data collection, Heiko Schütt, Matthias Kümmerer and Corey Ziemba for helpful comments on an earlier draft, Andrew Haun and Ben Balas for helpful comments on Twitter, and reviewer John Cass for suggesting the importance of junction information in explaining our results. TSAW was supported in part by an Alexander von Humboldt Postdoctoral Fellowship. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christina Funke.

Additional information

Competing interests

Leon A Gatys: This author now works for Apple, Inc. The author's contributions to this article were prior to commencing employment at Apple. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Bundesministerium für Bildung und Forschung	FKZ: 01GQ1002	Felix A Wichmann Matthias Bethge
Deutsche Forschungsgemeinschaft	BE 3848/2-1	Matthias Bethge
Deutsche Forschungsgemeinschaft	Priority program 1527	Felix A Wichmann Matthias Bethge
Deutsche Forschungsgemeinschaft	276693517; SFB 1233	Thomas SA Wallis Christina M Funke Felix A Wichmann Matthias Bethge
Alexander von Humboldt-Stiftung	Thomas Wallis	Thomas SA Wallis

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Thomas SA Wallis, Conceptualization, Resources, Data curation, Software, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing; Christina M Funke, Conceptualization, Resources, Software, Formal analysis, Validation, Investigation, Methodology, Writing—original draft, Writing—review and editing; Alexander S Ecker, Conceptualization, Supervision, Methodology, Project administration, Writing—review and editing; Leon A Gatys, Resources, Software, Methodology, Writing—review and editing; Felix A Wichmann, Conceptualization, Supervision, Funding acquisition, Methodology, Project administration, Writing—review and editing; Matthias Bethge,

Conceptualization, Supervision, Funding acquisition, Project administration, Writing—review and editing

Author ORCIDs

Thomas SA Wallis https://orcid.org/0000-0001-7431-4852
Alexander S Ecker http://orcid.org/0000-0003-2392-5105
Felix A Wichmann https://orcid.org/0000-0002-2592-634X
Matthias Bethge http://orcid.org/0000-0002-6417-7812

Ethics

Human subjects: All participants provided informed consent to participate in the study and for their anonymised data to be made publicly available. The study adhered to Standard 8 of the American Psychological Association's "Ethical Principles of Psychologists and Code of Conduct" (2010). The experiments were approved by the Ethics Commission of the University Clinics Tübingen (Nr. 222/2011B02).

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.42512.030
Author response https://doi.org/10.7554/eLife.42512.031

Additional files

Supplementary files

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.42512.010

Data availability

All raw data, processed data, model files, stimulus materials, and analysis code are provided for download in a Zenodo database at http://dx.doi.org/10.5281/zenodo.1475111.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Wallis TSA, Funke CM	2018	Materials to reproduce Wallis, Funke et al. "Image content is more important than Bouma's Law for scene metamers"	http://dx.doi.org/10.5281/zenodo.1475111	Zenodo, 10.5281/zenodo.1475111

References

Adelson EH. 2001. On seeing stuff: the perception of materials by humans and machines. *Human Vision and Electronic Imaging* **4299**:1–12. DOI: https://doi.org/10.1117/12.429489

Agaoglu MN, Chung ST. 2016. Can (should) theories of crowding be unified? *Journal of Vision* **16**:10. DOI: https://doi.org/10.1167/16.15.10, PMID: 27936273

Ariely D. 2001. Seeing sets: representation by statistical properties. *Psychological Science* **12**:157–162. DOI: https://doi.org/10.1111/1467-9280.00327, PMID: 11340926

Arnold JB. 2016. ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. *4.0*.https://rdrr.io/cran/ggthemes/

Auguie B. 2016. gridExtra: Miscellaneous Functions for "Grid" Graphics. 2.3. https://rdrr.io/cran/gridExtra/

Balas B, Nakano L, Rosenholtz R. 2009. A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision* **9**:13. DOI: https://doi.org/10.1167/9.12.13

Balas B. 2012. Contrast negation and texture synthesis differentially disrupt natural texture appearance. *Frontiers in Psychology* **3**:29–39. DOI: https://doi.org/10.3389/fpsyg.2012.00515, PMID: 23181049

Bex PJ. 2010. (In) Sensitivity to spatial distortion in natural scenes. *Journal of Vision* **10**:1–15. DOI: https://doi.org/10.1167/10.2.23

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological Review* **94**:115–147. DOI: https://doi.org/10.1037/0033-295X.94.2.115, PMID: 3575582

Block N. 2013. Seeing and windows of integration. *Thought: A Journal of Philosophy* **2**:29–39. DOI: https://doi.org/10.1002/tht3.62

Bouma H. 1970. Interaction effects in parafoveal letter recognition. *Nature* **226**:177–178. DOI: https://doi.org/10.1038/226177a0, PMID: 5437004
Brainard DH. 1997. The psychophysics toolbox. *Spatial Vision* **10**:433–436. DOI: https://doi.org/10.1163/156856897X00357, PMID: 9176952
Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**:434. DOI: https://doi.org/10.2307/1390675
Bürkner P-C. 2017. Brms: an R package for bayesian multilevel models using stan. *Journal of Statistical Software* **80**:1–28. DOI: https://doi.org/10.18637/jss.v080.i01
Bürkner P-C. 2018. Advanced bayesian multilevel modeling with the R package brms. *The R Journal* **10**:395–411. DOI: https://doi.org/10.32614/RJ-2018-017
Chang H, Rosenholtz R. 2016. Search performance is better predicted by tileability than presence of a unique basic feature. *Journal of Vision* **16**:13. DOI: https://doi.org/10.1167/16.10.13, PMID: 27548090
Clarke AM, Herzog MH, Francis G. 2014. Visual crowding illustrates the inadequacy of local vs. global and feedforward vs. feedback distinctions in modeling visual perception. *Frontiers in Psychology* **5**. DOI: https://doi.org/10.3389/fpsyg.2014.01193, PMID: 25374554
Cohen MA, Dennett DC, Kanwisher N. 2016. What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences* **20**:324–335. DOI: https://doi.org/10.1016/j.tics.2016.03.006, PMID: 27105668
Cornelissen FW, Peters EM, Palmer J. 2002. The eyelink toolbox: eye tracking with MATLAB and the psychophysics toolbox. *Behavior Research Methods, Instruments, & Computers* **34**:613–617. DOI: https://doi.org/10.3758/BF03195489
Craven BJ. 1992. A table of d' for M-alternative odd-man-out forced-choice procedures. *Perception & Psychophysics* **51**:379–385. DOI: https://doi.org/10.3758/BF03211631, PMID: 1603651
Croner LJ, Kaplan E. 1995. Receptive fields of P and M ganglion cells across the primate retina. *Vision Research* **35**:7–24. DOI: https://doi.org/10.1016/0042-6989(94)E0066-T, PMID: 7839612
Dacey DM, Petersen MR. 1992. Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *PNAS* **89**:9666–9670. DOI: https://doi.org/10.1073/pnas.89.20.9666, PMID: 1409680
Dakin SC, Bex PJ, Cass JR, Watt RJ. 2009. Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision* **9**:28. DOI: https://doi.org/10.1167/9.11.28, PMID: 20053091
Dakin SC, Baruch NJ. 2009. Context influences contour integration. *Journal of Vision* **9**:13. DOI: https://doi.org/10.1167/9.2.13, PMID: 19271923
Dakin SC, Watt RJ. 1997. The computation of orientation statistics from visual texture. *Vision Research* **37**:3181–3192. DOI: https://doi.org/10.1016/S0042-6989(97)00133-8, PMID: 9463699
Deza A, Jonnalagadda A, Eckstein M. 2017. Towards metamerism via foveated style transfer. *arXiv*. https://arxiv.org/abs/1705.10041.
Ehinger KA, Rosenholtz R. 2016. A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision* **16**:13. DOI: https://doi.org/10.1167/16.2.13, PMID: 27893077
Faivre N, Berthet V, Kouider S. 2012. Nonconscious influences from emotional faces: a comparison of visual crowding, masking, and continuous flash suppression. *Frontiers in Psychology* **3**. DOI: https://doi.org/10.3389/fpsyg.2012.00129, PMID: 22563325
Fischer J, Whitney D. 2011. Object-level visual information gets through the bottleneck of crowding. *Journal of Neurophysiology* **106**:1389–1398. DOI: https://doi.org/10.1152/jn.00904.2010, PMID: 21676930
Francis G, Manassi M, Herzog MH. 2017. Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review* **124**:483–504. DOI: https://doi.org/10.1037/rev0000070, PMID: 28437128
Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. 2013. A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience* **16**:974–981. DOI: https://doi.org/10.1038/nn.3402, PMID: 23685719
Freeman J, Simoncelli EP. 2011. Metamers of the ventral stream. *Nature Neuroscience* **14**:1195–1201. DOI: https://doi.org/10.1038/nn.2889, PMID: 21841776
Freeman J, Simoncelli E, . 2013. The radial and tangential extent of spatial metamers. *Journal of Vision* **13**:573. DOI: https://doi.org/10.1167/13.9.573
Gatys LA, Ecker AS, Bethge M. 2015. Texture synthesis using convolutional neural networks. 2016 23rd International Conference on Pattern Recognition (ICPR).
Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. 2019. *ImageNet-Trained CNNs are biased towards texture; Increasing shape bias Improves Accuracy and robustness*. International Conference on Learning Representations.
Gelman A, Hwang J, Vehtari A. 2014. Understanding predictive information criteria for bayesian models. *Statistics and Computing* **24**:997–1016. DOI: https://doi.org/10.1007/s11222-013-9416-2
Haun AM, Tononi G, Koch C, Tsuchiya N. 2017. Are we underestimating the richness of visual experience? *Neuroscience of Consciousness* **2017**. DOI: https://doi.org/10.1093/nc/niw023, PMID: 30042833
Herzog MH, Sayim B, Chicherov V, Manassi M. 2015. Crowding, grouping, and object recognition: a matter of appearance. *Journal of Vision* **15**:5. DOI: https://doi.org/10.1167/15.6.5, PMID: 26024452
Hoffman MD, Gelman A. 2014. The No-U-Turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* **15**:1593–1623.
Jäkel F, Singh M, Wichmann FA, Herzog MH. 2016. An overview of quantitative approaches in gestalt perception. *Vision Research* **126**:3–8. DOI: https://doi.org/10.1016/j.visres.2016.06.004, PMID: 27353224

Jones E, Oliphant T, Peterson P. 2001. *SciPy: Open Source Scientific Tools for Python*. https://www.researchgate.net/publication/213877848_SciPy_Open_Source_Scientific_Tools_for_Python
Judd T, Ehinger KA, Durand F, Torralba A. 2009. Learning to predict where humans look. IEEE 12th International Conference on Computer Vision 2106–2113. https://ieeexplore.ieee.org/document/5459462.
Judd T, Durand F, Torralba A. 2012. *A Benchmark of Computational Models of Saliency to Predict Human Fixations*: CSAIL Technical Reports.
Keshvari S, Rosenholtz R. 2016. Pooling of continuous features provides a unifying account of crowding. *Journal of Vision* **16**:39. DOI: https://doi.org/10.1167/16.3.39, PMID: 26928055
Kleiner M, Brainard DH, Pelli DG. 2007. What's New in Psychtoolbox-3. *Perception* **36**.
Koenderink J, Valsecchi M, van Doorn A, Wagemans J, Gegenfurtner K. 2017. Eidolons: novel stimuli for vision research. *Journal of Vision* **17**:7. DOI: https://doi.org/10.1167/17.2.7, PMID: 28245489
Koffka K. 1935. *Principles of Gestalt Psychology*. Oxford, UK: Harcourt Brace.
Kruschke J. 2015. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press.
Kubilius J, Bracci S, Op de Beeck HP. 2016. Deep neural networks as a computational model for human shape sensitivity. *PLOS Computational Biology* **12**:e1004896. DOI: https://doi.org/10.1371/journal.pcbi.1004896, PMID: 27124699
Landy MS. 2013. The New Visual Neurosciences. In: Werner J. S, Chalupa L. M (Eds). *Texture Analysis and Perception*. MIT Press. p. 639–652.
Lettvin JY. 1976. On seeing sidelong. *The Sciences* **16**:10–20. DOI: https://doi.org/10.1002/j.2326-1951.1976.tb01231.x
Levi DM, Carney T. 2009. Crowding in peripheral vision: why bigger is better. *Current Biology* **19**:1988–1993. DOI: https://doi.org/10.1016/j.cub.2009.09.056, PMID: 19853450
Long B, Konkle T, Cohen MA, Alvarez GA. 2016. Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General* **145**:95–109. DOI: https://doi.org/10.1037/xge0000130
Loschky LC, Hansen BC, Sethi A, Pydimarri TN. 2010. The role of higher order image statistics in masking scene gist recognition. *Attention, Perception, & Psychophysics* **72**:427–444. DOI: https://doi.org/10.3758/APP.72.2.427, PMID: 20139457
Mack A, Rock I. 1998. *Inattentional Blindness*. **33** Cambridge, MA: MIT press . DOI: https://doi.org/10.7551/mitpress/3707.001.0001
Macmillan NA, Creelman CD. 2005. *Detection Theory: A User's Guide*. Mahwah, NJ: Lawrence Erlbaum.
Manassi M, Sayim B, Herzog MH. 2012. Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision* **12**:13. DOI: https://doi.org/10.1167/12.10.13, PMID: 23019118
Manassi M, Sayim B, Herzog MH. 2013. When crowding of crowding leads to uncrowding. *Journal of Vision* **13**:10. DOI: https://doi.org/10.1167/13.13.10, PMID: 24213598
McElreath R. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. New York: CRC Press, Taylor & Francis Group.
Movshon JA, Simoncelli EP. 2014. Representation of naturalistic image structure in the primate visual cortex. *Cold Spring Harbor Symposia on Quantitative Biology* **79**:115–122. DOI: https://doi.org/10.1101/sqb.2014.79.024844, PMID: 25943766
Neri P. 2017. Object segmentation controls image reconstruction from natural scenes. *PLOS Biology* **15**:e1002611. DOI: https://doi.org/10.1371/journal.pbio.1002611, PMID: 28827801
O'Regan JK, Rensink RA, Clark JJ. 1999. Change-blindness as a result of 'mudsplashes'. *Nature* **398**:34. DOI: https://doi.org/10.1038/17953, PMID: 10078528
Okazawa G, Tajima S, Komatsu H. 2015. Image statistics underlying natural texture selectivity of neurons in macaque V4. *PNAS* **112**:E351–E360. DOI: https://doi.org/10.1073/pnas.1415146112, PMID: 25535362
Parkes L, Lund J, Angelucci A, Solomon JA, Morgan M. 2001. Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience* **4**:739–744. DOI: https://doi.org/10.1038/89532, PMID: 11426231
Pelli DG. 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision* **10**:437–442. DOI: https://doi.org/10.1163/156856897X00366, PMID: 9176953
Pelli DG, Tillman KA. 2008. The uncrowded window of object recognition. *Nature Neuroscience* **11**:1129–1135. DOI: https://doi.org/10.1038/nn.2187, PMID: 18828191
Petrov Y, Meleshkevich O. 2011. Asymmetries and idiosyncratic hot spots in crowding. *Vision Research* **51**:1117–1123. DOI: https://doi.org/10.1016/j.visres.2011.03.001, PMID: 21439309
Portilla J, Simoncelli EP. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision* **40**:49–70. DOI: https://doi.org/10.1023/A:1026553619983
R Core Team. 2017. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Vienna, Austria,
Rensink RA, O'Regan JK, Clark JJ. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* **8**:368–373. DOI: https://doi.org/10.1111/j.1467-9280.1997.tb00427.x
Robol V, Casco C, Dakin SC. 2012. The role of crowding in contextual influences on contour integration. *Journal of Vision* **12**:3. DOI: https://doi.org/10.1167/12.7.3, PMID: 22776847
Rosen S, Chakravarthi R, Pelli DG. 2014. The bouma law of crowding, revised: critical spacing is equal across parts, not objects. *Journal of Vision* **14**:10. DOI: https://doi.org/10.1167/14.6.10, PMID: 25502230

Rosenholtz R, Huang J, Ehinger KA. 2012a. Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology* **3**:13. DOI: https://doi.org/10.3389/fpsyg.2012.00013, PMID: 22347200

Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L. 2012b. A summary statistic representation in peripheral vision explains visual search. *Journal of Vision* **12**:14. DOI: https://doi.org/10.1167/12.4.14, PMID: 22523401

Rosenholtz R. 2016. Capabilities and limitations of peripheral vision. *Annual Review of Vision Science* **2**:437–457. DOI: https://doi.org/10.1146/annurev-vision-082114-035733, PMID: 28532349

Saarela TP, Sayim B, Westheimer G, Herzog MH. 2009. Global stimulus configuration modulates crowding. *Journal of Vision* **9**:5. DOI: https://doi.org/10.1167/9.2.5, PMID: 19271915

Seth AK. 2014. A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience* **5**:97–118. DOI: https://doi.org/10.1080/17588928.2013.877880, PMID: 24446823

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for Large-Scale image recognition. *Arxiv*. https://arxiv.org/abs/1409.1556.

Stan Development. 2015. *Stan Modeling Language Users Guide and Reference Manual*. 2.10.0.

Stan Development Team. 2017. Stan: A C++ Library for Probability and Sampling. 2.14.0.

Teufel C, Dakin SC, Fletcher PC. 2018. Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports* **8**. DOI: https://doi.org/10.1038/s41598-018-28845-5, PMID: 30022033

Thaler L, Schütz AC, Goodale MA, Gegenfurtner KR. 2013. What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision Research* **76**:31–42. DOI: https://doi.org/10.1016/j.visres.2012.10.012, PMID: 23099046

Valsecchi M, Koenderink J, van Doorn A, Gegenfurtner KR. 2018. Prediction shapes peripheral appearance. *Journal of Vision* **18**:21. DOI: https://doi.org/10.1167/18.13.21, PMID: 30593064

Van der Burg E, Olivers CN, Cass J. 2017. Evolving the keys to visual crowding. *Journal of Experimental Psychology: Human Perception and Performance* **43**:690–699. DOI: https://doi.org/10.1037/xhp0000337, PMID: 28182476

Vehtari A, Gelman A, Gabry J. 2016. Practical bayesian model evaluation using Leave-One-Out Cross-Validation and WAIC. *arXiv*. https://arxiv.org/abs/1507.04544.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*.. Fourth Edition. New York: Springer. DOI: https://doi.org/10.1007/978-0-387-21706-2

Vickery TJ, Shim WM, Chakravarthi R, Jiang YV, Luedeman R. 2009. Supercrowding: weakly masking a target expands the range of crowding. *Journal of Vision* **9**:12. DOI: https://doi.org/10.1167/9.2.12, PMID: 19271922

Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, von der Heydt R. 2012a. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological Bulletin* **138**:1172–1217. DOI: https://doi.org/10.1037/a0029333, PMID: 22845751

Wagemans J, Feldman J, Gepshtein S, Kimchi R, Pomerantz JR, van der Helm PA, van Leeuwen C. 2012b. A century of gestalt psychology in visual perception: ii. conceptual and theoretical foundations. *Psychological Bulletin* **138**:1218–1252. DOI: https://doi.org/10.1037/a0029334, PMID: 22845750

Wallis TSA, Bethge M, Wichmann FA. 2016. Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision* **16**:4. DOI: https://doi.org/10.1167/16.2.4, PMID: 26968866

Wallis TSA, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M. 2017. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision* **17**:5. DOI: https://doi.org/10.1167/17.12.5, PMID: 28983571

Wallis TSA, Bex PJ. 2012. Image correlates of crowding in natural scenes. *Journal of Vision* **12**:6. DOI: https://doi.org/10.1167/12.7.6, PMID: 22798053

Walther DB, Shen D. 2014. Nonaccidental properties underlie human categorization of complex natural scenes. *Psychological Science* **25**:851–860. DOI: https://doi.org/10.1177/0956797613512662, PMID: 24474725

Wandell BA. 1995. *Foundations of Vision*. Sinauer Associates.

Watson AB. 2014. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision* **14**:15–17. DOI: https://doi.org/10.1167/14.7.15, PMID: 24982468

Whitney D, Haberman J, Sweeny TD. 2014. From textures to crowds: multiple levels of summary statistical perception. In: *The New Visual Neurosciences*. MIT Press. p. 695–710.

Wickham H. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. DOI: https://doi.org/10.1007/978-0-387-98141-3

Wickham H. 2011. The Split-Apply-Combine strategy for data analysis. *Journal of Statistical Software* **40**:1–29. DOI: https://doi.org/10.18637/jss.v040.i01

Wickham H, Francois R. 2016. *Dplyr: A Grammar of Data Manipulation*. https://rdrr.io/cran/dplyr/

Wilder J, Rezanejad M, Dickinson S, Siddiqi K, Jepson A, Walther DB. 2019. Local contour symmetry facilitates scene categorization. *Cognition* **182**:307–317. DOI: https://doi.org/10.1016/j.cognition.2018.09.014, PMID: 30415132

Xia G-S, Delon J, Gousseau Y. 2014. Accurate junction detection and characterization in natural images. *International Journal of Computer Vision* **106**:31–56. DOI: https://doi.org/10.1007/s11263-013-0640-1

Xie Y. 2013. Knitr: A Comprehensive Tool for Reproducible Research in R. BT - Implementing Reproducible Computational Research. In: Stodden V, Leisch F, Peng R. D (Eds). *Implementing Reproducible Computational Research*. Chapman & Hall/CRC.

Xie Y. 2016. *Dynamic Documents with R and Knitr*. Chapman & Hall/CRC. DOI: https://doi.org/10.1201/b15166

Zhang X, Huang J, Yigit-Elliott S, Rosenholtz R. 2015. Cube search, revisited. *Journal of Vision* **15**:9. DOI: https://doi.org/10.1167/15.3.9, PMID: 25780063

Ziemba CM, Freeman J, Movshon JA, Simoncelli EP. 2016. Selectivity and tolerance for visual texture in macaque V2. *PNAS* **113**:E3140–E3149. DOI: https://doi.org/10.1073/pnas.1510847113, PMID: 27173899

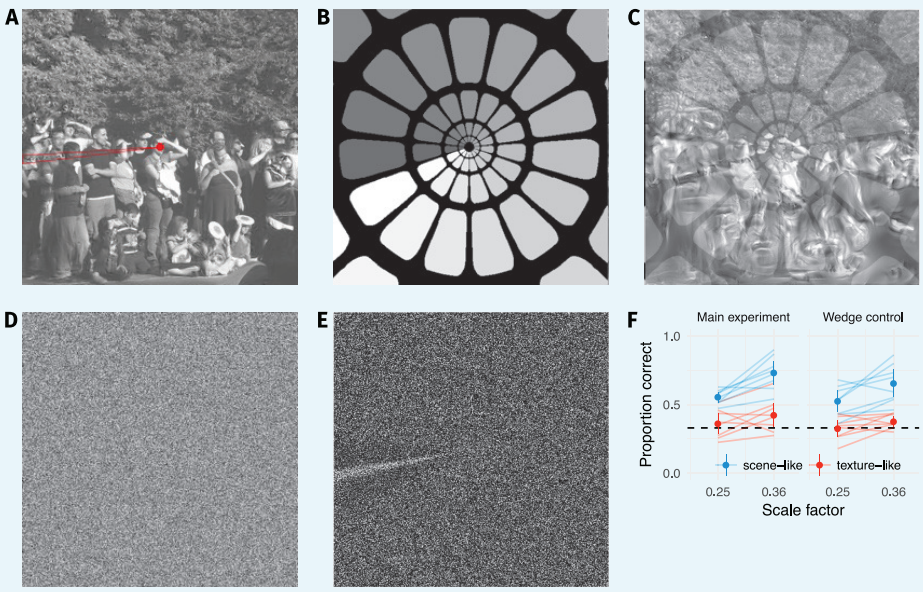
Appendix 1

DOI: https://doi.org/10.7554/eLife.42512.011

Additional experiments with the Freeman and Simoncelli model

Stimulus artifact control

During the course of our testing we noticed that synthesised images generated with the code from http://github.com/freeman-lab/metamers contained an artifact, visible as a wedge in the lower-left quadrant of the synthesised images in which the phases of the surrounding image structure were incorrectly matched (***Appendix 1—figure 1A***). The angle and extent of the wedge changed with the scale factor, and corresponded to the region where angular pooling regions wrapped from 0–2π (***Appendix 1—figure 1B–C***). The visibility of the artifact depended on image structure, but was definitely due to the synthesis procedure itself because it also occurred when synthesising matches to a white noise source image (***Appendix 1—figure 1D–E***). The artifact was not peculiar to our hardware or implementation because it is also visible in the stimuli shown in ***Deza et al. (2017)***.



Appendix 1—figure 1. Our results do not depend on an artifact in the synthesis procedure. (**A**) During our pilot testing, we noticed a wedge-like artifact in the synthesis procedure of Freeman and Simoncelli (highlighted in red wedge; image from https://github.com/freeman-lab/metamers and shared under a CC-BY license (https://creativecommons.org/licenses/by/3.0/)). The artifact occurred where the angular pooling regions wrapped from 0 to 2π (**B**) pooling region contours shown with increasing greyscale to wrap point, (**C**) overlayed on scene with artifact. (**D**) The artifact was not driven by image content, because it also occurred when synthesising to match white noise (shown with enhanced contrast in (**E**)). If participants' good performance at small scale factors was due to taking advantage of this wedge, removing it by masking out that image region should drop performance to chance. (**F**) Performance at the two smallest scale factors replotted from the main experiment (left) and with a wedge mask overlayed (right) in the orig vs synth comparison. Points show average (±2SE) over participants; faint lines link individual participant means. Performance remains above chance

for the scene-like images, indicating that the low critical scales we observed were not due to the wedge artifact.

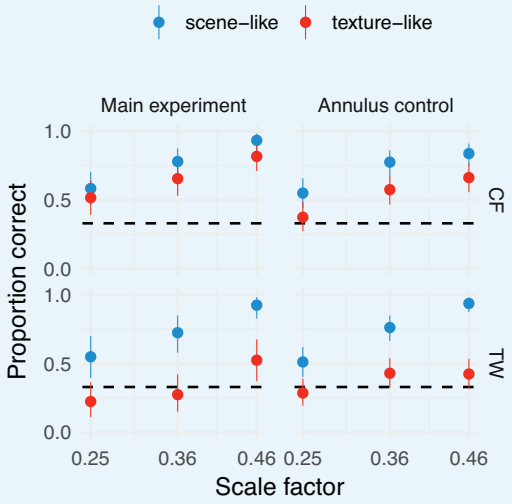
DOI: https://doi.org/10.7554/eLife.42512.012

Participants in our experiment could have learned to use the artifact to help discriminate images, particularly synthesised images from original images (since only synthesised images contain the artifact). This may have boosted their sensitivity more than might be expected from the model described by Freeman and Simoncelli, leading to the lower critical scales we observed. To control for this, we re-ran the original vs synth condition with the same participants, with the exception that the lower-left quadrant of the image containing the artifact was masked by a grey wedge (in both original and synthesised images) with angular subtense of 60 degrees. We used only the lowest two scale factors from the main experiment, and participants completed this control experiment after the main experiment reported in the paper. We discarded trials for which participants made no response (N = 9) or broke fixation (N = 57), leaving a total of 1014 trials for further analysis. If the high sensitivity at low scale factors we observed above were due to participants using the artifact, then their performance with the masked stimuli should fall to chance for low scale factors.

This is not what we observed: while performance with the wedge was slightly worse (perhaps because a sizable section of the image was masked), the scene-like images remained above chance performance for the lowest two scale factors (***Appendix 1—figure 1F***). This shows that the low critical scale factors we observed in the main experiment are not due to the wedge artifact.

We performed one additional artifact control experiment. The FS algorithm preserves a small central region of the image exactly in order to match foveal appearance. If there is any image artifact produced by the synthesis procedure at the border of this region, participants could have used this artifact to discriminate the stimuli in the original vs synth condition. Authors TW and CF performed new oddity discrimination trials in which a grey annular occluding zone (inner radius 0.4 deg, outer radius 1.95 deg) was presented over all images. If the low scale factors we find are because participants used a stimulus artifact, then performance at the low scales should drop to chance.

The results of this additional experiment are shown in Figure (***Appendix 1—figure 2***). Both participants can still discriminate real and synthesised scene-like images better than chance even after superposition of the occluding annulus, indicating that any central artifact is not a crucial determinant of discriminability.



Appendix 1—figure 2. Our results do not depend on any potential annular artifact resulting from the synthesis procedure. Performance at the three smallest scale factors replotted from the main experiment (left) and with an annular mask overlayed (right) in the orig vs synth

comparison for authors TW and CF. Points show average performance (error bars show 95% beta distribution confidence limits). Performance remains above chance for the scene-like images, indicating that the low critical scales we observed were not due to a potential annular artifact.

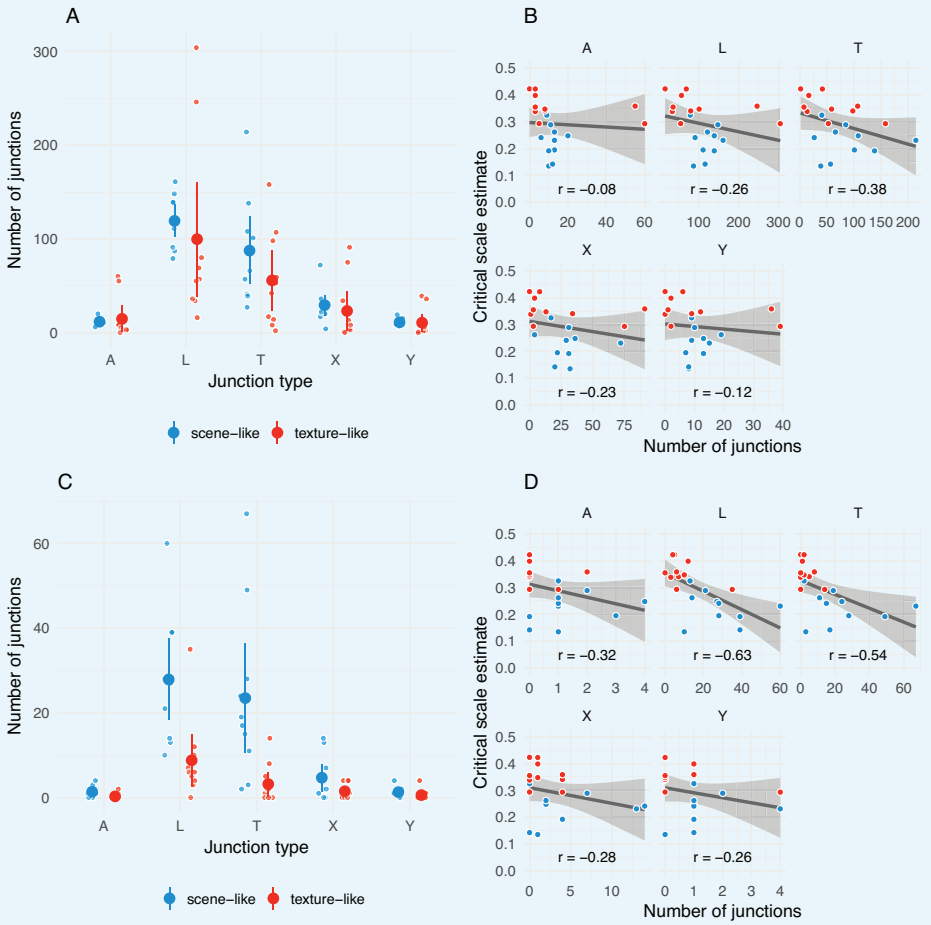
DOI: https://doi.org/10.7554/eLife.42512.013

Junctions in original images

For each of the 20 original images used in our first experiment, we used the junction detection algorithm of ***Xia et al. (2014)*** to identify junctions in the image (with algorithm parameter $r_{max} = 36$). We subdivided all three-edge junctions into T-, Y- and arrow-junctions according to the angle criteria used in ***Walther and Shen (2014)***, and excluded all junctions that fell outside the circular region of the image shown in our experiment.

We find that scene-like images tend to contain more junctions than texture-like images (***Appendix 1—figure 3A***). This relationship became stronger when we excluded 'less meaningful' junctions (using a 'meaningfulness' cutoff of $\log(\mathrm{NFA}) = -20$, ***Xia et al. (2014)***; ***Appendix 1—figure 3C***). Images with smaller critical scales are associated with the presence of junctions (***Appendix 1—figure 3B***), and this association gets stronger when small and weak junctions are excluded (***Appendix 1—figure 3D***).

If junction information is important for scene appearance and the FS-model fails to adequately capture this information, we would expect such a negative relationship between junctions and critical scales. Of course, the analysis above does not support a specific causal role for junction information: for example, it may be correlated with simple edge density. Future studies could confirm (or reject) this relationship using a larger and more diagnostic image set.



