How well can we predict where people look in images?

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls-Universität Tübingen

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

vorgelegt
von

Matthias Kümmerer
aus Tübingen

Juli 2019

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

Tag der mündlichen Prüfung:	13. 03. 2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Matthias Bethge
2. Berichterstatter:	Prof. Felix A. Wichmann, Dphil.
3. Berichterstatter:	Prof. Dr. Peter König

Ich erkläre, dass ich die zur Promotion eingerichte Arbeit mit dem Titel
"How well can we predict where people look in images?"
selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. 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/Date Unterschrift/Signature

Acknowledgments

This thesis is the result of several years of work. I had the luck to spend these years in a great environment that made my work possible, fun and successful. Many people contributed to creating this motivating atmosphere and I want to thank them here.

First and foremost I want to thank my supervisor Matthias Bethge. He was the one who got me interested in machine learning, vision science and neuroscience in the first place and made sure the fascination grew ever larger over the course of time. I am really thankful for all the support he provided, both in terms of logistics and in shaping the scientific direction of my work. While he was able and willing to weigh in with his scientific expertise, he also was willing to let me work independently and follow my intuitions. I am very grateful for this trust.

Tom Wallis joined our lab shortly after me as a Postdoc and over the years he has become an invaluable consultant and dear colleague. Besides giving great scientific advice, he was the one who showed me how my sometimes quite theoretical and abstract ideas could be transformed into understandable and intuitive stories. If there is something to like about the style of writing in this thesis, chances are, it is due to Tom's teaching.

Doing a PhD is not only about doing research but also a lot about filling the right form at the right time and finding the way through the organizational jungle that is university. Without Heike König's everlasting help, I certainly would have taken one detour or the other.

I also want to thank all members of the Bethgelab. Everybody in the lab was always happy to discuss science, be it in the lab or over drinks in town. This is especially the case for my friend Alex Böttcher, who joined the lab the very same day that I did and who was always up for discussions about science, the universe and everything.

Finally I want to thank my parents for raising me in a way that showed me that so many things are worth being curious about and my girlfriend Caroline Krauter for all her love and for making my life even better.

Summary

Understanding where people look in images is an interesting quest, both for the scientific implications in terms of visual and cognitive processing and behaviour as well as due to its potential applications such as smart image cropping. The field has a long history of computational modeling, resulting in a substantial number of so-called "saliency models" trying to predict where people look in images.

Here, two aspects of the problem are being considered: the one of benchmarking and comparing models and the one of building better models. The previous state of model benchmarking in the field was somewhat problematic: researchers used many different metrics to asses the quality of so-called saliency maps and depending on the chosen metric, the same model could be state-of-the-art or close to chance. This inconsistency was generally attributed to the metrics measuring substantially different things. Depending on the intended application, one would need to decide for the metric. This made it very hard to asses progress and state-of-the-art in the field. Here, we show that the underlying cause for the disagreement between saliency metrics is actually that they interpret saliency maps in highly different ways. By formulating saliency models as probabilistic models of fixation density prediction and optimizing them using suitable loss functions like log-likelihood, one can encode the model predictions into different metric-specific saliency maps that account for how the metric interprets the saliency maps. Doing this results in highly consistent metric scores and ranks and mostly solves the benchmarking problem in saliency, allowing for a clearer picture of state-of-the-art and what's still missing.

Besides benchmarking, this thesis focuses on building better models of fixation prediction and on understanding which features are relevant for predicting fixations well. We introduce transfer learning from deep convolutional features to the field of saliency modeling to create saliency models that utilize recent advances in the field of deep learning. With our saliency models "DeepGaze I" and "DeepGaze II", we were able to increase the percentage of explained information on the MIT1003 dataset from previously 34% first to 46.1% and subsequently to 80.3%. This sets a new state-of-the-art in the MIT Saliency Benchmark and shows the importance of high-level features for fixation prediction.

The model architecture of DeepGaze II allows for a principled comparison of the predictive power of different features for fixation locations. We show that while complex deep features are crucial to reach high performance, even very simple intensity-contrast features still can perform better than all previous models that don't use transfer-learning.

Zusammenfassung

Welche Bereiche von Bildern Menschen anschauen ist eine interessante Fragestellung. Sie hat wissenschaftliche Implikationen im Bereich der visuellen und kognitiven Verarbeitung und des Verhaltens, ermöglicht aber auch Anwendungen wie zum Beispiel das intelligente Zuschneiden von Bildern. Es ist bereits seit langer Zeit üblich, sogenannte "Salienz-Modelle" zu konstruieren, die versuchen vorherzusagen, wo Menschen in Bildern hinschauen.

Die vorliegende Arbeit beschäftigt sich mit zwei verschiedenen Aspekten des Problems: zum einen mit dem Aspekts des Bechmarkings und Vergleichs von Modellen und zum anderen mit dem Aspekt der Konstruktion besserer Modelle. Der bisherige Stand des Benchmarkings im Feld der Salienzforschung war teilweise problematisch: Forscher nutzten viele verschiedene Metriken um die Qualität der Modellvorhersagen in Form sogenannter "Salienzkarten" zu bewerten. Je nach verwendeter Metrik kann das selbe Modell oft sowohl als State-of-the-art als auch als nahe an Chance erscheinen. Diese Inkonsistenz wurde üblicherweise mit der Annahme erklärt, dass die Metriken grundverschiedene Dinge messen. Dementsprechend müsste man sich je nach geplanter Anwendung eines Modelles für die eine oder andere Metrik entscheiden und schlechtes Abschneiden in anderen Metriken akzeptieren. Dadurch wurde es aber sehr schwer, Fortschritt und Stand der Forschung zu bewerten. Hier zeigen wir, dass der eigentliche Grund für die inkonsistenten Modellbewertungen darin liegt, dass die Metriken Salienzkarten auf sehr verschiedene Art und Weise interpretieren. Wenn man Salienzmodelle als probabilistische Modelle formuliert, die Fixationsdichten für Bilder vorhersagen, und diese Modelle mit geeigneten Kostenfunktionen wie zum Beispiel log-likelihood trainiert, ist es möglich, die Modellvorhersagen in verschiedenen metrik-spezifischen Salienzkarten zu kodieren. Dadurch können die Salienzkarten berücksichtigen, wie einzelne Metriken die Salienzkarten interpretieren. Im Ergebnis erhält man Modellbewertungen, die über verschiedene Metriken hinweg sehr konsistent sind. Dies löst das Problem des Benchmarkings im Bereich der Salienzmodelle weitgehend und erlaubt einen klareren Blick darauf, wie gut Modelle bereits sind und wieviel besser sie noch werden könnten.

Neben dem Benchmarking beschäftigt sich diese Arbeit auch damit, bessere Modelle zur Vorhersage von Fixationen zu finden und zu verstehen, welche Features relevant sind, um Fixationen gut vorherzusagen. Dazu führen wir das Konzept des Transfer-Learnings mit Features von tiefen neuronalen Netzen in das Feld der Salienzmodellierung ein. Dies erlaubt uns Modelle zu konstruieren, die die Fortschritte im Bereich des tiefen Lernens ausnutzen können, die in den letzten Jahren erziehlt wurden. Mit unseren Salienzmodellen "DeepGaze I" und "DeepGaze

II" konnten wir den Anteil der erklärten Information im MIT1003-Datensatz von vorher 34% erst auf 46.1% und schließlich auf 80.3% erhöhen. Dies setzt einen neuen State-of-the-art im MIT Saliency Benchmark und zeigt, wie wichtig high-level-features für die Vorhersage von Fixationen sind.

Die Architektur von DeepGaze II erlaubt es, verschiedene Features daraufhin zu vergleichen, wie gut sie sich zur Vorhersage von fixierten Orten in Bildern eignen. Wir zeigen, dass komplexe tiefe Features zwar notwendig sind um hohe Performanz zu erreichen, gleichzeitig aber sogar einfachste Intensitäts- und -Kontrast-Features zu höherer Performanz führen können als sie von allen vorigen Modellen erreicht wurde, die noch nicht Transfer-Learning verwendet haben.

Contents

hic, ne deficeret, metuens avidusque videndi
flexit amans oculos; et protinus illa relapsa est,
bracchiaque intendens prendique et prendere certans
nil nisi cedentes infelix arripit auras.

To dowt him lest shee followed not, and through an eager love
Desyrous for to see her he his eyes did backward move.
Immediatly shee slipped backe. He retching out his hands,
Desyrous to bee caught and for to ketch her grasping stands.

Jetzo besorgt, sie bleibe zurück, und begierig des Anschauns,
Wandt' er die Augen voll Lieb'; und sogleich war jene versunken.
Streckend die Arm', und ringend, gefaßt zu sein und zu fassen,
Haschte der Unglückselige nichts, als weichende Lüfte.

Ovid's Metamorphes, Book X, Orpheus et Eurydice
quoted after Magnus 1892 (Latin), Golding 1567 (English) and Voss 1798 (German)

1 Introduction

1.1 Vision and Eye Movements

From all known senses, the visual sense stands out: it is the only sense that allows one to locate objects very precisely from a distance (Palmer 1999). It is easy to imagine that having this possibility gives a substantial evolutionary advantage and indeed most types of eyes that we know today evolved within only 5 million years, during the Cambrian explosion around 530 million years ago. This development was paralleled by increasing body size and speed and together enabled visually-guided predation (Land and Nilsson 2012, p. 21). Additional support for the evolutionary benefit of vision is given by the fact that while only six of the more than 30 animal phyla have developed eyes with some acuity, these contribute about 96% of the known species (Fernald 2008). It is estimated that 40% of human cortex receive at least some visual input, underlining how crucial vision is for humans (Felleman and Van Essen 1991).

The retinas of many species don't sample the full field of view at a constant resolution, but instead have some part where the receptor density is higher than in other parts. Most likely, this is due to an information bottleneck problem: the amount of information that the eye can transmit to the brain is limited by the capacity of the visual nerve. Humans have around one million retinal ganglion cells which are estimated to transmit roughly 10 bit/s, putting the capacity of the human optical nerve at around 10Mbit/s (Curcio and Allen 1990; Koch et al. 2006). It can be useful to use most of this capacity for the part of the field of view that is behaviourally most relevant. For example the retina of rabbits has higher resolution in those parts that perceive the horizon, while the parts directed towards the ground and towards the sky have much lower resolution (Land and Nilsson 2012, Figure 5.13).

Instead of adapting the density of the retina to the average relevance of a corresponding gaze direction in the visual environment such as sky, horizon or ground, other animals, including primates, evolved a different solution to the bottleneck problem: their retinas have a very small area central in the field of view (the *fovea*) with much higher resolution than the peripheral parts of the retina (receptor density is up to a factor of 100 higher than in the periphery, Curcio et al. 1990). This allows the organism to perceive the central part of the field of view in very high detail while maintaining a broad gist of a much larger visual angle. The information of the peripheral part can be used to direct the fovea sequentially at whatever is considered most relevant at that moment. This can either be done by moving the head (the dominant case, e.g., for owls; Steinbach and Money 1973) or additionally by moving the eyes inside the head (e.g., humans). Interestingly, these changes of gaze direction usually don't happen in a smooth way. Instead, the fovea is fixated on some point

in the visual field for some time and then changed rapidly to a different point in a so-called *saccade*. The reason for this is most likely that it takes some time to process the visual input: it takes the photoreceptors around 10ms to fully adapt to the visual input, (Land and Nilsson 2012, p. 218) and moving the gaze during this time would induce motion blur. Besides fixations and saccade there are some other types of eye movements, which include smooth pursuit where the eyes track a moving target and vergence movements where the eyes move in opposite direction to adapt to closer or farther visual targets.

Understanding how gaze is directed is of central importance for understanding visually-guided behaviour in a large number of organisms. Eye movements are a form of visual attention that is, in contrast to covert attention, comparatively easy to measure, and usually closely tied to the latter in natural behaviour (Henderson 2003). Therefore, eye movements can be used to better understand cognitive processes. Especially, eye movements allow to observe decision making in natural behaviour, e.g., the strategies employed to solve a task and what is considered relevant for solving that task. Examples of such studies have explored how people make sandwiches (Hayhoe 2000) and tea (Land et al. 1999). Besides attention and decision making, there are more reasons to be interested in eye movements. The efficient coding hypothesis (Barlow 2012) suggests that information processing is adapted to statistics of the input. For the visual system this means that in order to understand the visual system, one needs to understand the statistics of natural images (Hyvärinen et al. 2009). However, the statistics of the data processed by the visual system might be different from the statistics of natural images due to selective sampling (Reinagel and Zador 1999). Finally, there are many applications where understanding eye movements can help, e.g., when optimizing warning signs or advertisements, cropping images, scene understanding, or in robotics.

1.2 Eye Movement Research

Eye movement research is a large and diverse field that has produced a substantial body of literature, not least due to the many implications of eye movements, some of which have been detailed above. The field of eye movement research has a long history (see Wade (2010) for an extensive overview). As early as in the fourth century BC Aristotle observed that the movement of both eyes is tightly linked and that only certain movement combinations are possible (Aristotle et al. 1908–1952, pp. 957b-960a, Problemata book XXXI "Problems connected with the eyes"). In the second century AD, Ptolemy did experiments on binocular vision, while the Greek physician and philosopher Galen dissected rhesus monkeys to understand the muscles of the oculomotor system (Galen et al. 1956) Perhaps unsurprisingly,

Leonardo da Vinci, too, tried to understand anatomy and movement of the eyes (MacCurdy 1938, p. 186). The fact that eye movements are usually not smooth but discontinuous was already observed by Porterfield (1737) and described in more detail by Brown (1878). The term "saccade" for the fast eye movements between phases of very little movement was first used by Émile Javal in 1879 in his research on reading and is the french term for "jerk".

The quest of understanding eye movements in reading also spurred most progress in the recording of eye movements. Javal tried unsuccessfully to attach a pointer to the eye that was supposed to record eye movements on a kymograph. In 1879, both Lamare and Hering attached a tube to the eye lid that generated sounds when the eye moved (Wade and Tatler 2009). Finally, at the beginning of the 20th century, Dodge developed the photographic eye tracker that didn't need any attachment to the eye and allowed much more natural viewing behaviour.

This progress in recording eye movements is what eventually allowed the exploration of not only how we move our eyes but also where we actually look. For inferring the gaze position relative to a stimulus from eye position, it is necessary to precisely measure the eye position and transform this position into image coordinates. The question where we look first gained interest in the context of reading and was then extended to picture viewing by George Malcom Stratton. He recorded eye movements while subjects viewed simple geometric patterns and noticed that symmetric patterns do not result in symmetric eye movements (Stratton 1906). The question how fixation locations are selected gained substantial momentum in 1935 with Guy Buswell's foundational work in *How people look at pictures* (Buswell 1935). He already looked into many effects that still interest the field to the present day, including where people look spatially, how the spatial fixation distribution changes over presentation time, fixation durations, inter-observer consistency and the influence of instructions given to the subjects. The question about the influence of tasks on eye movements was made famous by Yarbus, who explored this effect in great depth (Yarbus 1967).

1.3 Saliency and the Effect of Visual Features on Eye Movements

The locations of fixations are influenced by many factors, including oculomotor biases and tasks. One factor gained particular interest over time: the influence of the observed visual stimulus. Already Buswell noticed that people attract fixations (Buswell 1935) and Yarbus showed that depending on the task, different areas of an image are fixated, while confirming the overall tendency towards looking at persons (Yarbus 1967). As opposed to those semantic image features, people later investigated the low-level properties of fixated image locations. For example, Mannan found that

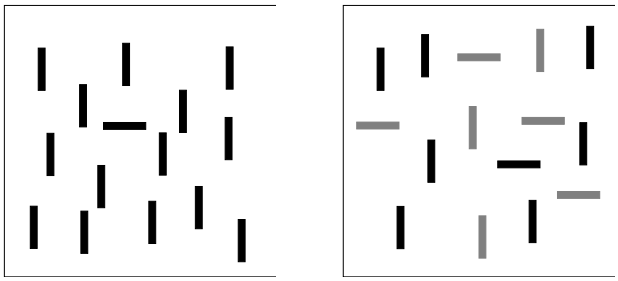


Figure 1: In visual search experiments, there are cases where the targets seem to pop out among the distractors (left: horizontal bar among vertical bars) while in other cases the target doesn't immediately pop-out and requires serial search (right: horizontal black bar among horizontal and vertical black and gray bars). These differences motivated the Feature Integration Theory.

fixated locations typically have higher spatial contrast (Mannan et al. 1996; Mannan et al. 1997) and other researchers explored similar statistical properties of fixation locations (Reinagel and Zador 1999; Krieger et al. 2000; Parkhurst and Niebur 2003).

The term *saliency* is often used in the context of the interaction between image features and fixation placement. The notion of saliency originated in the field of attention and visual search, and subsequently heavily influenced the field of fixation modeling as will be outlined in the following.

When visually searching for targets among distractors, the search durations behave qualitative differently for different combinations of targets and distractors. There are certain combinations of targets and distractors where the average search duration doesn't seem to depend on the number of distractors. For other combinations the search takes the longer, the more distractors are present. The former kinds of targets seem to pop out immediately (e.g., a horizontal bar among vertical bars, Figure 1 left) while the latter kinds of targets require a serial search (e.g., a black horizontal bar among many horizontal or vertical gray or black bars, Figure 1 right). Treisman and Galade proposed an explanation for this effect in their *Feature Integration Theory* (Treisman and Gelade 1980). They hypothesized a two stage attentional system, where a first stage computes feature maps, registering elementary features highly parallel over the full field of view while objects being defined as a conjunction of multiple features can only be detected with focused attention and therefore require serial shifts of this focused attention over candidate locations.

In 1985, Koch and Ullmann suggested a computational mechanism of how the feature integration theory could be implemented. They proposed that for many elementary features like color and orientation of motion, *conspicuity* maps are computed. These conspicuity maps single out locations that differ significantly

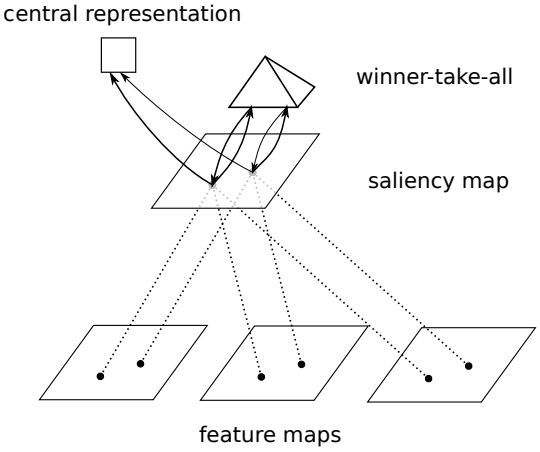


Figure 2: The model of Koch and Ullman proposes to combine multiple feature maps into a single saliency map where a winner-take-all mechanism selects the most salient location which is then fed to a central representation for further processing. Figure reproduced and adapted from Koch and Ullman (1985) with permission from Springer-Nature.

from their surrounding locations with respect to that feature and are subsequently combined into one global conspicuity map, which they termed a *saliency map*. How exactly different elementary features contribute to the saliency map is not specified and might be influenced by the current task. They propose that a winner-take-all mechanism selects the most conspicuous location in the saliency map. After a location has been selected, the visual information of that location is copied into a central representation where, e.g., it can be checked whether it is the actual search target. After the location has been processed, the saliency map is suppressed at that location, leading to a "shift of attention". Koch and Ullman proposed a neural algorithm for the winner-take-all mechanism that is consistent with experimental findings such as the one that attentional shifts over a larger spatial distance take longer time.

After Koch and Ullman outlined the described computational mechanism for the feature integration theory, it was implemented by Itti et al. (1998). The model extracts color, intensity and edge filter responses at different scales, takes center-surround differences and normalizes to detect feature pop-out. The results over different features and scales are then combined linearly to yield the final saliency map. The model of Itti et al. is an image-computable model, i.e., a model that can not only process artificial stimuli for which the elementary features were known *a priori* (such as Didday and Arbib 1975), but any arbitrary image. This allowed a wide range of applications and is one of the reasons for the large impact this model had on the

field. Itti and Koch (2000) evaluated the model on search durations when searching for military vehicles in natural images. The feature integration theory and the model of Itti et al. were proposed to understand covert shifts of attention and were not intended to predict eye movements. However, the assumption that eye movements are closely related to covert attention (e.g., Henderson 2003) motivated to compare the model predictions to eye movements. This comparison was suggested by Itti and Koch (2001a) and Itti and Koch (2001b) and eventually done by Parkhurst et al. (2002): they recorded fixations on natural images, computed saliency maps with the model of Itti and Koch for those images and compared the saliency of fixated locations with the average saliency of the image. They found that fixated locations have above-average model saliency and that earlier fixations have higher model saliency than later fixations (although the latter finding was shown to be an artifact by Tatler et al. 2005).

The model of Itti and Koch together with its evaluation on fixations by Parkhurst et al. established saliency maps as the standard framework for modeling the effects of image features on spatial fixation placement. Until the present day, most models of the spatial fixation distribution compute a saliency map that is then evaluated on fixations. What changes from model to model is how the saliency map was proposed to be computed from the image input. Since the model of Itti and Koch, an ever growing zoo of models propose computational mechanisms for computing saliency maps. With the influential MIT saliency benchmark listing around 90 models and new models being published monthly it is not easy to keep track of the field. Here we give a brief overview, focusing on the developments and concepts which are most important for this thesis. For a more extensive overview of existing models see, e.g., Borji and Itti (2013) and Borji et al. (2013a).

The oldest category of saliency models could be called "classic saliency models". In the tradition of the original model of Itti and Koch, these models are usually inspired by the classic idea of saliency as feature pop-out. They use hand-crafted low-level features for computing saliency maps and either operate purely locally (Itti et al. 1998; Zhang et al. 2008), or more often add some global features or statistics (Torralba et al. 2006; Harel et al. 2007; Hou and Zhang 2007; Bruce and Tsotsos 2009; Goferman et al. 2012; Erdem and Erdem 2013; Riche et al. 2013a). Also some simple heuristics turned out to work surprisingly well: thresholding color channels and selecting components that are not connected to the image border (Zhang and Sclaroff 2013) gave rise to the best model on the MIT Saliency Benchmark until late 2014.

Besides handcrafted models an increasing number of models apply machine learning to learn the relevant features from fixation data. Likely the first example of this class of models was proposed by Itti and Koch (2001b) via extending the model

of Itti and Koch with learned linear weights. Baddeley and Tatler (2006) modeled fixation density as a generalized linear model (GLM) with different feature maps such as high and low-frequency edges and contrast and trained the model with the maximum a posteriori estimator. Shortly after, Kienzle et al. (2007) trained an kernel support vector machine (SVM) with radial basis functions (RBF) to distinguish fixated locations from nonfixated locations. Subsequently they showed that the model can massively be simplified to just four basis units with center-surround features (Kienzle et al. 2009). Peters and Itti (2007) used a least squares regression to predict low resolution binary fixation maps and Zhao and Koch (2011) learned linear weights for the features of the original model of Itti and Koch with an additional face detector by minimizing the l2 distance to empirical saliency maps (i.e., blurred fixation maps). Judd et al. (2009) trained an SVM that combines the saliency model of Torralba et al. (2006) and several preexisting object detectors for, e.g., faces, people, cars and the horizon. Finally, deep learning reached the field of saliency modeling when Vig et al. (2014) trained three layer neural networks to distinguish fixated locations from non-fixated locations. In our model DeepGaze I described in Section 2.2 of this thesis (Kümmerer et al. 2015a) we used transfer learning for fixation prediction by training a linear readout on top of the last convolutional layer of the object detection model AlexNet (Krizhevsky et al. 2012). Since DeepGaze I, most high-performing saliency models are using transfer learning, such as SALICON (Huang et al. 2015), DeepFix (Kruthiventi et al. 2017), SalGAN (Pan et al. 2017) and our model DeepGaze II (Kümmerer et al. 2017) presented in Section 2.3 of this thesis.

At this point, it is crucial to point out that the term "saliency" is heavily overloaded. The term changed and multiplied its meaning over the last two decades, starting from the classic low-level feature pop-out in attention and in fact by now there are at least four different notions of saliency that are used in closely related fields. In vision science and psychology, saliency usually still denotes low-level feature pop-out. In computer vision, saliency or saliency prediction is equated with the prediction of fixations in images with whatever features are helpful and explicitly allowing high-level semantic features. Additionally, computer vision uses the term saliency in the context of *salient object detection*, which is essentially a foreground/background object segmentation task (Borji et al. 2015) that in most cases does not involve any eye movement data. Finally, in deep learning, the term "saliency map" has been established for visualization techniques that try to quantify which areas of an input image are how relevant for the final network decision, e.g., in object recognition (Simonyan et al. 2013; Zhou et al. 2016; Kindermans et al. 2018). In this thesis, saliency will be used in the meaning of "whatever drives fixations" and use the term "classical saliency" to refer to the concept of low-level feature pop-out in vision science.

Also, the distinction between bottom-up/top-down and low-level/high-level is used with different meanings. Sometimes these terms are used to differentiate between involuntary and task-driven effects, sometimes they are used to differentiate between simple, mostly linear features that could be computed early in the visual hierarchy and complex potentially semantic features. In this thesis, bottom-up and top-down will be used to differentiate between involuntary and task-driven effects, while low-level and high-level will be used to differentiate between simple and complex features, as suggested in Schütt (2018).

1.4 Evaluating and Benchmarking Saliency Models

As detailed above, the research community working on saliency modeling has grown large studying a variety of diverse models. As a consequence, the topic of model comparison has become increasingly important to develop a precise notion of how to judge the usefulness of different models. As it turns out, for the field of saliency models, this problem is surprisingly complex. A large part of this thesis is dedicated to organizing this complexity and suggesting principled solutions.

In general, there are multiple ways to judge whether a model is "good". Originally, researchers often checked whether the model reproduces certain qualitative effects, e.g., constant duration in elementary feature search versus linear search duration in conjunction search. Alternatively, especially in computer vision there is often an explicit task and objective function that allows one to measure model performance quantitatively on this task, e.g., search performance in the search for military vehicles in images (Itti and Koch 2000).

In the absence of a specific task, discriminative modeling is not possible. One usually resorts to generative modeling and one can try to quantify how well the model predicts the ground truth data that it is supposed to model, such as human eye movement data on images. This could be advantageous because it should result in just one performance number instead of having to decide which effects or tasks are most important, and one would hope that a model that is good at predicting the actual data is also good on all sorts of tasks and applications. Also, information theory provides very principled notions of what it means to predict data well. The field of fixation prediction usually chooses the approach of quantifying how well models predict fixation data since there is no obvious effect or application.

Quantifying how well a saliency map predicts ground truth fixation data is not straight-forward. The concept of saliency maps was originally not intended to be evaluated on human fixation data at all but was supposed to explain certain effects found in visual search (see Section 1.3). Saliency maps, i.e., two-dimensional arrays of values or scalar fields, and fixation data, i.e., sets of locations on images, live in

very different spaces and therefore many classic measures of prediction performance cannot directly be applied to them, such as mean squared error, variance explained or correlation. Additionally, there is not even a clear notion of what the saliency value of an image location is supposed to mean – again, because saliency maps were originally not intended to be evaluated on fixation data. The only real common ground among researchers is that areas of higher salience should be associated with more fixations.

Given this lack of an obvious performance measure, researchers came up with many different solutions to this question. With more than twenty metrics, there are too many metrics to go into detail about all of them. Instead I will introduce the ones that are either very relevant for historical or other reasons or that are commonly used. For more details, see Wilming et al. (2011), Riche et al. (2013b), Meur and Baccino (2013), Borji and Itti (2013), and Judd et al. (2012), discussing a total of close to 20 metrics, and two papers included in this thesis (Kümmerer et al. 2015b; Kümmerer et al. 2018).

In the first ever comparison of saliency maps and fixation data, Parkhurst et al. (2002) compared the saliency of fixated locations to the average image saliency. However, they only tested for significance but didn't establish a performance measure. In 2005, Tatler et al. established a metric that is now usually called the "shuffled AUC" metric.

The area under the ROC (receiver operating characteristic) curve is a standard metric from signal detection theory and measures the performance of a binary classifier. Given a classifier that assigns to each data point x a score $c(x)$ where high scores should correspond to class A and low scores to class B, for a threshold θ one classifies datapoints with $c(x) < \theta$ into class B and datapoints with $c(x) \geq \theta$ into class B. This yields a hit rate h_θ and a false positive rate f_θ and one would hope to have a high hit rate and a low false positive rate. In order to remove the dependency on the threshold θ, one can compute the integral over the ROC curve given by the points (f_θ, h_θ) for all θ which goes from $(0,0)$ to $(1,1)$. This integral is called the "area under the ROC curve" or "AUC" and is a number between 0 and 1. An AUC of 0.5 corresponds to a classifier operating at chance performance. An AUC of 1 is a perfect classifier and 0 a classifier that is always wrong. The AUC metric can seem hard to interpret, but actually it has a very intuitive meaning: One can use the classifier score not to classify single datapoints into the classes A and B but instead to classify pairs of datapoints (x, y) into classes (A, B) and (B, A). The AUC is exactly the accuracy of this classifier. This task is effectively a 2AFC (two alternative forced choice, Fechner 1860) task where the classifier has to decide which one of two image locations has been fixated.

Tatler et al. applied the AUC score to saliency metrics by treating the saliency map as binary classifier score for image locations with one class being the fixated locations and the other class ("nonfixations") being locations fixated in another random image from the dataset at hand.

By now there are multiple versions of the AUC metric in use for saliency model evaluation. Most prominent are the one originally used by Tatler et al. (2005) and another version where instead of fixated locations from other images, all pixels in the same image that haven't been fixated are used as second class (Cerf et al. 2009; Wilming et al. 2011; Judd et al. 2012).

Another popular metric is the *normalized scanpath salience* (NSS). It is essentially the z-score of saliency values at fixated locations compared to all locations and was introduced by Peters et al. (2005) as a refinement of the approach used by Parkhurst et al. (2002).

The AUC-type metrics and NSS can be categorized as *value based metrics* since they operate on the saliency values of fixated locations. There is a second class of metrics, often called *distribution based metrics*. Instead of bringing saliency maps into the space of fixation locations like AUC and NSS, these metrics transform the fixation locations into the space of saliency maps and apply metrics there. The transformation from fixation locations to a saliency map is usually done by counting the number of fixations for each image pixel and convolving the resulting map with a Gaussian of usually one degree of visual angle. This essentially computes a Gaussian kernel density estimate of the fixation distribution. The convolved map is often called empirical saliency map and is treated as ground truth in all distribution based metrics.

The most prominent distribution based metric is the *correlation coefficient* (CC, Ouerhani et al. 2003; Pomplun 2006; Hwang et al. 2007): it uses the pearson correlation coefficient to compare a model saliency map with an empirical saliency map. The perfect model would exactly reproduce the empirical saliency map and score a correlation coefficient of 1. Another frequently used distribution based metric is *similarity* (SIM, Judd et al. 2012). This metric treats saliency map and empirical saliency map as probability distributions, takes for each pixel the smaller value of both maps and computes the sum of those values (essentially computing an ℓ_1 distance).

KL divergence (KLDiv, Rajashekar et al. 2004; Meur and Baccino 2013) also treats saliency map and empirical saliency map as probability distributions and computes the KL divergence between both distributions. Unfortunately, there is another completely different metric that computes the KL divergence between histograms of saliency values at fixated and non-fixated locations and that is also usually just

referred to as KL divergence (Itti and Baldi 2006; Borji and Itti 2013) which often leads to some confusion. Finally, occasionally the *Earth Mover's Distance* (EMD, Zhao and Koch 2011; Judd et al. 2012) is used to compare saliency maps to empirical saliency maps.

Besides value based and distribution based metrics, there are also completely different approaches. For example, Li et al. (2015) and Xia et al. (2018) learn saliency metrics to reproduce human similarity judgments between model saliency maps and empirical saliency maps as a way to quantify the common practice of visually comparing saliency maps (e.g., done in Cornia et al. 2018; Borji et al. 2013a; Borji and Itti 2013).

Metrics are not the only important ingredient for evaluating models. For making model comparisons fair, it is vital to compare models on the same dataset. In order to do so, the field of saliency modeling uses benchmarks to assess progress in the field. This serves to exclude differences in subjects, choice of images or the experimental setup as confounding factors for the model performance. Benchmarks in saliency modeling usually mainly consist of a hold-out dataset of fixations where the actual fixations are not published to keep models from overfitting to the dataset. Often there is a second training dataset collected in the same experiment that is made public to allow matching experiment specific biases. Researchers can submit model predictions to the benchmark and be informed about how well their model does on the hold-out dataset. Additionally, often there is a leaderboard where the performance of submitted models is listed as a quick way to asses progress and state of the art.

By far the most widely used and accepted saliency benchmark is the MIT Saliency Benchmark (`mit.saliency.edu`). Its main benchmarking dataset *MIT300* (Judd et al. 2012) consists of 300 natural images with indoor scenes, outdoor scenes, natural scenes, portraits and more together with free-viewing fixation data from 45 observers over three seconds presentation time. Its leaderboard shows the performances of around 100 models. Additionally, since 2015 there is a second benchmark dataset *CAT2000* (Borji and Itti 2015) with 2000 images from 20 different categories like natural scenes, fractals or satellite images. Other relevant benchmarks are the SALICON challenge and the LSUN challenge.

Despite these benchmarking efforts, keeping track of progress and state of the art in the field of saliency prediction is a huge problem. There is no clear agreement about where the field stands. For example, Einhäuser and König (2010) wrote "Recent elaborations of such stimulus-driven models are now approaching the limits imposed by intersubject variability" while Borji et al. (2013a) wrote "the main conclusion of this study is that a significant gap still exists between the best models and human

inter-observer agreement". The reason for having such strongly opposed opinions in a field that puts great value on quantitative benchmarking is that the many different commonly used metrics in the field (see above) give rise to highly inconsistent model rankings. Different metrics do not necessarily pose a problem on their own. For example, in image segmentation, both per-pixel accuracy and intersection-over-union are commonly used metrics, however, model rankings are roughly correlated between those metrics. Models being substantially better in one metric usually are also substantially better in the other metric. Rankings might be different only for models with similar performance. This is what researchers intuitively expect from different metrics. For example, Borji et al. (2013a) write "A model that works well should score high (if not the best) at almost any score". This intuition fails for saliency models: it is quite common that one model is state-of-the-art in one metric and below baseline in another metric while the same holds in reverse for another model ("To the disappointment of the authors, many recent models overall perform worse than the Itti-CIO2 model published in 1998", (Borji and Itti 2013)). The field is very aware of this "metric confusion" and there is a substantial body of literature seeking to address it. Le Meur et al. (2007), Wilming et al. (2011) and Borji and Itti (2013) catalog metrics by conceptual similarities. Other researchers have tried to analyze weaknesses of some metrics, such as the saturation of AUC metrics (Zhao and Koch 2011) or edge effects of AUC and KL-Div (Zhang et al. 2008) and check extreme cases (Bylinskii et al. 2018).

On a more general level, metrics have been ranked and classified based on certain principles: Emami and Hoberock (2013) selected a best metric based on how well different metrics can differentiate between empirical saliency maps and random saliency maps, Xia et al. (2018) compared saliency metrics to human similarity judgments between model saliency maps and empirical saliency maps and Wilming et al. (2011) classify metrics based on properties like intuitive scales and robustness. Riche et al. (2013b) analyze the statistical differences between metrics over a large number of models and apply PCA to suggest three metrics that explain most of the variance, but do not discuss whether using multiple inconsistent metrics is a worthwhile goal in the first place.

There is one common element in all existing research on saliency metrics: all works accept the massive differences between saliency metrics as a given and suggest different ways to cope with them. Interestingly, there is no attempt to mitigate the disagreement between metrics altogether.

2 Papers

This chapter will summarize the papers that resulted from the research done in my PhD. For each paper, I will detail what motivated the work, state the main results and discuss those results shortly. The following chapter 3 will combine the papers into a bigger picture to discuss the overall outcome of the research presented here.

2.1 Information-theoretic Model Comparison Unifies Saliency Metrics

Matthias Kümmerer, Thomas S.A. Wallis & Matthias Bethge
PNAS 2015

2.1.1 Motivation

Despite there being a long tradition of modeling spatial fixation placement, the field of fixation prediction is very unsure about its own progress. While some researchers viewed the problem as basically solved (Einhäuser and König 2010), other researchers showed disappointment about bad model performances (Borji et al. 2013a). The main consensus in the field seems to be that this problem is due to disagreeing metrics and that this problem cannot really be solved (Wilming et al. 2011; Borji and Itti 2013; Li et al. 2015; Li et al. 2015; Bylinskii et al. 2018).

Starting from this state of the field, we decided to go back to first principles to think about how saliency models should be formulated and evaluated. To us it seemed natural to model fixation placement as a probabilistic process since the fixation data is inherently probabilistic: even the same person won't make the same fixations twice. While in theory the biological fixation placement process might be deterministic, most likely it depends on many unknown internal state variables and noise, making the data at least appear stochastically. While modeling fixation data in a probabilistic framework is not novel (Baddeley and Tatler 2006; Barthelmé et al. 2013), it has never been applied in the setting of benchmarking existing saliency models. This is not straightforward since existing models have to be converted to probabilistic models in a fair way.

When it comes to comparing models, probabilistic models have a striking advantage because they come with a natural performance metric. *Information gain* quantifies how much better a posterior predicts data than a prior (Shannon 1948). It is essentially the difference in log-likelihood between the prior distribution (a baseline model) and the posterior distribution (a model to be evaluated) and can be expressed in bits/fixation.

The goal of this study was to establish a fair way to convert existing saliency map models into probabilistic models predicting fixation densities, to establish

information gain as a principled performance metric for such models, to test how our proposed measure of information gain relates to commonly used existing saliency metrics and their inconsistency when applied to a variety of existing models and to quantify the progress of the field in a principled way.

2.1.2 Results

We converted fifteen existing influential saliency model into probabilistic models that predict fixation densities for images. This conversion was done by postprocessing saliency maps with a Gaussian blur, a pixelwise monotone nonlinearity and adding a centerbias. For each model, the parameters of these postprocessing steps were jointly optimized over the full dataset. Additionally, we build two baseline models to estimate lower and upper limits of how well models should and could perform: the first was a centerbias model: a non-parametric model predicting the fixations for one image from the fixations on all other images. The second was a gold standard model: a non-parametric model that predicted each subject's fixations on a given image from all other subjects fixations on the same image.

We evaluated all models using the proposed information gain metric and found that the eDN model (Vig et al. 2014) performed best by explaining 0.41 bit/fix more than the centerbias model. To estimate the overall state of the field, we put the model performances into relation to this estimate of the explainable information gain: the gold standard explained 1.21 bit/fix more than the centerbias model and therefore the best model explained 34% of the explainable information gain. Additionally we showed how the information gain metric can be used to pinpoint on different levels from dataset to individual fixations where models fail to predict fixations which should be helpful for model analyses and future research.

To understand the relation of information gain to the existing saliency metrics, we evaluated all included models with several common saliency metrics. When evaluating the saliency models on their original saliency maps as produced from their original implementation, the saliency metrics showed the usual strong inconsistency that motivated us in the first place to work on this project. However, when evaluating existing saliency metrics using the log-density predictions of the postprocessed models as saliency maps and putting the model scores into relation to the corresponding performances of the centerbias model and the gold standard model, we found the metrics to be highly correlated both in ranks and values[1].

[1]This is slightly simplified. Please check the paper for details

2.1.3 Discussion

When we started this study, our goal was to establish what we considered to be "the right" metric for evaluating models of fixation prediction, not to unify saliency metrics. Only later on and to our own surprise, we noticed that formulating saliency models as probabilistic models that are optimized for information gain and evaluating existing saliency metrics on the log-density predictions of those models removes most of the disagreement between different saliency metrics. While previously researchers mostly agreed that the different existing saliency metrics measure substantially different things and therefore large differences in metric scores are unavoidable, this pointed to a different underlying reason: it appeared that the metric inconsistencies are mainly due to metrics interpreting saliency maps differently and models encoding their predictions into saliency maps in different ways. By evaluating all metrics on log densities post-processed for information gain we made sure that all models encoded their predictions in the same way and subsequently avoided most of the metric disagreement. Saliency metrics still interpret saliency maps differently and therefore some saliency metrics penalized the log densities. For example, the sAUC metric penalized the log densities for including a center bias. But those penalties were applied to all compared models and this turned out to keep the model rankings and relative performances quite consistent.

A crucial step to reach this unification of saliency metrics was the model specific post processing via blur, nonlinearity and centerbias: depending on which metric the models where originally intended for, the absolute saliency values might be meaningless besides their rank (AUC, sAUC), they might not model the very dominant centerbias (sAUC) or they might have been overconfident in their spatial predictions (visual comparison of fixations and saliency maps) while all these factors are important for a probabilistic model evaluated under information gain. Opposed to that, previously researchers usually just normalized saliency maps to have unit sum when requiring fixation distributions (e.g., the KLDiv and SIM metrics in Judd et al. 2012). This imposes a lot of meaning on the absolute saliency values that they might have never had, resulting in potentially arbitrary model scores.

Our results showed that most existing saliency metrics behave very similar if evaluated as we suggested. Nevertheless we argued that information gain is the most natural way to evaluate probabilistic models for fixation prediction, i.e., comparing the average log-likelihood of the model to that of a baseline model. Besides being based on first principles from information theory, we considered this measure much more intuitive than many other metrics since it defines a ratio scale (Stevens 1946)

with a well defined zero and well-defined performances differences, hence allowing to reason about the explained ratio of the explainable information gain.

Our suggested approach of unifying saliency metrics has two main drawbacks. Firstly, the approach requires considerable computational efforts: it is necessary to compute saliency maps for all relevant saliency models (if the model implementation is available at all), optimize the parameters of the post-processing and compute centerbias and gold standard models. Secondly, our approach is mostly incompatible with the existing benchmarking practices and literature results. It is impossible to compare model scores as resulting from evaluating a model's log density under a certain saliency metric with literature results on that metric. Published model scores will usually be computed on saliency maps that are not log densities and therefore might give rise to different penalties on different metrics. Together, these drawbacks made it very hard for the saliency community to adopt our approach and we addressed this problem in our follow up paper (Kümmerer et al. 2018, see below).

2.2 Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet

Matthias Kümmerer, Lucas Theis & Matthias Bethge
ICLR Workshop Track, ICLR 2015

2.2.1 Motivation

From our work on saliency model evaluation (Section 2.1 and Kümmerer et al. 2015b) we learned that existing saliency models failed to explain a substantial part of the inter-observer consistency present in the spatial distribution of fixations. One potential reason for that is stated by the hypothesis that fixations in free-viewing are mainly driven by objects and semantic content as put forward by Henderson et al. (1999) and Einhäuser (2013). Until 2012, existing computational approaches for detecting objects were mostly far from human performance. Einhäuser manually annotated objects in their study. In 2012 this changed: With AlexNet (Krizhevsky et al. 2012), for the first time a deep convolutional neural network won the ImageNet Large Scale Visual Recognition Challenge (Deng et al. 2009), reducing the remaining error substantially and kicking off what is now often referred to as the deep learning revolution: since then deep learning based models have outperformed classic computer vision models on many tasks.

Deep learning had already been applied to saliency modeling in the eDN model (Vig et al. 2014) that used ensembles of many convolutional networks with up to three layers. While the model performed at state-of-the-art, it was not significantly

better than existing models. We attributed this to a lack of training data: while AlexNet had been trained on the ImageNet dataset with more than a million images, eDN was trained on the MIT1003 dataset with fixation data for 1003 images.

Shortly after AlexNet won the ImageNet Challenge, it became evident that the internal image representations learned by deep neural networks for object recognition is of much more general use than being just applicable to object recognition. Instead the representations can be used for many other tasks by so called *transfer learning* (Donahue et al. 2014). Here, one usually trains a model on object recognition using the large available datasets like ImageNet or downloads a pretrained model. Then the top layers of the DNN, which are very specific to the object recognition task, are removed. Instead, some new layers are added that enable the model to predict the task of interest. Finally, the full model or just the added layers are finetuned on the new task using the potentially much smaller dataset available. Often this results in better performance than training the model from scratch on the new task: the deep features generalize to the new task. Transfer learning enabled deep learning to outperform classic computer vision algorithms on many other tasks such as object detection or semantic image segmentation and even gave rise to better predictors of neural activity than previous models in neuroscience (Yamins et al. 2014).

This made us wonder whether transfer learning might be a promising technique for improving models of fixation prediction. The goal of this study was to test whether deep features from neural networks trained on object recognition hold information relevant to spatial fixation prediction. This way we intended to test whether a model with access to high-level features and object information performs better at fixation prediction and to understand which features are most important for model performance.

2.2.2 Results

Our proposed model *DeepGaze I* processed an in input image with the AlexNet deep neural network up to the last convolutional layer. After that, a weighted linear sum converted the deep feature maps given by the activations from the last layer of AlexNet into a single saliency map. Subsequently, the saliency map was blurred, combined with a centerbias and passed through a softmax to yield a predicted fixation density for the input image. The linear weights, the weight of the centerbias and the blur size were learned by training the model on a subset of the MIT1003 dataset for maximum likelihood.

DeepGaze I explained 54% of the explainable information gain in the spatial fixation structure on the validation set compared to 34% for the previously best model eDN. On the MIT Saliency Benchmark, DeepGaze I set a new state-of-the-art,

raising the performance w.r.t AUC from 82.6% to 84.4%. Additionally we found that deep layers from AlexNet yield better performance than early layers and that convolutional and relu layers generalize better than pooling and normalization layers. The three features from AlexNet that DeepGaze I put most weight on were sensitive to faces, text and something best described as pop-out effect.

2.2.3 Discussion

With DeepGaze I, we could show that deep features trained on object recognition are useful for predicting fixations and can substantially improve model performance. The high-level semantic information present in the deep layers seemed to be an important factor for this performance gain since deep layers that contain more semantic information yielded better performance than early layers. Further evidence for this conclusion is that the two most important features from AlexNet were sensitive to faces and text. Nevertheless, low-level factors might also contribute to fixation selection: the third most important feature seemed to detect some kind of pop-out that could be either low-level or high-level.

2.3 Understanding Low- and High-Level Contributions to Fixation Prediction

Matthias Kümmerer, Thomas S.A. Wallis, Leon Gatys & Matthias Bethge
ICCV 2017

2.3.1 Motivation

With DeepGaze I (Kümmerer et al. 2015a), we showed that transfer learning holds great potential for predicting free-viewing fixations. Nevertheless, the prediction performance still left a substantial margin to the limit of inter-observer consistency and in the mean time new transfer learning models significantly outperformed DeepGaze I in the MIT Saliency Benchmark. The better performance of these new models seemed to stem from two main advances: First, since AlexNet a new generation of DNNs substantially improved object recognition accuracy in the ImageNet challenge and served as transfer basis for recent saliency models. Second, an new dataset was published: In the SALICON dataset (Huang et al. 2015) subjects viewed blurred images on the screen and could use the mouse to move a high-resolution "fovea" over the image. The recorded mouse traces were shown to be highly correlated with the spatial free-viewing fixation distribution but, unlike recording eye movements, this task could easily be scaled up on Amazon mechanical turk to 10000 images, collecting much more data than in most public

fixation datasets. The SALICON dataset had been used for pretraining the recent saliency models before finetuning them on actual fixation data.

The generation of top-performing saliency models after DeepGaze I had an important difference to DeepGaze I: in training, they finetuned the full neural network, including the transferred part. Opposed to that, we preferred to keep the transferred features unchanged to be able to use what is known about them, e.g., in terms of the extracted information since we were interested in understanding which features contribute to fixation selection. It could have well been the case that the features of more recent DNNs generalized worse to fixation prediction than AlexNet without finetuning them. So we were interested to test whether that is the case.