Appendix 1—figure 3. The number of junctions present in original images may be related to critical scale estimates. (**A**) Distribution of arrow (A), L-, T-, X- and Y-junctions at all scales and all levels of 'meaningfulness' (***Xia et al., 2014***) in scene-like and texture-like images. Each small point is one image; larger points with error bars show mean ±2 SE. Points have been jittered to aid visibility. (**B**) Correlations between number of junctions of each type with critical scale estimates for that image from the main experiment. Grey line shows linear model fit with shaded region showing 95% confidence area. Pearson correlation coefficient shown below. Note that the x-axis scales in the subplots differ. (**C**) Same as A but for junctions defined with a more strict 'meaningfulness' cutoff of $\log(\mathrm{NFA}) = -20$ (***Xia et al., 2014***). (**D**) Same as B for more 'meaningful' junctions as in C.

DOI: https://doi.org/10.7554/eLife.42512.014

ABX replication

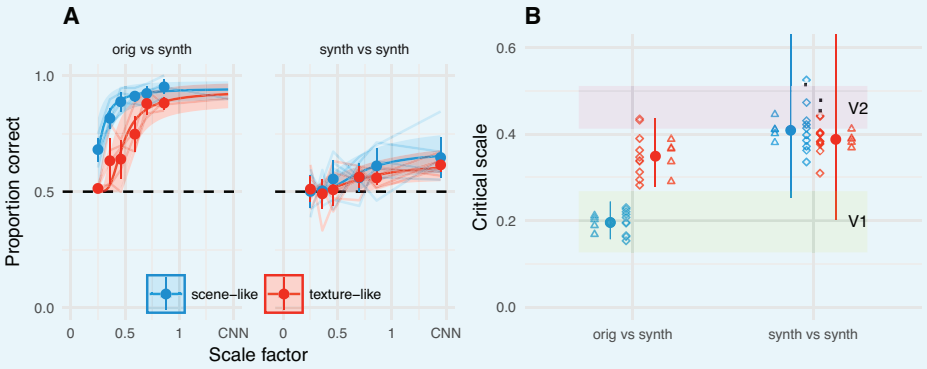
Participants in our experiment showed poor performance in the synth vs synth condition even for large scale factors (highest accuracy for a participant at the largest scale of 1.45 was 0.8, average accuracy 0.58), leading to relatively flat psychometric functions (***Figure 2F*** of main manuscript). In contrast, most participants in ***Freeman and Simoncelli (2011)*** achieved accuracies above 90% correct for the highest scale factor they test (1.45 as in our experiment). One difference between our experiment and ***Freeman and Simoncelli (2011)*** is that they used an ABX task, in which participants saw two images A and B, followed by image X, and had to report whether image X was the same as A or B. Perhaps our oddity task is simply harder: due to greater memory load or the cognitive demands of the comparison, participants in our experiment were unable to perform consistently well.

To assess whether the use of an oddity task lead to our finding of lower critical scales and/or poorer asymptotic performance in the synth vs synth condition, we re-ran our experiment as an ABX task. We used the same timing parameters as in Freeman and Simoncelli. Six participants participated in the experiment, including a research assistant (the same as in the main experiment), four naÃve participants and author AE (who only participated in the synth vs synth condition). We discarded trials for which participants made no response (N = 61) or broke fixation (N = 442), leaving a total of 7537 trials for further analysis. The predicted proportion correct in the ABX task was derived from d' using the link function given by ***Macmillan and Creelman (2005)***, (229–33) for a differencing model in a roving design:

$$\mathrm{p(correct)} = \Phi\left(\frac{d'(s)}{\sqrt{2}}\right)\Phi\left(\frac{d'(s)}{\sqrt{6}}\right) + \Phi\left(\frac{-d'(s)}{\sqrt{2}}\right)\Phi\left(\frac{-d'(s)}{\sqrt{6}}\right)$$

where Φ is the standard cumulative Normal distribution.

As in our main experiment with the oddity task, we find that participants could easily discriminate scene-like syntheses from their original at all scales we could generate (***Appendix 1—figure 4***). Critical scale factor estimates were similar to those in the main experiment, indicating that the ABX task did not make a large difference to these results. Critical scale estimates were slightly larger, but much more uncertain, in the synth vs synth condition. This uncertainty is largely driven by the even poorer asymptotic performance than in the main experiment. This shows that the results we report in the primary manuscript are not particular to the oddity task.



Appendix 1—figure 4. Results from the main paper replicated under an ABX task. (**A**) Performance in the ABX task as a function of scale factor. Points show grand mean ±2 SE over participants; faint lines link individual participant performance levels. Solid curves and shaded regions show the fit of a nonlinear mixed-effects model estimating the critical scale and gain. (**B**) When comparing original and synthesised images, estimated critical scales (scale at which performance rises above chance) are lower for scene-like than for texture-like images. Points with error bars show population mean and 95% credible intervals. Triangles show posterior means for participants; diamonds show posterior means for images. Black squares show critical scale estimates of the four participants from Freeman and Simoncelli reproduced from that paper (x-position jittered to reduce overplotting); shaded regions denote the receptive field scaling of V1 and V2 estimated by Freeman and Simoncelli.

DOI: https://doi.org/10.7554/eLife.42512.015

What explains the discrepancy between asymptotic performance in our experiment vs Freeman and Simoncelli? One possibility is that the participants in Freeman and Simoncelli's experiment were more familiar with the images shown, and that good asymptotic performance in the synth vs synth condition requires strong familiarity. Freeman and Simoncelli used four original (source) images, and generated three unique synthesised images for each source image at each scale, compared to our 20 source images with four syntheses.

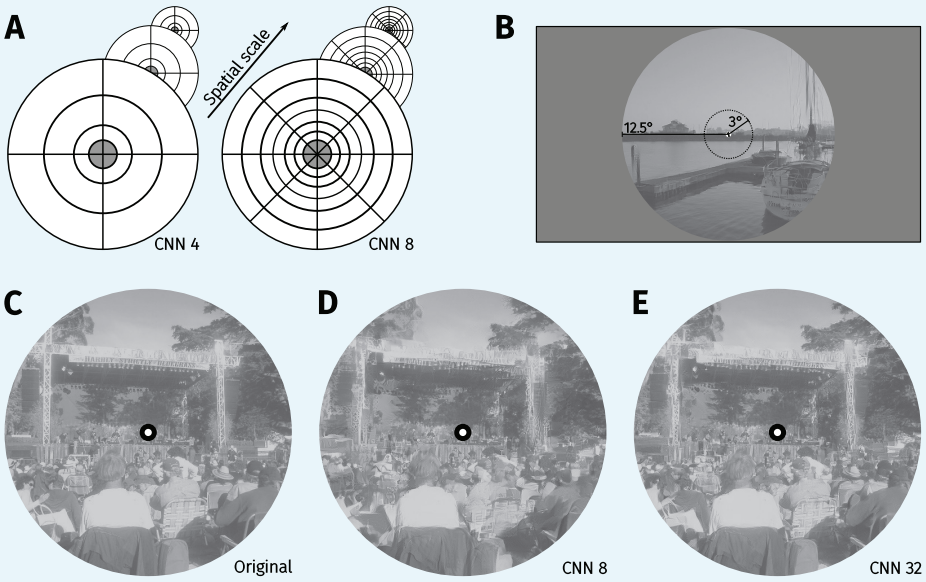
Appendix 2

DOI: https://doi.org/10.7554/eLife.42512.011

CNN scene appearance model

Here we describe the CNN scene appearance model presented in the paper in more detail, as well as additional experiments concerning this model.

To create a summary statistic model using CNN features, we compute the mean activation in a subset of CNN layers over a number of radial and angular spatial regions (see ***Appendix 2—figure 1***). Increasing the number of pooling regions (reducing the spatial area over which CNN features are pooled) preserves more of the structure of the original image. New images can be synthesised by minimising the difference between the model features for a given input image and a white noise image via an iterative gradient descent procedure (see below). This allows us to synthesise images that are physically different to the original but approximately the same according to the model. We did this for each of four pooling region sizes, named model 4, 8, 16 and 32 respectively after the number of angular pooling regions. These features were matched over three spatial scales, which we found improved the model's ability to capture long-range correlations.



Appendix 2—figure 1. Methods for the CNN scene appearance model. (**A**) The average activations in a subset of CNN feature maps were computed over non-overlapping radial and angular pooling regions that increase in area away from the image centre (not to scale), for three spatial scales. Increasing the number of pooling regions (CNN 4 and CNN 8 shown in this example) increases the fidelity of matching to the original image, restricting the range over which distortions can occur. Higher-layer CNN receptive fields overlap the pooling regions, ensuring smooth transitions between regions. The central 3˚ of the image (grey fill) is fixed to be the original. (**B**) The image radius subtended 12.5˚. (**C**) An original image from the MIT1003 dataset. (**D**) Synthesised image matched to the image from C by the CNN 8 pooling model. (**E**) Synthesised image matched to the image from E by the CNN 32 pooling model. Fixating the central bullseye, it should be apparent that the CNN 32 model preserves more information than the CNN 8 model, but that the periphery is nevertheless significantly distorted relative to the original. Images from the MIT 1003 dataset (***Judd et al., 2009***), (https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) and reproduced under a

DOI: https://doi.org/10.7554/eLife.42512.017

In Experiment 1, we tested the discriminability of syntheses generated from the four pooling models in a set of 400 images that were novel to the participants. Experiment 2 examines the effect of image familiarity by repeatedly presenting a small number of source images. Experiment 3 tested the effect of cueing spatial attention on performance.

CNN model methods

Radial and angular pooling

In the texture synthesis approach of ***Gatys et al. (2015)***, spatial information is removed from the raw CNN activations by computing summary statistics (the Gram matrices of correlations between feature maps) over the whole image. In the 'foveated' pooling model we present here, we compute and match the mean of the feature maps (i.e. not the full Gram matrices) over local image regions by dividing the image into a number of radial and angular pooling regions (***Appendix 2—figure 1***). The radius defining the border between each radial pooling region is based on a given number of angular regions N_θ (which divide the circle evenly) and given by

$$r_i = r_0\left(1-\frac{\sin(\frac{\pi}{N_\theta})\cdot 2}{\alpha}\right)^i,$$

where r_i is the radius of each region i, r_0 is the outermost radius (set to be half the image size), and α is the ratio between the radial and angular difference. Radial regions were created for all i for which $r_i \geq 64$~px, corresponding to the preserved central region of the image (see below). We set $\alpha = 4$ because at this ratio $N_\theta \approx N_e$ (where N_e is the number of radial regions) for most N_θ. The value of N_θ corresponds to the model name used in the paper (e.g. 'CNN 4' uses $N_\theta = 4$).

We now apply these pooling regions to the activations of the VGG-19 deep CNN (***Simonyan and Zisserman, 2015***). For a subset of VGG-19 layers (conv1_1, conv2_1, conv3_1, conv4_1, conv5_1) we compute the mean activation for each feature map j in each layer l within each (radial or angular) pooling region p as

$$w_{pj}^l = \frac{1}{N_l}\sum_{k\in p}(F_{kj}^l),$$

where is N_l the size of the feature map of layer l in pixels and k is the (vectorised) spatial position in feature map F_j^l. The set of all w_{pj}^l provides parameters that specify the foveated image at a given scale. Note that while the radial and angular pooling region responses are computed separately, because they are added together to the loss function during optimisation (see below) they effectively overlap (as depicted in ***Appendix 2—figure 1***).

In addition, while the borders of the pooling regions are hard-edged (i.e. pooling regions are non-overlapping), the receptive fields of CNN units (area of pixels in the image that can activate a given unit in a feature map) can span multiple pooling regions. This means that the model parameters of a given pooling region will depend on image structure lying outside the pooling region (particularly for feature maps in the higher network layers). This encourages smooth transitions between pooling regions in the synthesised images.

Multiscale model

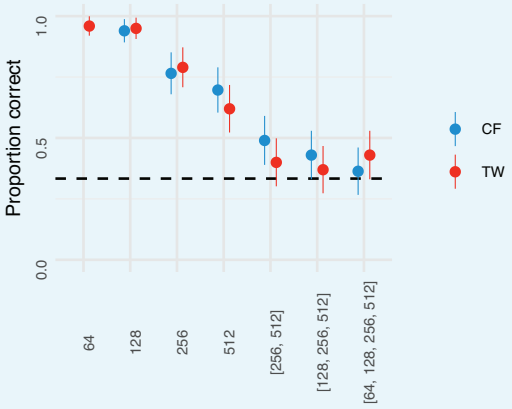
In the VGG-19 network, receptive fields of the units are squares of a certain size, and this size is independent of the input size of the image. That is, given a hypothetical receptive field centred in the image of size 128 ~ px square, the unit will be sensitive to one quarter of the image for input size 512 ~ px but half the image for input size 256. Therefore, the same unit in the network can receive image structure at a different scale by varying the input image size,

and in the synthesis process the low (high) frequency content can be reproduced with greater fidelity by using a small (large) input size.

We leverage this relationship to better capture long-range correlations in image structure (caused by for example edges that extend across large parts of the image) by computing and matching the model statistics over three spatial scales. This is not a controversial idea: for example, the model of ***Freeman and Simoncelli (2011)*** also computes features in a multiscale framework. How many scales is sufficient?

We evaluated the degree to which the number and combination of scales affected appearance in a psychophysical experiment on authors TW and CF. We matched 100 unique images using seven different models: four single-scale models corresponding to input sizes of 64, 128, 256 and 512 pixels, and three multiscale models in which features were matched at multiple scales ([256, 512 , 128, 256, 512] and [64, 128, 256, 512]). The foveated pooling regions corresponded to the CNN 32 model. Output images were upsampled to the final display resolution as appropriate. We discarded trials for which participants made no response (N = 2) or broke fixation (N = 5), leaving a total of 1393 trials for further analysis.

Appendix 2—figure 2 shows that participants are sensitive to the difference between model syntheses and original images when features are matched at only a single scale. However, using two or three scales appears to be sufficient to match appearance on average. As a compromise between fidelity and computational tractability, we therefore used three scales for all other experiments on the CNN appearance model. The final model used three networks consisting of the same radial and angular regions described above, computed over sizes 128, 256 and 512 ~ px square. The final model representation W therefore consists of the pooled feature map activations over three scales: $W = \{w^l_{pj,128}, w^l_{pj,256}, w^l_{pj,512}\}$.



Appendix 2—figure 2. Performance for discriminating model syntheses and original scenes for single- and multi-scale models (all with pooling regions corresponding to CNN 32) for participants CF and TW. Points show participant means (error bars show ±2 SEM), dashed line shows chance performance. The multiscale model with three scales produces close-to-chance performance.

DOI: https://doi.org/10.7554/eLife.42512.018

Gradient descent

As in ***Gatys et al. (2015)***, synthesised images are generated using iterative gradient descent, in which the mean-squared distance between the averaged feature maps of the original image and the synthesis is minimized. If T and W are the model representations for the synthesis and the original image respectively, then the loss for each layer is given by

$$L(\vec{x_t}, \vec{x_g}) = \sum_{l,s} \frac{1}{M_{ls} \cdot (N_\theta + N_e)} \sum_{p,j} (W^l_{pjs} - T^l_{pjs})^2,$$

where $\vec{x_t}$ and $\vec{x_g}$ are the vectorised pixels of the original and new image respectively, $M_{l,s}$ is the number of feature maps for layer l in scale s. A circular area in the middle of the image (radius 64 ~ px) is preserved to be the original image. Tiling pooling regions even for the centre of the image created reasonable syntheses but is prohibitively costly in generation time. To preserve the pixels in the circular area, the initialisation image of the gradient descent is identical to the original image. Outside the central area the gradient descent is initialised with Gaussian noise. The gradient descent used the L-BFGS optimiser (scipy implementation, ***Jones et al., 2001***) for 1000 iterations.

Experiment 1: Discriminability of CNN model syntheses for 400 unique images

This experiment measured whether any of the variants of the CNN scene appearance model could synthesise images that humans could not discriminate from their natural source images, and if so, identify the simplest variant of this model producing metamers. We chose a set of 400 images and had participants discriminate original and model-generated images in a temporal oddity paradigm.

Methods

The methods for this and the following psychophysical experiments were the same as in the main paper unless otherwise noted.

Participants

Thirteen participants participated in this experiment. Of these, ten participants were recruited via online advertisements and paid 15 Euro for a 1.5 hr testing session; the other three participants were authors AE, TW and CF. One session comprised one experiment using unique images (35 mins) followed by and one of repeated images (see below; 25 mins). All participants signed a consent form prior to participating. Participants reported normal or corrected-to-normal visual acuity. All procedures conformed to Standard 8 of the American Psychological Association's 'Ethical Principles of Psychologists and Code of Conduct' (2010).

Stimuli

We used 400 images (two additional images for authors, see below) from the MIT 1003 database (***Judd et al., 2012***; ***Judd et al., 2009***). One of the participants (TW) was familiar with the images in this database due to previous experiments. New images were synthesised using the multiscaled (512 px, 256 px, 128 px) foveated model described above, for four pooling region complexities (4, 8, 16 and 32). An image was synthesised for each of the 400 original images from each model (giving a total stimulus set including originals of 2000).

Procedure

Participants viewed the display from 60 cm; at this distance, pixels subtended approximately 0.024 degrees on average (approximately 41 pixels per degree of visual angle) – note that this is slightly further away than the experiment reported in the primary paper (changed to match the angular subtense used by Freeman and Simoncelli). Images therefore subtended $\approx 12.5^{\circ}$ at the eye. As in the main paper, the stimuli were presented for 200 ms, with an inter-stimulus interval of 1000 ms, followed by a 1200 ms response window. Feedback was provided by a 100 ms change in fixation cross brightness. Gaze position was recorded during the trial. If the participant moved the eye more than 1.5 degrees away from the fixation spot, feedback signifying a fixation break appeared for 200 ~ ms after the response feedback. Prior to the next trial, the state of the participant's eye position was monitored for 50 ms; if the eye position was reported as more than 1.5 degrees away from the fixation spot a recalibration was triggered. The inter-trial interval was 400 ms.

Each unique image was assigned to one of the four models for each participant (counterbalanced). That is, a given image might be paired with a CNN 4 synthesis for one participant and a CNN 8 synthesis for another. Showing each unique image only once ensures that the participants cannot become familiar with the images. For authors, images were divided into only CNN 8, CNN 16 and CNN 32 (making 134 images for each model and 402 trials in total for these participants). To ensure that the task was not too hard for naïve participants we added the easier CNN 4 model (making 100 images for each model version

and 400 trials in total). The experiment was divided into six blocks consisting of 67 trials (65 trials for the last block). After each block a break screen was presented telling the participant their mean performance on the previous trials. During the breaks the participants were free to leave the testing room to take a break and to rest their eyes. At the beginning of each block the eyetracker was recalibrated. Naïve participants were trained to do the task, first using a slower practice of 6 trials and second a correct-speed practice of 30 trials (using five images not part of the stimulus set for the main experiment).

Data analysis

We discarded trials for which participants made no response (N = 81) or broke fixation (N = 440), leaving a total of 4685 trials for further analysis.

Performance at each level of CNN model complexity was quantified using a logistic mixed-effects model. Correct responses were assumed to arise from a fixed effect factor of CNN model (with four levels) plus the random effects of participant and image. The model (in lme4-style notation) was

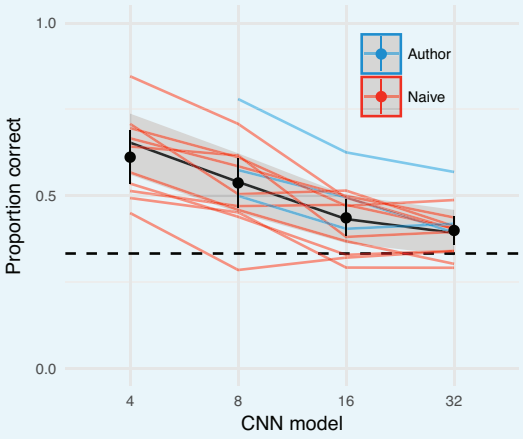
```
correct ~ model + (model | subj) + (model | im_code)
```

with `family = Bernoulli('logit')`, and using `contr.sdif` coding for the CNN model factor (***Venables and Ripley, 2002***).

The posterior distribution over model parameters was estimated using weakly-informative priors, which provide scale information about the setting of the model but do not bias effect estimates above or below zero. Specifically, fixed effect coefficients were given Cauchy priors with mean zero and SD 1, random effect standard deviations were given bounded Cauchy priors with mean 0.2 (indicating that we expect some variance between the random effect levels) and SD 1, with a lower-bound of 0 (variances cannot be negative), and correlation matrices were given LKJ(2) priors, reflecting a weak bias against strong correlations (***Stan Development, 2015***). The model posterior was estimated using MCMC implemented in the Stan language (version 2.16.2, ***Stan Development Team, 2017***; ***Hoffman and Gelman, 2014***), with the model wrapper package brms (version 1.10.2, ***Bürkner, 2017***) in the R statistical environment. We computed four chains of 15,000 steps, of which the first 5000 steps were used to tune the sampler; to save disk space we only saved every 5th sample.

Results and discussion

The CNN 32 model came close to matching appearance on average for a set of 400 images. Discrimination performance for ten naïve participants and three authors is shown in ***Appendix 2—figure 3*** (lines link individual participant means, based on at least 64 trials, median 94). All participants achieve above-chance performance for the simplest model (CNN 4), indicating that they understood and could perform the task. Performance deteriorates as models match the structure of the original image more precisely.



Appendix 2—figure 3. The CNN model comes close to matching appearance on average.

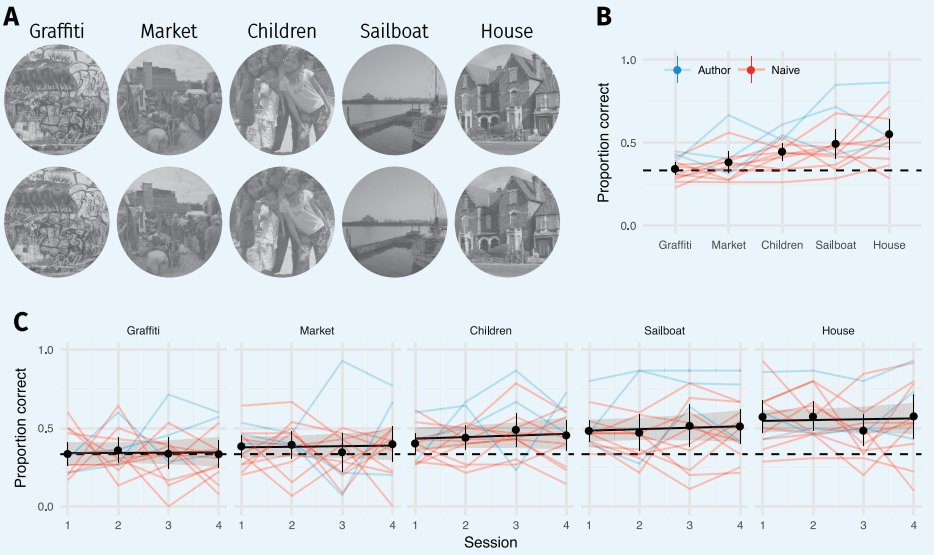
Oddity performance as a function of the CNN image model. Points show mean over participants (error bars ±2 SEM), coloured lines link the mean performance of each participant for each pooling model. For most participants, performance falls to approximately chance (dashed horizontal line) for the CNN 32 model. Black line and shaded regions show the mean and 95% credible intervals on the population mean derived from a mixed-effects model.

DOI: https://doi.org/10.7554/eLife.42512.019

To quantify the data, we estimated the posterior distribution of a logistic mixed-effects model with a population-level (fixed-effect) factor of CNN model, whose effect was nested within participants and image (i.e. random effects of participant and image). Regression coefficients coded the difference between successive CNN models, expressed using sequential difference coding from the MASS package (***Venables and Ripley, 2002***), and are presented below as the values of the linear predictor (corresponding to log odds in a logistic model). Mean performance had a greater than 0.99 posterior probability of being lower for CNN 8 than CNN 4 (-0.48, 95% CI [−0.74,–0.23], $\mathrm{p}(\beta<0)>0.999$), and for CNN 16 being lower than CNN 8 (-0.43, 95% CI [−0.68,–0.18], $\mathrm{p}(\beta<0)=0.999$); whereas the difference between the 16 and 32 models was somewhat smaller (−0.17, 95% CI [−0.37, 0.03], $\mathrm{p}(\beta<0)=0.951$). Most participants performed close to chance for the CNN 32 model (excluding authors, the population mean estimate had a 0.88 probability of being greater than chance; including authors this value was 0.96). Therefore, the model is capable of synthesising images that are indiscriminable from a large set of arbitrary scenes in our experimental conditions, on average, for naïve participants. However, one participant (author AE) performs noticably better than the others, even for the CNN 32 model. AE had substantial experience with the type of distortions produced by the model but had never seen this set of original images before. Therefore, the images produced by the model are not true metamers, because they do not encapsulate the limits of visible structure for all humans.

Experiment 2: Image familiarity and learning tested by repeated presentation

It is plausible that familiarity with the images played a role in the results above. That is, the finding that images become difficult on average to discriminate with the CNN 32 model may depend in part on participants having never seen the images before. To investigate the role that familiarity with the source images might play, the same participants as in the experiment above performed a second experiment in which five of the original images from the first experiment were presented 60 times, using 15 unique syntheses per image generated with the CNN 32 model (***Appendix 2—figure 4A***).



Appendix 2—figure 4. Familiarity with original image content did not improve discrimination performance. (**A**) Five original images (top) were repeated 60 times (interleaved over 4 blocks), and observers discriminated them from CNN 32 model syntheses (bottom). (**B**) Proportion of correct responses for each image from A. Some images are easier than others, even for the CNN 32 model. (**C**) Performance as a function of each 75-trial session reveals little evidence that performance improves with repeated exposure. Points show grand mean (error bars show bootstrapped 95% confidence intervals), lines link the mean performance of each observer for each pooling model (based on at least 5 trials; median 14). Black line and shaded region shows the posterior mean and 95% credible intervals of a logistic mixed-effects model predicting the population mean performance for each image. Images from the MIT 1003 dataset (***Judd et al., 2009***, https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) and reproduced under a CC-BY license (https://creativecommons.org/licenses/by/3.0/) with changes as described in the Materials and methods.

DOI: https://doi.org/10.7554/eLife.42512.020

Methods

Participants

The same thirteen participants participated as in Experiment 1.

Stimuli

We selected five images from the set of 400 and generated 15 new syntheses for each of these images from the CNN 32 model (yielding a stimulus set of 80 images).

Procedure

Each pairing of unique image (5) and synthesis (15) was shown in one block of 75 trials (pseudo-random order with the restriction that trials from the same source image could never follow one another). Participants performed four such blocks, yielding 300 trials in total (60 repetitions of each original image).

Data analysis

We discarded trials for which participants made no response (N = 63) or broke fixation (N = 294), leaving a total of 3543 trials for further analysis. Model fitting was as for Experiment 1 above, except that the final posterior was based on four chains of 16,000 steps, of which the first 8000 steps were used to tune the sampler; to save disk space we only saved every 4th sample.

The intercept-only model (assuming only random effects variation but no learning) was specified as

```
correct ~ 1 + (1 | subj) + (1 | im_name)
```

and the learning model was specified as

```
correct ~ session + (session | subj) + (session | im_name)
```

We compare models using an information criterion (LOOIC, ***Vehtari et al., 2016***; see also ***Gelman et al., 2014***; ***McElreath, 2016***) that estimates of out-of-sample prediction error on the deviance scale.

Results and discussion

While some images (e.g. House) could be discriminated quite well by most participants (***Appendix 2—figure 4B***), others (e.g. Graffiti) were almost indiscriminable from the model image for all participants (posterior probability that the population mean was above chance performance was 0.61 for Graffiti, 0.93 for Market, and greater than 0.99 for all other images). This image dependence shows that even the CNN 32 model is insufficient to produce metamers for arbitrary scenes.

Furthermore, there was little evidence that participants learned over the course of sessions (***Appendix 2—figure 4C***). The population-level linear slope of session number was 0.02, 95% CI [−0.1, 0.15], $p(\beta<0)=0.326$, and the LOOIC comparison between the intercept-only model and the model containing a learning term indicated equivocal evidence if random-effects variance was included (LOOIC difference 3.3 in favour of the learning model, SE = 6.1) but strongly favoured the intercept model if only fixed-effects were considered (LOOIC difference −23.3 in favour of the intercept model, SE = 1.7). The two images with the most evidence for learning were Children (median slope 0.04, 95% CI [−0.08, 0.17], $p(\beta<0)=0.247$) and Sailboat (0.04, 95% CI [−0.08, 0.17], $p(\beta<0)=0.269$). Two authors showed some evidence of learning: AE (0.17, 95% CI [−0.03, 0.37], $p(\beta<0)=0.047$), and CF (0.22, 95% CI [0.03, 0.44], $p(\beta<0)=0.008$). Overall, these results show that repeated image exposures with response feedback did not noticably improve performance.

Experiment 3: Spatial cueing of attention

The experiment presented in the primary paper showed that the discriminability of model syntheses depended on the source images, with scene-like images being easier to discriminate from model syntheses than texture-like images for a given image model. This finding was replicated in an ABX paradigm (above) and the general finding of source-image-dependence was corroborated by the data with repeated images (***Appendix 2—figure 3***). One possible reason for this image-dependence could be that participants found it easier to know where to attend in some images than in others, creating an image-dependence not due to the summary statistics per se. Relatedly, ***Cohen et al. (2016)*** suggest that the limits imposed by an ensemble statistic representation can be mitigated by the deployment of spatial attention to areas of interest. Can the discriminability of images generated by our model be influenced by focused spatial attention?

To probe this possibility we cued participants to a spatial region of the scene before the trial commenced. We computed the mean squared error (MSE) between the original and synthesised images within 12 partially-overlapping wedge-like regions subtending 60˚. We computed MSE in both the pixel space (representing the physical difference between the two images) and in the feature space of the fifth convolutional layer (conv5_1) of the VGG-19 network, with the hypothesis that this might represent more perceptually relevant information, and thus be a more informative cue.

We pre-registered the following hypotheses for this experiment (available at http://dx.doi.org/10.17605/OSF.IO/MBGSQ; click on 'View Registration Form'). For the overall effect of cueing (the primary outcome of interest), we hypothesised that

- performance in the Valid:Conv5 condition would be higher than the Uncued condition and
- performance in the Invalid condition would be lower than the Uncued condition

These findings would be consistent with the account that spatial attention can be used to overcome ensemble statistics in the periphery, providing that it is directed to an informative location. This outcome also assumes that our positive cues (Conv5 and Pixels) identify informative locations.

Alternative possibilities are

- if focussed spatial attention cannot influence the 'resolution' of the periphery in this task, then performance in the Valid:Conv5 and Invalid conditions will be equal to the Uncued condition.
- if observers use a global signal ('gist') to perform the task, performance in the Uncued condition would be higher than the Valid:Conv5 and Invalid conditions. That is, directing spatial attention interferes with a gist cue.

Our secondary hypothesis concerns the difference between Valid:Conv5 and Valid:Pixel cues. A previous analysis at the image level (see below) found that conv5 predicted image difficultly slightly better than the pixel space. We therefore predicted that Valid spatial cues based on Conv5 features (Valid:Conv5) should be more effective cues, evoking higher performance, than Valid:Pixel cues.

Methods

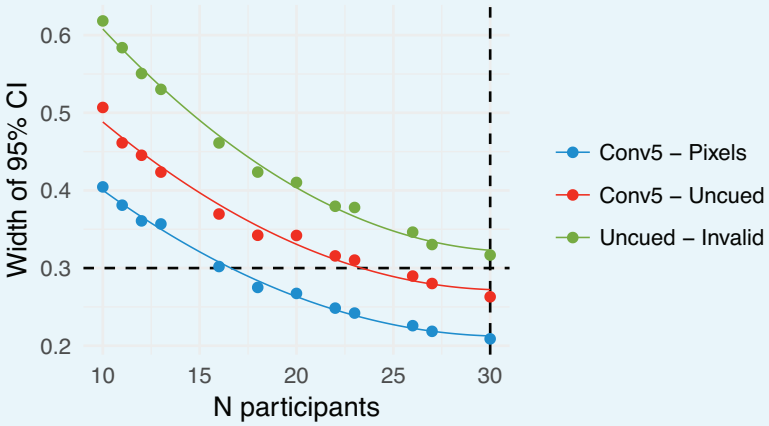
Participants

We pre-registered (http://dx.doi.org/10.17605/OSF.IO/MBGSQ) the following data collection plan with a stopping rule that depended on the precision (***Kruschke, 2015***). Specifically, we collected data from a minimum of 10 and a maximum of 30 participants, planning to stop in the intermediate range if the 95% credible intervals for the two parameters of interest (population fixed-effect difference between Valid and Uncued, and population fixed-effect difference between Invalid and Uncued) spanned a width of 0.3 or less on the linear predictor scale.