In addition to these straight-forward improvements to DeepGaze II, we had an additional idea that we hoped could improve our understanding of which features contribute to fixation selection. DeepGaze I used only a linear combination of deep feature maps. This made it very dependent on the scale of those used features and made it impossible to exploit interactions between different features. At the same time, overfitting problems showed the importance of keeping the number of trained parameters low. We hypothesized that replacing the linear readout with a few layers of 1×1 convolutions might be a good trade-off between high computational power and low number of parameters.

Besides improving prediction performance, this work had an additional motivation rooted in the long-going discussion on whether low-level or high-level features primarily drive fixations in free-viewing. Many researchers hypothesized free-viewing fixations to be mainly driven by classical saliency, i.e., low-level features and primarily feature pop-out (Itti et al. 1998; Parkhurst et al. 2002). Other researchers claimed that objects and semantic information play a more important role than local contrast (Einhäuser et al. 2008; Vincent et al. 2009).

The performance increase that DeepGaze I gained over classic saliency models by using deep features trained on object recognition seemed to support the second point of view. However, while the deep layers of neural networks contain a lot of semantic information, they still carry substantial low-level information too, such as size, color and rotation (Hong et al. 2016). It was possible that the performance of DeepGaze I was exclusively due to low-level information. On the other hand, there were very few low-level models that were trained on empirical data. Also, compared to recent models, they had very limited capacity. We decided that for getting a better understanding of the relevance of high-level information, we needed a better comparison model that uses simple low-level features but gives them the best changes at predicting fixations. To do so, we planned to complement our new model DeepGaze II with a model using the same architecture, but replacing the transferred high-level features with simple low-level features as a principled way of

comparing how well these different features predict fixation locations. The features we used encoded local intensity and intensity contrast (intensity contrast features, ICF) on different scales and therefore the essence of the classic saliency idea as, e.g., found in Kienzle et al. (2009).

To summarize, the goal of this study was to provide a fair comparison of the predictive power of low-level and high-level features for fixation prediction. Also we aimed to to quantify how much model performance is affected by using deep VGG features better at object recognition instead of AlexNet features, how much performance is affected by pretraining on the SALICON dataset and how much performance is affected by using a pixelwise nonlinear readout instead of a linear readout.

2.3.2 Results

The final model architecture used for DeepGaze II first extracted deep features for an input image from several of the conv5 layers from the VGG convolutional neural network (Simonyan and Zisserman 2014), processed them with a readout network of 4 layers of 1×1 convolutions yielding a single final saliency map that was then blurred, combined with a center bias prediction and converted into a fixation density prediction with a softmax. The ICF model replaced the VGG features with ICF features that were computed by converting the input image into a grayscale channel and two opponent color channels that then were blurred with Gaussians of five different size to yield local intensity maps. The local intensity contrast maps were computed by blurring the squared differences between input channels and local intensity maps with the same kernel again. Both local intensity and local intensity contrast maps were used as input features for the readout network. DeepGaze II and ICF were trained using maximum likelihood optimization, first on the SALICON dataset and then on the MIT1003 dataset using 10-fold crossvalidation over images.

DeepGaze II set a new state-of-the-art for free-viewing fixation prediction with 80.3% explainable information explained on MIT1003 and 88% AUC on the MIT Saliency Benchmark, while DeepGaze I explained 46.1% on MIT1003[2] and the best previous models on the MIT Saliency Benchmark all reached 87% AUC. Detailed evaluation showed that using VGG features instead of AlexNet, pretraining on SALICON, using a multi-layer readout network and crossvalidating over images instead of subjects all had substantial influence on the performance gain from DeepGaze I to DeepGaze II.

[2] in Kümmerer et al. 2015a we used only a subset of MIT1003, therefore the numbers differ slightly between the two papers

ICF set a new state-of-the-art for free-viewing fixation prediction without use of transfer learning from explaining 31.3% of the explainable information for eDN (Vig et al. 2014) to 37.2%.

Comparing the predictions of DeepGaze II and ICF in more detail, we found that there is a substantial number of images where the predictions of the low-level are better than those of the high-level model (about 10% of all images). By comparing model predictions on individual fixations we were able to find images where some clusters of fixations seemed to be driven by low-level features and other clusters seemed to be driven by high-level features.

2.3.3 Discussion

In this paper we were able to show that it is possible to improve state-of-the-art prediction performance with a model that is much simpler than all previous top-performers and has much fewer trainable parameters. Key to that was three factors: firstly, using VGG features, secondly, the readout network that could learn nonlinear pixelwise transformations and make use of interactions between features and finally, pretraining on the SALICON dataset.

Our comparison of ICF and DeepGaze II confirmed the intuition that high-level features are crucial for high-performing models of fixation prediction and outperform low-level features even in a fair comparison where as much information as possible is extracted from the low-level features. Nevertheless, low-level features can give rise to models that are substantially better than thought so far and still might contribute significantly to fixation placement in some images, especially in the absence of faces. It is an important caveat of the results in this work that the dataset itself makes it hard to differentiate between high-level and low-level effects since they are correlated. For example, faces are correlated with the colors of skin. To mitigate that problem, one would need to design a dataset that reduces the correlation between high-level features and low-level features. At the same time, the stimuli should not be too different to natural images to be able to draw conclusions about free-viewing in natural scenes.

The comparison of DeepGaze II and ICF showed that our architecture with fixed features and a readout network can be used as a feature testing framework for evaluating in a fair way which features are how predictive of the spatial fixation distribution. The readout network, while sufficiently flexible to learn nonlinearities and interactions, is still heavily constrained by the input features and therefore allows to conclude about their predictive power. For a more general discussion of readout networks, see Section 3.4.

2.4 Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics

Matthias Kümmerer, Thomas S.A. Wallis & Matthias Bethge
ECCV 2018

2.4.1 Motivation

In this paper, we follow up on the work summarized in Section 2.1 (Kümmerer et al. 2015b). In this previous work we showed that the disagreeing saliency metrics can be unified by optimizing models for information gain and computing saliency metrics using log density maps as input saliency maps. However, we failed to make a significant impact on how the community evaluates saliency models. We attributed this to the fact that our proposed solution for saliency benchmarking was not fully satisfactory. Evaluating saliency metrics on log density predictions makes saliency metrics consistent because all models incur similar penalties under a given metric. The log density of a very good model can perform very bad under certain metrics and therefore the results are not comparable with published results evaluated on saliency maps directly optimized for such a metric.

We wondered whether there is a way for a good model to reach state-of-the-art scores in each metric. It had been noticed before that given a certain fixation distribution and a certain nonfixation distribution, the quotient of both fixation densities yields the best possible sAUC score (Barthelmé et al. 2013). This motivated us to think about whether one should evaluate a model using *different* saliency maps for different metrics.

Given a certain fixation distribution, a saliency map and a saliency metric, one can compute (at least in theory) the expected score of the saliency map under the saliency metric, assuming that ground truth fixations are distributed according to the given fixation distribution. We reasoned that the correct saliency map to use for evaluating a model under a certain saliency metric is that saliency map which has *highest* expected metric score with respect to the model's fixation distribution: for each other saliency map, the model itself would expect to be penalized and therefore being treated unfair. At the same time, a model correctly predicting the true fixation distribution would actually on average yield the best scores in all metrics. Essentially this is an application of Bayesian Decision Theory where one uses a posterior distribution over all possible events in the world (fixation distribution) to make a decision (which saliency map to evaluate) that maximizes some expected utility (the metric score).

Coming from these considerations, the goal of this study was to first derive optimal saliency maps for different saliency metrics under arbitrary given fixation

distributions and to find out whether these saliency maps can even computed analytically from the fixation density or whether numerical approximations are required. Next, we wanted to understand how these derived saliency maps behave and look like in an experiment with toy data where the true fixation distribution was known. Finally, we wanted to understand to which degree these effects transfer to empirical data: in real life the true fixation distribution will be different from the model distribution and it is possible that the wrong model fixation distribution influences the model score more than the differences between saliency maps derived for different metrics It this would be the case, it would remove the practical value of using different saliency maps for different metrics.

2.4.2 Results

For the most influential saliency metrics (AUC, sAUC, NSS, CC, KL-DIV, SIM) as well as the IG metric we proposed in Kümmerer et al. (2015b), we derived which saliency map has highest expected metric score under any given fixation distribution. We found that for AUC, sAUC, NSS, KL-DIV these saliency maps can be computed analytically, for CC they can be approximated analytically up to high precision and for SIM they can be approximated iteratively. For a toy fixation distribution, we computed these optimal saliency maps and evaluated all saliency maps under all metrics with data sampled from the toy fixation distribution. It turned out that when evaluating saliency maps optimal for different metrics all under the same metric, the performances can be very different, easily on the scale that is considered to separate very bad models from very good models. Also, these saliency maps can appear very different visually. For example, CC and KL-Div require saliency maps to be smoother than NSS and AUC, the SIM metric expects it to have more zeros and sAUC requires the saliency map to not include a center bias while all other metrics do.

The final goal of the study was to test whether the results transfer from the case where the ground truth distribution is used to compute the saliency maps to the actual benchmarking case where the saliency maps are derived from model distributions that can be very different from the ground truth fixation distribution. To do so, we converted several existing saliency models into probabilistic models using the same method as we did in Kümmerer et al. (2015b). Then, for each model, we computed saliency maps for each metric. Finally, for each each model we evaluated the original saliency map and all derived saliency maps under each metric using actual human fixation data.

We found that when evaluating each saliency metric under the saliency map predicted to be optimal by the evaluated model, all metrics agreed in their model

ranking and the metric values were highly correlated. Also, for each metric, the saliency map predicted to be optimal by a model performed better than all other evaluated saliency maps for that model, including the original saliency map that might have been optimized for that metric – except for rare cases of old models where the predictions are far from the ground truth fixations. Opposed to that, when using other saliency maps than the ones predicted to be optimal for a certain metric, models often incurred heavy penalties that could be bigger than the difference between the best and worst evaluated model. Additionally in many cases the model rankings became inconsistent if not using the correct saliency maps for each metric.

2.4.3 Discussion

From our results in Kümmerer et al. (2015b), we concluded that main reason for saliency maps being inconsistent is that they interpret saliency maps in different ways. The results in this paper show that this is indeed the case: even when knowing the true fixation distribution, it is impossible to perform optimal in all metrics using the same saliency map. Even more, a saliency map that is optimal for one metric often performs worse than baselines in other metrics.

The results of evaluating actual models postprocessed to predict fixation densities optimized for information gain show that our proposed framework of separating saliency models from metric-specific saliency maps is not merely of theoretical interest: Evaluating a state-of-the-art model using the wrong saliency map can easily yield a worse score than evaluating a model with very bad predictions but using the right saliency map for the metric in question. This illustrates that the differences in how saliency metrics interpret saliency maps affect model scores more than the quality of the evaluated models themselves. Also the saliency optimal for different metrics maps look visually very different, hence visually comparing saliency maps (as, e.g., done in Cornia et al. 2018; Borji et al. 2013a; Borji and Itti 2013) is dangerous if one doesn't make sure that saliency maps are optimized for the same metric.

Formulating models as probabilistic models predicting fixation densities, optimizing them for information gain and evaluating each metric using the correct metric-specific saliency map allows a good model to perform at state-of-the-art in all metrics and additionally removes all inconsistencies between different metrics, even when including models whose predictions are far from ground truth. Using only one type of saliency maps for all metrics (as we did with log-densities in Kümmerer et al. 2015b) doesn't fully solve the problem since there are still inconsistencies in addition to the penalties in some metrics. Opposed to our proposal from Kümmerer et al. (2015b), the approach in this paper allows a researcher to compare their model to all classical models with their published metric results. There is no need to convert old

models into probabilistic models – only we had to do it here to show the validity of our approach for which we needed several saliency models of different quality.

It should be mentioned that by optimizing a saliency map model directly for a certain metric of interest, of course one should be able to get results at least as good as optimizing the model for information gain and evaluating the correct saliency map. However, one conclusion of our results is that this usually will give only a minor performance gain. Our study included models directly optimized for some metric and after converting these models into probabilistic models and evaluating the derived saliency maps these models did not perform worse than originally under the metric they had been optimized for.

3 Discussion

In the papers summarized in this thesis we investigated the question how well we can we predict where people look in images. By the nature of their contributions, these papers fall into two categories: the first category (summarized above in Sections 2.1 and 2.4; Kümmerer et al. 2015b; Kümmerer et al. 2018) investigates how to formulate and benchmark models of fixation prediction, while the second category (summarized above in Sections 2.2 and 2.3; Kümmerer et al. 2015a; Kümmerer et al. 2017) focuses on how to improve models of fixation prediction and what we can learn from the models. In this section, we will look at those results in a broader perspective.

3.1 Formulating and Benchmarking Models of Fixation Prediction

There is a general agreement that the field of saliency modeling has a substantial problem with evaluating models due to highly inconsistent saliency metrics. Researchers have stated that "studies use a wide variety of performance measures with markedly different properties, which makes a comparison difficult" (Wilming et al. 2011), "Regarding fair model comparison, results often disagree when using different evaluation metrics" (Borji and Itti 2013), "it becomes somehow confusing on which metrics should be used and which models should be compared with in designing new saliency models." (Li et al. 2015), and "The inconsistency in how different metrics rank saliency models can often leave performance up to interpretability" (Bylinskii et al. 2018).

We addressed this problem in two articles summarized above: Kümmerer et al. (2015b) and Kümmerer et al. (2018). Previously, the disagreement between saliency metrics was interpreted to be due to the metrics measuring fundamentally different things that cannot and should not be made consistent (e.g., Riche et al. 2013b). Consequently, depending on the intended application, one would have to decide which metric is the right one to use and it would be impossible to have an overall state-of-the-art. By looking at fixation prediction from the principled perspective of probabilistic modelling and Bayesian Utility Theory, our work pinpoints a fundamentally different reason for the benchmarking problem: the field is missing a clear definition of what a saliency model is. There are many different implicit definitions with contradicting behaviour, resulting in "apples to oranges" comparisons being common in the literature (see below for some examples).

The results presented in Kümmerer et al. (2015b) and Kümmerer et al. (2018) summarized in this thesis lay out a strategy which allows researchers to avoid problems with model evaluation that were common before while keeping backwards compatibility with the existing literature. The strategy can be separated into two

parts. The first part concerns how to formulate and train models and how to evaluate them in inhouse analyses. The second part concerns how these models can be benchmarked using existing saliency metrics that are common in the literature.

The first part of the strategy (mainly presented in Kümmerer et al. 2015b) suggests to formulate saliency models as models of fixation density prediction, train them using log-likelihood or similar loss functions and use information gain or log-likelihood, resp., for internal evaluations (i.e., when not comparing to existing saliency models). We consider fixation density prediction the right framework for spatial fixation prediction because the fixation density is a measurable physical quantity which underlies the fixation data collected in eye tracking experiments (see also Barthelmé et al. 2013). Information gain or other versions of log-likelihood are suitable metrics to train and evaluate such models of fixation density prediction. As variants of the Kullback-Leibler-Divergence and well-founded in information theory, they naturally quantify how well a model predicts the fixation data in bits per fixation. By comparing to a baseline model such as the center bias or a uniform model, information gain provides a ratio scale where both performance numbers as well as differences in performance are meaningful and can be compared. While several traditional saliency loss functions are invariant to properties of the model prediction such as monotone transformations and the center bias, log-likelihoods are sensitive to all these properties. Therefore, it is important to make models sufficiently complex to be able to capture these properties of the fixation density. In Kümmerer et al. (2015b) we suggest a simple but general way of accomplishing this by adding a pixelwise monotone nonlinearity and a center bias component to models.

We argue that, as long as possible, researchers should stay in the framework of probabilistic models and information theoretic model evaluation due to their advantages. For example, inhouse model comparisons can be done using information gain as we did when comparing the DeepGaze II and ICF models in Kümmerer et al. (2017).

However, sometimes one cannot stay within the framework of probabilistic models and information theoretic model comparison. Especially, this is the case when comparing to existing models that are not predicting fixation densities. We addressed this situation in Kümmerer et al. (2018): whenever researchers need to evaluate metrics that operate on classic saliency maps instead of fixation densities, one should apply Bayesian Utility Theory to select the saliency map that is predicted by the probabilistic model to have highest performance in the metric of interest. Evaluating the metric on this saliency map avoids penalties due to not adhering to how the metric interprets saliency maps. Following this approach for all metrics of interest results in adequate performance numbers in all metrics that can be compared to published results in the literature.

Many apparently confusing results can now be explained by taking into account that different saliency metrics require different saliency maps even when evaluating the same model. Previously, the community identified saliency models with their saliency maps and implicitly defined saliency models via the metrics they trained or evaluated a saliency model on. When benchmarking different models, this results in a clash of definitions. The definition according to which the saliency model was formulated by the authors and the definition according to which the saliency model is evaluated by a metric in the benchmark are likely not the same. Since we show in Kümmerer et al. (2018) that no single saliency map can perform optimally in all metrics, this creates the impression that saliency metrics are fundamentally different.

One case where this affected results is the debate between Einhäuser and Borji about whether low-level features or high-level objects are more relevant for fixation placement (Einhäuser et al. 2008; Borji et al. 2013b; Einhäuser 2013). Einhäuser et al. (2008) compare a saliency model without a centerbias with an object model which shows a substantial centerbias. Since they use AUC, the saliency model is penalized for not having a centerbias and performs worse than the object model. Borji et al. (2013b) instead use the sAUC metric, which penalizes the object model for having a centerbias and conclude that the saliency model performs better than the object model. In order to resolve this, one could either add a centerbias to the saliency model and evaluate with AUC or, alternatively, remove the centerbias from the object model and evaluate with sAUC.

As another example, it can be easily explained now why Riche et al. (2013b) found three distinct clusters of model rankings when comparing model rankings resulting from different metrics. The model ranking given by KL-DIV is a separate cluster because it is sensitive to the centerbias and to absolute saliency values. The sAUC metric also forms a separate cluster since it is invariant to monotone transformations and penalizes modeling a centerbias. The third, large cluster contains metrics that are invariant to global contrast and require a centerbias (NSS, AUC, CC, . . .). The three clusters could only emerge because saliency maps formulated for different saliency metrics have been compared. Had all evaluated models been optimized for IG, most likely no clusters would have appeared at all, even if all models had been evaluated using the fixation density as saliency map (or the log-density, as we did in Kümmerer et al. 2015b). Only by using different metrics *and* different kinds of saliency maps, the large disagreements between metrics become visible.

Finally, our results also allow explanation of the large diversity of opinions about progress of state-of-the-art. One overestimates state-of-the-art when comparing a model to a gold standard that doesn't adhere to the requirements of the metric used. For example, using the sAUC metric, Borji et al. (2013a) find that AWS performs at the level of inter-observer consistency. However they use empirical saliency

maps to estimate the inter-observer consistency. Those include the centerbias and consequently get penalized by sAUC, unlike AWS which doesn't include a centerbias. In the reversed case state-of-the-art can be underestimated: the gold standard adheres to the metric requirements, but the models don't.

The reasons that the field currently works with saliency maps in this somewhat confusing way are most likely historical in nature. Saliency maps were first conceived as a module of covert attention in visual search (see Section 1.3).They were not a model themselves, but part of a model which was not directly evaluated on whether the focus of attention did indeed move to the maximum of the saliency map. Rather, researchers were interested whether the model predicted certain effects like the dependency of search durations on different target types. Only later the question was raised whether saliency also influences fixation placement and therefore had to compare saliency maps to fixation locations. People suggested ways to do so (Parkhurst et al. 2002; Tatler et al. 2005) and it became common practice to formulate models of fixation prediction as saliency maps and evaluate them on one or multiple existing metrics without specifying what the saliency values in the saliency map precisely mean, leading to the current situation.

An advantage of our evaluation approach is that already existing models will be scored fairly – the approach is "backwards compatible". Previously, researchers usually tried to solve the benchmarking problem by deciding for one metric or proposing a new one, penalizing all models that had been trained with a metric requiring different saliency maps. Using our suggested approach, researchers can phrase their new saliency model as a probabilistic model and evaluate each metric on the correct saliency map, while they don't have to do the same for all models they want to compare to. Of course these existing models will be penalized in some metrics as they always were, but each existing model will be scored fairly at least in the metric the model was originally formulated for and here the comparison will be fair as well. Therefore, the researchers can compare their new probabilistic model to existing models optimized for any existing metrics without being penalized in any metric. Effectively this means that their model will perform better on many metrics and we hope that this encourages more and more researchers to adopt our approach, increasing the number of models that are phrased as models of fixation density prediction that can be fairly compared in any metric.

To facilitate the transition towards probabilistic models, we are teaming up with the MIT Saliency Benchmark. The benchmark will allow the submission of models as fixation densities instead of jpeg-encoded saliency maps as common so far. When evaluating models, we will compute the correct saliency maps for each metric and evaluate on them. This will result in more consistent model scores and therefore

a better assessment of progress and state-of-the-art in the field. We hope this has the chance to spur new progress in the field of fixation prediction, which some considered to be stagnating in the last years due to the mistaken impression that models were approaching the noise ceiling of predictable performance.

The presented strategy of formulating models probabilistically and deriving different saliency maps from predicted fixation densities is of course not without limitations. Firstly, there are certainly cases where it might make sense not to formulate models as models of fixation density prediction. This is most likely the case when a model is formulated for a specific application with quantifiable performance. In this case there is no need to use information-theoretic metrics like information gain or evaluate established metrics other than the performance on the application task. Besides the computational overhead, training the model on a different metric than the performance measure relevant for the application can make the model perform suboptimal on this performance measure.

But even in the setting of comparing models of fixation prediction on many different established metrics, at some point in the future the method of deriving different saliency maps for different metrics might not be able to remove all metric inconsistencies anymore. The result that saliency metrics agree in their ranking of models of fixation prediction when evaluated using the correct saliency map is only empirical in nature and there is no theoretical guarantee for it to hold. In fact it is very easy to construct counter examples. For example, assume that two models predict the real fixation distributions with two modifications: Model A changes the temperature of the predicted fixation density. This doesn't affect AUC scores but will affect NSS scores. Model B switches the fixation density values between a few fixated and unfixated pixels, which will slightly decrease both AUC and NSS scores. If the temperature change in model A is large enough, model A will score higher in AUC than model B but model B will score higher in NSS than model A. As model predictions get closer to the ground truth fixation distribution, they will be able to capture the most important properties of the fixation density and more subtle effects like the one just shown might become apparent. Since free-viewing conditions are by definition not very well controlled, the community might consider the problem solved and move on to more challenging tasks before effects like the one above become a problem. However, the approach of applying Bayesian Decision Theory to evaluating multiple metrics is not constrained to free-viewing. In more controlled settings models should be able to come very close to the true data distribution and at some point our approach might not be sufficient to make model rankings consistent across metrics. In this case researchers will again have to decide which metric is most important to them.

On the conceptual side we want to emphasize that we do not imply that the fixation density is implemented as such in the brain when we recommend that researchers should model spatial fixation placement via fixation densities. We do not suggest that the brain internally first computes a fixation density over the full field of view, subsequently samples one location from the density and then moves the gaze to this location, even less that it samples multiple fixation locations that are then attended sequentially. This is an important conceptual difference between our approach and the original concept of saliency maps (Koch and Ullman 1985; Itti et al. 1998), which were explicitly made to be "biologically plausible" and hypothesized to be implemented in the brain (Li 2002). They were usually thought of as being the input to a winner-take-all mechanism with inhibition of return that selects subsequent fixations. When we propose to model prediction with probabilistic models that predict fixation densities, this is merely a framework that allows one to make explicit how sure the model is about fixation locations. Besides many other advantages, this is what allows us to choose the best saliency map for each saliency metric and solve the problem of disagreeing metrics.

Although we just emphasized the differences between the approaches of probabilistic fixation prediction and biologically-plausible saliency maps, they are far from mutually exclusive. Biological fixation selection with saliency maps was always thought of as depending on previous fixations, e.g., via inhibition of return. A principled way to combine concepts like this with probabilistic models is to extend the probabilistic model from predicting scanpath independent fixation densities $p(x, y \mid I)$ to scanpath dependent fixation densities $p(x_{i+1}, y_{i+1} \mid I, x_0, y_0, \ldots, x_i, y_i)$, i.e., predictions that take into account where the subject has looked before. A scanpath dependent fixation distribution can model mechanisms such as inhibition of return and allows to formulate a model that computes an internal (potentially biological plausible) saliency map as an intermediate step before some kind of fixation selection mechanism chooses a new fixation location depending on previous fixations. This has been done previously, e.g., in the SceneWalk model (Engbert et al. 2015; Schütt et al. 2017). In ongoing work we are extending our model DeepGaze II to DeepGaze III, which is also such a model of scanpath prediction that consists of an internal saliency map and a subsequent fixation selection stage.

3.2 Improving Models of Fixation Prediction

In the first paper summarized in this thesis (Kümmerer et al. 2015b) we showed that models up to 2014 still left a substantial amount of inter-observer consistency unexplained. From the roughly 1.2 bit/fixation that constitute the difference in cross-entropy between an image-independent baseline centerbias model and a

nonparametric cross-validated gold standard model of inter-observer consistency, the best model in 2014 (eDN; Vig et al. 2014) explained 34% (see Section 2.1 and Kümmerer et al. 2015b for details on the used metrics). Two papers summarized in this thesis (Kümmerer et al. 2015a; Kümmerer et al. 2017) are dedicated to reducing this gap between model performance and inter-observer consistency. With our two models *DeepGaze I* and *DeepGaze II*, we were able to increase the percentage of explained information on the MIT1003 dataset first to 46.1% and eventually to 80.3%.

Our key contribution to reaching that performance gain was introducing transfer learning to the problem of fixation prediction. While transfer learning from deep features trained on ImageNet had previously been applied successfully to a variety of computer vision tasks (e.g., Donahue et al. 2014), the only previous model using deep learning for fixation prediction, eDN (Vig et al. 2014), trained shallow neural networks from scratch and had to invest substantial effort to avoid overfitting on the comparatively small datasets available with fixation data. Since we showed with DeepGaze I that transfer learning from features trained on object recognition can substantially boost fixation prediction performance, all subsequent high-performing saliency models such as SALICON (Huang et al. 2015), DeepFix (Kruthiventi et al. 2017) and SalGAN (Pan et al. 2017) use transfer learning. An important difference between the DeepGaze models and most subsequent transfer learning saliency models in the literature is that we don't retrain the transferred features themselves. Only the part of the models that computes fixation predictions from the pretrained deep features is fitted to empirical data. Instead, other models like the ones mentioned above usually retrain the full deep neural network, with a notable exception being *DeepFeat* (Mahdi and Qin 2017) which does not only keep the transferred features fixed, but has in fact no trainable parameters at all.

Our second crucial contribution towards improving prediction performance besides the use of transfer learning is the concept of readout networks. The small number of 1×1 convolutions on top of fixed features seems to constitute a "sweet spot" between the limitations of purely linear readout and the complexity of full convolutional layers (with their large number of parameters). Readout networks will be discussed in more detail below in Section 3.4.

Finally, the correct choice of cross-validation for early stopping is crucial for training reasonably deep models of fixation prediction. In DeepGaze I we used cross-validation over subjects, however, this turned out to lead to substantial overfitting to the images (see Kümmerer et al. 2015a, Figure 5b). Deep features from networks like AlexNet and VGG encode a great variety of high-level and semantic features and apparently there is more diversity in the features at locations fixated across different images (often by many subjects), than there are differences between the features that different subjects look at in one image (although recent results suggest that there

can be substantial differences in semantic preferences between subjects, cf. Haas et al. 2019).

DeepGaze II is, more than three years since its initial submission, still listed as the top-ranking saliency model in the MIT Saliency Benchmark (according to the AUC metric and – when using the correct saliency maps – also under sAUC). This is somewhat surprising, given that since then a substantial number of new models have been published in the benchmark. These models often have substantially larger model capacity and make use of the latest advances in deep learning. So what is the reason for the performance of DeepGaze II? We suggest several possibilities below.

First, it could be that the lower capacity of DeepGaze II is actually an advantage more than a disadvantage since it regularizes the model and might therefore help generalization. This seems likely, given that the most popular datasets like MIT1003 usually don't consist of substantially more than 1000 images. Additionally, we assume that using log-loss as a training objective is an important factor. In Kümmerer et al. (2018) we argue that information gain (a.k.a. log loss) is an optimal loss function that forces the model to extract as much information about the training distribution as possible, which in turn allows to perform well on all saliency metrics. Jetley et al. (2016) compared different probabilistic training objectives with respect to their generalization to other saliency metrics but without the Bayesian Decision Theory framework and also without including log-loss. Besides the DeepGaze models, log-loss has been rarely used as a training objective in the saliency field (see Baddeley and Tatler 2006 for a notable exception). Instead, many contemporary models use loss functions that compare to empirical saliency maps and therefore produce overly smooth predictions, such as the KL-DIV metric as training objective (e.g., Huang et al. 2015; Oyama and Yamanaka 2018) or euclidean loss (e.g., Kruthiventi et al. 2017).

The choice of centerbias might be an additional advantage of DeepGaze over other models. While many models either don't model the centerbias at all and therefore rely on extracting it from border artifacts in the convolutional layers (Huang et al. 2015; Pan et al. 2017) or give the model ways to learn it implicitly (Vig et al. 2014; Kruthiventi et al. 2017), we use a nonparametric model of the actual centerbias and build it explicitly into the model as an interaction term. Besides making the model more interpretable, this might help avoid overfitting.

As a last minor detail, we contrast normalize the saliency maps evaluated in the MIT Saliency Benchmark to minimize the effect of 8bit quantization and JPEG artifacts on the AUC scores.

It should be stressed that all listed reasons are just hypotheses for which we don't have any conclusive evidence. More research with controlled experiments would be

necessary to test which of those possibilities contribute most to the lasting top rank of DeepGaze II in the MIT Saliency Benchmark.

3.3 What Drives Fixations in Free-Viewing?

Saliency maps were introduced to the field of eye movements to test whether low-level feature pop-out drives fixations under free-viewing conditions. Starting from this point, different kinds of features were experimented with and a debate started whether it might actually be high-level features and semantic content that drive fixations even in free-viewing conditions (Einhäuser et al. 2008; Vincent et al. 2009; Borji et al. 2013b; Einhäuser 2013).

Over the course of our work, we accumulated additional evidence in support of both sides, adding to a more nuanced picture. In Kümmerer et al. (2015b) we showed that classic saliency models miss a substantial part of the inter-observer consistency present in free-viewing fixations on natural scenes. However, this could be due to low-level features only marginally contributing to fixation placements or to models not making use of those features in the correct way, not least due to the disagreeing saliency metrics. Our model DeepGaze I (Kümmerer et al. 2015a) showed that the three most relevant features from AlexNet are sensitive to faces, text and some kind of quite general pop-out that might include semantic pop-out.

Finally, in Kümmerer et al. (2017) we tried to contribute to the question of which kind of features drive fixations in a more principled way. We trained models to predict fixations by extracting features from an image and predict fixation densities from those features. By using fixed sets of low-level (intensity-contrast) and high-level (VGG) features instead of retraining them as most state-of-the-art saliency models do, we made sure that the models were constrained by those features. By using the same architecture for the low-level ICF model and the high-level DeepGaze II model we were able to compare those features in a fair way. Finally, the architecture of the readout network made sure that both models could make most use of the provided features but were not able to learn new spatial features. Altogether this allows us to reason about the predictive power of those features for fixation prediction.

The high-level features overall yielded much higher predictive performance than the low-level features, supporting the view that semantic content is crucial for fixation placement. However, the low-level features still performed substantially better than all classic saliency models. We found that if there is high-level content present in a scene (mainly in the form of faces and text), then low-level features mostly fail to predict fixation locations. However, in the absence of faces and text, low-level features seem to influence fixation placement.

These results have to be taken with a grain of salt, though. First, the VGG features are not purely high-level but still contain substantial low-level information (Hong et al. 2016). At the same time, many low-level features are strongly correlated with semantic content, e.g., skin color and faces. We could only test the predictive power of those features but not whether they causally drive fixations. More detailed experiments in future work are required to understand exactly how and when semantic content overwrites classic saliency. This will have to involve much more diverse datasets that include color images, grayscale images, line drawings and other image types that separate low-level features from content.

It is sometimes argued that deep learning based models cannot really contribute to a better understanding of the modeled processes (e.g., Marcus 2018; Henderson et al. 2019). While we agree that one has to be careful when interpreting results of deep learning methods in scientific contexts, we don't think that deep learning principally cannot create scientifically valuable results. The previously discussed results of the work with DeepGaze II on low-level and high-level features are a concrete example of how deep learning can yield scientific results. However, there are also more general ways in which complex and potentially uninterpretable models can advance understanding.

Since image-computable models can be probed on arbitrary input images, analyzing high-performing deep learning models can reveal subtle patterns that might be hard to find in the training dataset alone. These insights can allow one to formulate hypotheses about contributing mechanisms, and thereby simpler models. An example of this approach (using nonparametric models instead of deep learning) is the model of Kienzle et al. (2009). The authors first trained a nonparametric model on fixation data, then applied nonlinear system identification to this model to find attractors and repellors in the model space. This subsequently allowed them to formulate a much simpler and very intuitive model, which explained almost as much variance in the data as the more complex model.

Another useful application of deep learning models can be the estimation of explainable information in the modeled distribution. In the case of 2d spatial fixation prediction it is comparatively easy to estimate the level of explainable information by means of a gold standard model (often some kind of kernel density model) and therefore judge how much of that information is explained by models. However, in the case of scanpath prediction, estimating an upper limit of model performance becomes much harder: instead of estimating a 2d distribution $p(x, y)$, now a multidimensional distribution $p(x_0, y_0, \ldots, x_n, y_n)$ or $p(x_i, y_i \mid x_0, y_0, \ldots, x_{i-1}, y_{i-1})$ has to be estimated. Methods that work well in the 2d case quickly fail in the higher dimensional case, even when using huge amounts of data. Here, deep learning

based models such as the scanpath extension of DeepGaze (ongoing work) can be used as an alternative means of getting estimates or at least lower bounds of the explainable information. This can then be used to quantify how much the principled mechanisms of interpretable models such as SceneWalk (Engbert et al. 2015; Schütt et al. 2017) contribute to the structure of scanpaths.

3.4 Beyond Fixation Prediction: Readout Networks

While the research presented here focused on the question of how well we can predict where people look in images, some contributions from our work may apply more generally. One such contribution is the concept of readout networks that we established with the DeepGaze II and ICF models in Kümmerer et al. (2017). Readout networks, as we defined them here, are small convolutional neural networks consisting only of a few layers of 1×1 convolutions. We argue that they provide a very intuitive sweet spot between purely linear readouts and full DNNs. The strong constraints of a linear readout can be problematic, especially when used on top of potentially unintuitive features that might contribute to prediction on an unknown and nonlinear scale (e.g., logarithmic, squaring, exponential, or including mixing effects). Full DNNs on the other hand are often too underconstrained. In the worst case they could compute predictions by first reconstructing the original input image from the input features and therefore not depend on the used features at all.

Our readout networks try to combine the best of both worlds. Using only 1×1 convolutions, they are equivalent to a nonlinear function defined by a fully connected neural network that is applied pixel by pixel to a vector of input features. Therefore, they are able to learn the best nonlinear transformations adjusting the scale of the input features and make use of interactions between those features. But, since they are applied to each pixel individually, they cannot learn new spatial features. Essentially, given enough capacity, they can make use of all the pixelwise mutual information between input features and predicted quantities such as fixation density, and therefore provide a way to compare this mutual information between different sets of features.

This makes readout networks a powerful tool for testing the predictivity of arbitrary features for arbitrary tasks. One such example has been presented here: the comparison of low-level and high-level features with respect to their predictivity for fixation distributions. Other researchers have already applied readout networks to other tasks, such as comparing the ability of humans and DNNs to recognize closed contours in cluttered images (Funke et al. 2018). Because this task doesn't require an image-shaped prediction, a linear readout is added to the output of the last 1×1 convolution.

3.5 Applications

The work presented in this thesis has already had some applications within and beyond science. DeepGaze II has been tested for its applicability to comic reading (Laubrock et al. 2018) and was used in a study exploring the influence of scene content on gaze (Damiano et al. 2019). Rothkegel et al. (2019) used DeepGaze II when classifying different types of saccades in visual search. Engineers at Twitter used DeepGaze II as a starting point when building a saliency model used for smartly cropping images to required aspect ratios (*Speedy Neural Networks for Smart Auto-Cropping of Images* 2018). There is a public website of DeepGaze II[3] which allows to compute model predictions on uploaded images. Among other cases, it has been used in design courses at TU Delft to evaluate and change designs.

3.6 Outlook

After decades of research, we now have models that are able to predict spatial fixation patterns in free-viewing of natural scenes with performance approaching the inter-observer consistency. However, this performance boost came at the price of loosing a large part of the interpretability of classic models such as Itti et al. (1998), Bruce and Tsotsos (2009) and Kienzle et al. (2009). Therefore, now it makes sense to move from making interpretable models better at predicting fixations to making models that are good at predicting fixations more interpretable.

Starting from the work presented here, there are several ways to approach model interpretability. Instead of using VGG features as a deep feature space for predicting fixations, there now exist features that are much more straight-forward to interpret. One example is semantic segmentation prediction (Chen et al. 2018a; Chen et al. 2018b). While the intermediate layers of semantic segmentation networks are as hard to interpret as are VGG features, the semantic segmentation predictions themselves are interpretable by design and spatial. This makes them suitable for fixation prediction by, e.g., feeding logits or semantic masks into a readout network. Another interesting possibility is provided by BagNets (Brendel and Bethge 2019), which have very restricted receptive fields and therefore allow a very local computation of saliency.

While the models DeepGaze I and II predict the spatial distribution of fixation locations, it is known that a lot of relevant information is contained in the sequence of fixations and in their timing. This will be the case especially when moving from free-viewing to other tasks such as visual search. Therefore, extending the models to predict sequences of fixations could lead to new insights and is already the focus of ongoing work. Finally, static images themselves are restricted and do not represent

[3] `https://deepgaze.bethgelab.org`

the majority of natural behaviour. Moving on to model gaze in dynamic stimuli, i.e., videos, will allow to explore how visual patterns and temporal changes of those patterns interact in driving eye movements. This will also allow investigation of eye movements during natural tasks such as walking (Matthis et al. 2018).

The saliency community is – due to its connection to computer vision – to a certain degree motivated by benchmarks. This means the benchmarks can guide community efforts to some extent. As a result of the work on saliency benchmarking presented above (Kümmerer et al. 2018), the author is becoming part of the MIT Saliency Benchmark team and has the chance to add some of the ideas mentioned in this outlook to the benchmark which is most widely used in the community. In a first step, the submission of fixation densities as proposed above will be implemented while still also allowing for the submission of classical saliency maps. Since we expect this to make model scores for future submission more consistent, we hope that this spurs further progress in the field of free-viewing spatial fixation prediction. Furthermore, we are already planning to add a scanpath prediction track to the benchmark, where models of scanpath prediction can be evaluated and compared on the same well-known dataset that is already used for the spatial fixation prediction. In the longer term we are considering adding benchmark tracks for predicting eye movements in different tasks than free-viewing, such as visual search.

References

Aristotle, W. D. (William David) Ross, and J. A. (John Alexander) Smith (1908–1952). *The Works of Aristotele. Translated into English under the Editorship of W.D. Ross*. In collab. with Robarts - University of Toronto. Oxford Clarendon Press. 410 pp.

Baddeley, Roland J. and Benjamin W. Tatler (Sept. 1, 2006). "High Frequency Edges (but Not Contrast) Predict Where We Fixate: A Bayesian System Identification Analysis". In: *Vision Research* 46.18, pp. 2824–2833. ISSN: 0042-6989.

Barlow, H. B. (Sept. 28, 2012). "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Communication*. Ed. by Walter A. Rosenblith. The MIT Press, pp. 216–234. ISBN: 978-0-262-51842-0.

Barthelmé, Simon et al. (Oct. 1, 2013). "Modeling Fixation Locations Using Spatial Point Processes". In: *Journal of Vision* 13.12, pp. 1–1. ISSN: 1534-7362.

Borji, A. and L. Itti (Jan. 2013). "State-of-the-Art in Visual Attention Modeling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1, pp. 185–207. ISSN: 0162-8828.

Borji, A. et al. (Dec. 2015). "Salient Object Detection: A Benchmark". In: *IEEE Transactions on Image Processing* 24.12, pp. 5706–5722. ISSN: 1057-7149.

Borji, Ali and Laurent Itti (2015). "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research". In: *CVPR 2015 workshop on "Future of Datasets"*.

Borji, Ali, D. N. Sihite, and L. Itti (Jan. 2013a). "Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study". In: *IEEE Transactions on Image Processing* 22.1, pp. 55–69. ISSN: 1057-7149, 1941-0042.

Borji, Ali, Dicky N. Sihite, and Laurent Itti (Aug. 1, 2013b). "Objects Do Not Predict Fixations Better than Early Saliency: A Re-Analysis of Einhäuser et al.'s Data". In: *Journal of Vision* 13.10, pp. 18–18. ISSN: 1534-7362.

Brendel, Wieland and Matthias Bethge (2019). "Approximating CNNs with Bag-of-Local-Features Models Works Surprisingly Well on ImageNet". In: *International Conference on Learning Representations*.

Brown, Alex Crum (Oct. 17, 1878). "Cyon's Researches on the Ear. II". In: *Nature* 18, pp. 657–659. ISSN: 1476-4687.

Bruce, Neil D. B. and John K. Tsotsos (Mar. 1, 2009). "Saliency, Attention, and Visual Search: An Information Theoretic Approach". In: *Journal of Vision* 9.3, pp. 5–5. ISSN: 1534-7362.

Buswell, Guy Thomas (1935). *How People Look at Pictures*. The University of Chicago Press.

Bylinskii, Z. et al. (2018). "What Do Different Evaluation Metrics Tell Us about Saliency Models?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. ISSN: 0162-8828.

Cerf, Moran et al. (2009). "Decoding What People See from Where They Look: Predicting Visual Stimuli from Scanpaths". In: *Attention in Cognitive Systems 2008*. Ed. by Lucas Paletta and John K. Tsotsos. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 15–26. ISBN: 978-3-642-00582-4.

Chen, L. et al. (Apr. 2018a). "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4, pp. 834–848. ISSN: 0162-8828.

Chen, Liang-Chieh et al. (2018b). "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818.

Cornia, M. et al. (Oct. 2018). "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model". In: *IEEE Transactions on Image Processing* 27.10, pp. 5142–5154. ISSN: 1057-7149.

Curcio, Christine A. and Kimberly A. Allen (1990). "Topography of Ganglion Cells in Human Retina". In: *Journal of Comparative Neurology* 300.1, pp. 5–25. ISSN: 1096-9861.

Curcio, Christine A. et al. (1990). "Human Photoreceptor Topography". In: *Journal of Comparative Neurology* 292.4, pp. 497–523. ISSN: 1096-9861.

Damiano, Claudia, John Wilder, and Dirk B. Walther (Jan. 1, 2019). "Mid-Level Feature Contributions to Category-Specific Gaze Guidance". In: *Attention, Perception, & Psychophysics* 81.1, pp. 35–46. ISSN: 1943-393X.

Deng, J. et al. (June 2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Didday, Richard L. and Michael A. Arbib (July 1, 1975). "Eye Movements and Visual Perception: A "Two Visual System" Model". In: *International Journal of Man-Machine Studies* 7.4, pp. 547–569. ISSN: 0020-7373.

Donahue, Jeff et al. (Jan. 27, 2014). "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition". In: *International Conference on Machine Learning*. International Conference on Machine Learning, pp. 647–655.

Einhäuser, Wolfgang (Aug. 1, 2013). "Objects and Saliency: Reply to Borji et Al." In: *Journal of Vision* 13.10, pp. 20–20. ISSN: 1534-7362.

Einhäuser, Wolfgang and Peter König (June 1, 2010). "Getting Real—Sensory Processing of Natural Stimuli". In: *Current Opinion in Neurobiology*. Sensory Systems 20.3, pp. 389–395. ISSN: 0959-4388.

Einhäuser, Wolfgang, Merrielle Spain, and Pietro Perona (Oct. 2, 2008). "Objects Predict Fixations Better than Early Saliency". In: *Journal of Vision* 8.14, pp. 18–18. ISSN: 1534-7362.

Emami, Mohsen and Lawrence L. Hoberock (Oct. 1, 2013). "Selection of a Best Metric and Evaluation of Bottom-up Visual Saliency Models". In: *Image and Vision Computing* 31.10, pp. 796–808. ISSN: 0262-8856.

Engbert, Ralf et al. (Jan. 1, 2015). "Spatial Statistics and Attentional Dynamics in Scene Viewing". In: *Journal of Vision* 15.1, pp. 14–14. ISSN: 1534-7362.

Erdem, Erkut and Aykut Erdem (Mar. 1, 2013). "Visual Saliency Estimation by Nonlinearly Integrating Features Using Region Covariances". In: *Journal of Vision* 13.4, pp. 11–11. ISSN: 1534-7362.

Fechner, Gustav Theodor (1860). *Elemente der Psychophysik*. In collab. with Francis A. Countway Library of Medicine. 2nd ed. Vol. 2. 2 vols. Leipzig : Breitkopf und Härtel. 594 pp.

Felleman, D. J. and D. C. Van Essen (Jan. 1, 1991). "Distributed Hierarchical Processing in the Primate Cerebral Cortex". In: *Cerebral Cortex* 1.1, pp. 1–47. ISSN: 1047-3211, 1460-2199.

Fernald, R. D. (Jan. 1, 2008). "1.02 - Evolution of Vertebrate Eyes". In: *The Senses: A Comprehensive Reference*. Ed. by Richard H. Masland et al. New York: Academic Press, pp. 9–23. ISBN: 978-0-12-370880-9.

Funke, Christina et al. (Sept. 1, 2018). "Comparing the Ability of Humans and DNNs to Recognise Closed Contours in Cluttered Images". In: *Journal of Vision* 18.10, pp. 800–800. ISSN: 1534-7362.

Galen, Charles Singer, and Wellcome Historical Medical Museum (1956). *Galen on Anatomical Procedures [Electronic Resource] : De Anatomicis Administrationibus*. In collab. with Wellcome Library. London : Oxford University Press for the Wellcome Historical Medical Museum. 328 pp.

Goferman, S., L. Zelnik-Manor, and A. Tal (Oct. 2012). "Context-Aware Saliency Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.10, pp. 1915–1926. ISSN: 0162-8828.

Haas, Benjamin de et al. (May 27, 2019). "Individual Differences in Visual Salience Vary along Semantic Dimensions". In: *Proceedings of the National Academy of Sciences*, p. 201820553. ISSN: 0027-8424, 1091-6490. pmid: 31138705.

Harel, Jonathan, Christof Koch, and Pietro Perona (2007). "Graph-Based Visual Saliency". In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, pp. 545–552.

Hayhoe, Mary (Jan. 1, 2000). "Vision Using Routines: A Functional Account of Vision". In: *Visual Cognition* 7.1-3, pp. 43–64. ISSN: 1350-6285.

Henderson, John M. (Nov. 1, 2003). "Human Gaze Control during Real-World Scene Perception". In: *Trends in Cognitive Sciences* 7.11, pp. 498–504. ISSN: 1364-6613.

Henderson, John M., Phillip A. Weeks Jr., and Andrew Hollingworth (1999). "The Effects of Semantic Consistency on Eye Movements during Complex Scene Viewing". In: *Journal of Experimental Psychology: Human Perception and Performance* 25.1, pp. 210–228. ISSN: 1939-1277(Electronic),0096-1523(Print).