This value was determined as 75% of the width of our 'Region of Practical Equivalence' to zero effect (ROPE), pre-registered as [−0.2, 0.2] on the linear predictor scale (this corresponds to odds ratios of [0.82, 1.22]). We deemed any difference smaller than this value as being too small to be practically important.

As an example, if the performance in one condition is 0.5, then an increase of 0.2 in the linear predictor corresponds to a performance of 0.55. The target for precision was then determined as 75% of the ROPE width, in order to give a reasonable chance for the estimate to lie within the ROPE (***Kruschke, 2015***).

We tested these conditions by fitting the data model (see below) after every participant after the 10th, stopping if the above conditions were met. However, as shown in ***Appendix 2—figure 5***, this precision was not met with our maximum of 30 participants, and so we ceased data collection at 30, deeming further data collection beyond our resources for the experiment. Thus our data should be interpreted with the caveat that the desired precision was not reached (though we got close).



Appendix 2—figure 5. Parameter precision as a function of number of participants. (**A**) Width of the 95% credible interval on three model parameters as a function of the number of

participants tested. Points show model fit runs (the model was not re-estimated after every participant due to computation time required). We aimed to achieve a width of 0.3 (dashed horizontal line) on the linear predictor scale, or stop after 30 participants. The Uncued - Invalid parameter failed to reach the desired precision after 30 participants. Lines show fits of a quadratic polynomial as a visual guide.

DOI: https://doi.org/10.7554/eLife.42512.021

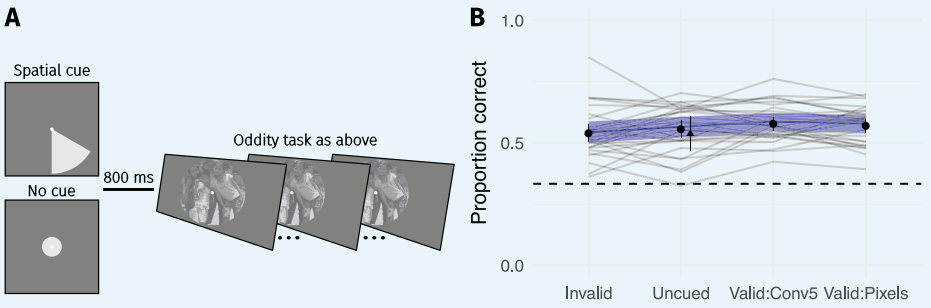
An additional five participants were recruited but showed insufficient eyetracking accuracy or training performance (criteria pre-registered). Of the 30, three were lab members unfamiliar with the purpose of the study, the other 27 were recruited online; all were paid 15 Euro for the 1.5 hr testing session. Of these, three participants did not complete the full session due to late arrival, and eyetracking calibration failed in the second last trial block for an additional participant.

Stimuli

This experiment used the same 400 source images and CNN 8 model syntheses as Experiment 1.

Procedure

The procedure for this experiment was as in Experiment 1 with the following exceptions. The same 400 original images were used as in Experiment 1, all with syntheses from the CNN 8 model. A trial began with the presentation of a bright wedge (60 degree angle, Weber contrast 0.25) or circle (radius 2 dva) for 400 ~ ms, indicating a spatial cue (85% of trials) or Uncued trial (15%) respectively (***Appendix 2—figure 6A***). A blank screen with fixation spot was presented for 800 ms before the oddity paradigm proceeded as above. On spatial cue trials, participants were cued to the wedge region containing either the largest pixel MSE between the original and synthesised images (35% of all trials), the largest conv5 MSE (35%), or the *smallest* pixel MSE (an invalid cue, shown on 15% of all trials). Thus, 70% of all trials were valid cues, encouraging participants to make use of the cues rather than learning to ignore them. Participants were also instructed to attend to the cued region on trials where a wedge was shown. For Uncued trials they were instructed to attend globally over the image. Cueing conditions were interleaved and randomly assigned to each unique image for each participant. The experiment was divided into eight blocks of 50 trials. Before the experiment we introduced participants to the task and fixation control with repeated practice sessions of 30 trials (using 30 images not used in the main experiment and with the CNN 4 model syntheses). Participants saw at least 60 and up to 150 practice trials, until they were able to get at least 50% correct and with 20% or fewer trials containing broken fixations or blinks.



Appendix 2—figure 6. Cueing spatial attention has little effect on performance. (**A**) Covert spatial attention was cued to the area of the largest difference between the images (70% of trials; half from conv5 feature MSE; half from pixel MSE) via a wedge stimulus presented before the trial. On 15% of trials the wedge cued an invalid location (smallest pixel MSE), and on 15% of trials no cue was provided (circle stimulus). (**B**) Performance as a function of cueing condition for 30 participants. Points show grand mean (error bars show ± 2 SE), lines link the mean performance of each observer for each pooling model (based on at least 30 trials; median 65). Blue lines and shaded area show the population mean estimate and 95% credible intervals from the mixed-effects model. Triangle in the Uncued condition replots the average

performance from CNN 8 in ***Figure 3*** for comparison. Images from the MIT 1003 dataset (***Judd et al., 2009***, https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) and reproduced under a CC-BY license (https://creativecommons.org/licenses/by/3.0/) with changes as described in the Materials and methods.

DOI: https://doi.org/10.7554/eLife.42512.022

Data analysis

We discarded trials for which participants made no response (N = 141) or broke fixation (N = 1398), leaving a total of 10261 trials for further analysis.

This analysis plan was pre-registered and is available at http://dx.doi.org/10.17605/OSF.IO/MBGSQ (click on 'view registration form'). We seek to estimate three performance differences:

1. The difference between Invalid and Uncued
2. The difference between Valid:Conv5 and Uncued
3. The difference between Valid:Conv5 and Valid:Pixels

The model formula (in lme4-style formula notation) is

```
correct ~ cue + (cue | subj) + (cue | im_code)
```

with `family = Bernoulli('logit')`. The 'cue' factor uses custom contrast coding (design matrix) to test the hypotheses of interest. Specifically, the design matrix for the model above was specified as

	β_0	β_1	β_2	β_3
Invalid	1	-1	0	0
Uncued	1	1	-1	0
Valid:Conv5	1	0	1	1
Valid:Pixels	1	0	0	-1

Therefore, β_1 codes Uncued - Invalid, β_2 codes Valid:Conv5 - Uncued, β_3 codes Valid:Conv5 - Valid:Pixels and β_0 codes the Intercept (average performance). Note that the generalised inverse of this matrix was passed to brms (***Venables and Ripley, 2002***).

Each of these population fixed-effects is offset by the random effects of participant (`subj`) and image (`im_code`). We also assume that the offsets for each fixed effect can be correlated (denoted by the single pipe character |). The model thus estimates:

1. Four fixed-effect coefficients. The coefficients coding Valid:Conv5 – Uncued and Uncued – Invalid constitute the key outcome measures of the study. The final coefficient is the analysis of secondary interest.
2. Eight random-effects standard-deviations (four for each fixed-effect, times two for the two random effects).
3. Twelve correlations (six for each pairwise relationship between the fixed-effects, times two for the two random effects).

These parameters were given weakly-informative prior distributions as for Experiment 1 (above): fixed-effects had Cauchy(0, 1) priors, random effect SDs had bounded Cauchy(0.2, 1) priors, and correlation matrices had LKJ(2) priors.

To judge the study outcome we pre-defined a region of practical equivalence (ROPE) around zero effect (0) of $[-0.2, 0.2]$ on the linear predictor scale. This corresponds to odds ratios of [0.82, 1.22]. Our decision rules were then:

- If the 95% credible interval of the parameter value falls outside the ROPE, we consider there to be a credible difference between the conditions.
- If the 95% credible interval of the parameter value falls fully within the ROPE, we consider there to be no practical difference between the conditions. This does not mean that there is no effect, but only that it is unlikely to be large.
- If the 95% credible interval overlaps the ROPE, the data are ambiguous as to the conclusion for our hypothesis. This does not mean that the data give no insight into the direction and

magnitude of any effect, but only that they are ambiguous with respect to our decision criteria.

For more discussion of this approach to hypothesis testing, see (***Kruschke, 2015***).

Results and discussion

The results of this experiment are shown in ***Appendix 2—figure 6B***. While mean performance across conditions was in the expected direction for all effects, no large differences were observed. Specifically, the population-level coefficient estimate on the linear predictor scale for the difference between the Valid:Conv5 cueing condition and the uncued condition was 0.09, 95% CI [−0.05, 0.22], $p(\beta < 0) = 0.1$. Given our decision rules above, the coefficient does not fall wholely within the ROPE and therefore this result is somewhat inconclusive; in general the difference is rather small and so large 'true' effects of spatial cueing are quite unlikely. Similarly, we find no large difference between uncued performance and the invalid cues (0.09, 95% CI [−0.07, 0.25], $p(\beta < 0) = 0.141$). Based on our pre-registered cutoff for a meaningful effect size we conclude that cueing spatial attention in this paradigm results in effectively no performance change.

We further hypothesised that the conv5 cue would be more informative (resulting in a larger performance improvement) than the pixel MSE cue. Note that for 269 of 400 images the conv5 and pixel MSE cued the same or neighbouring wedges, meaning that the power of this experiment to detect differences between these conditions is limited. Consistent with this and contrary to our hypothesis, we find no practical difference between the Valid:Conv5 and Valid:Pixels conditions, 0.04, 95% CI [−0.07, 0.14], $p(\beta < 0) = 0.253$. Note that for this comparison, the 95% credible intervals for the parameter fall entirely within the ROPE, leading us to conclude that there is no practical difference between these conditions in our experiment.

To conclude, our results here suggest that if cueing spatial attention improves the 'resolution' of the periphery, then the effect is very small. ***Cohen et al. (2016)*** have suggested that an ensemble representation serves to create phenomenal experience of a rich visual world, and that spatial attention can be used to gain more information about the environment beyond simple summary statistics. The results here are contrary to this idea, at least for the specific task and setting we measure here.

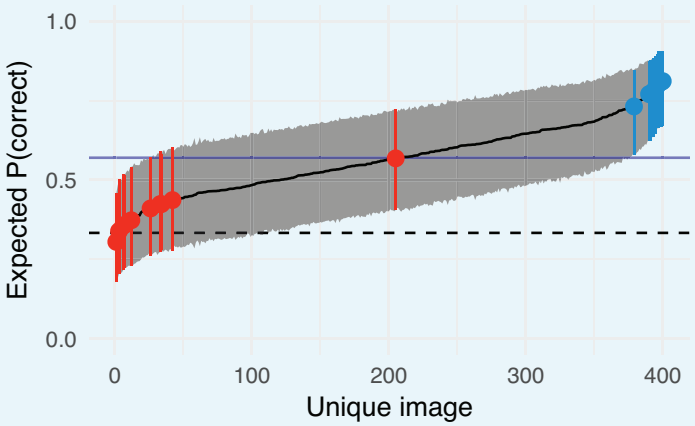
Note however that other experimental paradigms may in general be more suitable for assessing the influence of spatial attention than a temporal oddity paradigm. For example, in temporal oddity participants may choose to reallocate spatial attention after the first interval is presented (e.g. on invalid trials pointing at regions of sky). In this respect a single-interval yes-no design (indicating original/synthesis) might be preferable. However, analysis of such data with standard signal detection theory would need to assume that the participants' decision criteria remain constant over trials, whereas it seems likely that decision criteria would depend strongly on the image. To remain consistent with our earlier experiments we nevertheless employed a three-alternative temporal oddity task here; future work could assess whether our finding of minimal influence of spatial cueing depends on this choice.

Selection of scene- and texture-like images

As discussed in the main paper, we used the results of a pilot experiment (Experiment 3, above) to help select images to provide a strong test of the FS-model. Briefly, 30 observers discriminated 400 images from syntheses produced by the CNN 8 model. Each image was paired with only one unique synthesis (see Experiment 3 above for further details on the experiment).

In an exploratory analysis of that data, we found that there was a large range of difficulty for individual images (as in Experiment 2, above). ***Appendix 2—figure 7*** shows the image-specific intercepts estimated by the model described above. We examine this rather than the raw data because cueing conditions were randomly assigned to each image for each subject, meaning that the mean performance of the images will depend on this randomisation (though, given our results, the effects are likely to be small). The image-specific intercept from the model estimates the difficulty of each image, statistically marginalising over cueing

condition. While the posterior means for some images were close to chance, and the 95% credible intervals associated with about 100 images overlapped chance performance, approximately 30 images were easily discriminable from their model syntheses, lying above the mean performance for all images with the CNN 8 model.



Appendix 2—figure 7. Estimated difficulty of each image in Experiment 3 (syntheses with the CNN 8 model) and the images chosen to form the texture- and scene-like categories in the main experiment. Solid black line links model estimates of each image's difficulty (the posterior mean of the image-specific model intercept, plotted on the performance scale). Shaded region shows 95% credible intervals. Dashed horizontal line shows chance performance; solid blue horizontal line shows mean performance. Red and blue points denote the images chosen as texture- and scene-like images in the main experiment respectively. The red point near the middle of the range is the 'graffiti' image from the experiments above.

DOI: https://doi.org/10.7554/eLife.42512.024

Appendix 2—figure 8. The 50 easiest images from ***Appendix 2—figure 7*** where difficulty increases left-to-right, top-to-bottom. Images chosen for the main experiment as 'scene-like' are circled in blue. Images from the MIT 1003 dataset (***Judd et al., 2009***, https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) and reproduced under a CC-BY license (https://creativecommons.org/licenses/by/3.0/).

DOI: https://doi.org/10.7554/eLife.42512.025

Our final selection of ten images per category was made by examining the easiest and hardest images from this experiment (***Appendix 2—figure 7***) and selecting ten images we subjectively judged to contain scene-like or texture-like content. The final images used in the first experiment of the main paper are shown in ***Appendix 2—figure 7*** as coloured points. The 50 easiest and 50 hardest images are shown in ***Appendix 2—figures 8,9*** respectively.

Appendix 2—figure 9. The 50 hardest images from ***Appendix 2—figure 7*** where difficulty decreases left-to-right, top-to-bottom. Images chosen for the main experiment as 'texture-like' are circled in red. Images from the MIT 1003 dataset (***Judd et al., 2009***, https://people.csail.mit.edu/tjudd/WherePeopleLook/index.html) and reproduced under a CC-BY license (https://creativecommons.org/licenses/by/3.0/).

DOI: https://doi.org/10.7554/eLife.42512.026

Predicting the difficulty of individual images

As shown above, some images are easier than others. We assessed whether an image-based metric considering the difference between original and synthesised images could predict difficulty at the image level. Specifically, we asked whether the mean squared-error (MSE) between the original and synthesised images in two feature spaces (conv5 and pixels) could predict the relative difficulty of the source images. Note that we performed this analysis first on the results of Experiment 1 (***Appendix 2—figure 3***), and that these results were used to inform the hypothesis regarding the usefulness of conv5 vs pixel cues presented in Experiment 3, above. We subsequently performed the same analysis on the data from Experiment 3. We present both analyses concurrently here for ease of reading, but the reader should be aware of the chronological order.

Methods

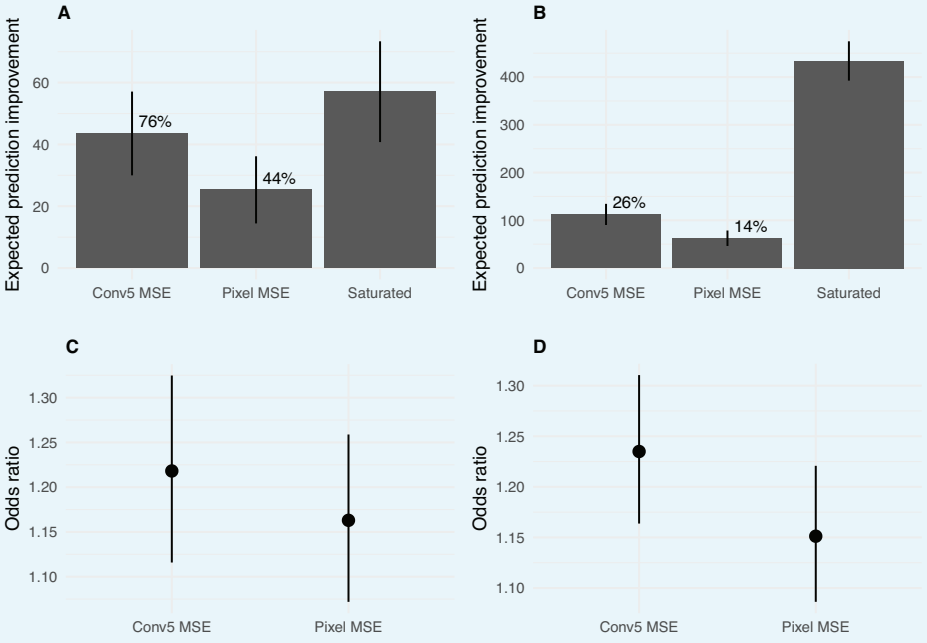
We computed the mean squared error between the original and synthesised images in two feature spaces. First, the MSE in the pixel space was used to represent the physical difference at all spatial scales. Second, the difference in feature activations in the conv5 layer of the VGG network was used as an abstracted feature space which may correspond to aspects of human perception (e.g. ***Kubilius et al., 2016***, see also ***Geirhos et al., 2019***). Both are also correlated with the final value of the loss function from our synthesis procedure. As a baseline we fit a mixed-effects logistic regression containing fixed-effects for the levels of the CNN model and a random effect of observer on all fixed effect terms. As a 'saturated' model (a weak upper bound) we added a random effect for image to the baseline model (that is, each image is uniquely predicted given the available data). Using the scale defined by the baseline and saturated models, we then compared models in which the image-level predictor (pixel or conv5 MSE, standardised to have zero mean and unit variance within each CNN model level) was added as an additional linear covariate to the baseline model. That is, each

image was associated with a scalar value of pixel/conv5 MSE with each synthesis. Additional image-level predictors were compared but are not reported here because they performed similarly or worse than the conv5 or pixel MSE.

As above, we compared the models using the LOOIC information criterion that estimates out-of-sample prediction error on the deviance scale. Qualitatively similar results were found using ten-fold crossvalidation for models fit with penalised maximum-likelihood in lme4.

Results

For the dataset from Experiment 1, the LOOIC favoured the model containing conv5 MSE over the pixel MSE (LOOIC difference 18.2, SE = 8.3) and the pixel MSE over the baseline model (LOOIC difference 25.3, SE = 10.9)—see ***Appendix 2—figure 10A***. The regression weight of the standardised pixel MSE feature fit to all the data was 0.04 (95% credible interval = 0.15–0.07), and the weight of the standardised conv5 feature was 0.04 (0.2–0.11; presented as odds ratios in ***Appendix 2—figure 10C***). Therefore, a one standard deviation increase in the conv5 feature produced a slightly larger increase in the linear predictor (and thus the expected probability) than the pixel MSE, in agreement with the model comparison.



Appendix 2—figure 10. Predicting image difficulty using image-based metrics. (**A**) Expected prediction improvement over a baseline model for models fit to the data from Experiment 1 (***Appendix 2—figure 3***), as estimated by the LOOIC (***Vehtari et al., 2016***). Values in deviance units (−2 * log likelihood; higher is better). Error bars show ±2 1 SE. Percentages are expected prediction improvement relative to the saturated model. (**B**) Same as A but for the data from Experiment 3 (***Appendix 2—figure 6***). (**C**) Odds of a success for a one SD increase in the image predictor for data from Experiment 1. Points show mean and 95% credible intervals on odds ratio (exponentiated logistic regression weight). (**D**) As for C for Experiment 3.

DOI: https://doi.org/10.7554/eLife.42512.027

Applying this analysis to the data from Experiment 3 lead to similar results (***Appendix 2—figure 10B,D***). The LOOIC favoured the model containing conv5 MSE over the pixel MSE (LOOIC difference 49.9, SE = 13.3) and the pixel MSE over the baseline model (LOOIC difference 62.4, SE = 16.2). Note that the worse performance of the

image metric models relative to the saturated model (compared to ***Appendix 2—figure 10A***) is because the larger data mass in this experiment provides a better constraint for the random effects estimates of image. The regression weight of the standardised pixel MSE feature fit to all the data was 0.03 (95% credible interval = 0.14–0.08), and the weight of the standardised conv5 feature was 0.03 (0.21–0.15).

These results show that the difficulty of a given image can be to some extent predicted from the pixel differences or conv5 differences, suggesting these might prove useful full-reference metrics, at least with respect to the distortions produced by our CNN model.

P4: Five points to check when comparing visual perception in humans and machines

Christina M. Funke*, Judy Borowski*, Karolina Stosio, Wieland Brendel†, Thomas S.A. Wallis†, Matthias Bethge†

* joint first authors, † joint senior authors

Published in *Journal of Vision, 21(3), 16–16*.

Contributions The closed contour case study was designed by CMF, JB, TSAW, and MB and later with WB. The code for the stimuli generation was developed by CMF. The neural networks were trained by CMF and JB. The psychophysical experiments were performed and analyzed by CMF, TSAW, and JB. The SVRT case study was conducted by CMF under supervision of TSAW, WB, and MB. KS designed and implemented the recognition gap case study under the supervision of WB and MB; JB extended and refined it under the supervision of WB and MB. The initial idea to unite the three projects was conceived by WB, MB, TSAW, and CMF, and further developed including JB. The first draft was jointly written by JB and CMF with input from TSAW and WB. All authors contributed to the final version and provided critical revisions.

Five points to check when comparing visual perception in humans and machines

Christina M. Funke[*] University of Tübingen, Tübingen, Germany

Judy Borowski[*] University of Tübingen, Tübingen, Germany

Karolina Stosio University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen and Berlin, Germany
Volkswagen Group Machine Learning Research Lab, Munich, Germany

Wieland Brendel[†] University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen and Berlin, Germany
Werner Reichardt Centre for Integrative Neuroscience, Tübingen, Germany

Thomas S. A. Wallis[†] University of Tübingen, Tübingen, Germany
Present address: Amazon.com, Tübingen

Matthias Bethge[†] University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen and Berlin, Germany
Werner Reichardt Centre for Integrative Neuroscience, Tübingen, Germany

With the rise of machines to human-level performance in complex recognition tasks, a growing amount of work is directed toward comparing information processing in humans and machines. These studies are an exciting chance to learn about one system by studying the other. Here, we propose ideas on how to design, conduct, and interpret experiments such that they adequately support the investigation of mechanisms when comparing human and machine perception. We demonstrate and apply these ideas through three case studies. The first case study shows how human bias can affect the interpretation of results and that several analytic tools can help to overcome this human reference point. In the second case study, we highlight the difference between necessary and sufficient mechanisms in visual reasoning tasks. Thereby, we show that contrary to previous suggestions, feedback mechanisms might not be necessary for the tasks in question. The third case study highlights the importance of aligning experimental conditions. We find that a previously observed difference in object recognition does not hold when adapting the experiment to make conditions more equitable between humans and machines. In presenting a checklist for comparative studies of visual reasoning in humans and machines, we hope to highlight how to overcome potential pitfalls in design and inference.

Introduction

Until recently, only biological systems could abstract the visual information in our world and transform it into a representation that supports understanding and action. Researchers have been studying how to implement such transformations in artificial systems since at least the 1950s. One advantage of artificial systems for understanding these computations is that many analyses can be performed that would not be possible in biological systems. For example, key

Citation: Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S. A., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3):16, 1–23, https://doi.org/10.1167/jov.21.3.16.

https://doi.org/10.1167/jov.21.3.16 Received April 21, 2020; published March 16, 2021 ISSN 1534-7362

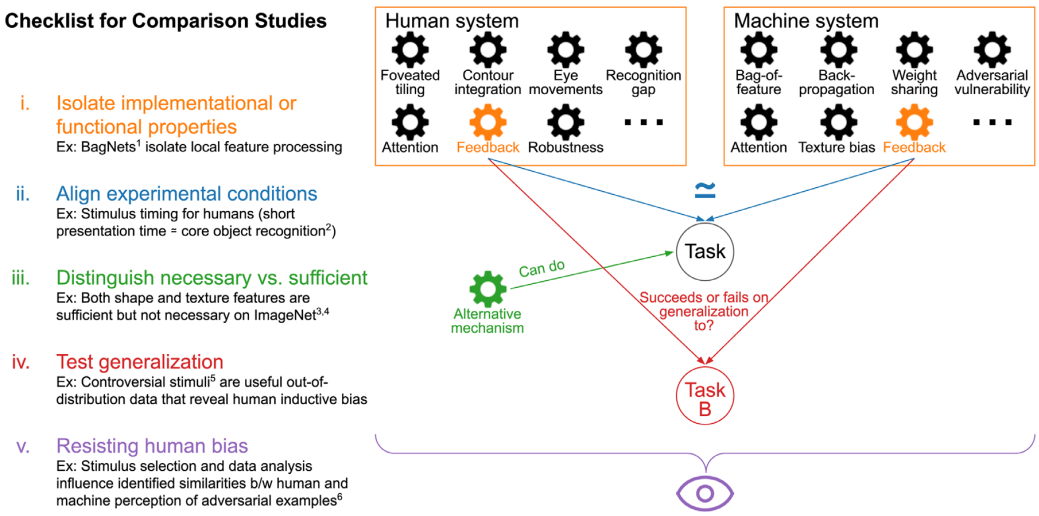


Figure 1. i: The human system and a candidate machine system differ in a range of properties. Isolating a specific mechanism (for example, feedback) can be challenging. ii: When designing an experiment, equivalent settings are important. iii: Even if a specific mechanism was important for a task, it would not be clear if this mechanism is necessary, as there could be other mechanisms (that might or might not be part of the human or machine system) that can allow a system to perform well. iv: Furthermore, the identified mechanisms might depend on the specific experimental setting and not generalize to, for example, another task. v: Overall, our human bias influences how we conduct and interpret our experiments. [1]Brendel and Bethge (2019); [2]DiCarlo et al. (2012); [3]Geirhos, Rubisch, et al. (2018); [4]Kubilius et al. (2016); [5]Golan et al. (2019); [6]Dujmović et al. (2020).

components of visual processing, such as the role of feedback connections, can be investigated, and methods such as ablation studies gain new precision.

Traditional models of visual processing sought to explicitly replicate the hypothesized computations performed in biological visual systems. One famous example is the hierarchical HMAX-model (Fukushima, 1980; Riesenhuber & Poggio, 1999). It instantiates mechanisms hypothesized to occur in primate visual systems, such as template matching and max operations, whose goal is to achieve invariance to position, scale, and translation. Crucially, though, these models never got close to human performance in real-world tasks.

With the success of learned approaches in the past decade, and particularly that of convolutional deep neural networks (DNNs), we now have much more powerful models. In fact, these models are able to perform a range of constrained image understanding tasks with human-like performance (Krizhevsky et al., 2012; Eigen & Fergus, 2015; Long et al., 2015).

While matching machine performance with that of the human visual system is a crucial step, the inner workings of the two systems can still be very different. We hence need to move beyond comparing accuracies to understand how the systems' mechanisms differ (Geirhos et al., 2020; Chollet, 2019; Ma & Peters, 2020; Firestone, 2020).

The range of frequently considered mechanisms is broad. They not only concern the architectural level (such as feedback vs. feed-forward connections, lateral connections, foveated architectures or eye movements, …), but also involve different learning schemes (back-propagation vs. spike-timing-dependent plasticity/Hebbian learning, …) as well as the nature of the representations themselves (such as reliance on texture rather than shape, global vs. local processing, …). For an overview of comparison studies, please see Appendix A.

Checklist for psychophysical comparison studies

We present a checklist on how to design, conduct, and interpret experiments of comparison studies that investigate relevant mechanisms for visual perception. The diagram in Figure 1 illustrates the core ideas that we elaborate on below.

i. **Isolating implementational or functional properties.** Naturally, the systems that are being compared often differ in more than just one aspect, and hence pinpointing one single reason for an observed difference can be challenging. One approach is to design an artificial network constrained such that the mechanism of interest will show its effect as clearly as possible. An example of such an attempt

is Brendel and Bethge (2019), which constrained models to process purely local information by reducing their receptive field sizes. Unfortunately, in many cases, it is almost impossible to exclude potential side effects from other experimental factors such as architecture or training procedure. Therefore, making explicit if, how, and where results depend on other experimental factors is important.

ii. **Aligning experimental conditions for both systems.** In comparative studies (whether humans and machines, or different organisms in nature), it can be exceedingly challenging to make experimental conditions equivalent. When comparing the two systems, any differences should be made as explicit as possible and taken into account in the design and analysis of the study. For example, the human brain profits from lifelong experience, whereas a machine algorithm is usually limited to learning from specific stimuli of a particular task and setting. Another example is the stimulus timing used in psychophysical experiments, for which there is no direct equivalent in stateless algorithms. Comparisons of human and machine accuracies must therefore be considered with the temporal presentation characteristics of the experiment. These characteristics could be chosen based on, for example, a definition of the behavior of interest as that occurring within a certain time after stimulus onset (as for, e.g., "core object recognition"; DiCarlo et al., 2012). Firestone (2020) highlights that as aligning systems perfectly may not be possible due to different "hardware" constraints such as memory capacity, unequal performance of two systems might still arise despite similar competencies.

iii. **Differentiating between necessary and sufficient mechanisms.** It is possible that multiple mechanisms allow good task performance — for example, DNNs can use either shape or texture features to reach high performance on ImageNet (Geirhos, Rubisch, et al., 2018; Kubilius et al., 2016). Thus, observing good performance for one mechanism does not imply that this mechanism is strictly necessary or that it is employed by the human visual system. As another example, Watanabe et al. (2018) investigated whether the rotating snakes illusion (Kitaoka & Ashida, 2003; Conway et al., 2005) could be replicated in artificial neural networks. While they found that this was indeed the case, we argue that the mechanisms must be different from the ones used by humans, as the illusion requires small eye movements or blinks (Hisakata & Murakami, 2008; Kuriki et al., 2008), while the artificial model does not emulate such biological processes.

iv. **Testing generalization of mechanisms.** Having identified an important mechanism, one needs to make explicit for which particular conditions (class of tasks, data sets, …) the conclusion is intended to hold. A mechanism that is important for one setup may or may not be important for another one. In other words, whether a mechanism works under generalized settings has to be explicitly tested. An example of outstanding generalization for humans is their visual *robustness* against various variations in the input. In DNNs, a mechanism to improve robustness is to "stylize" (Gatys et al., 2016) training data. First presented as raising performance on parametrically distorted images (Geirhos, Rubisch, et al., 2018), this mechanism was later shown to also improve performance on images suffering from common corruptions (Michaelis et al., 2019) but would be unlikely to help with adversarial robustness. From a different perspective, the work of Golan et al. (2019) on controversial stimuli is an example where using stimuli outside of the training distribution can be insightful. Controversial stimuli are synthetic images that are designed to trigger distinct responses for two machine models. In their experimental setup, the use of these out-of-distribution data allows the authors to reveal whether the inductive bias of humans is similar to one of the candidate models.

v. **Resisting human bias.** Human bias can affect not only the design but also the conclusions we draw from comparison experiments. In other words, our human reference point can influence, for example, how we interpret the behavior of other systems, be they biological or artificial. An example is the well-known Braitenberg vehicles (Braitenberg, 1986), which are defined by very simple rules. To a human observer, however, the vehicles' behavior appears as arising from complex internal states such as fear, aggression, or love. This phenomenon of anthropomorphizing is well known in the field of comparative psychology (Romanes, 1883; Köhler, 1925; Koehler, 1943; Haun et al., 2010; Boesch, 2007; Tomasello & Call, 2008). Buckner (2019) specifically warns of human-centered interpretations and recommends to apply the lessons learned in comparative psychology to comparing DNNs and humans. In addition, our human reference point can influence how we design an experiment. As an example, Dujmović et al. (2020) illustrate that the selection of stimuli and labels can have a big effect on finding similarities or differences between humans and machines to adversarial examples.

In the remainder of this article, we provide concrete examples of the aspects discussed above using three case studies[1]:

(1) **Closed contour detection**: The first case study illustrates how tricky overcoming our human bias

can be and that shedding light on an alternative decision-making mechanism may require multiple additional experiments.

(2) **Synthetic Visual Reasoning Test**: The second case study highlights the challenge of isolating mechanisms and of differentiating between necessary and sufficient mechanisms. Thereby, we discuss how human and machine model learning differ and how changes in the model architecture can affect the performance.

(3) **Recognition gap**: The third case study illustrates the importance of aligning experimental conditions.