Henderson, John M. et al. (June 2019). "Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach". In: *Vision* 3.2, p. 19.

Hong, Ha et al. (Apr. 2016). "Explicit Information for Category-Orthogonal Object Properties Increases along the Ventral Stream". In: *Nature Neuroscience* 19.4, pp. 613–622. ISSN: 1546-1726.

Hou, X. and L. Zhang (June 2007). "Saliency Detection: A Spectral Residual Approach". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Huang, Xun et al. (Dec. 2015). "SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, pp. 262–270. ISBN: 978-1-4673-8391-2.

Hwang, Alex D, Emily C Higgins, and Marc Pomplun (2007). "How Chromaticity Guides Visual Search in Real-World Scenes". In: *Proceedings of the 29th Annual Cognitive Science Society*. Cognitive Science Society. Austin, TX, pp. 371–378.

Hyvärinen, Aapo, Jarmo Hurri, and Patrik O. Hoyer (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Computational Imaging and Vision volume 39. Dordrecht: Springer. 448 pp. ISBN: 978-1-84882-490-4 978-1-84882-491-1.

Itti, Laurent and Pierre F. Baldi (2006). "Bayesian Surprise Attracts Human Attention". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press, pp. 547–554.

Itti, Laurent and Christof Koch (June 1, 2000). "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention". In: *Vision Research* 40.10, pp. 1489–1506. ISSN: 0042-6989.

— (Mar. 2001a). "Computational Modelling of Visual Attention". In: *Nature Reviews Neuroscience* 2.3, pp. 194–203. ISSN: 1471-0048.

— (Jan. 2001b). "Feature Combination Strategies for Saliency-Based Visual Attention Systems". In: *Journal of Electronic Imaging* 10.1, pp. 161–170. ISSN: 1017-9909, 1560-229X.

Itti, Laurent, Christof Koch, and Ernst Niebur (1998). "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11, pp. 1254–1259.

Jetley, Saumya, Naila Murray, and Eleonora Vig (June 2016). "End-to-End Saliency Mapping via Probability Distribution Prediction". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, pp. 5753–5761. ISBN: 978-1-4673-8851-1.

Judd, Tilke, Frédo Durand, and Antonio Torralba (Jan. 13, 2012). "A Benchmark of Computational Models of Saliency to Predict Human Fixations". In: *MIT Tech Report*.

Judd, Tilke et al. (2009). "Learning to Predict Where Humans Look". In: *Computer Vision, 2009 IEEE 12th International Conference On*. IEEE, pp. 2106–2113.

Kienzle, Wolf et al. (2007). "A Nonparametric Approach to Bottom-Up Visual Saliency". In: *Advances in Neural Information Processing Systems 19*. Ed. by B. Schölkopf, J. C. Platt, and T. Hoffman. MIT Press, pp. 689–696.

Kienzle, Wolf et al. (May 1, 2009). "Center-Surround Patterns Emerge as Optimal Predictors for Human Saccade Targets". In: *Journal of Vision* 9.5, pp. 7–7. ISSN: 1534-7362.

Kindermans, Pieter-Jan et al. (Feb. 15, 2018). "Learning How to Explain Neural Networks: PatternNet and PatternAttribution". In: *ICLR 2018*. ICLR 2018.

Koch, Christof and S Ullman (1985). "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry". In: *Human Neurobiology* 4, pp. 219–227.

Koch, Kristin et al. (July 25, 2006). "How Much the Eye Tells the Brain". In: *Current Biology* 16.14, pp. 1428–1434. ISSN: 0960-9822.

Krieger, Gerhard et al. (2000). *Object and Scene Analysis by Saccadic Eye-Movements: An Investigation with Higher-Order Statistics*.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., pp. 1097–1105.

Kruthiventi, S. S. S., K. Ayush, and R. V. Babu (Sept. 2017). "DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations". In: *IEEE Transactions on Image Processing* 26.9, pp. 4446–4456. ISSN: 1057-7149.

Kümmerer, Matthias, Lucas Theis, and Matthias Bethge (2015a). "Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet". In: *ICLR Workshop Track*. arXiv: 1411.1045.

Kümmerer, Matthias, Thomas S. A. Wallis, and Matthias Bethge (Dec. 29, 2015b). "Information-Theoretic Model Comparison Unifies Saliency Metrics". In: *Proceed-*

ings of the National Academy of Sciences 112.52, pp. 16054–16059. ISSN: 0027-8424, 1091-6490. pmid: 26655340.

Kümmerer, Matthias, Thomas S. A. Wallis, and Matthias Bethge (2018). "Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics". In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Lecture Notes in Computer Science. Springer International Publishing, pp. 798–814. ISBN: 978-3-030-01270-0.

Kümmerer, Matthias et al. (2017). "Understanding Low- and High-Level Contributions to Fixation Prediction". In: The IEEE International Conference on Computer Vision (ICCV), pp. 4789–4798.

Land, Michael, Neil Mennie, and Jennifer Rusted (Nov. 1, 1999). "The Roles of Vision and Eye Movements in the Control of Activities of Daily Living". In: *Perception* 28.11, pp. 1311–1328. ISSN: 0301-0066.

Land, Michael F. and Dan-Eric Nilsson (2012). *Animal Eyes*. 2nd ed. Oxford Animal Biology Series. Oxford ; New York: Oxford University Press. 271 pp. ISBN: 978-0-19-958114-6 978-0-19-958113-9.

Laubrock, Jochen, Sven Hohenstein, and Matthias Kümmerer (2018). "Attention to Comics: Cognitive Processing during the Reading of Graphic Literature". In: *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. Ed. by A. Dunst, J. Laubrock, and J. Wildfeuer. New York: Routledge, pp. 239–263.

Le Meur, Olivier, Patrick Le Callet, and Dominique Barba (Sept. 1, 2007). "Predicting Visual Fixations on Video Based on Low-Level Visual Features". In: *Vision Research* 47.19, pp. 2483–2498. ISSN: 0042-6989.

Li, Jia et al. (2015). "A Data-Driven Metric for Comprehensive Evaluation of Saliency Models". In: Proceedings of the IEEE International Conference on Computer Vision, pp. 190–198.

Li, Zhaoping (Jan. 1, 2002). "A Saliency Map in Primary Visual Cortex". In: *Trends in Cognitive Sciences* 6.1, pp. 9–16. ISSN: 1364-6613.

MacCurdy, Edward (1938). *The Notebooks of Leonardo Da Vinci.* Vol. I. London: Cape.

Mahdi, Ali and Jun Qin (Sept. 7, 2017). "DeepFeat: A Bottom Up and Top Down Saliency Model Based on Deep Features of Convolutional Neural Nets". In: arXiv: 1709.02495 [cs].

Mannan, S. K., K. H. Ruddock, and D. S. Wooding (Jan. 1, 1996). "The Relationship between the Locations of Spatial Features and Those of Fixations Made during Visual Examination of Briefly Presented Images". In: *Spatial Vision* 10.3, pp. 165–188. ISSN: 1568-5683.

— (Jan. 1, 1997). "Fixation Sequences Made during Visual Examination of Briefly Presented 2D Images". In: *Spatial Vision* 11.2, pp. 157–178. ISSN: 1568-5683.

Marcus, Gary (Jan. 2, 2018). "Deep Learning: A Critical Appraisal". In: arXiv: 1801.00631 [cs, stat].

Matthis, Jonathan Samir, Jacob L. Yates, and Mary M. Hayhoe (Apr. 23, 2018). "Gaze and the Control of Foot Placement When Walking in Natural Terrain". In: *Current Biology* 28.8, 1224–1233.e5. ISSN: 0960-9822.

Meur, Olivier Le and Thierry Baccino (Mar. 1, 2013). "Methods for Comparing Scanpaths and Saliency Maps: Strengths and Weaknesses". In: *Behavior Research Methods* 45.1, pp. 251–266. ISSN: 1554-3528.

Ouerhani, Nabil et al. (Dec. 18, 2003). "Empirical Validation of the Saliency-Based Model of Visual Attention". In: *ELCVIA Electronic Letters on Computer Vision and Image Analysis* 3.1, pp. 13–24. ISSN: 1577-5097.

Ovid (1567). *The XV Bookes of P. Ouidius Naso, Entytuled Metamorphosis*. Trans. by Arthur Golding. London: Willyam Seres.

— (1798). *Verwandlungen nach Publius Ovidius Naso*. Trans. by Heinrich Voss. Berlin: Friedrich Vieweg der Ältere.

— (1892). *Die Metamorphosen Des P. Ovidius Naso*. Ed. by Hugo Magnus. Gotha: F.A. Perthes.

Oyama, Taiki and Takao Yamanaka (July 24, 2018). "Influence of Image Classification Accuracy on Saliency Map Estimation". In: *CAAI Transactions on Intelligence Technology* 3.3, pp. 140–152. ISSN: 2468-2322.

Palmer, Stephen E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, Mass: MIT Press. 810 pp. ISBN: 978-0-262-16183-1.

Pan, Junting et al. (Jan. 4, 2017). "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks". In: arXiv: `1701.01081 [cs]`.

Parkhurst, Derrick, Klinton Law, and Ernst Niebur (Jan. 1, 2002). "Modeling the Role of Salience in the Allocation of Overt Visual Attention". In: *Vision Research* 42.1, pp. 107–123. ISSN: 0042-6989.

Parkhurst, Derrick and Ernst Niebur (June 1, 2003). "Scene Content Selected by Active Vision". In: *Spatial Vision* 16.2, pp. 125–154. ISSN: 0169-1015, 1568-5683.

Peters, R. J. and L. Itti (June 2007). "Beyond Bottom-up: Incorporating Task-Dependent Influences into a Computational Model of Spatial Attention". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8.

Peters, Robert J. et al. (Aug. 1, 2005). "Components of Bottom-up Gaze Allocation in Natural Images". In: *Vision Research* 45.18, pp. 2397–2416. ISSN: 0042-6989.

Pomplun, Marc (June 1, 2006). "Saccadic Selectivity in Complex Visual Search Displays". In: *Vision Research* 46.12, pp. 1886–1900. ISSN: 0042-6989.

Porterfield, W (1737). "An Essay Concerning the Motions of Our Eyes. Part I. Of Their External Motions". In: *Edinburgh Medical Essays and Observations* 3, pp. 160–263.

Rajashekar, Umesh, Lawrence K. Cormack, and Alan C. Bovik (June 7, 2004). "Point-of-Gaze Analysis Reveals Visual Search Strategies". In: *Human Vision and Electronic Imaging IX*. Human Vision and Electronic Imaging IX. Vol. 5292. International Society for Optics and Photonics, pp. 296–307.

Reinagel, Pamela and Anthony M. Zador (Jan. 1, 1999). "Natural Scene Statistics at the Centre of Gaze". In: *Network: Computation in Neural Systems* 10.4, pp. 341–350. ISSN: 0954-898X. pmid: 10695763.

Riche, Nicolas et al. (July 1, 2013a). "RARE2012: A Multi-Scale Rarity-Based Saliency Detection with Its Comparative Statistical Analysis". In: *Signal Processing: Image Communication* 28.6, pp. 642–658. ISSN: 0923-5965.

Riche, Nicolas et al. (Dec. 2013b). "Saliency and Human Fixations: State-of-the-Art and Study of Comparison Metrics". In: IEEE, pp. 1153–1160. ISBN: 978-1-4799-2840-8.

Rothkegel, Lars O. M. et al. (Feb. 7, 2019). "Searchers Adjust Their Eye-Movement Dynamics to Target Characteristics in Natural Scenes". In: *Scientific Reports* 9.1, p. 1635. ISSN: 2045-2322.

Schütt, Heiko H. et al. (2017). "Likelihood-Based Parameter Estimation and Comparison of Dynamical Cognitive Models." In: *Psychological Review* 124.4, pp. 505–524. ISSN: 1939-1471, 0033-295X.

Schütt, Heiko Herbert (Aug. 2, 2018). "Modelling Early Spatial Vision and Its Influence on Eye Movements in Natural Scenes". Dissertation. Universität Tübingen.

Shannon, Claude Elwood (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.1, pp. 379–423, 623–656.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (Dec. 20, 2013). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: arXiv: 1312.6034 [cs].

Simonyan, Karen and Andrew Zisserman (Sept. 4, 2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: arXiv: 1409.1556 [cs].

Speedy Neural Networks for Smart Auto-Cropping of Images (Jan. 24, 2018). URL: https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html (visited on 07/13/2019).

Steinbach, Martin J. and K. E. Money (Apr. 1, 1973). "Eye Movements of the Owl". In: *Vision Research* 13.4, pp. 889–891. ISSN: 0042-6989.

Stevens, S. S. (June 7, 1946). "On the Theory of Scales of Measurement". In: *Science* 103.2684, pp. 677–680. ISSN: 0036-8075, 1095-9203. pmid: 17750512.

Stratton, G. M. (1906). "Symmetry, Linear Illusions, and the Movements of the Eye". In: *Psychological Review* 13.2, pp. 82–96. ISSN: 1939-1471(Electronic),0033-295X(Print).

Tatler, Benjamin W., Roland J. Baddeley, and Iain D. Gilchrist (Mar. 1, 2005). "Visual Correlates of Fixation Selection: Effects of Scale and Time". In: *Vision Research* 45.5, pp. 643–659. ISSN: 0042-6989.

Torralba, Antonio et al. (2006). "Contextual Guidance of Eye Movements and Attention in Real-World Scenes: The Role of Global Features in Object Search". In: *Psychological Review* 113.4, pp. 766–786. ISSN: 1939-1471(Electronic),0033-295X(Print).

Treisman, Anne M and Garry Gelade (Jan. 1, 1980). "A Feature-Integration Theory of Attention". In: *Cognitive Psychology* 12.1, pp. 97–136. ISSN: 0010-0285.

Vig, Eleonora, Michael Dorr, and David Cox (2014). "Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2798–2805.

Vincent, Benjamin T. et al. (Aug. 1, 2009). "Do We Look at Lights? Using Mixture Modelling to Distinguish between Low- and High-Level Factors in Natural Image Viewing". In: *Visual Cognition* 17.6-7, pp. 856–879. ISSN: 1350-6285.

Wade, Nicholas J (Jan. 1, 2010). "Pioneers of Eye Movement Research". In: *i-Perception* 1.2, pp. 33–68. ISSN: 2041-6695.

Wade, Nicholas J. and Benjamin W. Tatler (May 14, 2009). "Did Javal Measure Eye Movements during Reading?" In: *Journal of Eye Movement Research* 2.5. ISSN: 1995-8692.

Wilming, Niklas et al. (Dec. 9, 2011). "Measures and Limits of Models of Fixation Selection". In: *PLOS ONE* 6.9, e24038. ISSN: 1932-6203.

Xia, Changqun et al. (June 26, 2018). "Learning a Saliency Evaluation Metric Using Crowdsourced Perceptual Judgments". In: arXiv: `1806.10257 [cs]`.

Yamins, Daniel L. K. et al. (June 10, 2014). "Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex". In: *Proceedings of the National Academy of Sciences* 111.23, pp. 8619–8624. ISSN: 0027-8424, 1091-6490. pmid: `24812127`.

Yarbus, Alfred L. (1967). *Eye Movements and Vision*. Plenum Press.

Zhang, Jianming and Stan Sclaroff (Dec. 2013). "Saliency Detection: A Boolean Map Approach". In: IEEE, pp. 153–160. ISBN: 978-1-4799-2840-8.

Zhang, Lingyun et al. (May 3, 2008). "SUN: A Bayesian Framework for Saliency Using Natural Statistics". In: *Journal of Vision* 8.7, pp. 32–32. ISSN: 1534-7362.

Zhao, Qi and Christof Koch (Mar. 2, 2011). "Learning a Saliency Map Using Fixated Locations in Natural Scenes". In: *Journal of Vision* 11.3, pp. 9–9. ISSN: 1534-7362.

Zhou, Bolei et al. (June 2016). "Learning Deep Features for Discriminative Localization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, pp. 2921–2929. ISBN: 978-1-4673-8851-1.

Appendix

Information-Theoretic Model Comparison Unifies Saliency Metrics

Matthias Kümmerer, Thomas S.A. Wallis and Matthias Bethge

Published in Proceedings of the National Academy of Sciences 112.52, pp. 16054–16059

Abstract

Learning the properties of an image associated with human gaze placement is important both for understanding how biological systems explore the environment and for computer vision applications. There is a large literature on quantitative eye movement models that seeks to predict fixations from images (sometimes termed "saliency" prediction). A major problem known to the field is that existing model comparison metrics give inconsistent results, causing confusion. We argue that the primary reason for these inconsistencies is because different metrics and models use different definitions of what a "saliency map" entails. For example, some metrics expect a model to account for image-independent central fixation bias whereas others will penalize a model that does. Here we bring saliency evaluation into the domain of information by framing fixation prediction models probabilistically and calculating information gain. We jointly optimize the scale, the center bias, and spatial blurring of all models within this framework. Evaluating existing metrics on these rephrased models produces almost perfect agreement in model rankings across the metrics. Model performance is separated from center bias and spatial blurring, avoiding the confounding of these factors in model comparison. We additionally provide a method to show where and how models fail to capture information in the fixations on the pixel level. These methods are readily extended to spatiotemporal models of fixation scanpaths, and we provide a software package to facilitate their use.

Contributions

The idea of converting existing saliency map models into probabilistic models of fixation density prediction by optimizing a pointwise monotone nonlinearity, a center bias and a blur radius was my own (center bias and blur radius have been applied earlier, but without a nonlinearity and not for converting saliency maps into probabilistic models, see Judd et al. 2012). The concept of evaluating these models using information gain was developed in joint discussions with Thomas Wallis and Matthias Bethge. I did all the experiments and analyses. The paper was written jointly by Thomas Wallis and me. All authors contributed to scientific discussions and paper revisions.

Information-theoretic model comparison unifies saliency metrics

Matthias Kümmerer[a,1], Thomas S. A. Wallis[a,b], and Matthias Bethge[a,c,d]

[a]Werner-Reichardt-Centre for Integrative Neuroscience, University Tübingen, 72076 Tübingen, Germany; [b]Department of Computer Science, University Tübingen, 72076 Tübingen, Germany; [c]Bernstein Center for Computational Neuroscience, 72076 Tübingen, Germany; and [d]Max-Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

Edited by Wilson S. Geisler, The University of Texas at Austin, Austin, TX, and approved October 27, 2015 (received for review May 28, 2015)

Learning the properties of an image associated with human gaze placement is important both for understanding how biological systems explore the environment and for computer vision applications. There is a large literature on quantitative eye movement models that seeks to predict fixations from images (sometimes termed "saliency" prediction). A major problem known to the field is that existing model comparison metrics give inconsistent results, causing confusion. We argue that the primary reason for these inconsistencies is because different metrics and models use different definitions of what a "saliency map" entails. For example, some metrics expect a model to account for image-independent central fixation bias whereas others will penalize a model that does. Here we bring saliency evaluation into the domain of information by framing fixation prediction models probabilistically and calculating information gain. We jointly optimize the scale, the center bias, and spatial blurring of all models within this framework. Evaluating existing metrics on these rephrased models produces almost perfect agreement in model rankings across the metrics. Model performance is separated from center bias and spatial blurring, avoiding the confounding of these factors in model comparison. We additionally provide a method to show where and how models fail to capture information in the fixations on the pixel level. These methods are readily extended to spatiotemporal models of fixation scanpaths, and we provide a software package to facilitate their use.

visual attention | eye movements | probabilistic modeling | likelihood | point processes

Humans move their eyes about three times/s when exploring the environment, fixating areas of interest with the high-resolution fovea. How do we determine where to fixate to learn about the scene in front of us? This question has been studied extensively from the perspective of "bottom–up" attentional guidance (1), often in a "free-viewing" task in which a human observer explores a static image for some seconds while his or her eye positions are recorded (Fig. 1*A*). Eye movement prediction is also applied in domains from advertising to efficient object recognition. In computer vision the problem of predicting fixations from images is often referred to as "saliency prediction," while to others "saliency" refers explicitly to some set of low-level image features (such as edges or contrast). In this paper we are concerned with predicting fixations from images, taking no position on whether the features that guide eye movements are "low" or "high" level.

The field of eye movement prediction is quite mature: Beginning with the influential model of Itti et al. (1), there are now over 50 quantitative fixation prediction models, including around 10 models that seek to incorporate "top–down" effects (see refs. 2–4 for recent reviews and analyses of this extensive literature). Many of these models are designed to be biologically plausible whereas others aim purely at prediction (e.g., ref. 5). Progress is measured by comparing the models in terms of their prediction performance, under the assumption that better-performing models must capture more information that is relevant to human behavior.

How close are the best models to explaining fixation distributions in static scene eye guidance? How close is the field to understanding image-based fixation prediction? To answer this question requires a principled distance metric, yet no such metric exists. There is significant uncertainty about how to compare saliency models (3, 6–8). A visit to the well-established MIT Saliency Benchmark (saliency.mit.edu) allows the reader to order models by seven different metrics. These metrics can vastly change the ranking of the models, and there is no principled reason to prefer one metric over another. Indeed, a recent paper (7) compared 12 metrics, concluding that researchers should use 3 of them to avoid the pitfalls of any one. Following this recommendation would mean comparing fixation prediction models is inherently ambiguous, because it is impossible to define a unique ranking if any two of the considered rankings are inconsistent.

Because no comparison of existing metrics can tell us how close we are, we instead advocate a return to first principles. We show that evaluating fixation prediction models in a probabilistic framework can reconcile ranking discrepancies between many existing metrics. By measuring information directly we show that the best model evaluated here (state of the art as of October 2014) explains only 34% of the explainable information in the dataset we use.

Results

Information Gain. Fixation prediction is operationalized by measuring fixation densities. If different people view the same image, they will place their fixations in different locations. Similarly, the same person viewing the same image again will make different eye movements than they did the first time. It is therefore natural to consider fixation placement as a probabilistic process.

The performance of a probabilistic model can be assessed using information theory. As originally shown by Shannon (9), information theory provides a measure, information gain, to quantify how much better a posterior predicts the data than a prior.

Significance

Where do people look in images? Predicting eye movements from images is an active field of study, with more than 50 quantitative prediction models competing to explain scene viewing behavior. Yet the rules for this competition are unclear. Using a principled metric for model comparison (information gain), we quantify progress in the field and show how formulating the models probabilistically resolves discrepancies in other metrics. We have also developed model assessment tools to reveal where models fail on the database, image, and pixel levels. These tools will facilitate future advances in saliency modeling and are made freely available in an open source software framework (www.bethgelab.org/code/pysaliency).

Author contributions: M.K., T.S.A.W., and M.B. designed research; M.K. performed research; M.K. analyzed data; and M.K., T.S.A.W., and M.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: matthias.kuemmerer@bethgelab.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1510393112/-/DCSupplemental.

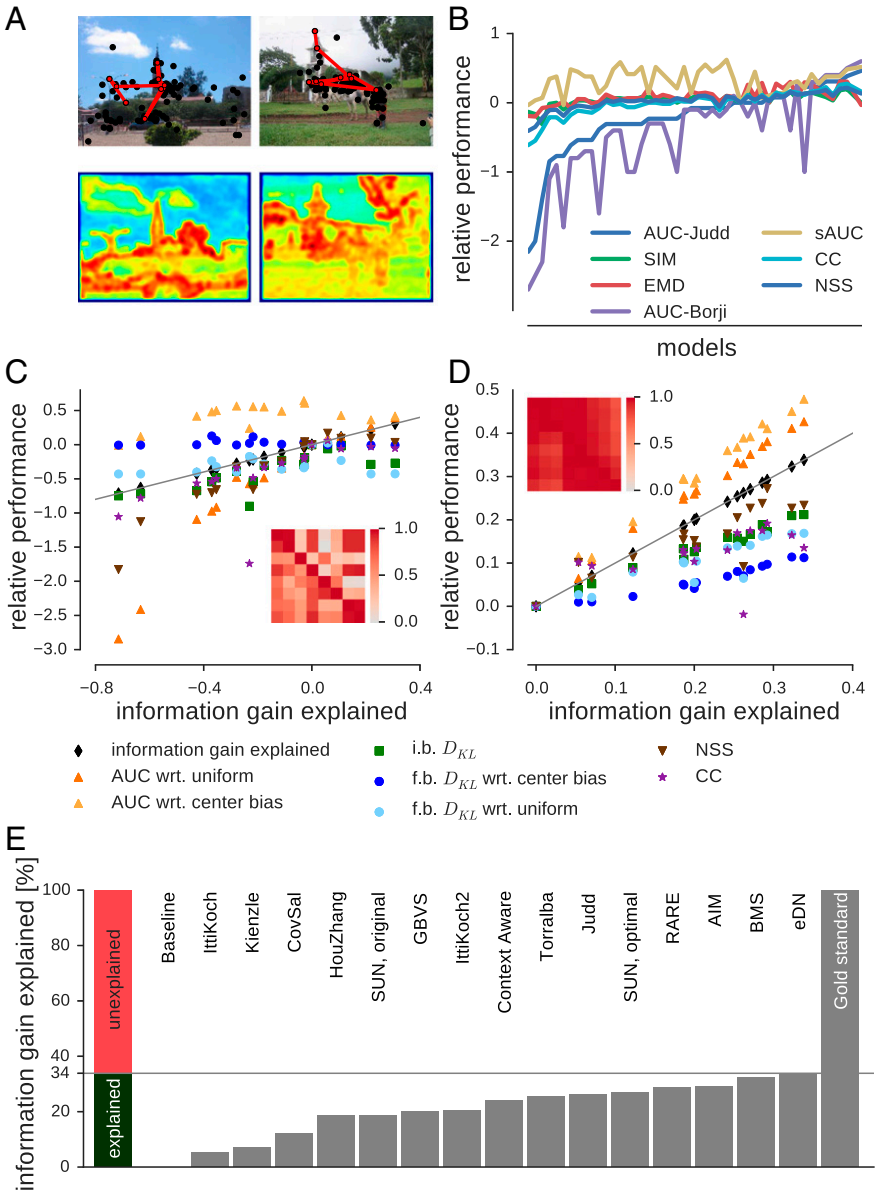


Fig. 1. Evaluation of fixation prediction models in terms of information. (*A*, *Upper*) Two example images with fixation locations (black points) and scanpaths (red). (*A*, *Lower*) Corresponding fixation predictions from an example model (AIM). Warmer colors denote more expected fixations. (*B*) Model rankings by seven metrics on the MIT Saliency Benchmark. Models are arranged along the *x* axis, ordered by "AUC-Judd" performance (highest-performing model to the right). Relative performance (*y* axis) shows each metric rescaled by baseline (0) and gold standard (1; higher is better). If the metrics gave consistent rankings, all colored lines would monotonically increase. (*C*) Different model comparison metrics evaluated on the raw model predictions (as in the MIT Benchmark), compared with information gain explained. Each color corresponds to a different metric (see key); each model forms a distinct column. The gray diagonal line shows where a metric would lie if it was linear in information. Many metrics are nonmonotonically related to information, explaining ranking inconsistencies in *B*. (*C*, *Inset*) Pearson (below diagonal) and Spearman (above diagonal) correlation coefficients in relative performance under the different metrics. (*D*) The same as *C* but for model predictions converted to probability densities, accounting for center bias and blurring. All metrics are now approximately monotonically related to information gain explained; correlations in relative performance between metrics are now uniformly high (*D*, *Inset*). Note that information gain is the only metric that is linear, because all metrics must converge to the gold standard model at (1, 1). (*E*) How close is the field to understanding image-based fixation prediction? Each model evaluated in the current study is arranged on the *x* axis in order of information gain explained. The best-performing model (eDN) explains about one-third of the information in the gold standard.

In the context of fixation prediction, this quantifies how much better an image-based model predicts the fixations on a given image than an image-independent baseline.

Information gain is measured in bits. To understand this metric intuitively, imagine a game of 20 questions in which a model is asking yes/no questions about the location of a fixation in the data. The model's goal is to specify the location of the fixation. If model A needs one question less than model B on average, then model A's information gain exceeds model B's information gain by one bit. If a model needs exactly as many questions as the baseline, then its information gain is zero bits. The number of questions the model needs is related to the concept of code length: Information gain is the difference in the average code length between a model and the baseline. Finally, information gain can also be motivated from the perspective of model comparison: It is the logarithm of the Bayes factor of the model and the baseline, divided by the number of data points. That is, if the information gain exceeds zero, then the model is more likely than the baseline.

Formally, if $\hat{p}_{\mathrm{A}}(x_i, y_i | I_i)$ is the probability that model A assigns to a fixation in location (x_i, y_i) when image I_i is viewed, and $p_{\mathbf{bl}}(x_i, y_i)$ is the probability of the baseline model for this fixation, then the information gain of model A with respect to the image-independent baseline is $(1/N)\sum_i^N \log \hat{p}_{\mathrm{A}}(x_i, y_i | I_i) - \log p_{\mathbf{bl}}(x_i, y_i)$ (to be precise, this value is the estimated expected information gain). Although information gain can be rewritten in terms of Kullback–Leibler (KL) divergence, our approach is fundamentally different from how KL divergence has previously been used to compare saliency models (*SI Text, Kullback–Leibler Divergence*).

For image-based fixation prediction, information gain quantifies the reduction in uncertainty (intuitively, the scatter of predicted fixations) in where people look, given knowledge of the image they are looking at. To capture the image-independent structure in the fixations in a baseline model, we use a 2D histogram of all fixations cross-validated between images: How well can the fixations on one image be predicted from fixations on all other images?

In addition to being principled, information gain is an intuitive model comparison metric because it is a ratio scale. Like the distance between two points, in a ratio-scaled metric "zero" means the complete absence of the quantity (in this case, no difference in code length from baseline). Second, a given change in the scale means the same thing no matter the absolute values. That is, it is meaningful to state relationships such as "the difference in information

gain between models A and B is twice as big as the difference between models C and D." Many existing metrics, such as the area under the ROC curve (AUC), do not meet these criteria.

To know how well models predict fixation locations, relative to how they could perform given intersubject variability, we want to compare model information gain to some upper bound. To estimate the information gain of the true fixation distribution, we use a nonparametric gold standard model: How well can the fixations of one subject be predicted by all other subjects' fixations? This gold standard captures the explainable information gain for image-dependent fixation patterns for the subjects in our dataset, ignoring additional task- and subject-specific information (we examine this standard further in *SI Text, Gold Standard Convergence* and Fig. S1). By comparing the information gain of models to this explainable information gain, we determine the proportion of explainable information gain explained. Like variance explained in linear Gaussian regression, this quantity tells us how much of the explainable information gain a model captures. Negative values mean that a model performs even worse than the baseline.

Reconciling the Metrics. Now that we have defined a principled and intuitive scale on which to compare models we can assess to what extent existing metrics align with this scale. In Fig. 1*B* we show the relative performance on all metrics for all saliency models listed on the MIT Saliency Benchmark website as of February 25, 2015. If all metrics gave consistent rankings, all colored lines would monotonically increase. They clearly do not, highlighting the problem with existing metrics.

Fig. 1*C* shows how the fixation prediction models we evaluate in this paper perform on eight popular fixation prediction metrics (colors) and information gain explained. As in Fig. 1*B*, the metrics are inconsistent with one another. This impression is confirmed in Fig. 1*C*, *Inset*, showing Pearson (below the diagonal) and Spearman (above the diagonal) correlation coefficients. If the metrics agreed perfectly, this plot matrix would be red. When considered relative to information gain explained, the other metrics are generally nonmonotonic and inconsistently scaled.

Why is this the case? The primary reason for the inconsistencies in Fig. 1 *B* and *C* is that both the models and the metrics use different definitions of the meaning of a saliency map (the spatial fixation prediction). For example, the "AUC wrt. uniform" metric expects the model to account for the center bias (a bias in free-viewing tasks to fixate near the center of the image), whereas "AUC wrt. center bias" expects the model to ignore the center bias (10). Therefore, a model that accounts for the center bias is penalized by AUC wrt. center bias whereas a model that ignores the center bias is penalized by AUC wrt. uniform. The rankings of these models will likely change between the metrics, even if they had identical knowledge about the image features that drive fixations.

To overcome these inconsistencies we phrased all models probabilistically, fitting three independent factors. We transformed the (often arbitrary) model scale into a density, accounted for the image-independent center bias in the dataset, and compensated for overfitting by applying spatial blurring. We then reevaluated all metrics on these probabilistic models. This yields far more consistent outcomes between the metrics (Fig. 1*D*). The metrics are now monotonically related to information gain explained, creating mostly consistent model rankings (compare the correlation coefficient matrices in Fig. 1 *C* and *D*, *Insets*).

Nevertheless, Fig. 1*D* also highlights one additional, critical point. All model relative performances must reconverge to the gold standard performance at (1, 1). That all existing metrics diverge from the unity diagonal means that these metrics remain nonlinear in information gain explained. This creates problems in comparing model performance. If we are interested in the information that is explained, then information gain is the only metric that can answer this question in an undistorted way.

How Close Is the Field to Understanding Image-Based Fixation Prediction? We have shown above that a principled definition of fixation prediction serves to reconcile ranking discrepancies between existing metrics. Information gain explained also tells us how much of the information in the data is accounted for by the models. That is, we can now provide a principled answer to the question, "How close is the field to understanding image-based fixation prediction?".

Fig. 1*E* shows that the best-performing model we evaluate here, ensemble of deep networks (eDN), accounts for about 34% of the explainable information gain, which is 1.21 bits per fixation (bits/fix) in this dataset (*SI Text, Model Performances as Log-Likelihoods* and Fig. S2). These results highlight the importance of using an intuitive evaluation metric: As of October 2014, there remained a significant amount of information that image-based fixation prediction models could explain but did not.

Information Gain in the Pixel Space. The probabilistic framework for model comparison we propose above has an additional advantage over existing metrics: The information gain of a model can be evaluated at the level of pixels (Table S1). We can examine where and by how much model predictions fail.

This procedure is schematized in Fig. 2. For an example image, the model densities show where the model predicts fixations to occur in the given image (Fig. 2*A*). This prediction is then divided by the baseline density, yielding a map showing where and by how much the model believes the fixation distribution in a given image is different from the baseline ("image-based prediction"). If the ratio is greater than one, the model predicts there should be more fixations than the center bias expects. The "information gain" images in Fig. 2 quantify how much a given pixel contributes to the model's performance relative to the baseline (code length saved in bits/fix). Finally, the difference between the model's information gain and the possible information gain, estimated by the gold standard, is shown in "difference to real information gain": It shows where and how much (bits) the model wastes information that could be used to describe the fixations more efficiently.

The advantage of this approach is that we can see not only how much a model fails (on an image or dataset level), but also exactly where it fails, in individual images. This can be used to make informed decisions about how to improve fixation prediction models. In Fig. 2*B*, we show an example image and the performance of the three best-performing models [eDN, Boolean map-based saliency (BMS), and attention based on information maximization (AIM)]. The pixel space information gains show that the eDN model correctly assigns large density to the boat, whereas the other models both underestimate the saliency of the boat.

To extend this pixel-based analysis to the level of the entire dataset, we display each image in the dataset according to its possible information gain and the percentage of that information gain explained by the eDN model (Fig. 3). In this space, points to the bottom right represent images that contain a lot of explainable information in the fixations that the model fails to capture. Points show all images in the dataset, and for a subset of these we have displayed the image itself. The images in the bottom right of the plot tend to contain human faces. See *SI Text, Pixel-Based Analysis on Entire Dataset* for an extended version of this analysis including pixel-space information gain plots and a model comparison.

Discussion

Predicting where people look in images is an important problem, yet progress has been hindered by model comparison uncertainty. We have shown that phrasing fixation prediction models probabilistically and appropriately evaluating their performance cause the disagreement between many existing metrics to disappear. Furthermore, bringing the model comparison problem into the principled domain of information allows us to assess the progress of the field, using an intuitive distance metric. The best-performing model we evaluate here (eDN) explains about

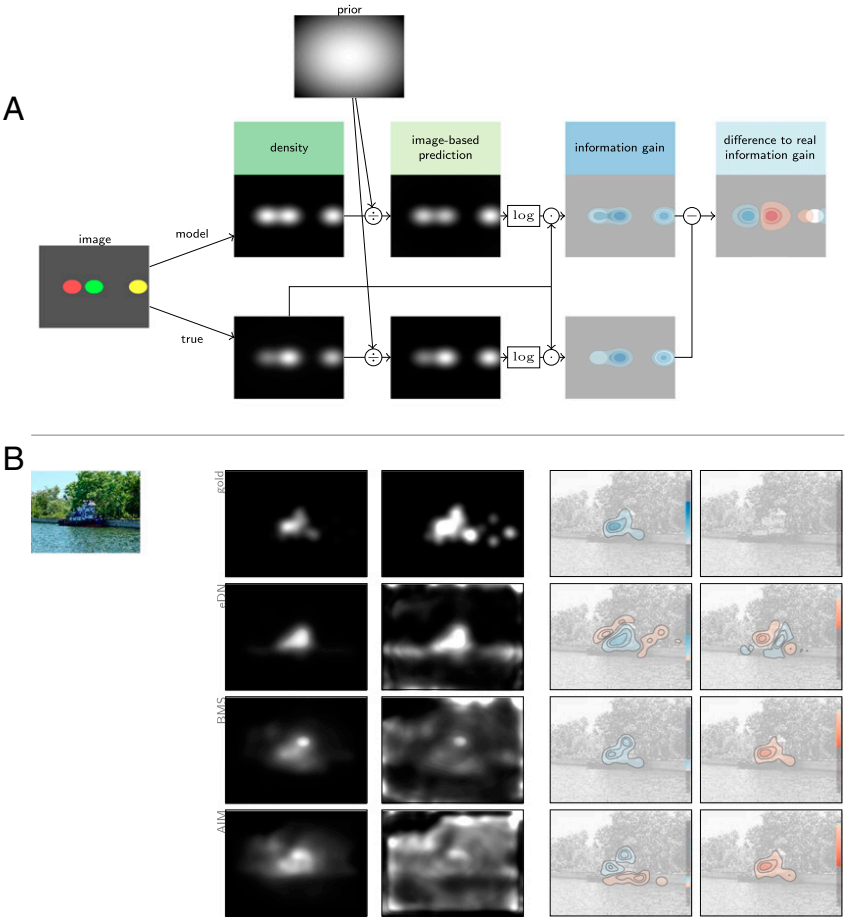


Fig. 2. Calculation of information gain in the pixel space. (*A*) For the hypothetical example image shown (*Left*), hypothetical fixation densities of the gold standard ("true") and model predictions are shown in the "density" column. These are divided by the baseline model (prior) to get the "image-based prediction" map. Both maps are then log-transformed and multiplied by the gold standard density to calculate information gain for each pixel. Subtracting the gold standard information gain from the model's information gain yields a difference map of the possible information gain: that is, where and by how much the model's predictions fail. In this case, the model overestimates (blue contours) the fixation density in the left (red) spot in the image, underestimates (red contours) the center (green) spot, and predicts the rightmost (yellow) spot almost perfectly. (*B*) For an example image from the Judd dataset (*Left*), the pixel space information gains are shown as in *A* for the gold standard (first row), eDN (second row), BMS (third row), and AIM (fourth row). eDN performs best for the image overall (3.12 bits/fix compared with 2.59 bits/fix and 2.28 bits/fix). By examining the pixel space information gains, we see this is because it correctly assigns large density to the boat, whereas the other models both underestimate the saliency of the boat. For the eDN model, the difference plot shows that it slightly overestimates the saliency of the front of the boat relative to the back.

34% of the explainable information gain. More recent model submissions to the MIT Benchmark have significantly improved on this number (e.g., ref. 11). This highlights one strength of information gain as a metric: As model performance begins to approach the gold standard, the nonlinear nature of other metrics (e.g., AUC) causes even greater distortion of apparent progress. The utility of information gain is clear.

To improve models it is useful to know where in images this unexplained information is located. We developed methods not only to assess model performance on a database level, but also to show where and by how much model predictions fail in individual images, on the pixel level (Figs. 2 and 3). We expect these tools will be useful for the model development community, and we provide them in our free software package.

Many existing metrics can be understood as evaluating model performance on a specific task. For example, the AUC is the performance of a model in a two-alternative forced-choice (2AFC) task, "Which of these two points was fixated?". If this is the task of interest to the user, then AUC is the right metric. Our results do not show that any existing metric is wrong. The metrics do not differ because they capture fundamentally different properties of fixation prediction, but mainly because they do not agree on the definition of "saliency map." The latter case requires only minor adjustments to move the field forward. This also serves to explain the three metric groups found by Riche et al. (7): One group contains among others AUC with uniform nonfixation distribution (called AUC-Judd by Riche), another group contains AUC with center bias nonfixation distribution (AUC-Borji), and the last group contains image-based KL divergence (KL-Div). We suggest that the highly uncorrelated results of these three groups are due to the fact that one group penalizes models without center bias, another group penalizes models with center bias, and the last group depends on the absolute saliency values. Compensating for these factors appropriately makes the metric results correlate almost perfectly.

Although existing metrics are appropriate for certain use cases, the biggest practical advantage in using a probabilistic framework is its generality. First, once a model is formulated in a probabilistic way many kinds of "task performance" can be calculated, depending on problems of applied interest. For example, we might be interested in whether people will look at an advertisement on a website or whether the top half of an image is more likely to be fixated than the bottom half. These predictions are a simple matter of integrating over the probability distribution. This type of evaluation is not well defined for other metrics that do not define the scale of saliency values. Second, a probabilistic model allows the examination of any statistical moments of the probability distribution that might be of practical interest. For example, Engbert et al. (12) examine the properties of second-order correlations between fixations in scanpaths. Third, information gain allows the contribution of different factors in explaining data variance to be quantified. For example, it is possible to show how much the center bias contributes to explaining fixation data independent of image-based saliency contributions (10) (*SI Text, Model Performances as Log-Likelihoods* and Fig. S2). Fourth, the information gain is differentiable in the probability density, allowing models to be numerically optimized using gradient techniques. In fact, the optimization is equivalent to maximum-likelihood estimation, which is ubiquitously used for density estimation and fulfills a few simple desiderata for density metrics (13). In some cases other loss functions may be preferable.

If we are interested in understanding naturalistic eye movement behavior, free viewing static images is not the most representative condition (14–18). Understanding image-based fixation behavior is not only a question of "where?", but of "when?" and "in what order?". It is the spatiotemporal pattern of fixation selection that is

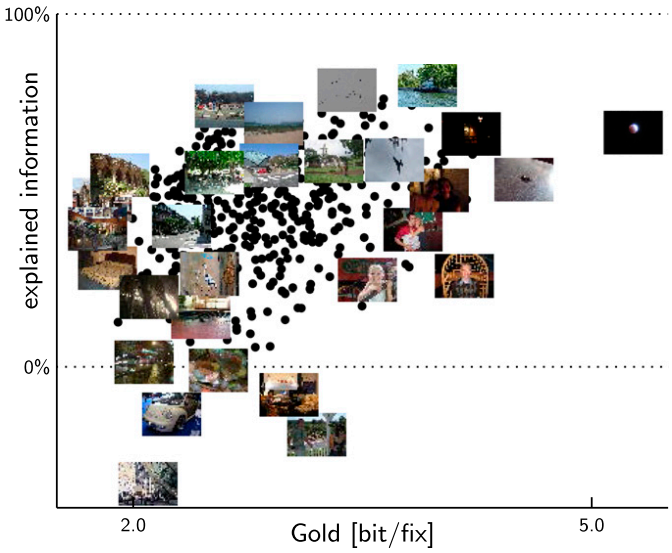


Fig. 3. Distribution of information gains and explained information (both relative to a uniform baseline model) over all images in the dataset for the eDN model. Each black circle represents an image from the dataset. These plots allow model performance to be assessed on all images in the dataset. Points in the lower right of the scatterplots are images where a lot of information could be explained but is not; these are where the model could be best improved for a given dataset. See Fig. S5 for an extended version of this plot, including an additional model and pixel-space information gain plots showing where the model predictions fail in individual images.

increasingly of interest to the field, rather than purely spatial predictions of fixation locations. The probabilistic framework we use in this paper (10, 19) is easily extended to study spatiotemporal effects, by modeling the conditional probability of a fixation given previous fixations (*Materials and Methods* and ref. 12).

Accounting for the entirety of human eye movement behavior in naturalistic settings will require incorporating information about the task, high-level scene properties, and mechanistic constraints on the eye movement system (12, 15–17, 20–22). Our gold standard contains the influence of high-level (but still purely image-dependent) factors to the extent that they are consistent across observers. Successful image-based fixation prediction models will therefore need to use such higher-level features, combined with task-relevant biases, to explain how image features are associated with the spatial distribution of fixations over scenes.

Materials and Methods

Image Dataset and Fixation Prediction Models. We use a subset of a popular benchmarking dataset (MIT-1003) (23) to compare and evaluate fixation prediction models. We used only the most common image size (1,024 × 768 px), resulting in 463 images included in the evaluation. We have verified our results in a second dataset of human fixations (24) (*SI Text, Kienzle Dataset* and Fig. S3).

We evaluated all models considered in ref. 25 and the top-performing models added to the MIT Saliency Benchmarking website (saliency.mit.edu) up to October 2014. For all models, the original source code and default parameters have been used unless stated otherwise. The included models are Itti et al. (1) [here, two implementations have been used: one from the Saliency Toolbox and the variant specified in the graph-based visual saliency (GBVS) paper], Torralba et al. (26), GBVS (27), saliency using natural statistics (SUN) (28) (for "SUN, original" we used a scale parameter of 0.64, corresponding to the pixel size of 2.3′ of visual angle of the dataset used to learn the filters; for "SUN, optimal" we did a grid search over the scale parameter; this resulted in a scale parameter of 0.15), Kienzle et al. (24, 29) (patch size 195 pixels corresponding to their reported optimal patch size of 5.4°), Hou and Zhang (30), AIM (31), Judd et al. (23), context-aware saliency (32, 33), visual saliency estimation by nonlinearly integrating features using region covariances (CovSal) (34), multiscale rarity-based saliency detection (RARE2012) (35), BMS (5, 36), and finally eDN (37). Table S2 specifies the source code used for each model.

Information Gain and Comparison Models. Given fixations (x_i, y_i) on images I_i and predictions of a probabilistic model $\hat{p}(x, y|I)$, the average log-likelihood for the data is $(1/N)\sum_i^N \log \hat{p}(x_i, y_i|I_i)$ and the information gain with respect to an image-independent baseline model $p_{\mathbf{bl}}(x, y)$ is

$$IG(\hat{p}\|p_{\mathbf{bl}}) = \frac{1}{N}\sum_i^N \log \hat{p}(x_i, y_i|I_i) - \log p_{\mathbf{bl}}(x_i, y_i).$$

The explainable information gain is the information gain of the gold standard model $IG(p_{\mathbf{gold}}\|p_{\mathbf{bl}})$. Finally, explainable information gain explained (called simply "information gain explained" in the paper) for model $\hat{p}$ is $IG(\hat{p}\|p_{\mathbf{bl}})/IG(p_{\mathbf{gold}}\|p_{\mathbf{bl}})$.

In this paper we use the logarithm to base 2, meaning that information gain is in bits. Model comparison within the framework of likelihoods is well defined and the standard of any statistical model comparison enterprise.

The baseline model is a 2D histogram model with a uniform regularization (to avoid zero bin counts) cross-validated between images (trained on all fixations for all observers on all other images). That is, reported baseline performance used all fixations from other images to predict the fixations for a specific image: It captures the image-independent spatial information in the fixations. Bin width and regularization parameters were optimized by gridsearch. If a saliency model captured all of the behavioral fixation biases but nothing about what causes parts of an image to attract fixations, it would do as well as the baseline model.