Case study 1: Closed contour detection

Closed contours play a special role in human visual perception. According to the Gestalt principles of prägnanz and good continuation, humans can group distinct visual elements together so that they appear as a "form" or "whole." As such, closed contours are thought to be prioritized by the human visual system and to be important in perceptual organization (Koffka, 2013; Elder & Zucker, 1993; Kovacs & Julesz, 1993; Tversky et al., 2004; Ringach & Shapley, 1996). Specifically, to tell if a line closes up to form a closed contour, humans are believed to implement a process called "contour integration" that relies at least partially on global information (Levi et al., 2007; Loffler et al., 2003; Mathes & Fahle, 2007). Even many flanking, open contours would hardly influence humans' robust closed contour detection abilities.

Our experiments

We hypothesize that, in contrast to humans, closed contour detection is difficult for DNNs. The reason is that this task would presumably require long-range contour integration, but DNNs are believed to process mainly local information (Geirhos, Rubisch, et al., 2018; Brendel & Bethge, 2019). Here, we test how well humans and neural networks can separate closed from open contours. To this end, we create a custom data set, test humans and DNNs on it, and investigate the decision-making process of the DNNs.

DNNs and humans reach high performance

We created a data set with two classes of images: The first class contained a closed contour; the second one did not. In order to make sure that the statistical properties of the two classes were similar, we included a main contour for both classes. While this contour line closed up for the first class, it remained open for the second class. This main contour consisted of 3–9 straight-line segments. In order to make the task more difficult, we added several flankers with either one or two line segments that each had a length of at least 32 pixels (Figure 2A). The size of the images was 256 × 256 pixels. All lines were black and the background was uniformly gray. Details on the stimulus generation can be found in Appendix B.

Humans identified the closed contour stimulus very reliably in a two-interval forced-choice task. Their performance was 88.39% ($SEM = 2.96\%$) on stimuli whose generation procedure was identical to the training set. For stimuli with white instead of black lines, human participants reached a performance of 90.52% ($SEM = 1.58\%$). The psychophysical experiment is described in Appendix B.

We fine-tuned a ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) on the closed contour data set. Similar to humans, it performed very well and reached an accuracy of 99.95% (see Figure 2A [i.i.d. to training]).

We found that both humans and our DNN reach high accuracy on the closed contour detection task. From a human-centered perspective, it is enticing to infer that the model had learned the concept of open and closed contours and possibly that it performs a similar contour integration-like process as humans. However, this would have been overhasty. To better understand the degree of similarity, we investigated how our model performs on variations of the data sets that were not used during the training procedure.

Generalization tests reveal differences

Humans are expected to have no difficulties if the number of flankers, the color, or the shape of lines would differ. We here test our model's robustness on such variants of the data set. If our model used similar decision-making processes as humans, it should be able to generalize well without any further training on the new images. This procedure is another perspective to shed light on whether our model really understood the concept of closedness or just picked up some statistical cues in the training data set.

We tested our model on 15 variants of the data set (out of distribution test sets) without fine-tuning on these variations. As shown in Figure 2A, B, our trained model generalized well to many but not all modified stimulus sets.

On the following variations, our model achieved high accuracy: Curvy contours (1, 3) were easily distinguishable for our model, as long as the diameter remained below 100 pixels. Also, adding a dashed, closed flanker (2) did not lower its performance. The classification ability of the model remained similarly high for the no-flankers (4) and the asymmetric

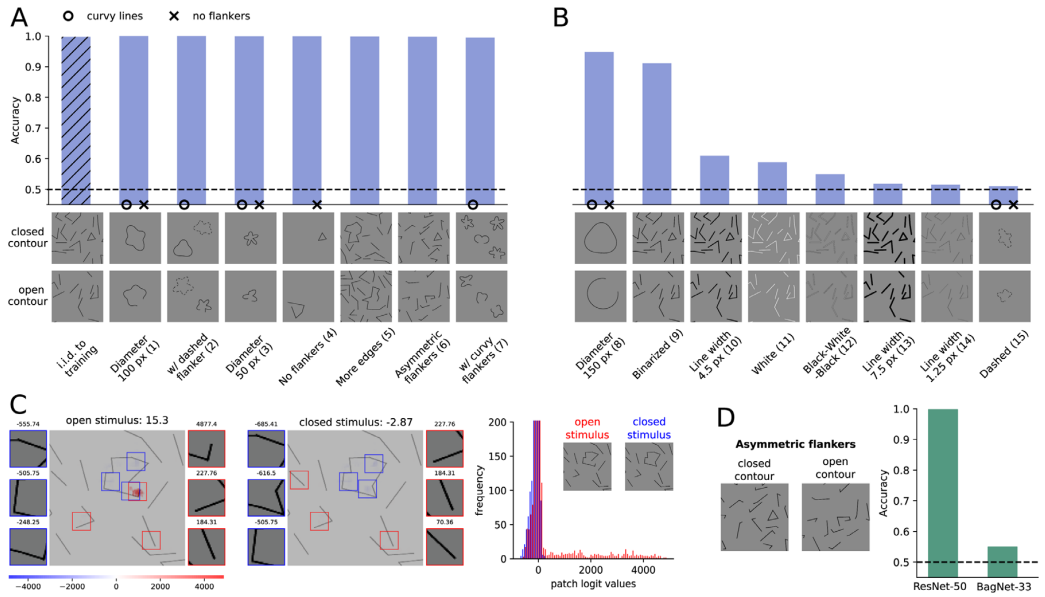


Figure 2. (A) Our ResNet-50-model generalized well to many data sets without further retraining, suggesting it would be able to distinguish closed and open contours. (B) However, the poor performance on many other data sets showed that our model did *not* learn the concept of closedness. (C) The heatmaps of our BagNet-33-based model show which parts of the image provided evidence for closedness (blue, negative values) or openness (red, positive values). The patches on the sides show the most extremely, nonoverlapping patches and their logit values. The logit distribution shows that most patches had logit values close to zero (y-axis truncated) and that many more patches in the open stimulus contributed positive logit values. (D) Our BagNet- and ResNet-models showed different performances on generalization sets, such as the asymmetric flankers. This indicates that the local decision-making process of the substitute model BagNet is not used by the original model ResNet. Figure best viewed electronically.

flankers condition (6). When testing our model on main contours that consisted of more edges than the ones presented during training (5), the performance was also hardly impaired. It remained high as well when multiple curvy open contours were added as flankers (7).

The following variations were more difficult for our model: If the size of the contour got too large, a moderate drop in accuracy was found (8). For binarized images, our model's performance was also reduced (9). And finally, (almost) chance performance was observed when varying the line width (14, 10, 13), changing the line color (11, 12), or using dashed curvy lines (15).

While humans would perform well on all variants of the closed contour data set, the failure of our model on some generalization tests suggests that it solves the task differently from humans. On the other hand, it is equally difficult to prove that the model does not understand the concept. As described by Firestone (2020), models can "perform differently despite similar underlying competences." In either way, we argue that it is important to openly consider alternative mechanisms to the human approach of global contour integration.

Our closed contour detection task is partly solvable with local features

In order to investigate an alternative mechanism to global contour integration, we here design an experiment to understand how well a decision-making process based on purely local features can work. For this purpose, we trained and tested BagNet-33 (Brendel & Bethge, 2019), a model that has access to local features only. It is a variation of ResNet-50 (He et al., 2016), where most 3×3 kernels are replaced by 1×1 kernels and therefore the receptive field size at the top-most convolutional layer is restricted to 33×33 pixels.

We found that our restricted model still reached close to 90% performance. In other words, contour integration was not necessary to perform well on the task.

To understand which local features the model relied on mostly, we analyzed the contribution of each patch to the final classification decision. To this end, we used the log-likelihood values for each 33×33 pixels patch from BagNet-33 and visualized them as a *heatmap*. Such a straightforward interpretation of the

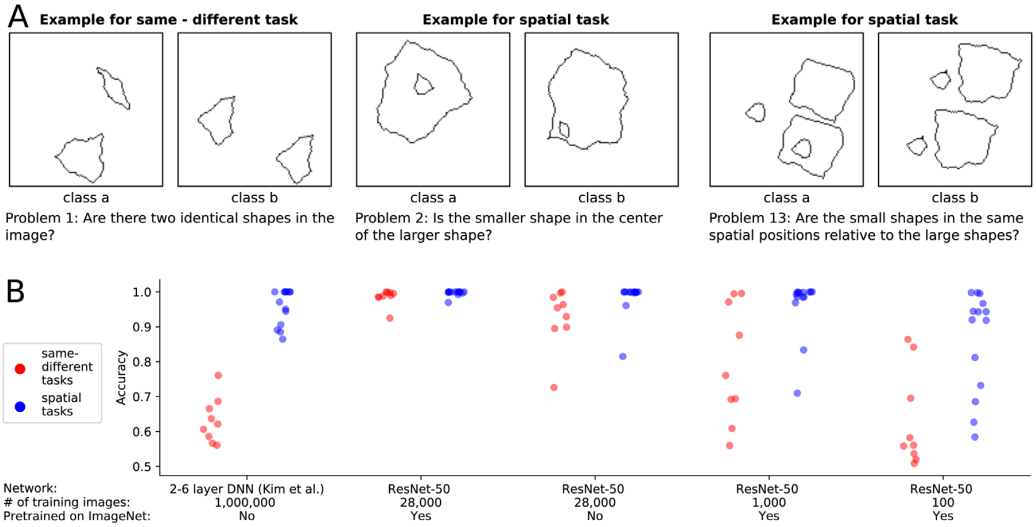


Figure 3. (A) For three of the 23 SVRT problems, two example images representing the two opposing classes are shown. In each problem, the task was to find the rule that separated the images and to sort them accordingly. (B) Kim et al. (2018) trained a DNN on each of the problems. They found that same-different tasks (red points), in contrast to spatial tasks (blue points), could not be solved with their models. Our ResNet-50-based models reached high accuracies for all problems when using 28,000 training examples and weights from pretraining on ImageNet.

contributions of single image patches is not possible with standard DNNs like ResNet (He et al., 2016) due to their large receptive field sizes in top layers.

The heatmaps of BagNet-33 (see Figure 2C) revealed which local patches played an important role in the decision-making process: An open contour was often detected by the presence of an endpoint at a short edge. Since all flankers in the training set had edges larger than 33 pixels, the presence of this feature was an indicator of an open contour. In turn, the absence of this feature was an indicator of a closed contour.

Whether the ResNet-50-based model used the same local feature as the substitute model was unclear. To answer this question, we tested BagNet on the previously mentioned generalization tests. We found that the data sets on which it showed high performance were sometimes different from the ones of ResNet (see Figure 7 in the Appendix B). A striking example was the failure of BagNet on the "asymmetric flankers" condition (see Figure 2D). For these images, the flankers often consisted of shorter line segments and thus obscured the local feature we assumed BagNet to use. In contrast, ResNet performed well on this variation. This suggests that the decision-making strategy of ResNet did not heavily depend on the local feature found with the substitute BagNet model.

In summary, the generalization tests, the high performance of BagNet as well as the existence of a distinctive local feature provide evidence that our human-biased assumption was misleading. We saw that other mechanisms for closed contour detection besides global contour integration do exist (see Introduction, *"Differentiating between necessary and sufficient mechanisms"*). As humans, we can easily miss the many statistical subtleties by which a task can be solved. In this respect, BagNets proved to be a useful tool to test a purportedly "global" visual task for the presence of local artifacts. Overall, various experiments and analyses can be beneficial to understand mechanisms and to overcome our human reference point.

Case study 2: Synthetic Visual Reasoning Test

In order to compare human and machine performance at learning abstract relationships between shapes, Fleuret et al. (2011) created the Synthetic Visual Reasoning Test (SVRT) consisting of 23 problems (see Figure 3A). They showed that humans need only few examples to understand the underlying concepts. Stabinger et al. (2016) as well as Kim et al. (2018) assessed the performance of deep convolutional neural networks on these problems. Both studies found a dichotomy between two task categories: While high accuracy was reached on spatial problems, the performance on same-different problems was poor. In order to compare the two types of tasks more systematically, Kim et al. (2018) developed a parameterized version of the SVRT data set called

PSVRT. Using this data set, they found that for same-different problems, an increase in the complexity of the data set could quickly strain their models. In addition, they showed that an attentive version of the model did not exhibit the same deficits. From these results, the authors concluded that feedback mechanisms as present in the human visual system such as attention, working memory, or perceptual grouping are probably important components for abstract visual reasoning. More generally, these studies have been perceived and cited with the broader claim of feed-forward DNNs not being able to learn same-different relationships between visual objects (Serre, 2019; Schofield et al., 2018) – at least not "efficiently" (Firestone, 2020).

We argue that the results of Kim et al. (2018) cannot be taken as evidence for the importance of feedback components for abstract visual reasoning:

(1) While their experiments showed that same-different tasks are harder to *learn* for their models, this might also be true for the human visual system. Normally sighted humans have experienced lifelong visual input; only looking at human performance with this extensive learning experience cannot reveal differences in learning difficulty.
(2) Even if there is a difference in learning complexity, this difference is not necessarily due to differences in the inference mechanism (e.g., feed-forward vs. feedback)—the large variety of other differences between biological and artificial vision systems could be critical causal factors as well.
(3) In the same line, small modifications in the learning algorithm or architecture can significantly change learning complexity. For example, changing the network depth or width can greatly improve learning performance (Tan & Le, 2019).
(4) Just because an attentive version of the model can learn both types of tasks does not prove that feedback mechanisms are necessary for these tasks (see Introduction, *"Differentiating between necessary and sufficient mechanisms"*).

Determining the necessity of feedback mechanisms is especially difficult because feedback mechanisms are not clearly distinct from purely feed-forward mechanisms. In fact, any finite-time recurrent network can be unrolled into a feed-forward network (Liao & Poggio, 2016; van Bergen & Kriegeskorte, 2020).

For these reasons, we argue that the importance of feedback mechanisms for abstract visual reasoning remains unclear.

In the following paragraph we present our own experiments on the SVRT data set and show that standard feed-forward DNNs can indeed perform well on same-different tasks. This confirms that feedback mechanisms are not strictly necessary for same-different tasks, although they helped in the specific experimental setting of Kim et al. (2018). Furthermore, this experiment highlights that changes of the network architecture and training procedure can have large effects on the performance of artificial systems.

Our experiments

The findings of Kim et al. (2018) were based on rather small neural networks, which consisted of up to six layers. However, typical network architectures used for object recognition consist of more layers and have larger receptive fields. For this reason, we tested a representative of such networks, namely, ResNet-50. The experimental setup can be found in Appendix C.

We found that our feed-forward model can in fact perform well on the same-different tasks of SVRT (see Figure 3B; see also concurrent work of Messina et al., 2019). This result was not due to an increase in the number of training samples. In fact, we used fewer images (28,000 images) than Kim et al. (2018) (1 million images) and Messina et al. (2019) (400,000 images). Of course, the results were obtained on the SVRT data set and might not hold for other visual reasoning data sets (see Introduction, *"Testing generalization of mechanisms"*).

In the very low-data regime (1,000 samples), we found a difference between the two types of tasks. In particular, the overall performance on same-different tasks was lower than on spatial reasoning tasks. As for the previously mentioned studies, this cannot be taken as evidence for systematic differences between feed-forward neural networks and the human visual system. In contrast to the neural networks used in this experiment, the human visual system is naturally pretrained on large amounts of visual reasoning tasks, thus making the low-data regime an unfair testing scenario from which it is almost impossible to draw solid conclusions about differences in the internal information processing. In other words, it might very well be that the human visual system trained from scratch on the two types of tasks would exhibit a similar difference in sample efficiency as a ResNet-50. Furthermore, the performance of a network in the low-data regime is heavily influenced by many factors other than architecture, including regularization schemes or the optimizer, making it even more difficult to reach conclusions about systematic differences in the network structure between humans and machines.

Case study 3: Recognition gap

Ullman et al. (2016) investigated the minimally necessary visual information required for object recognition. To this end, they successively cropped or reduced the resolution of a natural image until more than 50% of all human participants failed to

identify the object. The study revealed that recognition performance drops sharply if the minimal recognizable image crops are reduced any further. They referred to this drop in performance as the "recognition gap." The gap is computed by subtracting the proportion of people who correctly classify the largest unrecognizable crop (e.g., 0.2) from that of the people who correctly classify the smallest recognizable crop (e.g., 0.9). In this example, the recognition gap would evaluate to $0.9 - 0.2 = 0.7$. On the same human-selected image crops, Ullman et al. (2016) found that the recognition gap is much smaller for machine vision algorithms (0.14 ± 0.24) than for humans (0.71 ± 0.05). The researchers concluded that machine vision algorithms would not be able to "explain [humans'] sensitivity to precise feature configurations" and "that the human visual system uses features and processes that are not used by current models and that are critical for recognition." In a follow-up study, Srivastava et al. (2019) identified "fragile recognition images" (FRIs) with an exhaustive machine-based procedure whose results include a subset of patches that adhere to the definition of minimal recognizable configurations (MIRCs) by Ullman et al. (2016). On these machine-selected FRIs, a DNN experienced a moderately high recognition gap, whereas humans experienced a low one. Because of the differences between the selection procedures used in Ullman et al. (2016) and Srivastava et al. (2019), the question remained open whether machines would show a high recognition gap on machine-selected minimal images, if the selection procedure was similar to the one used in Ullman et al. (2016).

Our experiment

Our goal was to investigate if the differences in recognition gaps identified by Ullman et al. (2016) would at least in part be explainable by differences in the experimental procedures for humans and machines. Crucially, we wanted to assess machine performance on *machine*-selected, and not *human*-selected, image crops. We therefore implemented the psychophysics experiment in a machine setting to search the smallest recognizable images (or MIRCs) and the largest unrecognizable images (sub-MIRCs). In the final step, we evaluated our machine model's recognition gap using the *machine*-selected MIRCs and sub-MIRCs.

Methods

Our machine-based search algorithm used the deep convolutional neural network BagNet-33 (Brendel & Bethge, 2019), which allows us to straightforwardly analyze images as small as 33×33 pixels. In the first step, the classification accuracy was evaluated for the whole image. If it was above 0.5, the image was successively cropped and reduced in resolution. In each step, the best-performing crop was taken as the new parent. When the classification probability of all children fell below 0.5, the parent was identified as the MIRC, and all its children were considered sub-MIRCs. In order to evaluate the recognition gap, we calculate the difference in accuracy between the MIRC and the *best-performing* sub-MIRC. This definition is more conservative than the one from Ullman et al. (2016), who evaluated the difference in accuracy between the MIRC and the *worst-performing* sub-MIRC. For more details on the search procedure, please see Appendix D.

Results

We evaluated the recognition gap on two data sets: the original images from Ullman et al. (2016) and a subset of the ImageNet validation images (Deng et al., 2009). As shown in Figure 4A, our model has an average recognition gap of 0.99 ± 0.01 on the machine-selected crops of the data set from Ullman et al. (2016). On the machine-selected crops of the ImageNet validation subset, a large recognition gap occurs as well. Our values are similar to the recognition gap in humans and differ from the machines' recognition gap (0.14 ± 0.24) between human-selected MIRCs and sub-MIRCs as identified by Ullman et al. (2016).

Discussion

Our findings contrast claims made by Ullman et al. (2016). The latter study concluded that machine algorithms are not as sensitive as humans to precise feature configurations and that they are missing features and processes that are "critical for recognition." First, our study shows that a machine algorithm *is* sensitive to small image crops. It is only the precise minimal features that differ between humans and machines. Second, by the word "critical," Ullman et al. (2016) imply that object recognition would not be possible without these human features and processes. Applying the same reasoning to Srivastava et al. (2019), the low human performance on machine-selected patches should suggest that humans would miss "features and processes critical for recognition." This would be an obviously overreaching conclusion. Furthermore, the success of modern artificial object recognition speaks against the conclusion that the purported processes are "critical" for recognition, at least within this discretely defined recognition task. Finally, what we can conclude from the experiments of Ullman et al. (2016) and from our own is that both the human and a machine visual system can recognize small image crops and that there is a sudden drop in recognizability when reducing the amount of information.

In summary, these results highlight the importance of testing humans and machines in as similar settings

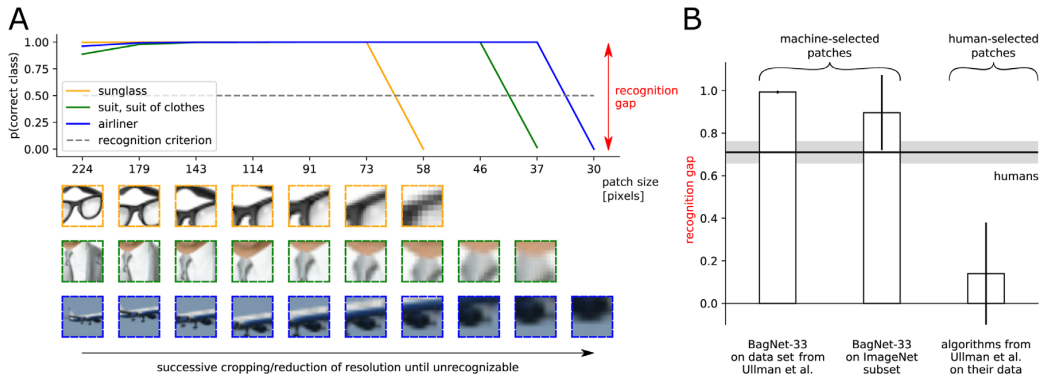


Figure 4. (A) BagNet-33's probability of correct class for decreasing crops: The sharp drop when the image becomes too small or the resolution too low is called the "recognition gap" (Ullman et al., 2016). It was computed by subtracting the model's predicted probability of the correct class for the sub-MIRC from the model's predicted probability of the correct class for the MIRC. As an example, the glasses stimulus was evaluated as $0.9999 - 0.0002 = 0.9997$. The crop size on the x-axis corresponds to the size of the original image in pixels. Steps of reduced resolution are not displayed such that the three sample stimuli can be displayed coherently. (B) Recognition gaps for machine algorithms (vertical bars) and humans (gray horizontal bar). A recognition gap is identifiable for the DNN BagNet-33 when testing machine-selected stimuli of the original images from Ullman et al. (2016) and a subset of the ImageNet validation images (Deng et al., 2009). Error bars denote standard deviation.

as possible, and of avoiding a human bias in the experiment design. All conditions, instructions, and procedures should be as close as possible between humans and machines in order to ensure that observed differences are due to inherently different decision strategies rather than differences in the testing procedure.

Conclusion

Comparing human and machine visual perception can be challenging. In this work, we presented a checklist on how to perform such comparison studies in a meaningful and robust way. For one, isolating a single mechanism requires us to minimize or exclude the effect of other differences between biological and artificial and to align experimental conditions for both systems. We further have to differentiate between necessary and sufficient mechanisms and to circumscribe in which tasks they are actually deployed. Finally, an overarching challenge in comparison studies between humans and machines is our strong internal human interpretation bias.

Using three case studies, we illustrated the application of the checklist. The first case study on closed contour detection showed that human bias can impede the objective interpretation of results and that investigating which mechanisms could or could not be at work may require several analytic tools. The second case study highlighted the difficulty of drawing robust conclusions about mechanisms from experiments. While previous studies suggested that feedback mechanisms might be important for visual reasoning tasks, our experiments showed that they are not necessarily required. The third case study clarified that aligning experimental conditions for both systems is essential. When adapting the experimental settings, we found that, unlike the differences reported in a previous study, DNNs and humans indeed show similar behavior on an object recognition task.

Our checklist complements other recent proposals about how to compare visual inference strategies between humans and machines (Buckner, 2019; Chollet, 2019; Ma & Peters, 2020; Geirhos et al., 2020) and helps to create more nuanced and robust insights into both systems.

Acknowledgments

The authors thank Alexander S. Ecker, Felix A. Wichmann, Matthias Kümmerer, Dylan Paiton, and Drew Linsley for helpful discussions. We thank Thomas Serre, Junkyung Kim, Matthew Ricci, Justus Piater, Sebastian Stabinger, Antonio Rodríguez-Sánchez, Shimon Ullman, Liav, Assif, and Daniel Harari for discussions and feedback on an earlier version of this manuscript. Additionally, we thank Nikolas Kriegeskorte for his detailed and constructive feedback, which helped us make our manuscript stronger. Furthermore, we thank Wiebke Ringels for helping with data collection for the psychophysical experiment.

We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for

supporting CMF and JB. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the competence center for machine learning (FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellence Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the Deutsche Forschungsgemeinschaft (DFG; Projektnummer 276693517 SFB 1233).

The closed contour case study was designed by CMF, JB, TSAW, and MB and later with WB. The code for the stimuli generation was developed by CMF. The neural networks were trained by CMF and JB. The psychophysical experiments were performed and analyzed by CMF, TSAW, and JB. The SVRT case study was conducted by CMF under supervision of TSAW, WB, and MB. KS designed and implemented the recognition gap case study under the supervision of WB and MB; JB extended and refined it under the supervision of WB and MB. The initial idea to unite the three projects was conceived by WB, MB, TSAW, and CMF, and further developed including JB. The first draft was jointly written by JB and CMF with input from TSAW and WB. All authors contributed to the final version and provided critical revisions.

Elements of this work were presented at the Conference on Cognitive Computational Neuroscience 2019 and the Shared Visual Representations in Human and Machine Intelligence Workshop at the Conference on Neural Information Processing Systems 2019.

The icon image is modified from the image by Gerd Leonhard, available under https://www.flickr.com/photos/gleonhard/33661762360 on December 17, 2020. The original license is CC BY-SA 2.0, and therefore so is the one for the icon image.

Commercial relationships: Matthias Bethge: Amazon scholar Jan 2019 – Jan 2021, Layer7AI, DeepArt.io, Upload AI; Wieland Brendel: Layer7AI.
Corresponding authors: Christina M. Funke; Judy Borowski.
Email: christina.funke@bethgelab.org; judy.borowski@bethgelab.org.
Address: Maria-von-Linden-Strasse 6, 72076, Tübingen, Germany.

*CMF and JB are both first authors on this work.
†WB, TSAW and MB are joint senior authors.

Footnote

[1]The code is available at https://github.com/bethgelab/notorious_difficulty_of_comparing_human_and_machine_perception.

References

Barrett, D. G., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning, 80*, 511–520. PMLR, http://proceedings.mlr.press/v80/barrett18a.html.

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology, 55*, 55–64.

Boesch, C. (2007). What makes us human (homo sapiens)? The challenge of cognitive cross-species comparison. *Journal of Comparative Psychology, 121*(3), 227.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT Press, https://mitpress.mit.edu/books/vehicles.

Brendel, W., & Bethge, M. (2019). Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint, arXiv:1904.00760.

Buckner, C. (2019). The comparative psychology of artificial intelligences, http://philsci-archive.pitt.edu/16128/.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., . . . Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Computational Biology, 15*(4), e1006897.

Chollet, F. (2019). The measure of intelligence. arXiv preprint, arXiv:1911.01547.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317, https://doi.org/10.1016/j.tics.2019.01.009.

Conway, B. R., Kitaoka, A., Yazdanbakhsh, A., Pack, C. C., & Livingstone, M. S. (2005). Neural basis for a powerful static motion illusion. *Journal of Neuroscience, 25*(23), 5651–5656.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2019). Crowding reveals fundamental

differences in local vs. global processing in humans and machines. bioRxiv, 744268.

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife, 9*, e55978. Retrieved from https://doi.org/10.7554/eLife.55978.

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems, 29*, 1100–1108. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2016/file/42e77b63637ab381e8be5f8318cc28a2-Paper.pdf.

Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision* (pp. 2650–2658), doi:10.1109/ICCV.2015.304.

Elder, J., & Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research, 33*(7), 981–991.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., . . . Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31*, 3910–3920. Curran Associates, Inc.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*. Retrieved from https://www.pnas.org/content/early/2020/10/13/1905334117.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences, 108*(43), 17621–17625.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414–2423), doi:10.1109/CVPR.2016.265.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. arXiv preprint, arXiv:2006.16736.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint, arXiv:1811.12231.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31*, 7538–7550. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2019). Controversial stimuli: Pitting neural networks against each other as models of human recognition. arXiv preprint, arXiv:1911.09288.

Gomez-Villa, A., Martin, A., Vazquez-Corral, J., & Bertalmio, M. (2019). Convolutional neural networks can be deceived by visual illusions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12301–12309), doi:10.1109/CVPR.2019.01259.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning, 70*, 1321–1330.

Han, C., Yoon, W., Kwon, G., Nam, S., & Kim, D. (2019). Representation of white-and black-box adversarial examples in deep neural networks and humans: A functional magnetic resonance imaging study. arXiv preprint, arXiv:1905.02422.

Haun, D. B. M., Jordan, F. M., Vallortigara, G., & Clayton, N. S. (2010). Origins of spatial, temporal, and numerical cognition: Insights from comparative psychology. *Trends in Cognitive Sciences, 14*(12), 552–560, https://doi.org/10.1016/j.tics.2010.09.006, http://www.sciencedirect.com/science/article/pii/S1364661310002135.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778), doi:10.1109/CVPR.2016.90.

Hisakata, R., & Murakami, I. (2008). The effects of eccentricity and retinal illuminance on the illusory motion seen in a stationary luminance gradient. *Vision Research, 48*(19), 1940–1948.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models

may explain it cortical representation. *PLoS Computational Biology, 10*(11), e1003915.

Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do neural networks show gestalt phenomena? An exploration of the law of closure. arXiv preprint, arXiv:1903.01069.

Kim, J., Ricci, M., & Serre, T. (2018). Not-so-clevr: Learning same–different relations strains feedforward neural networks. *Interface Focus, 8*(4), 20180011.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint, arXiv:1412.6980.

Kitaoka, A., & Ashida, H. (2003). Phenomenal characteristics of the peripheral drift illusion. *Vision, 15*(4), 261–262.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C., (2007). What's new in psychtoolbox-3. *Perception, 36*(14), 1.

Koehler, O. (1943). Counting experiments on a common raven and comparative experiments on humans. *Zeitschrift für Tierpsychologie, 5*(3), 575–712.

Koffka, K. (2013). *Principles of Gestalt psychology*. New York: Routledge.

Köhler, W. (1925). *The mentality of apes*. New York, NY: Kegan Paul, Trench, Trubner & Co.

Kovacs, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences, 90*(16), 7495–7497.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems, 25*, 1097–1105. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology, 12*(4), e1004896.

Kuriki, I., Ashida, H., Murakami, I., & Kitaoka, A. (2008). Functional brain imaging of the rotating snakes illusion by fMRI. *Journal of Vision, 8*(10), 16, https://doi.org/10.1167/8.6.64.

Levi, D. M., Yu, C., Kuai, S.-G., & Rislove, E. (2007). Global contour processing in amblyopia. *Vision Research, 47*(4), 512–524.

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv preprint, arXiv:1604.03640.

Loffler, G., Wilson, H. R., & Wilkinson, F. (2003). Local and global contributions to shape discrimination. *Vision Research, 43*(5), 519–530.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431–3440), doi:10.1109/CVPR.2015.7298965.

Luo, L., Xiong, Y., Liu, Y., & Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint, arXiv:1902.09843.

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. arXiv preprint, arXiv:2005.02181.

Majaj, N. J., & Pelli, D. G. (2018). Deep learning—using machine learning to study biological vision. *Journal of Vision, 18*(13), 2, https://doi.org/10.1167/18.13.2.

Mathes, B., & Fahle, M. (2007). Closure facilitates contour integration. *Vision Research, 47*(6), 818–827, https://doi.org/10.1167/18.13.2.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint, arXiv:1902.01007.

Messina, N., Amato, G., Carrara, F., Falchi, F., & Gennaro, C. (2019). Testing deep neural networks on the same-different task. In *International Conference on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6). IEEE, doi:10.1109/CBMI.2019.8877412.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., ... Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint, arXiv:1907.07484.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. arXiv preprint, arXiv:1907.07355.

Pelli, D. G., & Vision, S. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. arXiv preprint, arXiv:1608.02164.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Ringach, D. L., & Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Research, 36*(19), 3037–3050.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *Proceedings of the 34th International Conference on Machine Learning, 70*, 2940–2949.

Romanes, G. J. (1883). *Animal intelligence*. D. Appleton, https://books.google.de/books?hl=en&lr=&id=Vx8aAAAAYAAJ&oi=fnd&pg=PA1&dq=animal+intelligence+1883&ots=IUOqpa2YRA&sig=JIVJfeIN7HlireTKzBd2tdv8IzM&redir_esc=y#v=onepage&q=animal%20intelligence%201883&f=false.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., ... Lillicrap, T. (2017). A simple neural network module for relational reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, ... R. Garnett (Eds.), *Advances in Neural Information Processing Systems, 30*, 4967–4976. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2017/file/e6acf4b0f69f6f6e60e9a815938aa1ff-Paper.pdf.

Schofield, A. J., Gilchrist, I. D., Bloj, M., Leonardis, A., & Bellotto, N. (2018). Understanding images in biological and computer vision. *Interface Focus, 8*, https://doi.org/10.1098/rsfs.2018.0027

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... Schmidt, K. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? BioRxiv, 407007.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology, 8*, 1551.

Srivastava, S., Ben-Yosef, G., & Boix, X. (2019). Minimal images in deep neural networks: Fragile object recognition in natural images. arXiv preprint, arXiv:1902.03227.

Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of CNNs: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks*. (pp. 380–387). Cham: Springer.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., ... Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint, arXiv:1312.6199.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint, arXiv:1905.11946.

Tomasello, M., & Call, J. (2008). Assessing the validity of ape-human comparisons: A reply to boesch (2007). *Journal of Comparative Psychology, 122*(4), 449–452. American Psychological Association.

Tversky, T., Geisler, W. S., & Perry, J. S. (2004). Contour grouping: Closure effects are explained by good continuation and proximity. *Vision Research, 44*(24), 2769–2777.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences, 113*(10), 2744–2749.

van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: The role of recurrence in visual inference. arXiv preprint, arXiv:2003.12128.

Villalobos, K. M., Stih, V., Ahmadinejad, A., Dozier, J., Francl, A., Azevedo, F., Sasaki, T., ... Boix, X. (2020). Do deep neural networks for segmentation understand insideness? https://cbmm.mit.edu/publications/do-neural-networks-segmentation-understand-insideness.

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do deep neural networks suffer from crowding? In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, ... R. Garnett (Eds.), *Advances in Neural Information Processing Systems, 30*, 5628–5638. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2017/file/c61f571dbd2fb949d3fe5ae1608dd48b-Paper.pdf.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology, 9*, 345.

Wu, X., Zhang, X., & Du, J. (2019). Challenge of spatial cognition for deep learning. arXiv preprint, arXiv:1908.04396.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624.

Yan, Z., & Zhou, X. S. (2017). How intelligent are convolutional neural networks? arXiv preprint, arXiv:1709.06126.

Zhang, R., Wu, J., Zhang, C., Freeman, W. T., & Tenenbaum, J. B. (2016). A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. arXiv preprint, arXiv:1605.01138.

Zhang, X., Watkins, Y., & Kenyon, G. T. (2018). Can deep learning learn the principle of closed

contour detection? In G. Bebis, R. Boyle, B. Parvin, D. Koracin, M. Turek, S. Ramalingam, K. Xu, S. Lin, B. Alsallakh, J. Yang, E. Cuervo, . . . J. Ventura (Eds.), *International Symposium on Visual Computing* (pp. 455–460). Cham: Springer, https://doi.org/10.1007/978-3-030-03801-4_40.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*(1), 1334.

Appendix A: Literature overview of comparison studies

A growing body of work discusses comparisons of humans and machines on a higher level. Majaj and Pelli (2018) provide a broad overview how machine learning can help vision scientists to study biological vision, while Barrett et al. (2019) review methods on how to analyze representations of biological and artificial networks. From the perspective of cognitive science, Cichy and Kaiser (2019) stress that deep learning models *can* serve as scientific models that not only provide both helpful predictions and explanations but that can also be used for exploration. Furthermore, from the perspective of psychology and philosophy, Buckner (2019) emphasizes often-neglected caveats when comparing humans and DNNs such as human-centered interpretations and calls for discussions regarding how to properly align machine and human performance. Chollet (2019) proposes a general artificial intelligence benchmark and suggests to rather evaluate intelligence as "skill-acquisition efficiency" than to focus on skills at specific tasks.

In the following, we give a brief overview of studies that compare human and machine perception. In order to test if DNNs have similar cognitive abilities as humans, a number of studies test DNNs on abstract (visual) reasoning tasks (Barrett et al., 2018; Yan & Zhou, 2017; Wu et al., 2019; Santoro et al., 2017; Villalobos et al., 2020). Other comparison studies focus on whether human visual phenomena such as illusions (Gomez-Villa et al., 2019; Watanabe et al., 2018; Kim et al., 2019) or crowding (Volokitin et al., 2017; Doerig et al., 2019) can be reproduced in computational models. In the attempt to probe intuition in machine models, DNNs are compared to intuitive physics engines, that is, probabilistic models that simulate physical events (Zhang et al., 2016).

Other works investigate whether DNNs are sensible models of human perceptual processing. To this end, their prediction or internal representations are compared to those of biological systems, for example, to human and/or monkey behavioral representations (Peterson et al., 2016; Schrimpf et al., 2018; Yamins et al., 2014; Eberhardt et al., 2016; Golan et al., 2019), human fMRI representations (Han et al., 2019; Khaligh-Razavi & Kriegeskorte, 2014) or monkey cell recordings (Schrimpf et al., 2018; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Cadena et al., 2019).

A great number of studies focus on manipulating tasks and/or models. Researchers often use generalization tests on data dissimilar to the training set (Zhang et al., 2018; Wu et al., 2019) to test whether machines understood the underlying concepts. In other studies, the degradation of object classification accuracy is measured with respect to image degradations (Geirhos et al., 2018) or with respect to the type of features that play an important role for human or machine decision-making (Geirhos, Rubisch, et al., 2018; Brendel & Bethge, 2019; Kubilius et al., 2016; Ullman et al., 2016; Ritter et al., 2017). A lot of effort is being put into investigating whether humans are vulnerable to small, adversarial perturbations in images (Elsayed et al., 2018; Zhou & Firestone, 2019; Han et al., 2019; Dujmović et al., 2020), as DNNs are shown to be (Szegedy et al., 2013). Similarly, in the field of natural language processing, a trend is to manipulate the data set itself by, for example, negating statements to test whether a trained model gains an understanding of natural language or whether it only picks up on statistical regularities (Niven & Kao, 2019; McCoy et al., 2019).

Further work takes inspiration from biology or uses human knowledge explicitly in order to improve DNNs. Spoerer et al. (2017) found that recurrent connections, which are abundant in biological systems, allow for higher object recognition performance, especially in challenging situations such as in the presence of occlusions—in contrast to pure feed-forward networks. Furthermore, several researchers suggest (Zhang et al., 2018; Kim et al., 2018) or show (Wu et al., 2019; Barrett et al., 2018; Santoro et al., 2017) that designing networks' architecture or features with human knowledge is key for machine algorithms to successfully solve abstract (reasoning) tasks.

Appendix B: Closed contour detection

Data set

Each image in the training set contained a main contour, multiple flankers, and a background image. The main contour and flankers were drawn into an image of size $1,028 \times 1,028$ pixels. The main contour and flankers could be straight or curvy lines, for which the generation processes are respectively described in the next two subsections. The lines had a default thickness of 10 pixels. We then resized the image to 256×256 pixels using anti-aliasing to transform the black and white pixels into smoother lines that had

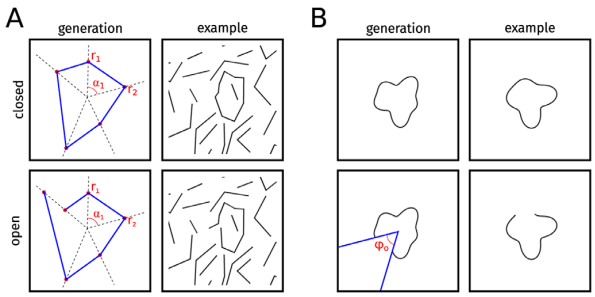


Figure 5. Closed contour data set. (A) Left: The main contour was generated by connecting points from a random sampling process of angles and radii. Right: Resulting line-drawing with flankers. (B) Left: Generation process of curvy contours. Right: Resulting line-drawing.

gray pixels at the borders. Thus, the lines in the resized image had a thickness of 2.5 pixels. In the following, all specifications of sizes refer to the resized image (i.e., a line described of final length 10 pixels extended over 40 pixels when drawn into the $1,028 \times 1,028$-pixel image). For the psychophysical experiments (see Appendix B, Psychophysical experiment), we added a white margin of 16 pixels on each side of the image to avoid illusory contours at the borders of the image.

Varying contrast of background. An image from the ImageNet data set was added as background to the line drawing. We converted the image into LAB color space and linearly rescaled the pixel intensities of the image to produce a normalized contrast value between 0 (gray image with the RGB values [118, 118, 118]) and 1 (original image) (see Figure 8A). When adding the image to the line drawing, we replaced all pixels of the line drawing by the values of the background image for which the background image had a higher grayscale value than the line drawing. For the experiments in the main body, the contrast of the background image was always 0. The other contrast levels were used only for the additional experiment described in Appendix B, Additional experiment: Increasing the task difficulty by adding a background image.

Generation of image pairs. We aimed to reduce the statistical properties that could be exploited to solve the task without judging the closedness of the contour. Therefore, we generated image pairs consisting of an "open" and a "closed" version of the same image. The two versions were designed to be almost identical and had the same flankers. They differed only in the main contour, which was either open or closed. Examples of such image pairs are shown in Figure 5. During training, either the closed or the open image of a pair was used. However, for the validation and testing, both versions were used. This allowed us to compare the predictions and heatmaps for images that differed only slightly but belonged to different classes.

Line-drawing with polygons as main contour

The data set used for training as well as some of the generalization sets consisted of straight lines. The main contour consisted of $n \in \{3, 4, 5, 6, 7, 8, 9\}$ line segments that formed either an open or a closed contour. The generation process of the main contour is depicted on the left side of Figure 5A. To get a contour with n edges, we generated n points, which were defined by a randomly sampled angle α_n and a randomly sampled radius r_n (between 0 and 128 pixels). By connecting the resulting points, we obtained the closed contour. We used the python PIL library (PIL 5.4.1, python3) to draw the lines that connect the endpoints. For the corresponding open contour, we sampled two radii for one of the angles such that they had a distance of 20 to 50 pixels from each other. When connecting the points, a gap was created between the points that share the same angle. This generation procedure could allow for very short lines with edges being very close to each other. To avoid this, we excluded all shapes with corner points closer to 10 pixels from nonadjacent lines.

The position of the main contour was random, but we ensured that the contour did not extend over the border of the image.

Besides the main contour, several flankers consisting of either one or two line segments were added to each stimulus. The exact number of flankers was uniformly sampled from the range [10,25]. The length of each line segment varied between 32 and 64 pixels. For the flankers consisting of two line segments, both lines had the same length, and the angle between the line segments was at least 45°. We added the flankers successively to the image and thereby ensured a minimal distance of 10 pixels between the line centers. To ensure that the corresponding image pairs would have the same flankers, the distances to both the closed and open version of the main contour were accounted for when re-sampling flankers. If a flanker did not fulfill this criterion, a new flanker was sampled of the same size and the same number of line segments, but it was placed somewhere else. If a flanker extended over the border of the image, the flanker was cropped.

Line-drawing with curvy lines as main contour

For some of the generalization sets, the contours consisted of curvy instead of straight lines. These were generated by modulating a circle of a given radius r_c with a radial frequency function that was defined by two sinusoidal functions. The radius of the contour was thus given by

$$r(\phi) = A_1 \sin(f_1(\phi + \theta_1)) + A_2 \sin(f_2(\phi + \theta_2)) + r_c, \quad (1)$$

with the frequencies f_1 and f_2 (integers between 1 and 6), amplitudes A_1 and A_2 (random values between 15

and 45), and phases θ_1 and θ_2 (between 0 and 2π). Unless stated otherwise, the diameter (diameter $= 2 \times r_c$) was a random value between 50 and 100 pixels, and the contour was positioned in the center of the image. The open contours were obtained by removing a circular segment of size $\phi_o = \frac{\pi}{3}$ at a random phase (see Figure 5B).

For two of the generalization data sets, we used dashed contours that were obtained by masking out 20 equally distributed circular segments each of size $\phi_d = \frac{\pi}{20}$.

Details on generalization data sets

We constructed 15 variants of the data set to test generalization performance. Nine variants consisted of contours with straight lines. Six of these featured varying line styles like changes in line width (10, 13, 14) and/or line color (11, 12). For one variant (5), we increased the number of edges in the main contour. Another variant (4) had no flankers, and yet another variant (6) featured asymmetric flankers. For variant 9, the lines were binarized (only black or gray pixels instead of different gray tones).

In another six variants, the contours as well as the flankers were curved, meaning that we modulated a circle with a radial frequency function. The first four variants did not contain any flankers and the main contour had a fixed size of 50 pixels (3), 100 pixels (1), and 150 pixels (8). For another variant (15), the contour was a dashed line. Finally, we tested the effect of different flankers by adding one additional closed, yet dashed contour (2) or one to four open contours (7).

Below, we provide more details on some of these data sets:

Black-white-black lines (12). For all contours, black lines enclosed a white one in the middle. Each of these three lines had a thickness of 1.5 pixels, which resulted in a total thickness of 4.5 pixels.

Asymmetric flankers (6). The two-line flankers consisted of one long and one short line instead of two equally long lines.

W/ dashed flanker (2). This data set with curvy contours contained an additional dashed, yet closed contour as a flanker. It was produced like the main contour in the dashed main contour set. To avoid overlap of the contours, the main contour and the flanker could only appear at four determined positions in the image, namely, the corners.

W/ multiple flankers (7). In addition to the curvy main contour, between one and four open curvy contours were added as flankers. The flankers were generated by the same process as the main contour. The circles that were modulated had a diameter of 50 pixels and could appear at either one of the four corners of the image or in the center.

Psychophysical experiment

To estimate how well humans would be able to distinguish closed and open stimuli, we performed a psychophysical experiment in which observers reported which of two sequentially presented images contained a closed contour (two-interval forced choice [2-IFC] task).

Stimuli

The images of the closed contour data set were used as stimuli for the psychophysical experiments. Specifically, we used the images from the test sets that were used to evaluate the performance of the models. For our psychophysical experiments, we used two different conditions: The images contained either black (i.i.d. to the training set) or white contour lines. The latter was one one of the generalization test sets.

Apparatus

Stimuli were displayed on a VIEWPixx 3D LCD (VPIXX Technologies; spatial resolution $1,920 \times 1,080$ pixels, temporal resolution 120 Hz, operating with the scanning backlight turned off). Outside the stimulus image, the monitor was set to mean gray. Observers viewed the display from 60 cm (maintained via a chinrest) in a darkened chamber. At this distance, pixels subtended approximately 0.024° on average (41 pixels per degree of visual angle). The monitor was linearized (maximum luminance 260 cd/m^2 using a Konica-Minolta LS-100 photometer. Stimulus presentation and data collection were controlled via a desktop computer (Intel Core i5-4460 CPU, AMD Radeon R9 380 GPU) running Ubuntu Linux (16.04 LTS), using the Psychtoolbox Library (Pelli & Vision, 1997; Kleiner et al., 2007; Brainard & Vision, 1997, version 3.0.12) and the iShow library (http://dx.doi.org/10.5281/zenodo.34217) under MATLAB (The Mathworks, Inc., R2015b).

Participants

In total, 19 naïve observers (4 male, 15 female, age: 25.05 years, $SD = 3.52$) participated in the experiment. Observers were paid 10€ per hour for participation. Before the experiment, all subjects had given written informed consent for participating. All subjects had normal or corrected-to-normal vision. All procedures conformed to Standard 8 of the American Psychological 405 Association's "Ethical Principles of Psychologists and Code of Conduct" (2010).

Procedure

On each trial, one closed and one open contour stimulus were presented to the observer (see Figure 6A).

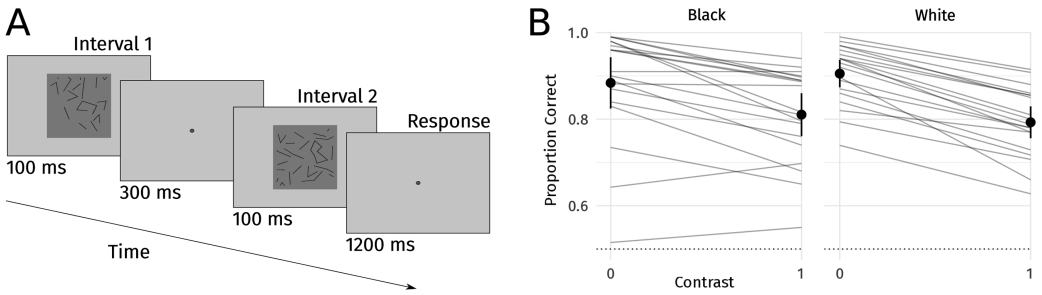


Figure 6. (A) In a 2-IFC task, human observers had to tell which of two images contained a closed contour. (B) Accuracy of the 20 naïve observers for the different conditions.

The images used for each trial were randomly picked, but we ensured that the open and closed images shown in the same trial were not the ones that were almost identical to each other (see Appendix B, Generation of image pairs). Thus, the number of edges of the main contour could differ between the two images shown in the same trial. Each image was shown for 100 ms, separated by a 300-ms interstimulus interval (blank gray screen). We instructed the observer to look at the fixation spot in the center of the screen. The observer was asked to identify whether the image containing a closed contour appeared first or second. The observer had 1,200 ms to respond and was given feedback after each trial. The intertrial interval was 1,000 ms. Each block consisted of 100 trials and observers performed five blocks. Trials with different line colors and varying background images (contrasts including 0, 0.4, and 1) were blocked. Here, we only report the results for black and white lines of contrast 0. Upon the first time that a block with a new line color was shown, observers performed a practice session with 48 trials of the corresponding line color.

Training of ResNet-50 model

We fine-tuned a ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) on the closed contour task. We replaced the last fully connected, 1,000-way classification layer by layer with only one output neuron to perform binary classification with a decision threshold of 0. The weights of all layers were fine-tuned using the optimizer Adam (Kingma & Ba, 2014) with a batch size of 64. All images were preprocessed to have the same mean and standard deviation and were randomly mirrored horizontally and vertically for data augmentation. The model was trained on 14,000 images for 10 epochs with a learning rate of 0.0003. We used a validation set of 5,600 images.

Generalization tests. To determine the generalization performance, we evaluated the model on the test sets without any further training. Each of the test sets contained 5,600 images. Poor accuracy could simply result from a suboptimal decision criterion rather than because the network would not be able to tell the stimuli apart. To account for the distribution shift between the original training images and the generalization tasks, we optimized the decision threshold (a single scalar) for each data set. To find the optimal threshold for each data set, we subdivided the interval, in which 95% of all logits lie, into 100 sub points and picked the threshold that would lead to the highest performance.

Training of BagNet-33 model

To test an alternative decision-making mechanism to global contour integration, we trained and tested a BagNet-33 (Brendel & Bethge, 2019) on the closed contour task. Like the ResNet-50 model, it was pretrained on ImageNet (Deng et al., 2009) and we replaced the last fully connected, 1,000-way classification layer by layer with only one output neuron. We fine-tuned the weights using the optimizer AdaBound (Luo et al., 2019) with an initial and final learning rate of 0.0001 and 0.1, respectively. The training images were generated on-the-fly, which meant that new images were produced for each epoch. In total, the fine-tuning lasted 100 epochs, and we picked the weights from the epoch with the highest performance.

Generalization tests. The generalization tests were conducted equivalently to the ones with ResNet-50. The results are shown in Figure 7.

Additional experiment: Increasing the task difficulty by adding a background image

We performed an additional experiment, where we tested if the model would become more robust and thus generalized better if we trained on a more difficult task. This was achieved by adding an image to the

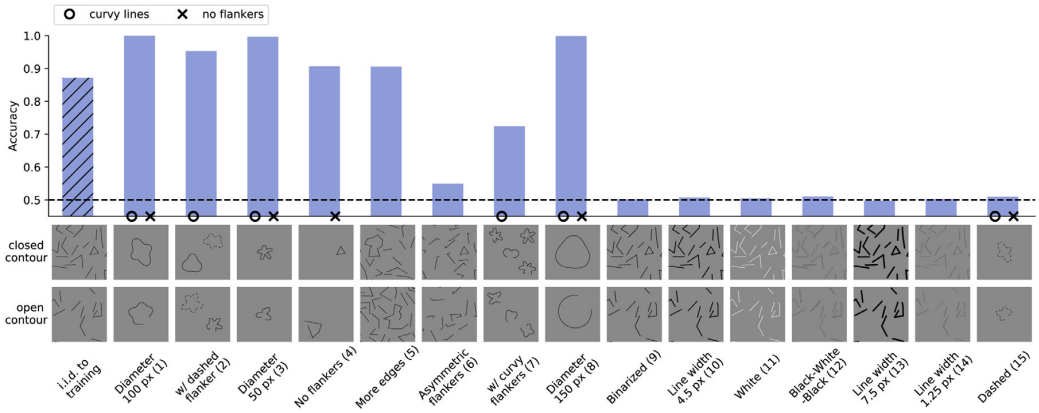


Figure 7. Generalization performances of BagNet-33.

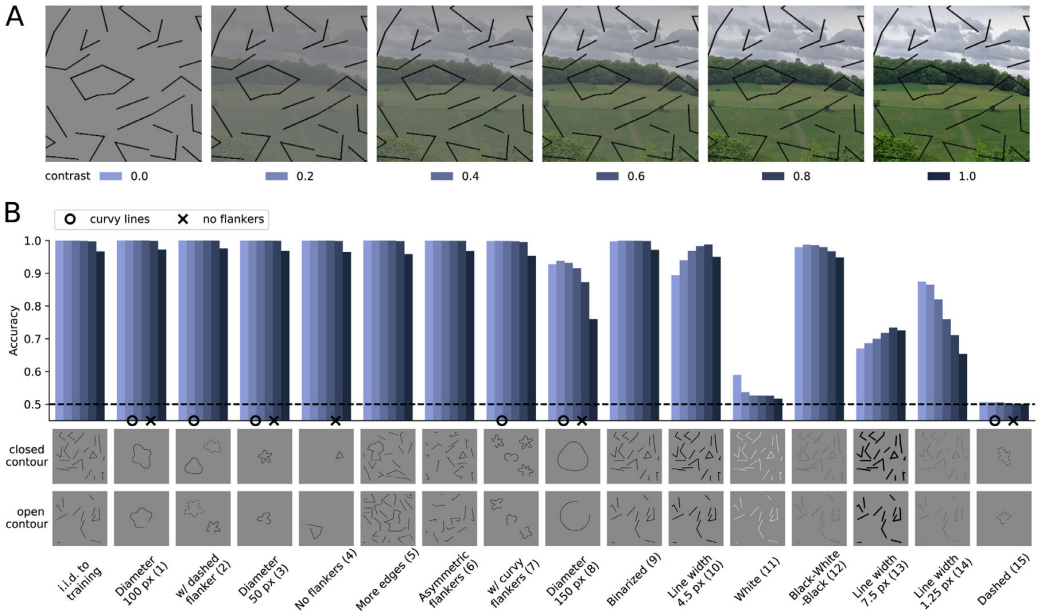


Figure 8. (A) An image of varying contrast was added as background. (B) Generalization performances of our models trained on random contrast levels and tested on single contrast levels.

background, such that the model had to learn how to separate the lines from the task-irrelevant background.

In our experiment, we fine-tuned our ResNet-50-based model on images with a background image of a uniformly sampled contrast. For each data set, we evaluated the model separately on six discrete contrast levels {0, 0.2, 0.4, 0.6, 0.8, 1} (see Figure 8A). We found that the generalization performance varied for some data sets compared to the experiment in the main body (see Figure 8B).

Appendix C: SVRT

Methods

Data set. We used the original C-code provided by Fleuret et al. (2011) to generate the images of the SVRT data set. The images had a size of 128 × 128 pixels. For each problem, we used up to 28,000 images for training,

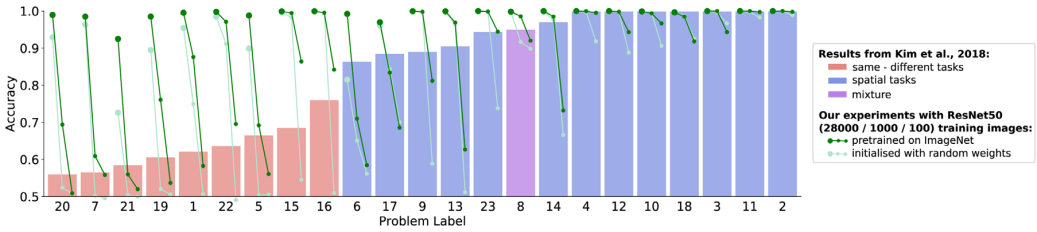


Figure 9. Accuracy of the models for the individual problems. Problem 8 is a mixture of same-different task and spatial task. In Figure 3, this problem was assigned to the spatial tasks. Bars replotted from Kim et al. (2018).

5,600 images for validation, and 11,200 images for testing.

Experimental procedures. For each of the SVRT problems, we fine-tuned a ResNet-50 that was pretrained on ImageNet (Deng et al., 2009) (as described in Appendix B, Training of ResNet-50 model). The same preprocessing, data augmentation, optimizer, and batch size as for the closed contour task were used.

For the different experiments, we varied the number of training images. We used subsets containing 28,000, 1,000, or 100 images. The number of epochs depended on the size of the training set: The model was fine-tuned for respectively 10, 280, or 2800 epochs. For each training set size and SVRT problem, we used the best learning rate after a hyper-parameter search on the validation set, where we tested the learning rates $[6 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}]$.

As a control experiment, we also initialized the model with random weights, and we again performed a hyper-parameter search over the learning rates $[3 \times 10^{-4}, 6 \times 10^{-4}, 1 \times 10^{-3}]$.

Results

In Figure 9, we show the results for the individual problems. When using 28,000 training images, we reached above 90% accuracy for all SVRT problems, including the ones that required same-different judgments (see also Figure 3B). When using less training images, the performance on the test set was reduced. In particular, we found that the performance on same-different tasks dropped more rapidly than on spatial reasoning tasks. If the ResNet-50 was trained from scratch (i.e., weights were randomly initialized instead of loaded from pretraining on ImageNet), the performance dropped only slightly on all but one spatial reasoning task. Larger drops were found on same-different tasks.

Appendix D: Recognition gap

Details on methods

Data set. We used two data sets for this experiment. One consisted of 10 natural, color images whose grayscale versions were also used in the original study by Ullman et al. (2016). We discarded one image from the original data set as it does not correspond to any ImageNet class. For our ground truth class selection, please see Table 1. The second data set consisted of 1,000 images from the ImageNet (Deng et al., 2009) validation set. All images were preprocessed like in standard training of ResNet (i.e., resizing to 256 × 256 pixels, cropping centrally to 224 × 224 pixels and normalizing).

Model. In order to evaluate the recognition gap, the model had to be able to handle small input images. Standard networks like ResNet (He et al., 2016) are not equipped to handle small images. In contrast, BagNet-33 (Brendel & Bethge, 2019) allows us to straightforwardly analyze images as small as 33 × 33 pixels and hence was our model of choice for this experiment. It is a variation of ResNet-50 (He et al., 2016), where most 3 × 3 kernels are replaced by 1 × 1 kernels such that the receptive field size at the top-most convolutional layer is restricted to 33 × 33 pixels.

Machine-based search procedure for minimal recognizable images. Similar to Ullman et al. (2016), we defined minimal recognizable images or configurations (MIRCs) as those patches of an image for which an observer—by which we mean an ensemble of humans or one or several machine algorithms—reaches $\geq 50\%$ accuracy, but any additional 20% cropping of the corners or 20% reduction in resolution would lead to an accuracy $< 50\%$. MIRCs are thus inherently observer-dependent. The original study only searched for MIRCs in humans. We implemented the following procedure to find MIRCs in our DNN: We passed each preprocessed image through BagNet-33 and selected the most predictive crop according to its

Image	WordNet Hierarchy ID	WordNet Hierarchy description	Neuron number in ResNet-50 (indexing starts at 0)
fly	n02190166	fly	308
ship	n02687172	aircraft carrier, carrier, flattop, attack aircraft carrier	403
	n03095699	container ship, containership, container vessel	510
	n03344393	fireboat	554
	n03662601	lifeboat	625
	n03673027	liner, ocean liner	628
eagle	n01608432	kite	21
	n01614925	bald eagle, American eagle, Haliaeetus leucocephalus	22
glasses	n04355933	sunglass	836
	n04356056	sunglasses, dark glasses, shades	837
bike	n02835271	bicycle-built-for-two, tandem bicycle, tandem	444
	n03599486	jinrikisha, ricksha, rickshaw	612
	n03785016	moped	665
	n03792782	mountain bike, all-terrain bike, off-roader	671
	n04482393	tricycle, trike, velocipede	870
suit	n04350905	suit, suit of clothes	834
	n04591157	windsor tie	906
plane	n02690373	airliner	404
horse	n02389026	sorrel	339
	n03538406	horse cart, horse-cart	603
car	n02701002	ambulance	407
	n02814533	beach wagon, station wagon, wagon estate car, beach waggon, station waggon, waggon	436
	n02930766	cab, hack, taxi, taxicab	468
	n03100240	convertible	511
	n03594945	jeep, landrover	609
	n03670208	limousine, limo	627
	n03769881	minibus	654
	n03770679	minivan	656
	n04037443	racer, race car, racing car	751
	n04285008	sports car, sport car	817

Table 1. Selection of ImageNet classes for stimuli of Ullman et al. (2016).

probability. See Appendix D, Selecting best crop when probabilities saturate on how to handle cases where the probability saturates at 100% and Appendix D, Analysis of different class selections and different number of descendants for different treatments of ground truth class selections. If this probability of the full-size image for the ground-truth class was $\geq$ 50%, we again searched for the 80% subpatch with the highest probability. We repeated the search procedure until the class probability for all subpatches fell below 50%. If the 80% subpatches would be smaller than 33×33 pixels, which is BagNet-33's smallest natural patch size, the crop was increased to 33×33 pixels using bilinear sampling. We evaluated the recognition gap as the difference in accuracy between the MIRC and the *best-performing* sub-MIRC. This definition was more conservative than the one from Ullman et al. (2016), who considered the maximum difference between a MIRC and its sub-MIRCs, that is, the difference between the MIRC and the *worst-performing* sub-MIRC. Please note that one difference between our machine procedure and the psychophysics experiment by Ullman et al. (2016) remained: The former was greedy, whereas the latter corresponded to an exhaustive search under certain assumptions.

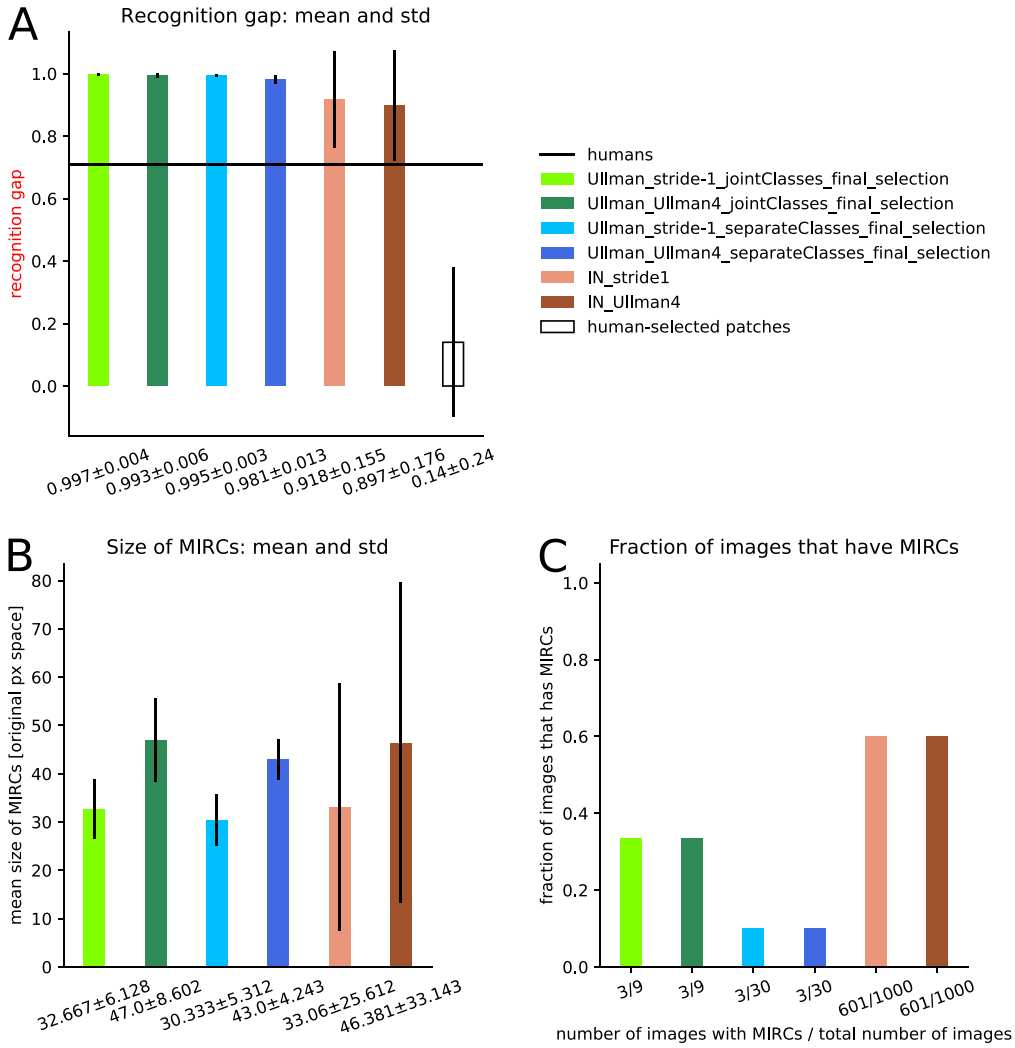


Figure 10. (A) Recognition gaps. The legend holds for all subplots. (B) Size of MIRCs. (C) Fraction of images that have MIRCS.