Fixation preferences that are inconsistent between observers are by definition unpredictable from fixations alone. If we have no additional knowledge about interobserver differences, the best predictor of an observer's fixation pattern on a given image is therefore to average the fixation patterns from all other observers and add regularization. This is our gold standard model. It was created by blurring the fixations with a Gaussian kernel and including a multiplicative center bias (*Phrasing Saliency Maps Probabilistically*), learned by leave-one-out cross-validation between subjects. That is, the reported gold standard performance (for information gain and AUCs) always used only fixations from other subjects to predict the fixations of a specific subject, therefore giving a conservative estimate of the explainable information. It accounts for the amount of information in the spatial structure of fixations to a given image that can be explained while averaging over the biases of individual observers. This model is the upper bound on prediction in the dataset (see ref. 8 for a thorough comparison of this gold standard and other upper bounds capturing different constraints).

Existing Metrics. We evaluate the models on several prominent metrics (Fig. 1*C*): AUC wrt. uniform, AUC wrt. center bias, image-based KL divergence, fixation-based KL divergence, normalized scanpath saliency, and correlation coefficient. For details on these metrics, their implementation, and their relationship to information gain see *SI Text, Existing Metrics*, Fig. S4, and Table S3.

Phrasing Saliency Maps Probabilistically. We treat the normalized saliency map [$s(x, y|I)$ denotes the saliency at point (x, y) in image I] as the predicted gaze density for the fixations: $\hat{p}(x, y|I) \propto s(x, y|I)$. This definition marginalizes over previous fixation history and fixation timings, which are not included in any evaluated models.

Because many of the models were optimized for AUC, and because AUC is invariant to monotonic transformations whereas information gain is not, we cannot simply compare the models' raw saliency maps to one another. The saliency map for each model was therefore transformed by a pointwise monotonic nonlinearity that was optimized to give the best log-likelihood for that model. This corresponds to picking the model with the best log-likelihood from all models that are equivalent (under AUC) to the original model.

Every saliency map was jointly rescaled to range from 0 to 1 (i.e., over all images at once, not per image, keeping contrast changes from image to image intact).

Then a Gaussian blur with radius σ was applied that allowed us to compensate in models that make overly precise, confident predictions of fixation locations (25).

Next, the pointwise monontonic nonlinearity was applied. This nonlinearity was modeled as a continuous piecewise linear function supported in 20 equidistant points x_i between 0 and 1 with values y_i with $0 \le x_0 \le \ldots \le x_{19}$: $p_{\text{nonlin}}(x, y) \propto f_{\text{nonlin}}(s(x, y))$ with $f_{\text{nonlin}}(x) = (y_{i+1} - y_i)/(x_{i+1} - x_i)(x - x_i) + y_i$ for $x_i \le x \le x_{i+1}$.

Finally, we included a center bias term (accounting for the fact that human observers tend to look toward the center of the screen) (25).

The center bias was modeled as $p_{\text{cb}}(x, y) \propto f_{\text{cb}}(d(x, y))p_{\text{nonlin}}(x, y)$.

Here, $d(x, y) = \sqrt{(x - x_{\mathbf{c}})^2 + \alpha(y - y_{\mathbf{c}})^2}/d_{\max}$ is the normalized distance of (x, y) to the center of the image $(x_{\mathbf{c}}, y_{\mathbf{c}})$ with eccentricity α, and $f_{\text{cb}}(d)$ is again a continuous piecewise linear function that was fitted in 12 points.

All parameters were optimized jointly, using the L-BFGS SLSQP algorithm from scipy.optimize (38).

Evaluating the Metrics on Probabilistic Models. To evaluate metrics described above on the probabilistic models (the results shown in Fig. 1*D*), we used the log-probability maps as saliency maps. All other computations were as described above. An exception is the image-based KL divergence. Because this metric operates on probability distributions, our model predictions were used directly.

The elements of Fig. 2 are calculated as follows: First, we plot the model density for each model (column "density" in Fig. 2). This is $\hat{p}(x, y|I)$. Then we plot the model's image-based prediction $\hat{p}(x, y|I)/p_{\text{bl}}(x, y)$. It tells us where and how much the model believes the fixation distribution in a given image is different from the prior $p(x, y)$ (baseline).

Now we separate the expected information gain (an integral over space) into its constituent pixels, as $p_{\text{gold}}(x, y|I)\log(\hat{p}(x, y|I)/p_{\text{bl}}(x, y))$ [using the gold standard as an approximation for the real distribution $p(x, y|I)$]. Weighting by the gold standard $p_{\text{gold}}(x, y|I)$ results in a weaker penalty for incorrect predictions in areas where there are fewer fixations. Finally, the last column in Fig. 2 shows the difference between the model's information gain and the possible information gain, estimated by the gold standard, resulting in $p(x, y|I)\log(\hat{p}(x, y|I)/p(x, y|I))$.

Note that this detailed evaluation is not possible with existing saliency metrics (Table S1).

Generalization to Spatiotemporal Scanpaths. The models we consider in this paper are purely spatial: They do not include any temporal dependencies. A complete understanding of human fixation selection would require an understanding of spatiotemporal behavior, that is, scanpaths. The model adaptation and optimization procedure we describe above can be easily generalized to account for temporal effects. For details see *SI Text, Generalization to Spatiotemporal Scanpaths*.

ACKNOWLEDGMENTS. We thank Lucas Theis for his suggestions and Eleonora Vig and Benjamin Vincent for helpful comments on an earlier draft of this manuscript. We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG) through the priority program 1527, research Grant BE 3848/2-1. T.S.A.W. was supported by a Humboldt Postdoctoral Fellowship from the Alexander von Humboldt Foundation. We further acknowledge support from the DFG through the Werner-Reichardt Centre for Integrative Neuroscience (EXC307) and from the BMBF through the Bernstein Center for Computational Neuroscience (FKZ: 01GQ1002).

1. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell* 20(11):1254–1259.
2. Borji A, Itti L (2013) State-of-the-art in visual attention modeling. *IEEE Trans Pattern Anal Mach Intell* 35(1):185–207.
3. Borji A, Sihite DN, Itti L (2013) Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study. *IEEE Trans Image Processing* 22(1):55–69.
4. Borji A, Tavakoli HR, Sihite DN, Itti L (2013) Analysis of scores, datasets, and models in visual saliency prediction. *Proceedings of the 2013 IEEE International Conference on Computer Vision* (IEEE Computer Society, Washington, DC), pp 921–928.
5. Zhang J, Sclaroff S (2013) Saliency detection: A Boolean map approach. *Proceedings of the 2013 IEEE International Conference on Computer Vision* (IEEE Computer Society, Washington, DC), pp 153–160.
6. Bruce ND, Wloka C, Frosst N, Rahman S, Tsotsos JK (2015) On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Res* 116(2):92–112.
7. Riche N, Duvinage M, Mancas M, Gosselin B, Dutoit T (2013) Saliency and human fixations: State-of-the-art and study of comparison metrics. *Proceedings of the 2013 IEEE International Conference on Computer Vision* (IEEE Computer Society, Washington, DC), pp 1153–1160.
8. Wilming N, Betz T, Kietzmann TC, König P (2011) Measures and limits of models of fixation selection. *PLoS One* 6(9):e24038.
9. Shannon CE, Weaver W (1949) *The Mathematical Theory of Communication* (Univ of Illinois Press, Urbana, IL).
10. Barthelmé S, Trukenbrod H, Engbert R, Wichmann F (2013) Modeling fixation locations using spatial point processes. *J Vis* 13(12):1–34.
11. Kümmerer M, Theis L, Bethge M (2015) Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. *arXiv*:1411.1045.
12. Engbert R, Trukenbrod HA, Barthelmé S, Wichmann FA (2015) Spatial statistics and attentional dynamics in scene viewing. *J Vis* 15(1):14.
13. Bernardo JM (1979) Reference posterior distributions for Bayesian inference. *J R Stat Soc Ser B Methodol* 41(2):113–147.
14. Tatler BW, Vincent BT (2008) Systematic tendencies in scene viewing. *J Eye Mov Res* 2(2):1–18.
15. Tatler BW, Vincent BT (2009) The prominence of behavioural biases in eye guidance. *Vis Cogn* 17(6–7):1029–1054.
16. Tatler BW, Hayhoe MM, Land MF, Ballard DH (2011) Eye guidance in natural vision: Reinterpreting salience. *J Vis* 11(5):5.
17. Ehinger KA, Hidalgo-Sotelo B, Torralba A, Oliva A (2009) Modeling search for people in 900 scenes: A combined source model of eye guidance. *Vis Cogn* 17(6–7):945–978.
18. Dorr M, Martinetz T, Gegenfurtner KR, Barth E (2010) Variability of eye movements when viewing dynamic natural scenes. *J Vis* 10(10):28.
19. Vincent BT, Baddeley RJ, Correani A, Troscianko T, Leonards U (2009) Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Vis Cogn* 17(6–7):856–879.
20. Najemnik J, Geisler WS (2009) Simple summation rule for optimal fixation selection in visual search. *Vision Res* 49(10):1286–1294.
21. Najemnik J, Geisler WS (2005) Optimal eye movement strategies in visual search. *Nature* 434(7031):387–391.
22. Morvan C, Maloney LT (2012) Human visual search does not maximize the post-saccadic probability of identifying targets. *PLoS Comput Biol* 8(2):e1002342.
23. Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. *Proceedings of the 2009 IEEE International Conference on Computer Vision* (IEEE Computer Society, Washington DC), pp 2106–2113.
24. Kienzle W, Franz MO, Schölkopf B, Wichmann FA (2009) Center-surround patterns emerge as optimal predictors for human saccade targets. *J Vis* 9(5):1–15.
25. Judd T, Durand F, Torralba A *A Benchmark of Computational Models of Saliency to Predict Human Fixations*. Cambridge, MA: MIT Computer Science and Artificial Intelligence Laboratory; 2012. Report No.: MIT-CSAIL-TR-2012-001.
26. Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychol Rev* 113(4):766–786.
27. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. *Advances in Neural Information Processing Systems 2006*, pp 545–552.
28. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: A Bayesian framework for saliency using natural statistics. *J Vis* 8(7):1–20.
29. Kienzle W, Wichmann FA, Schölkopf B, Franz MO (2007) *A Nonparametric Approach to Bottom-Up Visual Saliency*, eds Schölkopf B, Platt J, Hoffman T (MIT Press, Cambridge, MA), pp 689–696.
30. Hou X, Zhang L (2007) Saliency detection: A spectral residual approach. *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC), pp 1–8.
31. Bruce N, Tsotsos J (2009) Saliency, attention, and visual search: An information theoretic approach. *J Vis* 9(3): 1–24.
32. Goferman S, Zelnik-Manor L, Tal A (2010) Context-aware saliency detection. *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC), pp 2376–2383.
33. Goferman S, Zelnik-Manor L, Tal A (2012) Context-aware saliency detection. *IEEE Trans Pattern Anal Mach Intell* 34(10):1915–1926.
34. Erdem E, Erdem A (2013) Visual saliency estimation by nonlinearly integrating features using region covariances. *J Vis* 13(4):11–11.
35. Riche N, et al. (2013) RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Process Image Commun* 28(6):642–658.
36. Zhang J, Sclaroff S (2015) Exploiting surroundedness for saliency detection: A Boolean map approach. *IEEE Trans Pattern Anal Mach Intell*, in press.
37. Vig E, Dorr M, Cox D (2014) Large-scale optimization of hierarchical features for saliency prediction in natural images. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC), pp 2798–2805.
38. Jones E, Oliphant T, Peterson P, others (2001) SciPy: Open source scientific tools for Python. Available at www.scipy.org/. Accessed November 24, 2014.
39. Tatler BW, Baddeley RJ, Gilchrist ID (2005) Visual correlates of fixation selection: Effects of scale and time. *Vision Res* 45(5):643–659.
40. Itti L, Baldi PF (2005) Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems 2005*, pp 547–554.
41. Itti L, Baldi P (2005) A Principled Approach to Detecting Surprising Events in Video A principled approach to detecting surprising events in video. *Proceedings of the 2005 IEEE Conference on Computer Vision and Patter Recognition* (IEEE Computer Society, Washington, DC) Vol 1, pp 631–637.
42. Itti L (2005) Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Vis Cogn* 12(6):1093–1123.
43. Baldi P, Itti L (2005) Attention: Bits versus wows. *Proceedings of the 2005 IEEE International Conference on Neural Networks and Brain*, Vol 1, pp PL-56–PL-61.
44. Baldi P, Itti L (2010) Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Netw* 23(5):649–666.
45. Wang W, Wang Y, Huang Q, Gao W (2010) Measuring visual saliency by site entropy rate. *Proceedings of the 2010 IEEE Conferernce on Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC), pp 2368–2375.
46. Rajashekar U, Cormack LK, Bovik AC (2004) Point-of-gaze analysis reveals visual search strategies. *SPIE Proceedings Vol. 5292: Human Vision and Electronic Imaging IX* (International Society for Optics and Photonics, Bellingham WA), pp 296–306.
47. Le Meur O, Le Callet P, Barba D (2007) Predicting visual fixations on video based on low-level visual features. *Vision Res* 47(19):2483–2498.
48. Le Meur O, Baccino T (2013) Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behav Res Methods* 45(1):251–266.
49. Engbert R, Trukenbrod HA, Barthelmé S, Wichmann FA (2014) Spatial statistics and attentional dynamics in scene viewing. *arXiv*:1405.3270v2.

Supporting Information

Kümmerer et al. 10.1073/pnas.1510393112

SI Text

Model Performances as Log-Likelihoods

In Fig. S2, we report the average log-likelihoods of the tested models. All reported log-likelihoods are relative to the maximum entropy model predicting a uniform fixation distribution.

The gold standard model shows that the total mutual information between the image and the spatial structure of the fixations amounts to 2.1 bits/fix. To give another intuition for this number, a model that would for every fixation always correctly predict the quadrant of the image in which it falls would also have a log-likelihood of 2 bits/fix.

The lower-bound model is able to explain 0.89 bits/fix of this mutual information. That is, 42% of the information in spatial fixation distributions can be accounted for by behavioral biases (e.g., the bias of human observers to look at the center of the image).

The eDN model performs best of all of the saliency models compared, with 1.29 bits/fix, capturing 62% of the total mutual information. It accounts for 19% more than the lower-bound model or 34% of the possible information gain (1.21 bits/fix) between baseline and gold standard.

Fig. S2 also shows performances where only a subset of our optimization procedure was performed, allowing the contribution of different stages of our optimization to be assessed. Considering only model performance (i.e., without also including center bias and blur factors; the pink sections in Fig. S2) shows that many of the models perform worse than the lower-bound model. This means that the center bias is more important than the portion of image-based saliency that these models do capture (39). Readers will also note that the center bias and blurring factors account for very little of the performance of the Judd model and the eDN model relative to most other models. This is because these models already include a center bias that is optimized for the Judd dataset.

Gold Standard Convergence

The absolute performance level of the gold standard (the estimate of explainable information gain) depends on the size of the dataset. With fewer data points, the true gold standard performance will be underestimated because more regularization is required to generalize across subjects. With enough data, our estimate of the gold standard will converge to the true gold standard performance.

To examine the convergence of our gold standard estimate in the dataset we use, we repeated our cross-validation procedure using, for each subject, only a subset of the other 14 subjects. Fig. S1 shows the average gold standard performance (in bits per fixation) as a function of the number of other subjects used for cross-validation. The curve rapidly increases and then begins to flatten as we reach the full dataset size. This result indicates that more data would be required to gain a precise estimate of the true gold standard performance. Nevertheless, that the curve begins to saturate indicates that more data are unlikely to qualitatively change the results we report here. If anything, the gold standard performance would increase, reducing our estimate of the explainable information gain explained (34%) even further.

Kienzle Dataset

We repeated the full evaluation on the dataset of Kienzle et al. (24). It consists of 200 grayscale images of size $1{,}024 \times 678$ px and 15 subjects. This dataset is of special interest, as the authors removed the photographer bias by using random crops from larger images. The results are shown in Fig. S3.

In this dataset, with 22% even less of the possible information gain is covered by the best model (here, GBVS. Note that we were not able to include eDN into this comparision, as the source code was not yet released at the time of the analysis). Removing the photographer bias leads to a smaller contribution (34%) of the nonparametric model compared with the increase in log-likelihood by saliency map-based models. The possible information gain is with 0.92 bits/fix smaller than for the Judd dataset (1.21 bits/fix) There are multiple possible reasons for this. Primarily, this dataset contains no pictures of people, but a lot of natural images. In addition, the images are in grayscale.

Pixel-Based Analysis on Entire Dataset

In Fig. S5, we display each image in the dataset according to its possible information gain and the percentage of that information gain explained by the model. In this space, points to the bottom right represent images that contain a lot of explainable information in the fixations that the model fails to capture. Points show all images in the dataset, and for a subset of these we have displayed the image itself (Fig. S5 *A* and *C*) and the information gain difference to the gold standard (Fig. S5 *B* and *D*). For the eDN model (Fig. S5 *A* and *B*), the images in the bottom right of the plot tend to contain human faces. The Judd model contains an explicit face detection module, and as can be seen in Fig. S5 *C* and *D*, it tends to perform better on these images. In terms of the whole dataset, however, the eDN model performs better on images with a moderate level of explainable information (around 3 bits/fix).

Existing Metrics

We evaluate the models on several prominent metrics. The area under the curve (AUC) metrics are the most widely used. They calculate the performance of the model when using the saliency map as classifier score in a two-alternative forced-choice (2AFC) task where the model has to separate fixations from nonfixations. There are several variants of AUC scores, differing by the nonfixation distribution used and in approximations to speed up computation. We use all sample values as thresholds, therefore using no approximation. AUC wrt. uniform uses a uniform nonfixation distribution, i.e., the full saliency map as nonfixations [this corresponds to "AUC-Judd" in the MIT Benchmark (25)]. AUC wrt. center bias uses the fixations from all other images as nonfixations, thus capturing structure unrelated to the image [behavioral biases, primarily center bias (3, 4, 39)]. This corresponds to "sAUC" in the MIT benchmark ("shuffled AUC").

Confusingly, there are two completely independent measures referred to as "Kullback–Leibler divergence" used in the saliency literature. We discuss the precise definitions of these metrics and their relationship to information gain as used in this paper in *SI Text, KL Divergence*. What we refer to as image-based Kullback–Leibler (KL) divergence treats the saliency maps as 2D probability distributions and calculates the KL divergence between the model distribution and an approximated true distribution (8, 39). To compute this metric, the saliency maps were rescaled to have a maximum of 1 and a minimum of at least 10^{-20} over all maps. The saliency maps are then divided by the sum of their values to convert them into probability distributions. We use our gold standard as the true distribution.

The other variant of KL divergence, here called fixation-based KL divergence, calculates the KL divergence between the

distribution of saliency values at fixations and the distribution of saliency values at some choice of nonfixations (40). We use histograms with 10 bins to calculate the KL divergence. For the nonfixations, we use all saliency values [fixation-based (f.b.) D_{KL} wrt. uniform] or the saliency values at the fixation locations of the fixations from all other images (f.b. D_{KL} wrt. center bias).

Normalized scanpath saliency (NSS) normalizes each saliency map to have zero mean and unit variance and then takes the mean saliency value over all fixations.

The correlation coefficient (CC) metric normalizes the saliency maps of the model and the saliency maps of the approximated true distribution (gold standard) to have zero mean and unit variance and then calculates the correlation coefficient of these maps over all pixels.

Detailed Comparison of Log-Likelihoods, AUC, and KL Divergence

Here we consider the relationship between log-likelihoods and prominent existing saliency metrics: AUC and KL divergence.

AUC. The most prominent metric used in the saliency literature is the area under the receiver operating characteristic curve (AUC). The AUC is the area under a curve of model hit rate against false positive rate for each threshold. It is equivalent to the performance in a 2AFC task where the model is "presented" with two image locations: one at which an observer fixated and another from a nonfixation distribution. The thresholded saliency value is the model's decision, and the percentage correct of the model in this task across all possible thresholds is the AUC score. The different versions of AUC used in saliency research differ primarily in the nonfixation distribution used. This is usually either a uniformly selected distribution of not-fixated points across the image (e.g., in ref. 25) or the distribution of fixations for other images in the database [the shuffled AUC (3, 4, 39)]. The latter provides an effective control against center bias (a tendency for humans to look in the center of the screen, irrespective of the image content), by ensuring that both fixation and nonfixation distributions have the same image-independent bias. It is important to bear in mind that this measure will penalize models that explicitly try to model the center bias. The AUC therefore depends critically on the definition of the nonfixation distribution. In the case of the uniform nonfixation distribution, AUC is tightly related to area counts: Optimizing for AUC with uniform nonfixation distribution is equivalent to finding for each percentage $0 \leq r \leq 100$ the area consisting of $r\%$ of the image that includes most fixations (10).

One characteristic of the AUC that is often considered an advantage is that it is sensitive only to the rank order of saliency values, not their scale (i.e., it is invariant under monotonic pointwise transformations) (39). This allows the modeling process to focus on the shape (i.e., the geometry of iso-saliency points) of the distribution of saliency without worrying about the scale, which is argued to be less important for understanding saliency than the contour lines (39). However, in certain circumstances the insensitivity of AUC to differences in saliency can lead to counterintuitive behavior, if we accept that higher saliency values are intuitively associated with more fixations.

By using the likelihood of points as a classifier score, one can compute the AUC for a probabilistic model just as for saliency maps. This has a principled connection with the probabilistic model itself: If the model performed the 2AFC task outlined above using maximum-likelihood classification, then the model's performance is exactly the AUC. Given the real fixation distribution, it can also be shown that the best saliency map in terms of AUC with uniform nonfixation distribution is exactly the gaze density of the real fixation. However, this does not imply that a better AUC score will yield a better log-likelihood or vice versa. For more details and a precise derivation of these claims, see ref. 10.

Kullback–Leibler Divergence. KL divergence is tightly related to log-likelihoods. However, KL divergence as used in practice in the saliency literature is not the same as the approach we advocate.

In general, the KL divergence between two probability distributions p and q is given by

$$D_{\mathrm{KL}}[p\|q] = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)dx$$

and is a popular measure of the difference between two probability distributions. In the saliency literature, there are at least two different model comparison metrics that have been called Kullback–Leibler divergence. Thus, when a study reports a KL metric, one needs to check how this was computed. The first variant treats the saliency map as a 2D probability distribution and computes the KL divergence between this predicted distribution and the empirical density map of fixations (8, 39); we call this image-based KL-divergence. The second metric referred to as Kullback–Leibler divergence is the KL divergence between the distribution of saliency values at fixations and the distribution of saliency values at nonfixation locations; we call this fixation-based KL divergence (40). This is calculated by binning the saliency values at fixations and nonfixations into a histogram and then computing the KL divergence of these histograms. Like AUC, it depends critically on the definition of the nonfixation distribution and additionally on the histogram binning. In Table S3 we list a number of papers using one of these two definitions of KL divergence.