Analysis of different class selections and different number of descendants

Treating the 10 stimuli from Ullman et al. (2016) in our machine algorithm setting required two design choices: We needed to both pick suitable ground truth classes from ImageNet for each stimulus as well as choose if and how to combine them. The former is subjective, and using relationships from WordNet Hierarchy (Miller, 1995) (as Ullman et al. [2016] did in their psychophysics experiment) only provides limited guidance. We picked classes to our best judgment (for our final ground truth class choices, please see Table 1). Regarding the aspect of handling several ground truth classes, we extended our experiments: We tested whether considering all classes as one ("joint classes," i.e., summing the probabilities) or separately ("separate classes," i.e., rerunning the stimuli for each ground truth class) would have an effect on the recognition gap. As another check, we investigated whether the number of descendant options would alter the recognition gap: Instead of only considering the four corner crops as in the psychophysics experiment by Ullman et al. (2016) ("Ullman4"), we looked at every crop shifted by 1 pixel as a potential new parent ("stride-1"). The results reported in the main body correspond to joint classes and corner crops. Finally, besides analyzing the recognition gap, we also analyzed the sizes of MIRCs and the fractions of images that possess MIRCs for the mentioned conditions.

Figure 10A shows that all options result in similar values for the recognition gap. The trend of smaller

MIRC sizes for stride-1 compared to four corner crops shows that the search algorithm can find even smaller MIRCs when all crops are possible descendants (see. Figure 10B). The final analysis of how many images possess MIRCs (see Figure 10C) shows that recognition gaps only exist for fractions of the tested images: In the case of the stimuli from Ullman et al. (2016), three out of nine images, and in the case of ImageNet, about 60% of the images have MIRCs. This means that the recognition performance of the initial full-size configurations was $\geq$ 50% for those fractions only. Please note that we did not evaluate the recognition gap over images that did not meet this criterion. In contrast, Ullman et al. (2016) average only across MIRCs that have a recognition rate above 65% and sub-MIRCs that have a recognition rate below 20% (personal communication, 2019). The reason why our model could only reliably classify three out of the nine stimuli from Ullman et al. (2016) can partly be traced back to the oversimplification of single-class attribution in ImageNet as well as to the overconfidence of deep learning classification algorithms (Guo et al., 2017): They often attribute a lot of evidence to one class, and the remaining ones only share very little evidence.

Selecting best crop when probabilities saturate

We observed that several crops had very high probabilities and therefore used the "logit" measure $logit(p)$, where p is the probability. It is defined as the following: $logit(p) = \log(\frac{p}{1-p})$. Note that this measure is different from what the deep learning community usually refers to as "logits," which are the values before the softmax layer. In the following, we denote the latter values as $\mathbf{z}$. The logit $logit(p)$ is monotonic w.r.t. to the probability p, meaning that the higher the probability p, the higher the logit $logit(p)$. However, while p saturates at 100%, $logit(p)$ is unbounded. Therefore, it yields a more sensitive discrimination measure between image patches j that all have $p(\mathbf{z}^j) = 1$, where the superscript j denotes different patches.

In the following, we will provide a short derivation for the logit $logit(p)$. Consider a single patch with the correct class c. We start with the probability p_c of class c, which can be obtained by plugging the logits z_i into the softmax formula, where i corresponds to the classes [0, …, 1,000].

$$p_c(\mathbf{z}) = \frac{exp(z_c)}{exp(z_c) + \sum_{i \neq c} exp(z_i)} \quad (2)$$

Since we are interested in the probability of the correct class, it holds that $p_c(\mathbf{z}) \neq 0$. Thus, in the regime of interest, we can invert both sides of the equation. After simplifying, we get

$$\frac{1}{p_c(\mathbf{z})} - 1 = \frac{\sum_{i \neq c} exp(z_i)}{exp(z_c)}. \quad (3)$$

When taking the negative logarithm on both sides, we obtain

$$\Leftrightarrow -log\left(\frac{1}{p_c(\mathbf{z})} - 1\right) = -log\left(\frac{\sum_{i \neq c} exp(z_i)}{exp(z_c)}\right) \quad (4)$$

$$\Leftrightarrow - log\left(\frac{1 - p_c(\mathbf{z})}{p_c(\mathbf{z})}\right) = -log\left(\sum_{i \neq c} exp(z_i)\right) - (-log(exp(z_c))) \quad (5)$$

$$\Leftrightarrow log\left(\frac{p_c(\mathbf{z})}{1 - p_c(\mathbf{z})}\right) = z_c - log\left(\sum_{i \neq c} exp(z_i)\right) \quad (6)$$

The left-hand side of the equation is exactly the definition of the logit $logit(p)$. Intuitively, it measures in log-space how much the network's belief in the correct class outweighs the belief in all other classes taken together. The following reassembling operations illustrate this:

$$\begin{aligned} logit(p_c) &= log\left(\frac{p_c(\mathbf{z})}{1 - p_c(\mathbf{z})}\right) \\ &= \underbrace{log(p_c(\mathbf{z}))}_{\text{log probability of correct class}} \\ &\quad - \underbrace{log(1 - p_c(\mathbf{z}))}_{\text{log probability of all incorrect classes}} \end{aligned} \quad (7)$$

The above formulations regarding one correct class hold when adjusting the experimental design to accept several classes k as correct predictions. In brief, the logit $logit(p_C(z))$, where C stands for several classes, then states

$$\begin{aligned} logit(p_C(\mathbf{z})) &= -log\left(\frac{1}{p_{c_1}(\mathbf{z}) + p_{c_2}(\mathbf{z}) + ... + p_{c_k}(\mathbf{z})} - 1\right) \\ &= -log\left(\frac{1}{\sum_k p_k(\mathbf{z})} - 1\right) \\ &= \underbrace{log\left(\sum_k p_k(\mathbf{z})\right)}_{\text{log probability of all correct classes}} \end{aligned}$$

$$- \underbrace{log\left(1 - \sum_k p_k(\mathbf{z})\right)}_{\text{log probability of all incorrect classes}}$$

$$= log\left(\sum_k exp(z_k)\right) - log\left(\sum_{i \neq k} exp(z_i)\right) \quad (8)$$

Selection of ImageNet classes for stimuli of Ullman et al. (2016)

Note that our selection of classes is different from the one used by Ullman et al. (2016). We went through all classes for each image and selected the ones that we considered sensible. The 10th image of the eye does not have a sensible ImageNet class; hence, only nine stimuli from Ullman et al. (2016) are listed in Table 1.

P5: Disentanglement and generalization under correlation shifts

Christina M. Funke*, Paul Vicol*, Kuan-Chieh Wang, Matthias Kümmerer†, Richard Zemel†, Matthias Bethge†

* joint first authors, † joint senior authors

Presented at the *ICLR 2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality* and at the *Conference on Lifelong Learning Agents, 2022*.

Contributions The project was jointly led by CMF and PV with support from MK. All authors contributed in shaping the study at the conceptual level. MB, MK and CMF developed the crucial ideas for the basic example with correlated Gaussian source variables. The method for minimizing conditional mutual information for more complex tasks originated from PV, KW and RZ. The experiments were designed and implemented by CMF and PV with substantial help from MK and KW. CMF and MK took the lead on the toy tasks, while PV took the lead on the MNIST and CelebA datasets. The first draft of the paper was jointly written by PV, CMF, and MK with input from KW, RZ, and MB. All authors contributed to the final version and provided critical revisions.

DISENTANGLEMENT AND GENERALIZATION UNDER CORRELATION SHIFTS

Christina M. Funke*
University of Tübingen

Paul Vicol*
University of Toronto
Vector Institute

Kuan-Chieh Wang
University of Toronto
Vector Institute

Matthias Kümmerer†
University of Tübingen

Richard Zemel†
University of Toronto
Vector Institute

Matthias Bethge†
University of Tübingen

ABSTRACT

Correlations between factors of variation are prevalent in real-world data. Exploiting such correlations may increase predictive performance on noisy data; however, often correlations are not robust (e.g., they may change between domains, datasets, or applications) and models that exploit them do not generalize when correlations shift. Disentanglement methods aim to learn representations which capture different factors of variation in latent subspaces. A common approach involves minimizing the mutual information between latent subspaces, such that each encodes a single underlying attribute. However, this fails when attributes are correlated. We solve this problem by enforcing independence between subspaces conditioned on the available attributes, which allows us to remove only dependencies that are not due to the correlation structure present in the training data. We achieve this via an adversarial approach to minimize the conditional mutual information (CMI) between subspaces with respect to categorical variables. We first show theoretically that CMI minimization is a good objective for robust disentanglement on linear problems. We then apply our method on real-world datasets based on MNIST and CelebA, and show that it yields models that are disentangled and robust under correlation shift, including in weakly supervised settings.

1 INTRODUCTION

Disentangled representations can be useful for improving fairness (Locatello et al., 2019a), interpretability (Adel et al., 2018), controllable generative modeling (He et al., 2019), and transfer to downstream tasks (Van Steenkiste et al., 2019). In addition, they can improve robustness on out-of-distribution data (Higgins et al., 2017b) (e.g., for domain adaptation (Ilse et al., 2020) and domain generalization (Ben-Tal et al., 2009)). Most research on disentanglement has assumed that the underlying factors of variation in the data are *independent* (e.g., that factors are not correlated). However, this assumption is often violated in real-world settings: for example, in domain adaptation, the class distribution often shifts between domains (yielding a correlation between the class and domain); in natural images, there is often a strong correlation between the foreground and background (Beery et al., 2018), or between multiple foreground objects that tend to co-occur (e.g., a keyboard and monitor) (Tsipras et al., 2020; Beyer et al., 2020). Importantly, correlated data occur in areas that affect people's lives, including in healthcare (Chartsias et al., 2018) and fairness applications (Madras et al., 2018; Creager et al., 2019; Locatello et al., 2019a), and correlation shifts in these applications are common (e.g., demographics are likely to differ from one hospital to another).

The goal of disentanglement is to encode data into independent subspaces that preferably match the ground truth generative factors. A common approach to achieve this (used in ICA, PCA, and VAEs) is to ensure that the latent subspaces share as little information as possible, by minimizing the mutual information (MI) between subspaces. However, recently it has been shown that this fails to disentangle correlated factors (Träuble et al., 2020). Several works have sought to address this by introducing partial supervision (Träuble et al., 2020; Shu et al., 2019; Locatello et al., 2020b). Here, we show that even with *full* supervision, minimizing the MI can fail: it is impossible to encode generative factors into independent subspaces if they are correlated in the training data. To address this, we propose minimizing the MI between subspaces *conditioned* on the correlated attributes.

We compare three objective functions for learning disentangled representations: 1) standard supervised losses (such as mean-squared error or cross-entropy) that encourage each subspace to encode a specific attribute; 2) a supervised

* Equal contribution. † Shared senior authors.

loss plus *unconditional* MI minimization; and 3) a supervised loss plus *conditional* MI (CMI) minimization. We first show that approaches (1) and (2) fail on correlated and noisy data: minimizing a supervised loss cannot enforce that there is little information shared between subspaces; MI minimization is too strong a constraint to satisfy when the underlying factors of variation are correlated, and thus minimizing MI leads to decreased performance. We then show that minimizing CMI yields disentangled representations that are robust to correlation shifts.

Overall, we aim to establish conditional independence as the correct notion of independence between latent subspaces when disentangling data with correlated factors of variation.

Contributions.

- Most disentanglement metrics used in the literature assume that the attributes are uncorrelated, and thus are not directly applicable to correlated data. We propose to use the *predictive performance under correlation shift* as a *measure of disentanglement* applicable to settings with correlated factors of variation.
- We analyze the behavior of each objective function on a linear regression problem where all quantities of interest can be computed analytically (Section 3). We show that minimizing the CMI between latent subspaces yields a solution robust to test-time correlation shifts, while minimizing the unconditional MI (or only a supervised loss) does not.
- We describe an adversarial approach for learning conditionally disentangled representations (Section 4).
- Then, we apply our approach to CMI minimization to two tasks based on real-world datasets—a multi-digit occluded MNIST task and correlated CelebA—and demonstrate improved performance under correlation shift relative to baselines (Section 5).
- We investigate the interplay between correlation strength and noise level in the training data. When data are noisy and have strong correlations, the noise forces the model to rely on correlations when making a prediction; this leads to failures of the baseline approaches when correlations shift at test-time, and demonstrates the benefits of CMI minimization, which performs well across correlation strengths and noise levels.
- We show that CMI minimization can be applied in the weakly supervised setting, and show significant gains compared to baselines.

Our code is available on Github.

2 Background & Related Work

ICA/ISA. Disentanglement is related to blind source separation (BSS), as both problems revolve around the question of identifiability. A classic approach to BSS is Independent Component Analysis (ICA) (Comon, 1994; Jutten & Herault, 1991; Bell & Sejnowski, 1997; Olshausen & Field, 1996), which assumes statistical independence between the source variables (Jutten & Herault, 1991; Jutten & Karhunen, 2003). Independent Subspace Analysis (ISA) (Hyvärinen & Hoyer, 2000), or multidimensional ICA (Cardoso, 1998), is a generalization of ICA where each component is a k-dimensional subspace; dimensions within a subspace may have dependencies, while dimensions from different subspaces must be independent. Our work can be seen as a form of nonlinear ISA that enforces conditional independence between subspaces.

Correlations Between Features. With roots in ICA, most research on disentanglement focuses on data that was generated by independent factors, including synthetic benchmarks such as dSprites (Matthey et al., 2017), Shapes3D (Burgess & Kim, 2018), Cars3D (Reed et al., 2015), SmallNORB (LeCun et al., 2004), or MPI3D (Gondal et al., 2019). In real-world datasets on the other hand, factors are often correlated (Welinder et al., 2010; Lin et al., 2014). Träuble et al. (2020) pointed out the challenges that arise when attempting to learn disentangled representations on correlated data, and performed a large-scale empirical evaluation of the effect of correlations on widely-used VAE-based disentanglement models. They proposed two approaches to ameliorate the harmful effects of correlations: 1) introducing weak supervision during training, and 2) labeling data post-hoc to "correct" a pre-trained encoder. We show that even with full supervision, correlations are problematic when enforcing independence between latent subspaces. Causally-informed modeling (Zhang et al., 2020) is another approach to learning disentangled representations and extracting invariant features.To investigate the effect of correlations systematically, it is common to modify existing datasets to induce correlations, for example by subsampling the data, or generating synthetic datasets with the desired properties (Dittadi et al., 2020; Cimpoi et al., 2014; Jacobsen et al., 2018; Locatello et al., 2019b). We follow this approach in our experiments.

Unsupervised and Weakly-Supervised Disentanglement. Disentangled representation learning is often studied in the unsupervised setting, where the ground-truth factors of variation are unknown. Widely-used approaches for this include variational autoencoders (VAEs) (Kingma & Welling, 2013) and their variants (beta-VAE (Higgins

et al., 2017a), TC-beta-VAE (Chen et al., 2018), FactorVAE (Kim & Mnih, 2018), etc.). However, it was shown by Locatello et al. (2019b) that the assumption of independent source variables (e.g., attributes) is questionable, and that *purely unsupervised* disentanglement may not be possible. This spurred interest in *weakly-supervised* methods (Shu et al., 2019; Locatello et al., 2020b), where weak supervision is provided in the form of partial labels or grouping information (Bouchacourt et al., 2018; Nemeth, 2020; Klindt et al., 2020). In this paper, we focus on comparing MI and CMI minimization in the fully-supervised setting, as this is already challenging and provides useful insights.

Domain Adaptation/Generalization. We use predictive performance under correlation shift as a measure for the quality of disentanglement. This is closely related to the fields of domain adaptation and generalization, with the difference that we assume access to one source domain only. The goal of most related work in this field is to learn representations from multiple source domains that transfer to known (e.g., adaptation) or previously unseen (e.g., generalization) target domains. This is done by either learning domain-invariant representations which discard domain information (Tzeng et al., 2017) or by learning disentangled representations, with latent subspaces that correspond to the domain and the class, respectively (Peng et al., 2019; Ilse et al., 2020; Liu et al., 2018). For the latter approach, disentanglement is achieved by minimizing the mutual information between latent subspaces (Cheng et al., 2020; Gholami et al., 2020; Nemeth, 2020). Zhao et al. (2019) discuss fundamental problems inherent in learning domain-invariant representations when there are correlations between classes and domains (e.g., when the class distribution shifts in the target domain). The goal of Invariant Risk Minimization (Arjovsky et al., 2019) is to find correlations that are invariant over multiple training domains in order to improve generalization to out-of-distribution data.

Fairness. An important application of disentanglement is fairness. As machine learning systems are typically trained on historical data, they often inherit past biases (e.g., from human decision-makers). This may result in unfair treatment on the basis of sensitive properties such as ethnicity, gender, or disability. Typically, this can be addressed by modifying the training data to be unbiased or by adding a regularizer (e.g. based on mutual information) that quantifies and minimizes the degree of bias (Kamiran & Calders, 2009; Kamishima et al., 2011; Zemel et al., 2013; Hardt et al., 2016; Cho et al., 2020).

Mutual Information. The mutual information (MI) between two random variables $\mathbf{x}$ and $\mathbf{y}$, denoted $I(\mathbf{x};\mathbf{y})$, is the KL divergence between the joint distribution $p(\mathbf{x},\mathbf{y})$ and the product of the marginal distributions $p(\mathbf{x})p(\mathbf{y})$: $I(\mathbf{x};\mathbf{y}) = D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})]$. Minimization of MI has been used to implement an information bottleneck (Alemi et al., 2016) and to factorize representations (Jacobsen et al., 2018). MI minimization is at the heart of many approaches to disentanglement. The *conditional mutual information* (CMI) is defined as: $I(\mathbf{x};\mathbf{y} \mid \mathbf{z}) = \mathbb{E}_{\mathbf{z}}\left[D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{y} \mid \mathbf{z}) \mid\mid p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})\right]$. CMI measures the dependency between two variables given that we know the value of a third variable. For example, there is a dependency between a country's number of Nobel laureates per capita and chocolate consumption per capita (Prinz, 2020). However, this dependency is largely explained by the wealth of a country, thus $I(nobel;chocolate \mid wealth) < I(nobel;chocolate)$. In general, the CMI can be smaller or larger than the unconditional MI.

Estimating & Optimizing Mutual Information. Many approaches have been proposed for MI and CMI estimation and optimization. The Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018) uses a lower-bound of the MI based on the Donsker-Varadhan dual representation of the KL divergence (Donsker & Varadhan, 1983). Poole et al. (2019) provide an overview of variational bounds that can be used to estimate MI; most are *lower bounds*, which are useful in principle for *maximizing* MI, but which have also been used to minimize MI (even though minimizing a lower bound is not guaranteed to decrease MI). CLUB (Cheng et al., 2020) introduced a variational upper bound of MI, providing a more principled objective for minimizing MI. Several CMI estimators have been proposed, including conditional-MINE (Molavipour et al., 2020a), C-MI-GAN (Mondal et al., 2020), CCMI (Mukherjee et al., 2020), and an approach based on nearest neighbors (Molavipour et al., 2020b). Many approaches to MI minimization are based on batchwise shuffling of latent subspaces, sometimes referred to as metameric sampling (Belghazi et al., 2018; Nemeth, 2020; Feng et al., 2018; Park et al., 2020; Peng et al., 2019). The approach we use in Section 4 follows this paradigm of latent-space shuffling.

3 Disentanglement with Correlated Variables: Motivating CMI

A summary of notation is provided in Appendix A.

Problem Statement. Suppose we observe noisy data $\mathbf{x} \in \mathbb{R}^m$ obtained from an (unknown) generative process $\mathbf{x} = g(\mathbf{s})$ where $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ are the *underlying factors of variation*, also called source variables or attributes, which may be correlated with each other. We wish to find a mapping $f : \mathbb{R}^m \to \mathbb{R}^n$ to a latent space $f(\mathbf{x}) = \mathbf{z} =$

	Base	**Base + MI**	**Base + CMI**
Variance Explained, Training (Corr = 0.8)	91.9%	69.8%	90.9%
Variance Explained, Test (Corr = 0)	87.6%	65.0%	90.9%
Regression Matrix M (where $\hat{\mathbf{s}} = M\mathbf{x}$)	$\begin{pmatrix} 0.81 & 0.14 \\ 0.14 & 0.81 \end{pmatrix}$	$\begin{pmatrix} 1.07 & -0.46 \\ -0.46 & 1.07 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Table 1: Robustness of linear regression under correlation shift for each of the objectives *Base*, *Base+MI*, and *Base+CMI*. Here, the observations and predictions are in $\mathbb{R}^2$. The performance of the *Base* model drops under correlation shift. The optimal solution under the constraint of minimal MI, $I(z_1; z_2) = 0$, fails to model the in-distribution correlated training data. The solution with minimal *conditional* MI, $I(z_1; z_2 \mid s_1) = I(z_1; z_2 \mid s_2) = 0$, maintains consistent performance under correlation shift. Note that because the generative process is given by $g(\mathbf{s}) = \mathbf{A}\mathbf{s} = \mathbf{I}\mathbf{s}$, the inverse is $\mathbf{A}^{-1} = \mathbf{I}$. In the last row, we see that only Base + CMI recovers this true inverse.

$(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K)$ such that each attribute s_k can be recovered from the corresponding latent subspace $\mathbf{z}_k$ by a linear mapping $\mathbf{R}_k$, e.g., $\hat{s}_k = \mathbf{R}_k \mathbf{z}_k$ such that $\hat{s}_k \approx s_k$. We denote by $\mathbf{z}_{-i}$ the set of subspaces $\{\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_K\}$. We consider three different objectives for learning the latent subspaces: 1) minimizing a supervised loss L (e.g., mean squared error or cross-entropy), $\sum_{i=1}^{K} L(\hat{s}_i, s_i)$, denoted *"Base"*; 2) minimizing the *unconditional mutual information between subspaces* in addition to the supervised loss, $\sum_i L(\hat{s}_i, s_i) + I(\mathbf{z}_1, \ldots, \mathbf{z}_K)$, denoted *"Base+MI"*; and 3) minimizing the *conditional mutual information between subspaces conditioned on observed attributes*, in addition to the supervised loss, $\sum_i L(\hat{s}_i, s_i) + I(\mathbf{z}_i; \mathbf{z}_{-i} \mid s_i)$ denoted *"Base+CMI"*. We wish to learn a model that is robust to correlation shifts, e.g., if we train on data where $\text{corr}(s_i, s_j) > 0$, then we desire that the resulting model will perform similarly on uncorrelated data, $\text{corr}(s_i, s_j) = 0$, or anticorrelated data, $\text{corr}(s_i, s_j) < 0$.

In this section, we motivate the use of CMI minimization for learning robust disentangled representations. We use a linear regression task that can be solved analytically, and for which all quantities of interest, including MI and CMI, can be computed in closed form. This allows us to compare the solutions obtained via the vanilla mean-squared error objective (*Base*) to the solutions obtained by minimizing the MSE *under the constraint* that the MI or CMI between latent subspaces is minimized. This yields insight into the behavior of the objectives in the idealized case where the constraints they prescribe ($I(z_1; z_2) = 0$ for MI or $I(z_1; z_2 \mid s_1) = I(z_1; z_2 \mid s_2) = 0$ for CMI) are exactly satisfied.

First, we show that the supervised loss alone does not yield robust disentangled representations. Then, we show that additionally minimizing the unconditional MI forces the model to learn an *even worse solution*. Finally, we show that minimizing the conditional MI yields appropriately disentangled representations that are robust to correlation shift.

3.1 Full Supervision Does Not Yield Disentanglement

Here, we introduce a linear regression problem with correlated attributes. First, we analyze the solution obtained by optimizing only the *Base* objective, which in this case is the mean squared error. Consider a linear generative model with correlated Gaussian source variables $\mathbf{s}$, given by:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \quad , \quad \mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s}) \quad , \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$$

where $\mathbf{A}$ is the ground-truth mixing matrix and $\mathbf{C_s}$ and $\mathbf{C_n}$ are the covariance matrices for the source and noise variables, respectively. We assume that $\mathbf{x}$ is observed and wish to disentangle the underlying source variables $\mathbf{s}$; this corresponds to finding the mapping $\mathbf{A}^{-1}$ that inverts the data generating process. When we have access to the source variables, a natural approach is to minimize a supervised loss to ensure that each subspace contains information about its attribute. The optimal linear regression solution, both in the least squares sense and with respect to maximum likelihood, is given by the posterior mean:

$$\hat{\mathbf{s}}(\mathbf{x}) = \mathbb{E}\left[\mathbf{s} \mid \mathbf{x}\right] = \mathbf{C_{sx}}\mathbf{C_x^{-1}}\mathbf{x} \tag{1}$$

where $\mathbf{C_{sx}}$ and $\mathbf{C_x}$ are the following covariance matrices:

$$\mathbf{C_{sx}} = \mathbb{E}\left[\mathbf{s}(\mathbf{A}\mathbf{s} + \mathbf{n})^\top\right] = \mathbf{C_s}\mathbf{A}^\top \tag{2}$$

$$\mathbf{C_x} = \mathbf{A}\mathbf{C_s}\mathbf{A}^\top + \mathbf{C_n} \tag{3}$$

The least-squares optimal mapping $\mathbf{C_{sx}}\mathbf{C_x^{-1}}$ in Eq. 1 is not equal to the inverse $\mathbf{A}^{-1}$ of the generative model, as it is biased by the correlation structure $\mathbf{C_s}$ and $\mathbf{C_n}$ towards directions of maximal signal-to-noise ratio. Thus, regression is sensitive to noise, and this can lead to failures when evaluating the model on correlation-shifted data. For this Gaussian problem, we can compute the expected mean squared error (and therefore the expected variance explained) analytically:

$$\mathbb{E}\left[(\mathbf{s} - \hat{\mathbf{s}}(\mathbf{x}))^2\right] = \text{Var}\left(\mathbf{s}\right) = \text{Tr}(\mathbf{C_s}) \tag{4}$$

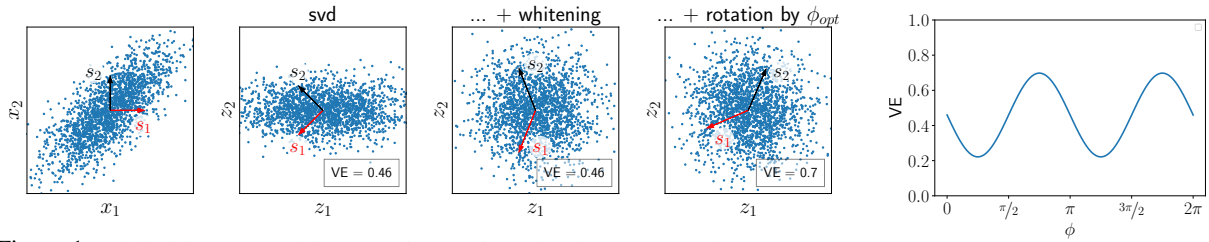


Figure 1: **Minimizing unconditional MI for the Gaussian linear regression task.** To enforce unconditional independence, we choose $\mathbf{W}$ such that $\mathrm{Cov}(\mathbf{z})$ is diagonal. In our case this is easy: the principal components of $\mathbf{x}$ are $x_1 + x_2$ and $x_1 - x_2$. The optimal regression loss with minimal MI is then given by whitening and rotating the result by angle ϕ_{opt} which leads to maximal variance explained ($\phi_{\text{opt}} = -\pi/4$ for positive correlations and $\mathbf{A} = \mathbf{I}$).

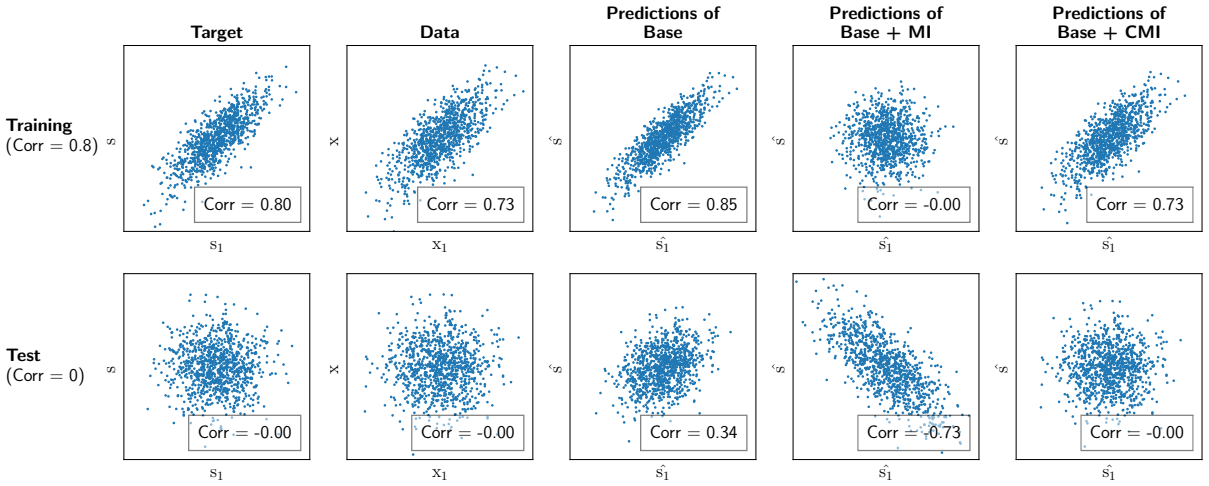


Figure 2: **Visualisation of targets $\mathbf{s}$, input data $\mathbf{x}$ and the predictions $\hat{\mathbf{s}}$ made by models using each of the different objectives {*Base*, *Base+MI*, *Base+CMI* }.** For *Base*, the predictions are more correlated than the data, revealing that the correlation in the training data is used to compensate for the noise. *Base+MI* leads to uncorrelated predictions. This cannot be the correct solution, as the targets are correlated. Only for *Base+CMI* does the correlation between the predictions and data match for both training and test data.

In Table 1, we see that in the two-dimensional case where $\mathbf{s} = (s_1, s_2)$ for $\mathbf{A} = \mathbf{I}$, $\mathbf{C_n} = 0.01 \cdot \mathbf{I}$ and the train-time correlation is $\mathrm{corr}(s_1, s_2) = 0.8$, $\hat{\mathbf{s}}$ explains 91.9% of the variance in $\mathbf{s}$ (column *"Base"*). However, when the correlation between s_1 and s_2 shifts at test time, such that $\mathrm{corr}(s_1, s_2) = 0$, then performance drops to 87.6%. This drop occurs because the estimator $\hat{\mathbf{s}}$ tries to make use of the assumed correlation between s_1 and s_2 to counteract the information lost due to noise, but this correlation is no longer present in the test data (see also Figure 2). The gap in performance between correlated and uncorrelated data indicates that s_1 and s_2 have not been correctly disentangled.

3.2 Unconditional Disentanglement Fails Under Correlation Shift

In the 2D linear case, we have:

$$\mathbf{z} = (z_1, z_2) = \mathbf{W}\mathbf{x}, \qquad \widehat{s}_1 = R_1 z_1, \qquad \widehat{s}_2 = R_2 z_2 \tag{5}$$

where the matrix $\mathbf{W}$ encodes the observation into the latent space. The linear regression example in Sec. 3.1 corresponds to $\mathbf{W} = \mathbf{C_{sx}}\mathbf{C_x^{-1}}$ and $R_k = 1$. In standard supervised objectives, there is no constraint preventing a subspace z_k from containing information about other source variables than s_k. A common approach to enforce independence is to minimize the MI between the latent subspaces z_1 and z_2 (Chen et al., 2018; Peng et al., 2019). In the Gaussian case, random variables are independent if and only if they are *uncorrelated*. The optimal linear regression weights $\mathbf{W}$ that yield $I(z_1; z_2) = 0$ (e.g., such that $\mathrm{Cov}(\mathbf{z})$ is diagonal) can be computed by whitening $\mathbf{x}$ and rotating the result by an angle ϕ_{opt} which leads to maximal variance explained. For our example in Table 1, where we have positive correlation and $\mathbf{A} = \mathbf{I}$, the optimal rotation is $\phi_{\text{opt}} = -\pi/4$ (see Figure 1). However, the resulting model no longer performs well on in-distribution data (Table 1, column *"Base+MI"*). There is correlation between the source variables s_1 and s_2 and therefore $I(s_1; s_2) > 0$. By enforcing independence, at least one of the subspaces cannot contain all relevant

information about its attribute and thus will have poor predictive performance. We make this precise in the following proposition.

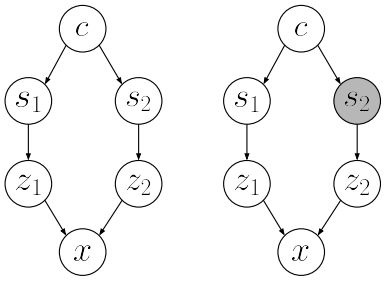
Proposition 3.1. *If $I(s_1; s_2) > 0$, then enforcing $I(z_1; z_2) = 0$ leads to $I(z_k; s_k) < H(s_k)$ for at least one k.*

Proof. The proof is provided in Appendix D. □

3.3 Conditional Disentanglement is Robust to Correlation Shift

We have seen that enforcing unconditional independence between the latent spaces does not solve the disentanglement problem. However, considering the graphical model in Figure 3, $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent *conditioned on either of s_1 or s_2*: assuming a common cause for the correlation between s_1 and s_2, there is a connection in the graphical model between $\mathbf{z}_1$ and $\mathbf{z}_2$ introducing a statistical dependence. Observing either s_1 or s_2 disconnects $\mathbf{z}_1$ and $\mathbf{z}_2$. Here, we show that enforcing independence *conditioned on each of the source variables* is also sufficient to yield a robust disentangled representation. For our 2D example, enforcing conditional independence corresponds to:

$$\mathrm{I}(\mathbf{z}_1; \mathbf{z}_2 \mid s_1) = 0 \qquad \text{and} \qquad \mathrm{I}(\mathbf{z}_1; \mathbf{z}_2 \mid s_2) = 0 \tag{6}$$



$I(z_1; z_2) > 0$ $\quad$ $I(z_1; z_2 | s_2) = 0$

Figure 3: The graphical model for two sources s_1, s_2 and corresponding latent subspaces z_1, z_2. We assume the source variables have a common cause c. In (a), when none of the sources are observed, there is a path from z_1 to z_2, so we have $I(z_1; z_2) > 0$; in (b) we observe s_2, which breaks the path, and thus $I(z_1; z_2 \mid s_2) = 0$.

Intuitively, if s_1 and s_2 are correlated, then $I(s_1; s_2) > 0$ and knowing s_1 gives us information about s_2. If we can predict s_1 from $\mathbf{z}_1$, and s_1 tells us about s_2, then it must be the case that $\mathbf{z}_1$ contains information about s_2.

We wish to ensure that $\mathbf{z}_1$ and $\mathbf{z}_2$ share *as little information as possible* (given the ground-truth correlation), to improve robustness to shifts. Since $\mathbf{z}_1$ necessarily contains some information about s_2, we enforce that it does not contain *any more information about $\mathbf{z}_2$ than necessary* via $I(\mathbf{z}_1; \mathbf{z}_2 | s_2)$, which states that if we know s_2, then knowing $\mathbf{z}_1$ does not give us more information about $\mathbf{z}_2$.

This does not penalize $\mathbf{z}_1$ for containing information about s_2 due to correctly predicting the correlated variable s_1 (and vice versa). In contrast to MI, this removes only the shared information which is not robust under correlation shift, but keeps the shared information which is necessary to account for the correlation between the source variables. The optimal solution under the conditional independence constraint (Eq. 6) is achieved by the mapping $\mathbf{W} = \mathbf{A}^{-1}$, successfully recovering the underlying generative model. This demonstrates the usefulness of minimizing CMI for generalization under correlation shifts in the case of linear regression with Gaussian variables and motivates us to investigate CMI minimization for larger-scale tasks.

4 Method: Minimizing CMI

For simple cases such as linear regression, we can compute and minimize the MI and CMI analytically; however, for most tasks, there is no closed form for the mutual information. In this section, we describe an approach to minimize the CMI for general classification tasks. Suppose we have a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{s}^{(i)})\}_{i=1}^{N}$ where $\mathbf{x}^{(i)}$ is an example and $\mathbf{s}^{(i)}$ is a vector of attribute labels — $\mathbf{s}_k^{(i)}$ is the label for the k^{th} attribute of the i^{th} example. We consider discrete attributes, $\mathbf{s}_k^{(i)} \in \mathbb{N}$. Let $f_{\boldsymbol{\theta}} : \mathbf{x} \mapsto \mathbf{z}$ denote an encoder parameterized by $\boldsymbol{\theta}$ that maps examples $\mathbf{x} \in \mathbb{R}^m$ to latent representations $\mathbf{z} \in \mathbb{R}^n$. We aim to learn one latent subspace per attribute, such that each subspace is independent from all other subspaces conditioned on the attribute it encodes.

We have $I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) = 0$ if $p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$. Our method enforces the latter condition using an adversarial discriminator. To obtain samples from $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \mathbf{s}_k)$ and $p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$, we loop over values of $\mathbf{s}_k$, and for each condition $\{\mathbf{s}_k = 0, \mathbf{s}_k = 1, \ldots\}$, we select examples from the minibatch that satisfy the condition, giving us samples from $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \mathbf{s}_k)$; then we shuffle the latent subspaces $\mathbf{z}_j, \forall j \neq k$ jointly batchwise (e.g., combining $\mathbf{z}_k$ from one example with $\mathbf{z}_{-k}$ from another) to obtain samples from $p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$. To enforce $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \mathbf{s}_k) = p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$, we train the encoder f adversarially against a discriminator trained to distinguish between these two distributions. The discriminator takes as input a representation and predicts whether it is "real" (e.g., drawn from the joint distribution) or "fake" (e.g., drawn from the product of marginals). One discriminator is trained for each attribute $\mathbf{s}_k$, which receives samples from the two distributions and the attribute value it is conditioned on. In practice, we use a conditional discriminator, effectively sharing parameters between the discriminators for each of the attributes. This process is illustrated in Figure 4. Algorithm 1 describes the encoder training loop; Algorithm 5

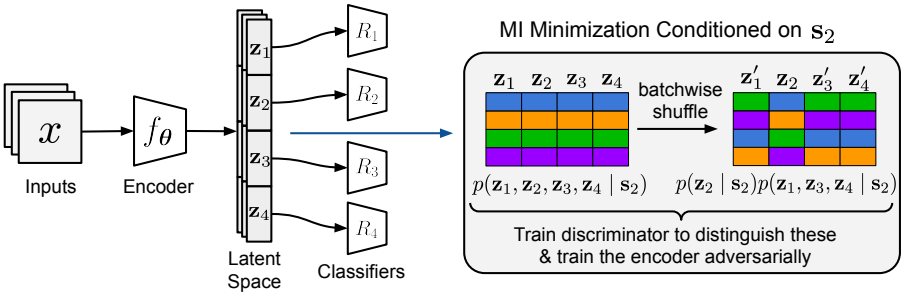


Figure 4: **Adversarial minimization of conditional mutual information via latent-space shuffling.** We minimize the CMI between latent subspaces, $I(\mathbf{z}_1; \cdots; \mathbf{z}_K \mid \mathbf{s}_k)$. Here, we illustrate the algorithm for four attributes with corresponding latent spaces $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$, where we condition on attribute $\mathbf{s}_2$. See Sec. 4 for a description of the method.

Algorithm 1 Adversarial Learning of Conditionally Disentangled Subspaces — Training the Encoder

Input: $\{\phi_1, \ldots, \phi_K\}$, initial parameters for K linear classifiers $R_1, \ldots, R_K$
Input: $\boldsymbol{\theta}$, initial parameters for the encoder f
Input: α, β learning rates for training the encoder and linear classifiers
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, K)$ ▷ Partition the latent space into K subspaces
  $L \leftarrow \sum_{k=1}^K L_{\text{cls}}(R_k(\mathbf{z}_k; \phi_k), \mathbf{s}_k)$ ▷ Cross-entropy for each attribute
  for $k \in \{1, \ldots, K\}$ **do** ▷ For each attribute/subspace
    $\mathbf{z}' \sim p(\mathbf{z}_1, \ldots \mathbf{z}_K \mid \mathbf{s}_k)$ ▷ Samples from the joint distribution
    $\mathbf{z}'' \sim p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
    $L \leftarrow L + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z}'')\right) + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z}')\right)$ ▷ Add adversarial loss
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L$ ▷ Update encoder parameters
  $\phi_k \leftarrow \phi_k - \beta \nabla_{\phi_k} L \quad , \quad \forall k \in \{1, \ldots, K\}$ ▷ Update classifier parameters
end while

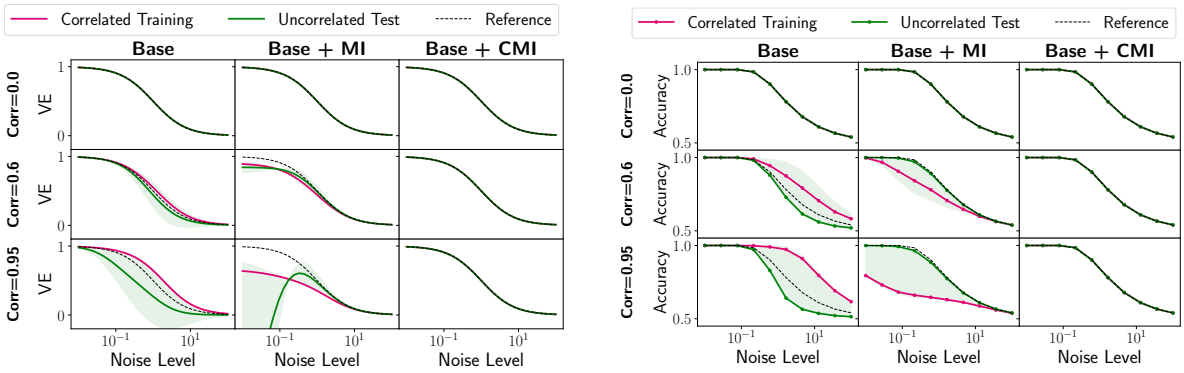
in Appendix C describes the corresponding discriminator training loop. We formally describe the algorithms for the baselines (*Base* and *Base + MI*) in Appendix C.

This approach is architecture-agnostic, and can be used to factorize the latent space of any classifier or generative model (e.g., VAEs (Joy et al., 2020) or flow-based models (Kingma & Dhariwal, 2018)). However, some models (such as VAEs) may have objectives that interfere with the goal of obtaining conditionally independent subspaces; for example, the ELBO encourages independence between all latent dimensions. In our experiments, we used linear and MLP encoders rather than VAEs to avoid this conflicting objective.

Because the latent space is typically low-dimensional, we have a choice of different distribution alignment techniques, including maximum mean discrepancy (MMD) (Gretton et al., 2006) and adversarial approaches (Goodfellow et al., 2014). Different GAN formulations can be interpreted as minimizing different divergences: the vanilla GAN (Goodfellow et al., 2014) minimizes the Jensen-Shannon divergence; WGAN (Arjovsky et al., 2017) minimizes the Wasserstein distance, which has been used to define an analogue of mutual information called the *Wasserstein dependency measure* (Ozair et al., 2019); f-GAN (Nowozin et al., 2016) minimizes an arbitrary f-divergence, etc. Each of these divergence measures will be 0 if and only if the subspaces are independent, however their training dynamics may differ. In practice, we found the vanilla GAN formulation to work well across our experiments.

5 Experiments

Our experiments aim to answer the following questions: 1) What is the effect of the train-time correlation strength and noise level on the solutions found by training with each objective, *Base*, *Base+MI*, and *Base+CMI*? 2) Can we successfully learn conditionally disentangled representations for classification tasks using Algorithm 1? and 3) Does CMI minimization lead to improved correlation-shift robustness on natural image datasets including MNIST and CelebA?



(a) Toy linear regression.

(b) Toy classification with ten attributes.

Figure 5: **Synthetic linear regression (left) and linear classification (right) tasks.** We measure the performance (variance explained for regression and accuracy for classification) on the correlated training data (magenta) and on test data with a range of correlation shifts (green, solid line is the uncorrelated test data). The performance of the *Base* model in the uncorrelated setting serves as a reference in each plot (dashed black line) and facilitates the comparison of the performance of the different objectives (columns). In both tasks, we find that, *Base+CMI* leads to robustness to correlation shift independent of the noise level (x-axis) and the strength of the correlation in the training data (rows), while the other approaches do not.

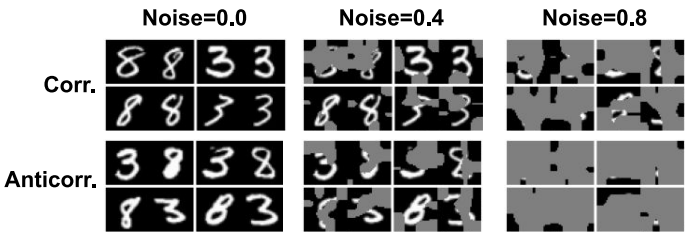
First, we present results on the analytically-solvable linear regression example, illustrating the effect of the correlation strength and noise level on the solution obtained by each objective. Then, we demonstrate that our findings also hold for a synthetic classification task with multiple attributes. Next, we employ the method described in Section 4 and investigate two realistic tasks, a multi-digit MNIST task with occlusions and correlated CelebA, and show that minimizing CMI can largely eliminate the gap in performance caused by test-time correlation shifts. Finally, we evaluate common disentanglement metrics and apply Algorithm 1 in weakly supervised settings. Experimental details and extended results are provided in Appendix B.

Linear Regression. Here, we revisit the linear regression problem from Section 3, to investigate the impact of the train-time correlation strength and noise level on the models learned with each of the objectives *Base*, *Base+MI*, and *Base+CMI*. The results are shown in Figure 5a. We found that *Base+CMI* yields robustness to correlation shift across all correlation strengths and noise levels, while the baselines do not. The performance of *Base* drops most severely under correlation shift for strong train-time correlations and intermediate noise levels; in this regime, *Base+CMI* improves performance substantially.

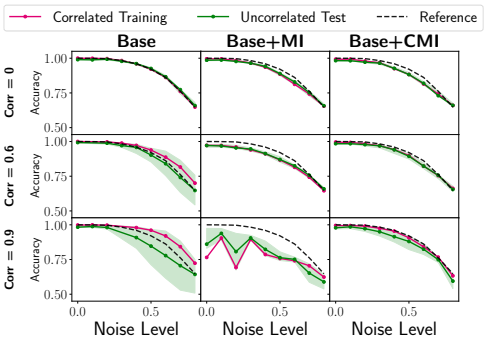
Toy Multi-Attribute Classification. Next, we investigated whether these findings hold for classification tasks with multiple attributes. Here, binary source attributes $s_k = \pm 1, \forall k \in \{1, \dots, K\}$ generate the observed data via $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$ (we set $\mathbf{A} = \mathbf{I}$ for simplicity) with normally distributed noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$. We induced correlations between the attributes a_k, such that the number of datapoints differs for the different combinations of attribute values. In the multi-attribute setting, the correlation strength refers to the pairwise correlation between all attributes. Similarly to the regression task, we find that *Base+CMI* leads to robustness under correlation shift (see Figure 5b and Appendix B.1).

Multi-Digit Occluded MNIST. Next, we designed a larger-scale task to investigate whether these properties hold in a more complex setting. We created a dataset by concatenating two MNIST digits side-by-side, where the aim is to predict both the left- and right-hand labels. We generated occlusion masks using the procedure used by Chai et al. (2021); examples from our synthetic dataset under a range of noise settings are shown in Figure 6a. We used a subset of MNIST consisting of classes 3 and 8 (which are visually similar and can become ambiguous under occlusions). This mimics multiple-object classification in a way that allows us to control the correlation strength and noise level (via the amount of occlusion), allowing for systematic analysis. This task is a more complex analogue of the synthetic classification task from Figure 5b. We added explicit occlusion noise because the MNIST data itself is simple, and has too little "natural" noise to clearly observe the predicted effects (e.g., for low noise levels, the supervised loss already does well). While this task would also be possible for colored MNIST and dSprites, one advantage of our task is its symmetry, which allows us to exclude potential side-effects: here, the attributes have the same type (the digit identity), whereas the attributes in colored MNIST (digit identity and color) and dSprites (shape, size, position, etc.) are more diverse.

Similarly to the toy tasks, we train an encoder to map images onto a D-dimensional latent space, which is partitioned in two equal-sized subspaces corresponding to the two digits; we train a linear classifier on each subspace to predict the



(a) Correlated training data and anti-correlated test data.

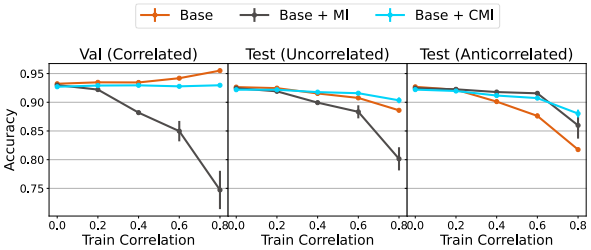


(b) Performance comparison.

Figure 6: **Multi-digit occluded MNIST.** (a) Examples of the correlated training data (where `3-3` and `8-8` pairs are frequent) and anticorrelated test data (where `3-8` and `8-3` pairs are frequent), under a range of occlusion strengths. (b) Accuracies under correlation shifts for different noise levels, achieved by training with each of the objective functions *Base*, *Base+MI*, and *Base+CMI*. *Base+CMI* achieves consistent performance across correlation shifts. Similarly to Figure 5, here we show the reference performance of the model trained on uncorrelated data (solid black line), the performance on correlated training data (magenta) and on a range of test-time correlations in $[0, 1]$ (shaded green region, where solid green denotes the uncorrelated test performance).

(a) Correlated train data.

(b) Anti-correlated test data.



(c) Performance comparison.

Figure 7: **Correlated CelebA.** (a) Training examples with correlation 0.8 between attributes `Male` and `Smiling`, such that the majority of men are smiling while the majority of women are not. (b) Anti-correlated test examples, where the majority of women are smiling. (c) Accuracies of each method under a range of correlation strengths, for validation data with the same correlation as the training data, uncorrelated test data, and anticorrelated test data.

respective class labels. We consider different correlation strengths between the left and right digits in the training set (where strong correlation means that the digits often match, e.g., `3-3` or `8-8` are more common than `3-8` or `8-3`). We evaluate each model on test data with correlation strengths ranging from $[-1, 1]$. The results are shown in Figure 6b. We found that the conclusions from the toy experiments hold in this setting: supervised learning with only the cross-entropy loss, as well as with unconditional MI minimization, fail under test-time correlation shift, while minimizing CMI is more robust. Experimental details and extended results are provided in Appendix B.2.

Correlated CelebA. Finally, we consider a realistic setting using the CelebA faces dataset (Liu et al., 2015). In contrast to the multi-digit MNIST task, here we do not add any artificial observation noise (as CelebA is a more complex dataset that naturally has noise in observations and/or labels). We selected two attributes that we know *a priori* are not causally related, `Male` and `Smiling`, and we created subsampled datasets with a range of training correlations $\{0, 0.2, 0.4, 0.6, 0.8\}$. We evaluated our models on both *anti-correlated* and *uncorrelated* test sets (Figures 7a and 7b). Figure 7c compares the performance of the baseline classifier, unconditional MI model, and conditionally disentangled model under a range of correlation strengths. We found that minimizing CMI has a larger effect for medium-to-high correlation; however, CMI minimization does not hurt performance at low correlation strengths. Note that while the unconditional model appears to have good performance on the anti-correlated test set, its performance is poor on the validation set (that has the same correlation structure as the training set), so this model does not perform well on in-distribution-data. In contrast, the *Base+CMI* model performs well on both in-distribution data and shifted test distributions. Also note that the problem of disentangling correlated attributes does not occur only under correlation shift, but is already present in the source domain where certain attribute combinations will reliably be treated incorrectly. For example, *Base* fails to recognize the rare non-smiling male faces in 49% of the cases, while *Base+CMI* fails only in 25% of the cases. Additional details are in Appendix B.3.

Disentanglement Metrics. Locatello et al. (2020a) showed that common disentanglement metrics are not suitable for the correlated setting. For this reason, we focused on comparing performance under correla-

tion shift, which we consider more suitable for correlated data: if a model cannot predict a factor of variation well for certain values of another factor, then the model did not successfully disentangle those factors. However, one can still make use of the disentanglement metrics by evaluating them on *uncorrelated data*, using models trained on correlated data. We performed this analysis for the toy classification and CelebA tasks, and found that *Base+CMI* leads to improved disentanglement scores across a wide range of metrics, compared to *Base* and *Base+MI* (Appendix B.4).

Extension to the Weakly Supervised Setting. Algorithm 1 can be applied directly to weakly supervised settings; it is not necessary for each datapoint to have labels for all attributes. We find that when reducing the number of labels, *Base+CMI* outperforms the other objectives under correlation shift (see Figure 8 and Appendix B.5).

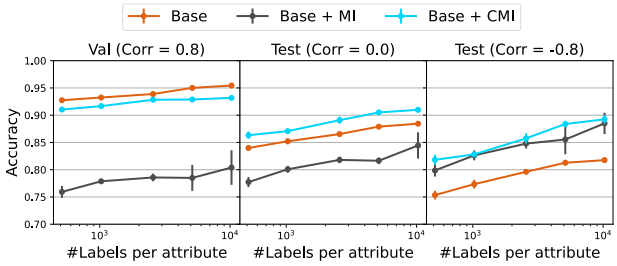


Figure 8: **Weakly-supervised CelebA.** The x-axis shows the number of labels per attribute used during training; the rightmost datapoint corresponds to full supervision. *Base+CMI* outperforms the other objectives under correlation shift.

6 Limitations & Future Work

Our study mainly concerns the setting where the underlying factors of variation are known. Practical applications of this setting can occur with respect to fairness, where one may wish to train a model such that correlations that exist in the training data are not relied upon for prediction. Nonetheless, full supervision is a strong assumption and an exciting goal for future work would be to look into relaxing this assumption. Our experiment with the weakly supervised version of the CelebA experiment is a first step in this direction.

We have shown that minimizing CMI yields predictions that disregard correlations between attributes in the training data, which is helpful when correlations shift between the training and test data. This approach relies on knowing a priori which correlations should not be used. This is the case, for example, for fairness applications where a person's race or gender should not affect the results. A direction for future work would be to automatically determine which correlations are more or less likely to shift in held-out data and to add this step before applying our approach of avoiding the unwanted correlations. One may incorporate ideas from IRM (Arjovsky et al., 2019), which leverages multiple environments at training time to discover which correlations tend to shift and which are stable—e.g., to distinguish between causal and spurious correlations, the latter of which we wish to avoid relying on. A fruitful direction for future work would be to combine IRM-style discovery of spurious correlations with our approach, which can be used to control for these correlations when learning disentangled representations. In a related vein, there has been recent work which aims to discover environments when none are given explicitly (Creager et al., 2021), which may be useful in combination with our work.

While CMI is defined for both continuous and discrete attributes, our method of shuffling the latent subspaces is only applicable to discrete attributes. Discrete attributes are prevalent in many settings: in domain adaptation, the class and domain are discrete; in multi-object classification, the class of each object is a discrete attribute; the foreground and background of natural images are discrete, etc. Nevertheless, finding methods to minimize CMI for continuous attributes is an interesting direction for future work. Another caveat of our method for minimizing the CMI via latent subspace shuffling is the increased computational cost relative to minimizing the unconditional MI: the cost for CMI scales linearly with the number of attributes and attribute values, while the cost for MI is constant.

7 Conclusion

Correlations are prevalent in real-world data, yet pose a substantial challenge for disentangled representation learning. Standard approaches learn to rely on these correlations, especially when data are noisy, as the correlations provide an easy-to-learn signal with predictive power. When the attributes are not causally related, this leads to poor performance under test-time correlation shift. Although for small correlations the effects may not be large, relying on these correlations and thereby systematically treating a subset of the data incorrectly, can be catastrophic for fairness. We first showed that supervised learning and *unconditional* mutual information minimization fail to learn representations robust to such shifts. We then argued that the correct notion of disentanglement in such cases is *conditional disentanglement*, and we proposed a simple approach to minimize the conditional mutual information between latent subspaces. We showed that conditionally disentangled representations improve robustness to correlation shift in analytically solvable linear tasks, as well as on natural images. Overall, we established CMI minimization as a more appropriate alternative to MI minimization, which sets the stage for the development of more powerful objective functions for disentanglement.

ACKNOWLEDGEMENTS

We thank Jörn-Henrik Jacobsen for his valuable contributions in the early stage of this work. We thank Steffen Schneider, Dylan Paiton, Lukas Schott, Elliot Creager, and Frederik Träuble for helpful discussions. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christina Funke. Paul Vicol was supported by a JP Morgan AI Fellowship.

We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the Competence Center for Machine Learning (FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellence Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the Deutsche Forschungsgemeinschaft (DFG; Projektnummer 276693517 – SFB 1233). Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/partners.

REFERENCES

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning (ICML)*, pp. 50–59, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

Anthony J Bell and Terrence J Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with Imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Chris Burgess and Hyunjik Kim. 3D Shapes Dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

J-F Cardoso. Multidimensional independent component analysis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pp. 1941–1944, 1998.

Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in GANs. *arXiv preprint arXiv:2103.10426*, 2021.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 490–498, 2018.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2610–2620, 2018.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning (ICML)*, pp. 1779–1788, 2020.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using mutual information. In *International Symposium on Information Theory (ISIT)*, pp. 2521–2526. IEEE, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407*, 2020.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Zunlei Feng, Xinchao Wang, Chenglong Ke, An-Xiang Zeng, Dacheng Tao, and Mingli Song. Dual swap disentangling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5894–5904, 2018.

Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29: 3993–4002, 2020.

Muhammad Waleed Gondal, Manuel Wüthrich, Đorđe Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: A new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample problem. *Advances in Neural Information Processing Systems (NeurIPS)*, 19:513–520, 2006.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017a.

Irina Higgins, Arka Pal, Andrei A Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: Improving zero-shot transfer in reinforcement learning. *arXiv preprint arXiv:1707.08475*, 2017b.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. DIVA: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348, 2020.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Tom Joy, Sebastian Schmon, Philip Torr, N Siddharth, and Tom Rainforth. Capturing label characteristics in VAEs. In *International Conference on Learning Representations (ICLR)*, 2020.

Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

Christian Jutten and Juha Karhunen. Advances in nonlinear blind source separation. In *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 245–256, 2003.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, 2009.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, pp. 2649–2658, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 11–104, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *arXiv preprint arXiv:1809.01361*, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14611–14624, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. *Journal of Machine Learning Research*, 21:1–62, 2020a.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning (ICML)*, 2020b.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Sina Molavipour, Germán Bassi, and Mikael Skoglund. Conditional mutual information neural estimator. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5025–5029, 2020a.

Sina Molavipour, Germán Bassi, and Mikael Skoglund. On neural estimators for conditional mutual information using nearest neighbors sampling. *arXiv preprint arXiv:2006.07225*, 2020b.

Arnab Kumar Mondal, Arnab Bhattacharya, Sudipto Mukherjee, Sreeram Kannan, Himanshu Asnani, and Prathosh AP. C-MI-GAN: Estimation of conditional mutual information using MinMax formulation. *arXiv preprint arXiv:2005.08226*, 2020.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. CCMI: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pp. 1083–1093, 2020.

Jozsef Nemeth. Adversarial disentanglement with grouped observations. In *34th AAAI Conference on Artificial Intelligence*, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. F-GAN: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *arXiv preprint arXiv:1903.11780*, 2019.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances on Neural Information Processing Systems (NeurIPS)*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Ken Perlin. Improving noise. In *29th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 681–682, 2002.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Aloys Leo Prinz. Chocolate consumption and Nobel laureates. *Social Sciences & Humanities Open*, 2(1):100082, 2020. ISSN 2590-2911. doi: 10.1016/j.ssaho.2020.100082. URL https://www.sciencedirect.com/science/article/pii/S2590291120300711.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in Neural Information Processing Systems (NeurIPS)*, 28:1252–1260, 2015.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning (ICML)*, pp. 6056–6065, 2019.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? On the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From Imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning (ICML)*, pp. 9625–9635, 2020.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7167–7176, 2017.

Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14245–14258, 2019.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pp. 325–333, 2013.

Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *arXiv preprint arXiv:2005.01095*, 2020.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

APPENDIX

This appendix is structured as follows:

- In Section A we provide an overview of the notation we use throughout the paper.
- In Section B we provide experimental details, as well as extended results.
- In Section C we provide the algorithms for the baseline methods, namely for classification-only training and unconditional mutual information minimization.
- In Section D we provide a proof of Proposition 3.1.

A NOTATION

Symbol	Meaning
$\mathbf{x}$	Observations
$\mathbf{s}$	Ground-truth latent factors
$\hat{\mathbf{s}}$	Predictions of factors
$\mathbf{z}$	Latent representation
$\mathbf{W}$	Linear regression weights
R_1, R_2	Linear readout from the latent space $\mathbf{z}$ to predictions $\hat{\mathbf{s}}$
$\mathbf{n}$	Isotropic Gaussian noise, $\mathbf{n} \sim \mathcal{N}(0, \mathbf{C_n})$ with $\mathbf{C_n} = \sigma^2 I$
$\mathbf{A}$	Square matrix used to generate observations for the linear task as $\mathbf{x} = \mathbf{As} + \mathbf{n}$
f	Encoder function
f_θ	Encoder function with parameters θ

Table 2: Summary of the notation used in this paper.

B EXPERIMENTAL DETAILS AND EXTENDED RESULTS

Method Details. Note that the dimensions m and n are arbitrary—in particular, n does not need to be smaller than m. In principle, each subspace can have different dimension (e.g., the linear readout layer for each attribute can have arbitrary dimensions $A \times S$ where A is the attribute dimensionality and S is the dimensionality of a particular subspace).

Compute Environment. Our experiments were implemented using PyTorch (Paszke et al., 2019), and were run on our internal clusters. The toy 2D experiments were run on a single NVIDIA RTX 2080 TI GPU, and took approximately 48 hours for all the results presented. The MNIST and CelebA experiments were run on NVIDIA Titan Xp GPUs. Each run of the multi-digit MNIST and CelebA tasks for a given method and correlation strength (and noise level in the MNIST case) took approximately 12 hours, and these were run in parallel.

B.1 TOY MULTI-ATTRIBUTE CLASSIFICATION

We performed this experiment with two, four and ten binary attributes. The results for varying numbers of attributes are shown in Figure 10. For two attributes we illustrated the data $\mathbf{x}$ for different correlation strength and noise levels (Figure 9). Here, increasing the correlation strength means that data points with $a_1 = a_2$ are increasingly more common relative to $a_1 \neq a_2$. The noise level on the other hand determines the overlap of the distributions and therefore the difficulty of the task.

Experimental Details. We used a PacGAN-style setup (Lin et al., 2018) for our toy experiments, where the discriminator takes as input a concatenation of 50 samples.

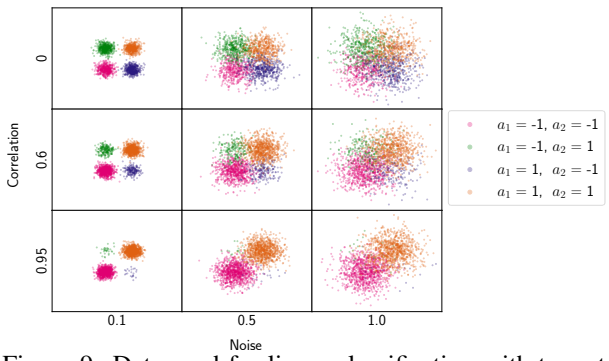
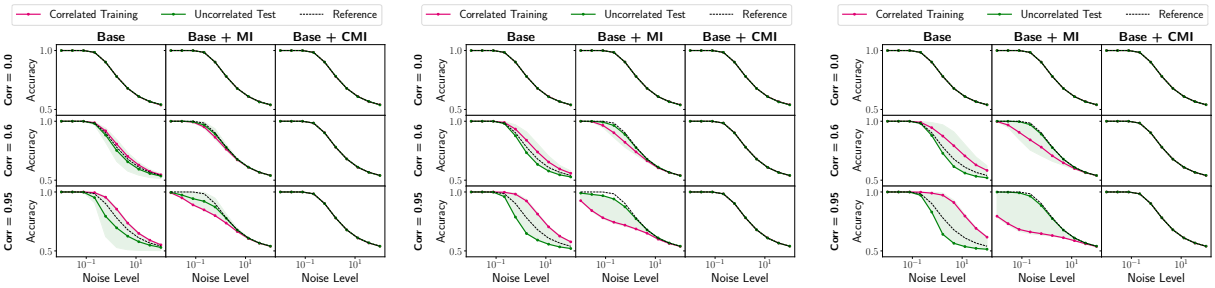


Figure 9: Data used for linear classification with two attributes (a_1 and a_2), visualized for a range of correlation strengths and noise levels.



(a) Two attributes. (b) Four attributes. (c) Ten attributes.

Figure 10: Toy classification with different numbers of attributes. Strong negative correlations could not be generated for multiple attributes; thus only positive test correlations were evaluated for (b) and (c).

- **Base**: We used Adam (Kingma & Ba, 2014) with a learning rate of 0.01.
- **Base + MI**: We used Adam to optimize the encoder, linear classifiers, and discriminators. After each step of optimizing the discriminator and encoder, we optimized the linear classifiers (R) for 10 steps. The disentanglement loss term was weighted by a factor of 100 relative to the classification loss. In preliminary tests, we found that the optimal learning rate depended on noise level, correlation strength, and number of attributes. The results in Figure 5b were obtained using one of the following learning rates for the discriminator $\{1e-4, 2e-4, 5e-4, 1e-3, 5e-3\}$. The learning rate of the generator and linear classifiers was chosen to be 10 times smaller than the discriminator learning rate.
- **Base + CMI**: For $\mathbf{A} = \mathbf{I}$, no optimization was necessary, as we already know the optimal solution to be $\mathbf{W} = \mathbf{A}^{-1} = \mathbf{I}$. We confirmed experimentally that the discriminator could not get above chance performance for this solution.

B.2 Multi-Object Occluded MNIST

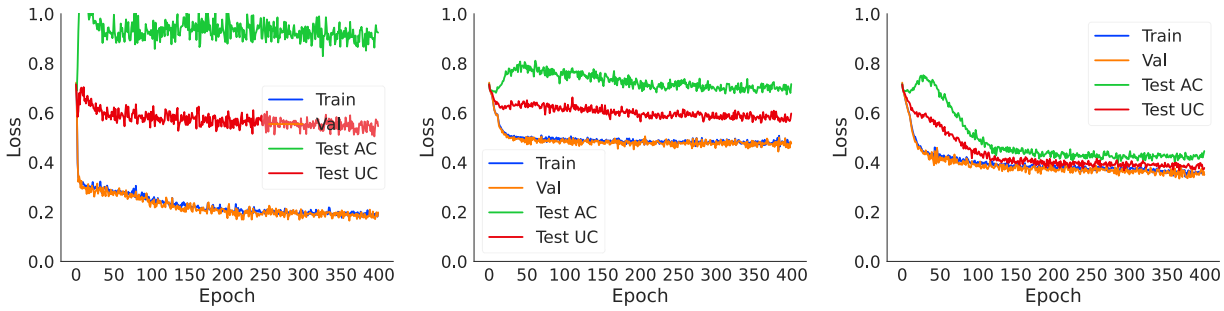
We used minibatch size 100, and latent dimension $D = 10$, yielding two subspaces each of dimension 5. As the encoder model, we used a three-layer MLP with 50 hidden units per layer and ReLU activations. We trained for 400 epochs, using Adam (Kingma & Ba, 2014) to optimize the encoder, linear classifiers, and discriminators, with separate learning rates for each component chosen via a grid search over $\{1e-5, 1e-4, 1e-3\}$.

Correlated Data Generation. We used the default MNIST training and test splits, and held out 10k of the original training examples to form a validation set, yielding 50k, 10k, and 10k examples in the training, validation, and test sets, respectively. Each digit is first rescaled to be 32×32 pixels. The correlated data was generated on-the-fly during training. Each example in a minibatch was created by: 1) sampling the left-right digit combination (e.g., { `3-3`, `3-8`, `8-3`, `8-8` }) from a joint distribution encoding the desired correlation; 2) choosing random instances of each of the selected classes (e.g., a random image of a `3` and a random image of an `8`); 3) applying occlusions separately to each image; and 4) concatenating the images, yielding a 32×64 example. This procedure was performed for each training and test minibatch, yielding a larger amount of data than would be possible with a fixed dataset generated *a priori*. To generate occlusions, we use the approach from (Chai et al., 2021), which produces contiguous masks similar to Perlin noise (Perlin, 2002). We used gray occlusions to remove a potential ambiguity that exists with black masks (which blend into the MNIST background): a masked `8` can become identical to a `3`, so one could not tell whether the image is a noisy `8` or a clean `3`.

B.3 CelebA

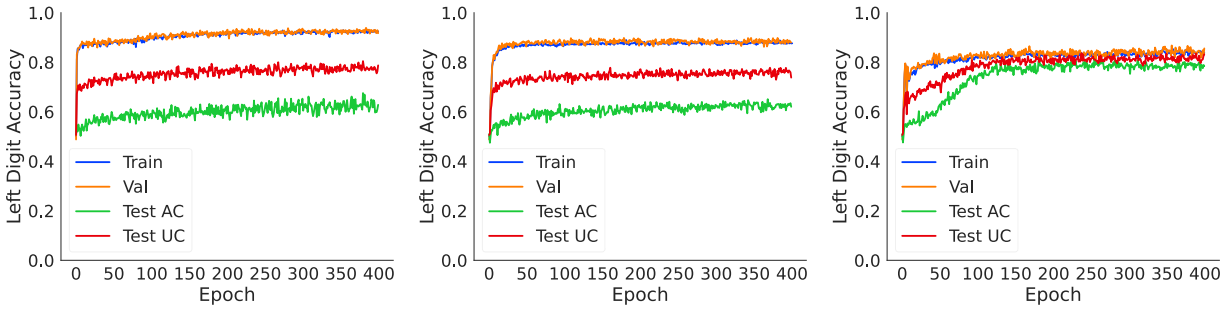
For all experiments, we used minibatch size 100, and latent dimension $D = 10$. As the encoder model, we used a three-layer MLP with 50 hidden units per layer and ReLU activations. Similarly to the MNIST setup, we trained for 200 epochs, using Adam to optimize the encoder, linear classifiers, and discriminators. For each method, we performed a grid search over learning rates $\{1e-5, 1e-4, 1e-3\}$ separately for each of the encoder, discriminator(s), and linear classification heads; we selected the best learning rates based on validation accuracy.

Correlated Data Generation. We first pre-processed all images by taking a 128×128 center crop, and then resizing to 64×64. Pixel values were normalized to the range $[0, 1]$. We used the original training, validation, and test splits



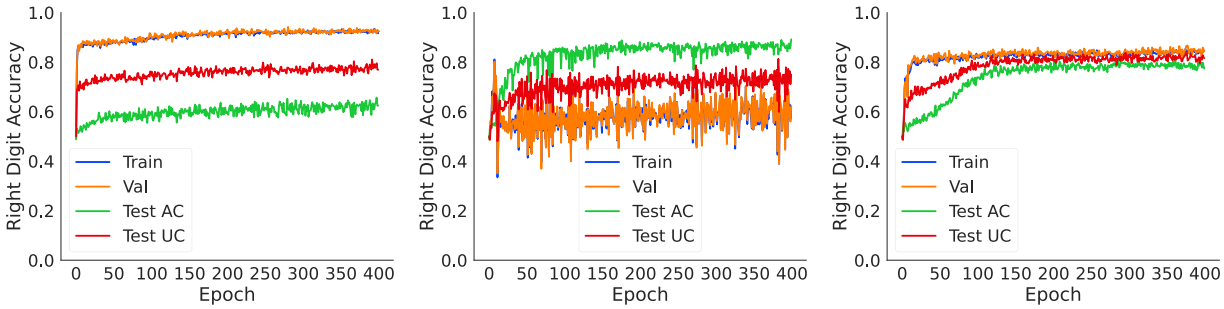
(a) Base (b) Base + MI (c) Base + CMI

Figure 11: Average cross-entropy loss for the left and right digit predictions, under the strongest correlation we consider, $c = 0.9$, at noise level 0.6 (where the noise is parameterized by a factor that has range $[0, 1]$).



(a) Base (b) Base + MI (c) Base + CMI

Figure 12: Accuracies for the left digit, under the strongest correlation we consider, $c = 0.9$, at noise level 0.6 (where the noise is parameterized by a factor that has range $[0, 1]$).



(a) Base (b) Base + MI (c) Base + CMI

Figure 13: Accuracies for the right digit, under the strongest correlation we consider, $c = 0.9$, at noise level 0.6 (where the noise is parameterized by a factor that has range $[0, 1]$).

provided with the CelebA dataset. In order to enforce arbitrary correlations between specific attributes, we subsampled the data such that we retained the maximum possible number of examples in each of the Train/Validation/Anticorrelated Test/Uncorrelated Test splits, while satisfying precisely the desired correlation. The validation set has the same correlation as the training set, and Figure 14 shows the number of examples in each of these sets for the strongest correlation we consider, $c = 0.8$. Figures 15, 16, and 17 show the cross-entropy loss and accuracies on each of the factors `Male` and `Smiling` (with training correlation 0.8) over the course of optimization, for each of the methods we compare (classification-only, unconditional disentanglement, and conditional disentanglement). We see that the conditional model substantially outperforms the baselines, with a much smaller gap between validation accuracy and both anti-correlated (AC) and uncorrelated (UC) test accuracies. Figures 18, 19 and 20 show confusion matrices for each method on the correlated validation set, anticorrelated test set, and uncorrelated test set, respectively. Finally,

Tables 3 and 4 show the prediction error of the models trained with the different objectives for both the combinations that were common and rare during training. These results shows that some attribute combinations (such as the rare non-smiling male faces) are reliably treated incorrectly.

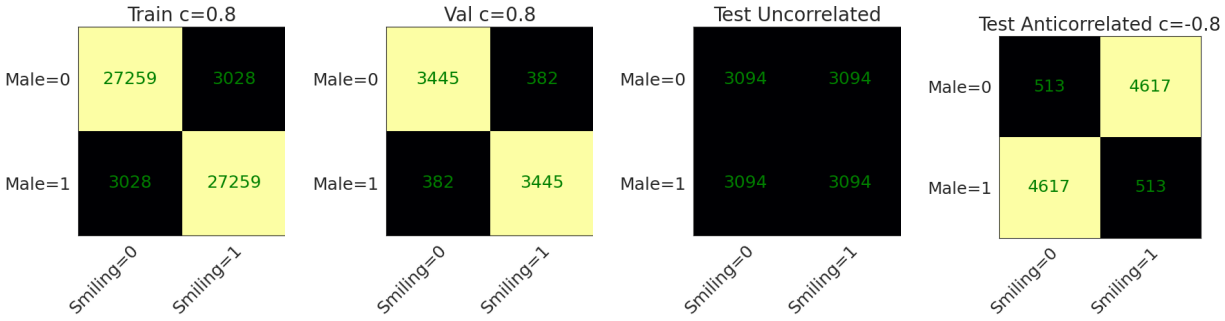
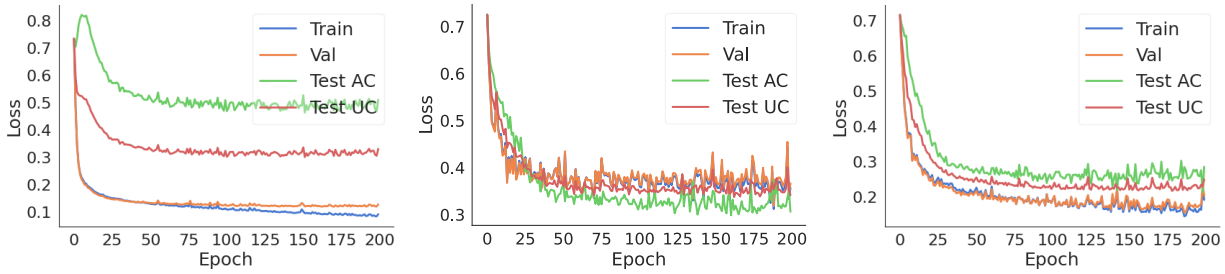
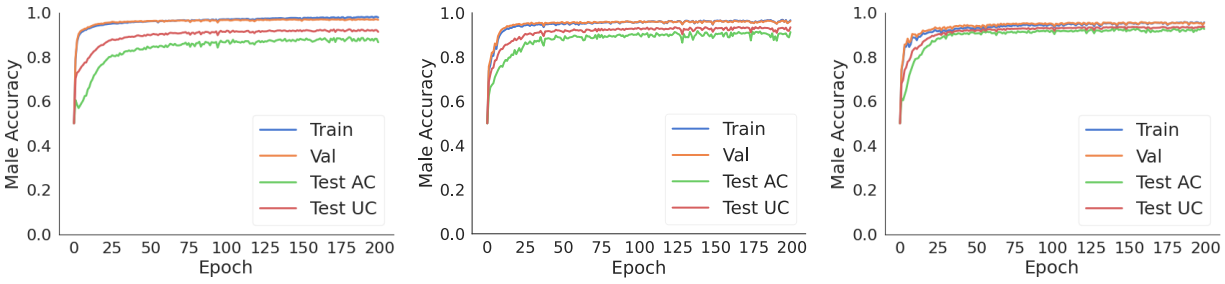


Figure 14: Numbers of examples in the subsampled CelebA datasets for the strongest correlation we consider, $c = 0.8$.



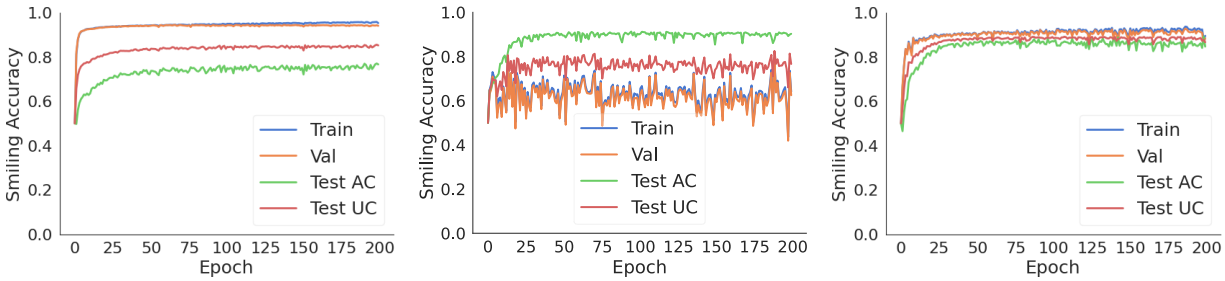
(a) Base (b) Base + MI (c) Base + CMI

Figure 15: Loss curves for each approach on the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



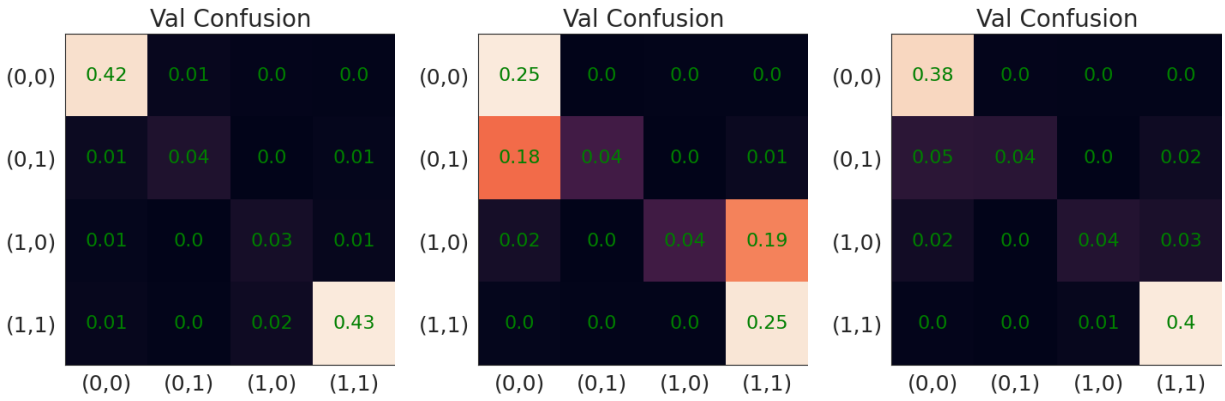
(a) Base (b) Base + MI (c) Base + CMI

Figure 16: Accuracies on the attribute `Male` for each approach on the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



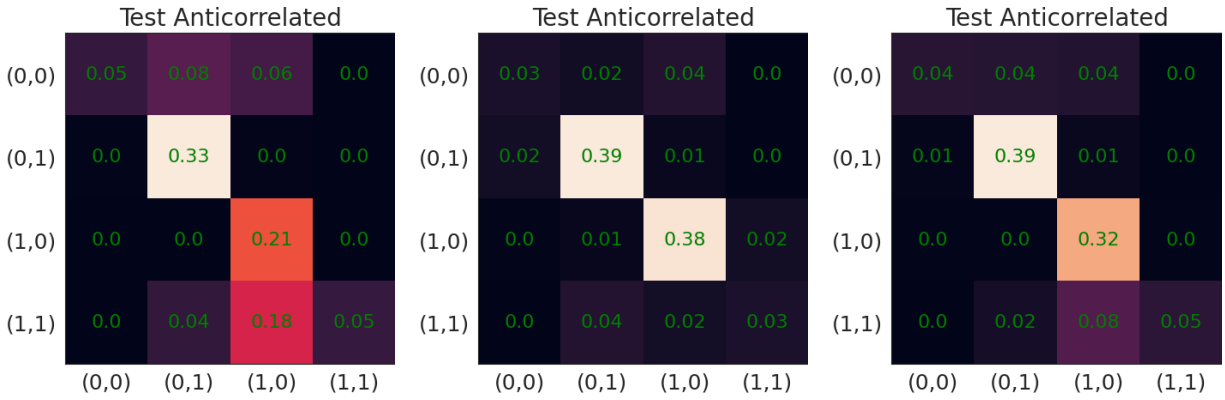
(a) Base (b) Base + MI (c) Baes + CMI

Figure 17: Accuracies on the attribute `Smiling` for each approach on the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



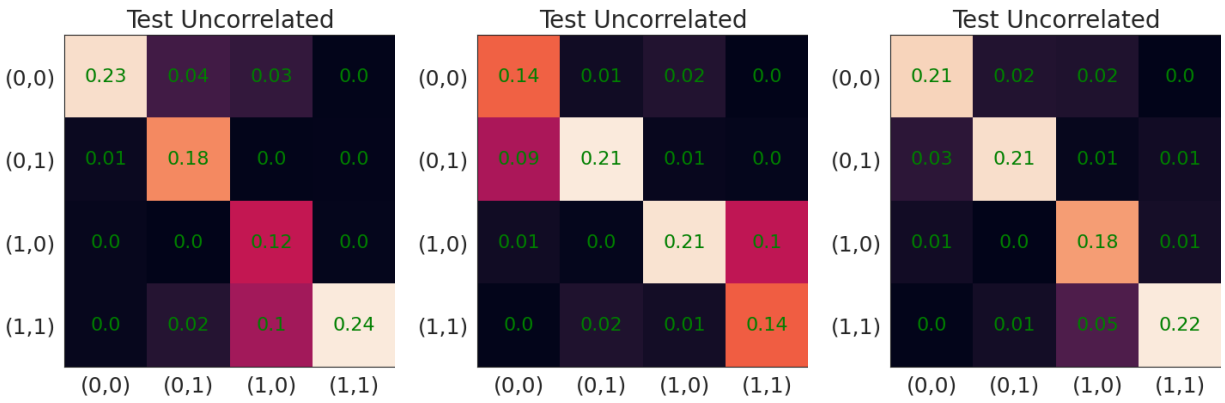
(a) Base (b) Base + MI (c) Base + CMI

Figure 18: Confusion matrices for each approach on the correlated validation set of the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



(a) Base (b) Base + MI (c) Base + CMI

Figure 19: Confusion matrices for each approach on the anti-correlated test set of the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



(a) Base (b) Base + MI (c) Base + CMI

Figure 20: Confusion matrices for each approach on the uncorrelated test set of the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.

	Common Combinations		**Rare Combinations**	
	Female + Non-Smiling	Male + Smiling	Female + Smiling	Male + Non-Smiling
Base	**4%**	**4%**	29%	51%
Base + MI	23%	28%	**12%**	31%
Base + CMI	10%	9%	20%	**29%**

Table 3: Percentage of incorrect predictions per subgroup for CelebA, evaluated on natural data (e.g., data with naturally-occurring correlations, that has not been subsampled to induce a specific correlation strength), using models trained on correlated data with $c = 0.8$.

	Common Combinations		**Rare Combinations**	
	Female + Non-Smiling	Male + Smiling	Female + Smiling	Male + Non-Smiling
Base	**4%**	**5%**	33%	49%
Base + MI	24%	28%	**11%**	26%
Base + CMI	9%	9%	19%	**25%**

Table 4: Percentage of incorrect predictions per subgroup for CelebA, evaluated on validation data ($c = 0.8$), using models trained on correlated data with $c = 0.8$.

B.4 DISENTANGLEMENT METRICS

We evaluated common disentanglement metrics (Locatello et al., 2019b) on uncorrelated test data using models trained on correlated data. We performed this analysis for two of our datasets and found in both cases that *Base+CMI* reached better scores compared to the other objectives for almost all metrics.

Toy Classification: Disentanglement results for the toy classification task with ten attributes are shown in Table 5. We obtained similar results for two and four attributes, which are not reported for brevity.

CelebA: Since the disentanglement metrics require that the factors of variation are each encoded in one-dimensional subspaces, we set latent dimension $D = 2$ for this experiment. In Table 6, we report the average and 68% confidence intervals for five models trained on data with correlation level 0.8.

Metric	Base	Base+MI	Base+CMI
IRS (Suter et al., 2019) ↑	0.377	0.573	**0.605**
SAP (Kumar et al., 2017) ↑	0.118	0.470	**0.477**
MIG (Chen et al., 2018) ↑	0.179	0.939	**0.975**
DCI Disentanglement (Eastwood & Williams, 2018) ↑	0.413	0.980	**0.998**
Beta-VAE (Higgins et al., 2017a) ↑	0.996	1	1
Factor-VAE (Kim & Mnih, 2018) ↑	1	1	1
Gaussian Total Correlation ↓	10.073	0.485	**0.025**
Gaussian Wasserstein Corr ↓	12.905	0.373	**0.027**
Gaussian Wasserstein Corr Norm ↓	0.866	0.037	**0.002**
Mutual Info Score ↓	0.975	0.197	**0.149**

Table 5: **Disentanglement metrics for toy classification with ten attributes.** Metrics are evaluated on the uncorrelated test set. Bold font indicates model with best disentanglement score.

Metric	Base	Base+MI	Base+CMI
IRS ↑	0.524 ± 0.043	$\mathbf{0.548 \pm 0.038}$	0.531 ± 0.041
SAP ↑	0.306 ± 0.003	0.296 ± 0.046	$\mathbf{0.389 \pm 0.005}$
MIG ↑	0.506 ± 0.01	0.455 ± 0.074	$\mathbf{0.674 \pm 0.007}$
DCI Disentanglement ↑	0.46 ± 0.009	0.596 ± 0.038	$\mathbf{0.807 \pm 0.023}$
Beta-VAE ↑	1.0 ± 0.0	1.0 ± 0.0	$\mathbf{1.0 \pm 0.0}$
Factor-VAE ↑	1.0 ± 0.0	0.999 ± 0.003	$\mathbf{1.0 \pm 0.0}$
Gaussian Total Correlation ↓	0.222 ± 0.012	0.056 ± 0.061	$\mathbf{0.011 \pm 0.003}$
Gaussian Wasserstein Corr ↓	0.351 ± 0.039	0.01 ± 0.009	$\mathbf{0.002 \pm 0.001}$
Gaussian Wasserstein Corr Norm ↓	0.098 ± 0.005	0.006 ± 0.004	$\mathbf{0.005 \pm 0.001}$
Mutual Info Score ↓	0.302 ± 0.022	0.111 ± 0.052	$\mathbf{0.042 \pm 0.006}$

Table 6: **Disentanglement metrics for CelebA.** Metrics are evaluated on the uncorrelated test set. Bold font indicates model with best disentanglement score.

B.5 WEAKLY SUPERVISED SETTING

For the fully supervised CelebA experiment, labels for both attributes were available for all 10260 images. For the weakly supervised setting, we reduced the number of labels to 5130 (50% of the labels of the fully supervised dataset), 2565 (25%), 1026 (10%), or 513 (5%) for each attribute. This implies that some images had both labels, some had only one label and some images had no labels (for example when using 50% of the labels the distinction is as follows: 25% of the images had both labels; 25% had only labels for attribute 1; 25% had only labels for attribute 2; and 25% had no labels). The three objectives can be applied to these weakly supervised settings. For *Base*, the cross-entropy loss for each attribute was computed only for the images that had labels for the corresponding attribute. For *Base+MI* no labels are required for the unconditional shuffling; thus this objective can be applied even for the images without labels. For *Base+CMI*, our method shuffles only images that have the same value for a given attribute. This also works if the labels of the other attribute are missing. We used the same training parameters as for the supervised experiment, except for increasing the number of training epochs (up to 1200 epochs) and adapting the minibatch size to the number of labels. In Figure 8 we report the average and 68% confidence intervals over three runs with different seeds.

C ALGORITHMS

In this section, we provide formal descriptions of the baseline approaches we use. Algorithm 2 describes the classification-only baseline, that trains separate linear classifiers to predict attributes $\mathbf{s}_k$ from the corresponding latent subspaces $\mathbf{z}_k$. Algorithm 3 and Algorithm 4 describe the unconditional disentanglement baseline, that adversarially minimizes the discrepancy between samples from the joint distribution $p(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ and the product of marginals $p(\mathbf{z}_1) \cdots p(\mathbf{z}_k)$. Algorithm 5 describes the discriminator training loop for the CMI minimization approach from Section 4.

Algorithm 2 Supervised Learning on Subspaces

Input: $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$, initial parameters for K linear classifiers $R_1, \ldots, R_K$
Input: $\boldsymbol{\theta}$, initial parameters for the encoder f
Input: α, β learning rates for training the encoder and linear classifiers
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into k subspaces
  $L \leftarrow \sum_{k=1}^K L_{\text{cls}}(R_k(\mathbf{z}_k; \boldsymbol{\phi}_k), \mathbf{s}_k)$ ▷ Cross-entropy for each attribute
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L$ ▷ Update encoder parameters
  $\boldsymbol{\phi}_k \leftarrow \boldsymbol{\phi}_k - \beta \nabla_{\boldsymbol{\phi}_k} L \quad , \quad \forall k \in \{1, \ldots, K\}$ ▷ Update classifier parameters
end while

Algorithm 3 Learning Unconditionally Disentangled Subspaces — Training the Encoder

Input: $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$, initial parameters for K linear classifiers $R_1, \ldots, R_K$
Input: $\boldsymbol{\theta}$, initial parameters for the encoder f
Input: α, β learning rates for training the encoder and linear classifiers
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into k subspaces
  $L \leftarrow \sum_{k=1}^K L_{\text{cls}}(R_k(\mathbf{z}_k; \boldsymbol{\phi}_k), \mathbf{s}_k)$ ▷ Cross-entropy for each attribute
  $\mathbf{z}' \sim p(\mathbf{z}_1)p(\mathbf{z}_2) \cdots p(\mathbf{z}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
  $L \leftarrow L + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z}')\right) + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z})\right)$ ▷ Add adversarial loss
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L$ ▷ Update encoder parameters
  $\boldsymbol{\phi}_k \leftarrow \boldsymbol{\phi}_k - \beta \nabla_{\boldsymbol{\phi}_k} L \quad , \quad \forall k \in \{1, \ldots, K\}$ ▷ Update classifier parameters
end while

Algorithm 4 Learning Unconditionally Disentangled Subspaces — Training the Discriminator

Input: $\boldsymbol{\omega}$, initial parameters for the discriminator D
Input: γ, learning rate for training the discriminator
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into k subspaces
  $\mathbf{z}' \sim p(\mathbf{z}_1)p(\mathbf{z}_2) \cdots p(\mathbf{z}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
  $L \leftarrow L + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z}')\right) + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z})\right)$ ▷ Add adversarial loss
  $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \gamma \nabla_{\boldsymbol{\omega}} L$ ▷ Update discriminator parameters
end while

Algorithm 5 Learning Conditionally Disentangled Subspaces Adversarially – Training the Discriminator

Input: $\boldsymbol{\omega}$, initial parameters for the discriminator D
Input: γ, learning rate for training the discriminator
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into K subspaces
  $L \leftarrow 0$ ▷ L will accumulate the losses over all subspaces
  for $k \in \{1, \ldots, K\}$ **do**
    $\mathbf{z}' \sim p(\mathbf{z}_1, \ldots \mathbf{z}_K \mid \mathbf{s}_k)$ ▷ Samples from the joint distribution
    $\mathbf{z}'' \sim p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
    $L \leftarrow L + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z}'')\right) + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z}')\right)$ ▷ Add adversarial loss
  end for
  $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \gamma \nabla_{\boldsymbol{\omega}} L$ ▷ Update discriminator parameters
end while

D PROOF OF PROPOSITION 3.1

Proposition 3.1 *If $I(s_1; s_2) > 0$, then enforcing $I(z_1; z_2) = 0$ means that $I(z_k; s_k) < H(s_k)$ for at least one k.*

Proof. Assume that $I(s_1; s_2) > 0$ and at the same time $I(z_k; s_k) = H(s_k)$ (i.e., we are proving by contradiction). Since $I(z_1; s_1) = H(s_1)$, we have $H(s_1 \mid z_1) = 0$ and with $H(s_1 \mid z_1) = H(s_1 \mid z_1, s_2) + I(s_1; s_2 \mid z_1)$ (both non-negative), it follows that $H(s_1 \mid z_1, s_2) = I(s_1; s_2 \mid z_1) = 0$. Since for the interaction information, by definition $I(s_1; s_2; z_1) = I(s_1; s_2) - I(s_1; s_2 \mid z_1)$, and $I(s_1; s_2 \mid z_1) = 0$, we have $I(s_1; s_2; z_1) = I(s_1; s_2) > 0$. Since we also assume $H(s_2 \mid z_2) = 0$, we also have $I(s_1; s_2; z_2) = I(s_1; s_2) > 0$.

We can use this to compute the fourth order interaction information $I(s_1; s_2; z_1; z_2)$. By definition, we have $I(s_1; s_2; z_1; z_2) = I(s_1; s_2; z_1) - I(s_1; s_2; z_1 \mid z_2)$. We just showed that $I(s_1; s_2; z_1) = I(s_1; s_2)$, and therefore we have $I(s_1; s_2; z_1 \mid z_2) = I(s_1; s_2 \mid z_2)$. Together it follows that:

$$
\begin{aligned}
I(s_1; s_2; z_1; z_2) &= I(s_1; s_2; z_1) - I(s_1; s_2; z_1 \mid z_2) && (7)\\
&= I(s_1; s_2) - I(s_1; s_2 \mid z_2) && (8)\\
&= I(s_1; s_2; z_2) && (9)\\
&= I(s_1; s_2) > 0 && (10)
\end{aligned}
$$

On the other hand, we know that $0 = H(s_1 \mid z_1) = H(s_1 \mid z_1; z_2) + I(s_1, z_2 \mid z_1)$ and therefore $I(s_1, z_2 \mid z_1) = 0$. Therefore, the interaction information $I(s_1; z_2; z_1) = I(s_1; z_2) - I(s_1; z_2 \mid z_1) = I(s_1; z_2) \geq 0$. At the same time, we assumed that $I(z_1; z_2) = 0$ and hence $I(z_1; z_2; s_1) + I(z_1; z_2 \mid s_1) = 0$, which shows that $I(z_1; z_2; s_1) \leq 0$. Together, we see that $I(z_1; z_2; s_1) = I(s_1; z_2) = 0$.

Now we can decompose $I(s_1; s_2; z_1; z_2)$ in a different way: $I(s_1; s_2; z_1; z_2) = I(s_1; z_1; z_2) - I(s_1; z_1; z_2 \mid s_2)$. We know that $I(s_1; z_1; z_2) = I(s_1; z_2)$ and therefore $I(s_1; z_1; z_2 \mid s_2) = I(s_1; z_2 \mid s_2) > 0$ and that $I(s_1; z_1; z_2) = 0$. Therefore, it follows that:

$$
\begin{aligned}
I(s_1; s_2; z_1; z_2) &= I(s_1; z_1; z_2) - I(s_1; z_1; z_2 \mid s_2) && (11)\\
&= 0 - I(s_1; z_2 \mid s_2) && (12)\\
&\leq 0 && (13)
\end{aligned}
$$

which is a contradiction with $I(s_1; s_2; z_1; z_2) = I(s_1; s_2) > 0$. Therefore, if $I(s_1; s_2) > 0$ and $I(z_1; z_2) = 0$, it must hold that $I(z_k; s_k) < H(s_k)$ for at least one k, which we wanted to show. □