We now precisely show the relationship between these measures and our information theoretic approach. Very generally, information theory can be derived from the task of assigning code words to different events that occur with different probabilities such that their average code word length becomes minimal. It turns out that the negative log-probability is a good approximation to the optimal code word length possible, which gives rise to the definition of the log-loss:

$$l(x) = -\log p(x).$$

In the case of a discrete uniform distribution $p(x) = \frac{1}{n}$ the log-loss for any possible x is simply $\log n$, i.e., the log of the number of possible values of x. Accordingly, the more ambiguous the possible values of a variable are, the larger its average log-loss, which is also known as its entropy:

$$H[X] = \mathbb{E}[-\log p(x)].$$

If $p(x)$ denotes the true distribution that accurately describes the variable behavior of x and we have a model $q(x)$ of that distribution, then we can think of assigning code words to different values of x that are of length $-\log q(x)$ and compute the average log-loss for the model distribution

$$\begin{aligned}\mathbb{E}[-\log q(x)] &= -\int p(x)\log q(x)dx \\ &= H[X] + D_{\mathrm{KL}}[p(x)\|q(x)].\end{aligned}$$

That is, the KL divergence measures how much the average log-loss of a model distribution $q(x)$ exceeds the average log-loss of the true distribution. The KL divergence is also used to measure the information gain of an observation if $p(x)$ denotes a posterior distribution that correctly describes the variability of x after the observation has been made whereas $q(x)$ denotes the prior distribution. In a completely analog fashion we can measure how much more or less information one model distribution $q_1(x)$ provides about x than an alternative model $q_2(x)$ does by computing how much the average log-loss of model 1 is reduced (or increased) relative to the average log-loss of model 2. This can

also be phrased as an expected log-likelihood ratio (ELLR; the concept of log-likelihood ratios is familiar to readers with knowledge of model comparison using, e.g., χ^2 tests):

$$\begin{aligned}\mathrm{ELLR} &:= [\mathbb{E} - \log q_2(x)] - \mathbb{E}[-\log q_1(x)] \\ &= \mathbb{E}[\log q_1(x)] - \mathbb{E}[\log q_2(x)] \\ &= \int p(x) \log \frac{q_1(x)}{q_2(x)} dx.\end{aligned}$$

In other words, very generally, the amount of information model 2 provides about a variable relative to model 1 can be measured by asking how much more efficiently the variable can be encoded when assuming the corresponding model distribution $q_2(x)$ instead of $q_1(x)$ for the encoding. Note that this reasoning does not require any of the two model distributions to be correct. For example, in the context of saliency maps we can ask what the best possible model distribution is that does not require any knowledge of the actual image content. This baseline model can capture general biases of the subjects such as the center bias. To evaluate the information provided by a saliency map that can be assigned to the specific content of an image we thus have to ask how much more the model distribution of that saliency model provides relative to the baseline model.

Our information gain metric reported in the main text is exactly the ELLR, where q_1 is the model, q_2 is the baseline, and we estimated the expectation value using the sampling estimator. The ELLR can be rewritten as a difference between KL divergences:

$$\begin{aligned}\mathrm{ELLR} &= \mathbb{E}[\log(q_1(x)/q_2(x))] \\ &= \mathbb{E}[\log q_1(x)] - \mathbb{E}[\log q_2(x)] \\ &= \mathrm{D_{KL}}[p(x)||q_2(x)] - \mathrm{D_{KL}}[p(x)||q_1(x)].\end{aligned}$$

This naturally raises the question: Is our measure equivalent to the KL divergence that has been used in the saliency literature? The answer is no.

It is crucial to note that in the past the scale used for saliency maps was only a rank scale. This was the case because AUC was the predominant performance measure and is invariant under such transformations. That is, two saliency maps $S_1(x)$ and $S_2(x)$ were considered equivalent if a strictly monotonic increasing function $g : \mathbb{R} \to \mathbb{R}$ exists such that $S_1(x) = g(S_2(x))$. In contrast, in the equation for ELLR, the two distributions q_1 and q_2 are directly proportional to the saliency map times the center bias distribution and well defined only if the scale used for saliency maps is meaningful. In other words, if one applies a nonlinear invertible function to a saliency map, the ELLR changes.

Fixation-based KL divergence is the more common variant in the literature: Researchers wanted to apply information theoretic measures to saliency evaluation while remaining consistent with the rank-based scale of AUC (40). Therefore, they did not interpret saliency maps themselves as probability distributions, but applied the KL divergence to the distribution of saliency values obtained when using the fixations to that obtained when using nonfixations. We emphasize that this measure has an important conceptual caveat: Rather than being invariant under only monotonic increasing transformations, KL divergence is invariant under any reparameterization. This implies that the measure cares only about which areas are of equal saliency, but does not care about which of any two areas is actually the more salient one. For illustration, for any saliency map $S(x,y)$, its negative counterpart $\overline{S}(x,y) := \sup(S) - S(x,y)$ is completely equivalent with respect to the fixation-based KL metric, even though for any two image regions $\overline{S}$ would always make the opposite prediction about their salience (see Fig. S4 for this as well as other examples). Furthermore, the measure is sensitive to the histogram binning used, and in the limit of small bin width all models have the same KL divergence: the model-independent KL divergence between $p(x_{\mathrm{fix}})$ and $p(x_{\mathrm{nonfix}})$.

Image-based KL-divergence requires that the saliency maps are interpreted as probability distributions. Previous studies using this method (Table S3) simply divided the saliency values by their sum to obtain such probability distributions. However, they did not consider that this measure is sensitive to the scale used for the saliency maps. Optimization of the pointwise nonlinearity (i.e., the scale) has a huge effect on the performance of the different models. More generally, realizing that image-based KL divergence treats saliency maps as probability distributions means that other aspects of density estimation, like center bias and regularization strategies (blurring), must also be taken into account.

The only conceptual difference between image-based KL divergence and log-likelihoods is that for estimating expected log-likelihood ratios, it is not necessary to have a gold standard. One can simply use the unbiased sample mean estimator (*SI Text, Estimation Considerations*). Furthermore, by conceptualizing saliency in an information-theoretic way, we can not only assign meaning to expected values (such as ELLR or DKL) but also know how to measure the information content of an individual event (here, a single fixation), using the notion of its log-loss (see our application on the individual pixel level in the main text). Thus, although on a theoretical level log-likelihoods and image-based KL divergence are tightly linked, on a practical level a fundamental reinterpretation of saliency maps as probability distributions is necessary.

Estimation Considerations

One principle advantage of using log-likelihoods instead of image-based KL divergence is that for all model comparisons except comparing against the gold standard we do not have to rely on the assumptions made for the gold standard but can simply use the unbiased sample mean estimator:

$$\hat{\mathbb{E}}[\log q_1(x)/q_2(x)] = \frac{1}{N} \sum_{k=1}^{N} \log q_1(x_k)/q_2(x_k).$$

This is why we used the sample mean estimator for all model comparisons rather than the gold standard to estimate the ELLR.

However, estimating the upper limit on information gain still requires a gold standard [an estimate of the true distribution $p(x)$]. Image-based KL divergence requires this not only for estimating the upper bound, but also for calculating the performance of any model. There, it has usually been done using a 2D histogram or Gaussian kernel density estimate (Table S3), and the hyperparameters (e.g., bin size, kernel size) have commonly been chosen based on fovea size or eye tracker precision. In our framework of interpreting saliency maps as probability distributions, a principled way of choosing these hyperparameters is to cross-validate over them to get the best possible estimate of the true distribution.

For our dataset, the optimal cross-validated kernel size was 27 pixels, which is relatively close to the commonly used kernel size of 1° (37 pixels). However, with more fixations in the dataset the optimal cross-validated kernel sizes will shrink, because the local density can be estimated more precisely. Therefore, choosing these hyperparameters on criteria other than cross-validation will produce inaccurate estimates of the ELLR in the large data limit.

Because we conclude that our understanding of image-based saliency is surprisingly limited, we have been using a conservative strategy for estimating the information gain of the gold standard that is downward biased such that we obtain a conservative upper bound on the fraction of how much we understand about image-based saliency. To this end, we not only used the unbiased sample estimator for averaging over the true distribution but also resorted to a cross-validation strategy for estimating the gold standard that

takes into account how well the distributions generalize across subjects,

$$\hat{\mathbb{E}}\left[p_{\text{gold}}\right] = \sum_{j=1}^{M} \frac{1}{N_j} \sum_{k=1}^{N_j} \log p_{\text{gold}}\left(x_{jk} \middle| j\right),$$

where the first sum runs over all subjects j and $p_{\text{gold}}(x_{jk}|j)$ denotes a kernel density estimator that uses all fixations but the one of subject j. For comparison, if one would simply use the plain sample mean estimator for the gold standard, the fraction explained would drop to an even smaller value of only 22%. Our approach guarantees that it is very likely that the true value falls into the range between 22% and 34%.

Generalization to Spatiotemporal Scanpaths

The models we consider in this paper are purely spatial: They do not include any temporal dependencies. A complete understanding of human fixation selection would require an understanding of spatiotemporal behavior, that is, scanpaths. The model adaptation and optimization procedure we describe above can be easily generalized to account for temporal effects, as follows.

A scanpath consists of N fixations with positions x_i, y_i, t_i, where x_i and y_i denote the spatial position of the fixation in the image and t_i denotes the time of the fixation. A scanpath can be viewed as a sample of a 3D point process (12). Conceiving of scanpaths as 3D point processes allows us to model the joint probability distribution of all fixations of a subject on an image. In general, a model's average log-likelihood is $\frac{1}{N}\sum_k \log \hat{p}(x_k)$, where $\hat{p}$ is the probability distribution of the model and x_k, $k = 1, \ldots, N$ are samples from the probabilistic process that we would like to model. Our likelihoods are therefore of the form $\hat{p}(x_1, y_1, t_1, \ldots, x_N, y_N, t_N, N|I)$, where N is part of the data distribution (not a fixed parameter) and I denotes the image for which the fixations should be predicted. By chain rule, this is decomposed into conditional likelihoods $\hat{p}(x_1, y_1, t_1, \ldots, x_N, y_N, t_N, N|I) = \hat{p}(N|I)\prod_{i=1}^{N} p(x_i, y_i, t_i|N, x_1, y_1, t_1, \ldots, x_{i-1}, y_{i-1}, t_{i-1}, I)$.

The above holds true for any 3D point process. In this way, the model comparison framework we propose in this paper is general in that it can account for spatiotemporal fixation dependencies (see ref. 12 for a recent application of spatiotemporal point processes to the study of scanpaths).

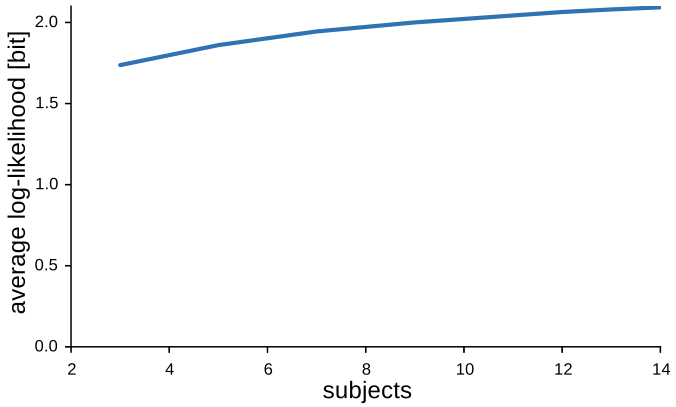


Fig. S1. Dependence of gold standard performance on the number of subjects used to predict one subject's data.

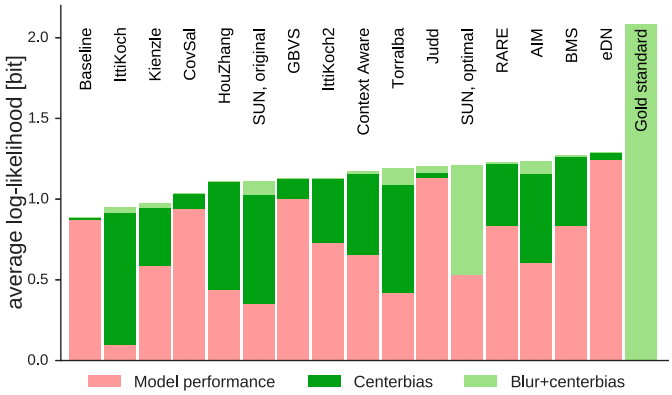


Fig. S2. Average log-likelihoods of all tested models as differences in log-likelihood compared with the maximum-entropy model predicting a uniform fixation distribution. Model performance indicates the model performance if only the nonlinearity has been fitted. Centerbias and blur+centerbias indicate the model performances if the centerbias alone or the blur and centerbias have been fitted together with the nonlinearity.

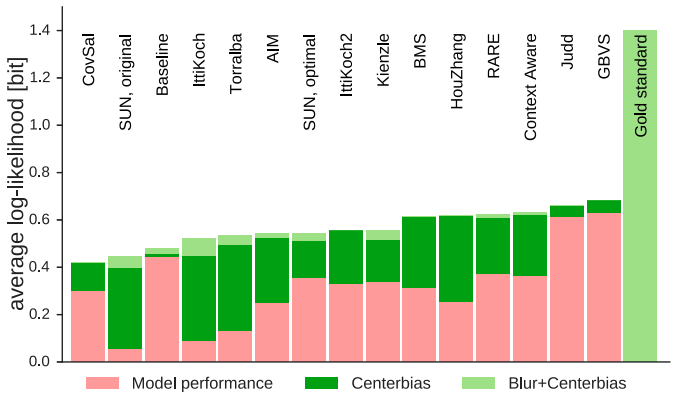


Fig. S3. Average log-likelihoods of all tested models on the Kienzle dataset as in Fig. S2.

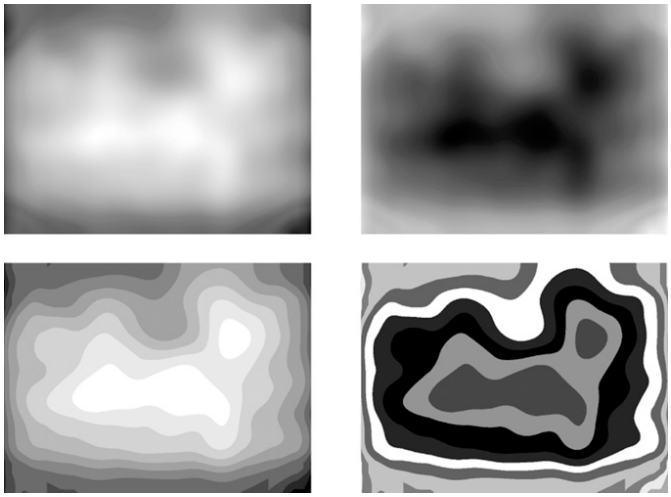


Fig. S4. Fixation-based Kullback–Leibler divergence for saliency maps. *Upper Left* shows a real saliency map (from eDN), *Upper Right* is inverted, *Lower Left* is the same map with binned saliency values, and in the *Lower Right* map, the saliency assigned to each bin is shuffled. These maps have identical fixation-based KL divergence (and very different log-likelihoods).

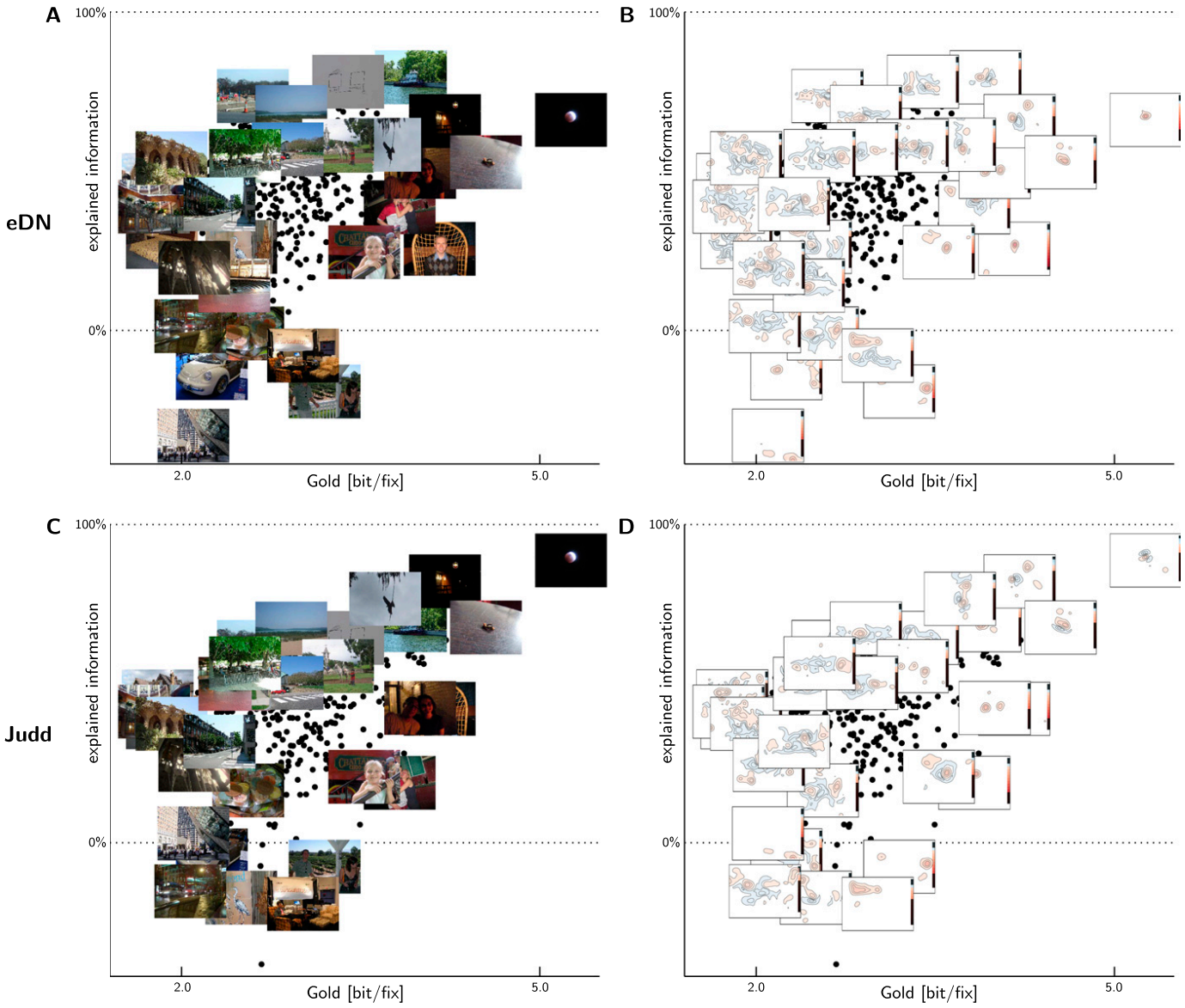


Fig. S5. Distribution of information gains and explained information (both relative to a uniform baseline model) over all images in the dataset. Each black dot represents an image from the dataset. For some images we show the actual image (*A* and *C*) and the information gain difference from the gold standard (*B* and *D*). These plots allow model performance to be assessed on all images in the dataset. Points in the lower right of the scatterplots are images where a lot of information could be explained but is not; these are where the model could be best improved for a given dataset. The pixel-space information gain scatter plots (*B* and *D*) show exactly where in the images the model predictions fail.

Table S1. Finest resolution for different metrics

Resolution	Metrics
Dataset	Fixation-based KL divergence
Image	Image-based KL divergence, CC
Fixation	AUC, NSS
Pixel	Information gain

Table S2. Evaluated models and their sources

Model	Source
Itti et al. (1)	www.saliencytoolbox.net (IttiKoch) www.vision.caltech.edu/~harel/share/gbvs.php (IttiKoch2)
Torralba et al. (26)	people.csail.mit.edu/tjudd/SaliencyBenchmark/Code/torralbaSaliency.m
GBVS (27)	www.vision.caltech.edu/~harel/share/gbvs.php
SUN (28)	cseweb.ucsd.edu/~l6zhang
Kienzle et al. (24, 29)	Code provided by Simon Barthelmé
Hou and Zhang (30)	www.klab.caltech.edu/~xhou/projects/spectralResidual/spectralresidual.html
AIM (31)	www-sop.inria.fr/members/Neil.Bruce/
Judd (23)	people.csail.mit.edu/tjudd/WherePeopleLook/index.html
Context-aware saliency (32, 33)	webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html
CovSal (34)	web.cs.hacettepe.edu.tr/~erkut/projects/CovSal/
RARE2012 (35)	www.tcts.fpms.ac.be/attention/?categorie17/rare2012
Boolean map-based saliency (BMS) (5, 36)	cs-people.bu.edu/jmzhang/BMS/BMS.html
Ensemble of deep networks (eDN) (37)	github.com/coxlab/edn-cvpr2014

Table S3. Papers using KL divergence to evaluate saliency models

Ref.	KL divergence	Estimate of true distribution
(40)	Fixation based	
(41)	Fixation based	
(42)	Fixation based	
(43)	Fixation based	
(28)	Fixation based	
(31)	Fixation based	
(44)	Fixation based	
(45)	Fixation based	
(2)	Fixation based	
(3)	Fixation based	
(4)	Fixation based	
(46)	Image based	Gaussian kernel, width of fovea
(39)	Image based	2D histograms, bins of $2° \times 2°$ and 10^{-5} added as prior
(47)	Image-based	Precision of the eye tracking
(8)	Image based	Gaussian with 2°, motivated by fovea + eye tracker
(48)	Image based, fixation based	Gaussian kernel density estimate, kernel size 1° of visual angle
(7)	Image based	Not stated
(49)	Image based	"Kernel-density estimates with bandwidth parameters chosen according to Scott's rule"

Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet

Matthias Kümmerer, Lucas Theis and Matthias Bethge

Published in ICLR Workshop Track 2015, arXiv 1411.1045

Abstract

Recent results suggest that state-of-the-art saliency models perform far from optimal in predicting fixations. This lack in performance has been attributed to an inability to model the influence of high-level image features such as objects. Recent seminal advances in applying deep neural networks to tasks like object recognition suggests that they are able to capture this kind of structure. However, the enormous amount of training data necessary to train these networks makes them difficult to apply directly to saliency prediction. We present a novel way of reusing existing neural networks that have been pretrained on the task of object recognition in models of fixation prediction. Using the well-known network of Krizhevsky et al. (2012), we come up with a new saliency model that significantly outperforms all state-of-the-art models on the MIT Saliency Benchmark. We show that the structure of this network allows new insights in the psychophysics of fixation selection and potentially their neural implementation. To train our network, we build on recent work on the modeling of saliency as point processes.

Contributions

The original idea of using pretrained deep features to predict fixations was suggested by Lucas Theis. I designed the precise model, ran all the experiments and all analyses. The paper draft was written by me. All authors contributed to scientific discussions and paper revisions.

DEEP GAZE I: BOOSTING SALIENCY PREDICTION WITH FEATURE MAPS TRAINED ON IMAGENET

Matthias Kümmerer, Lucas Theis & Matthias Bethge
Werner Reichardt Centre for Integrative Neuroscience
University Tübingen, Germany
{matthias.kuemmerer,lucas,matthias}@bethgelab.org

ABSTRACT

Recent results suggest that state-of-the-art saliency models perform far from optimal in predicting fixations. This lack in performance has been attributed to an inability to model the influence of high-level image features such as objects. Recent seminal advances in applying deep neural networks to tasks like object recognition suggests that they are able to capture this kind of structure. However, the enormous amount of training data necessary to train these networks makes them difficult to apply directly to saliency prediction. We present a novel way of reusing existing neural networks that have been pretrained on the task of object recognition in models of fixation prediction. Using the well-known network of Krizhevsky et al. (2012), we come up with a new saliency model that significantly outperforms all state-of-the-art models on the MIT Saliency Benchmark. The structure of this network allows new insights in the psychophysics of fixation selection and potentially their neural implementation. To train our network, we build on recent work on the modeling of saliency as point processes.

By understanding how humans choose eye fixations, we can hope to understand and explain human behaviour in a number of vision-related tasks. For this reason human eye movements have been studied for more than 80 years (e. g. Buswell, 1935). During the last 20 years, many models have been developed trying to explain fixations in terms of so called "saliency maps".

Recently, it has been suggested to model saliency maps probabilistically using point processes (Barthelmé et al., 2013) and to evaluate them using log-likelihood (Kümmerer et al., 2014). This evaluation revealed that state-of-the-art models of saliency explain only one third of the explainable information in the spatial fixation structure (Kümmerer et al., 2014).

Most of the existing models use low-level cues like edge-detectors and color filters (Itti et al., 1998) or local image statistics (Zhang et al., 2008; Bruce & Tsotsos, 2009). However, human fixations are largely clustered around objects (see Figure 1 for examples). This has led to some models trying to incorporate more high level features: Cerf et al. (2008) combined existing saliency map models with a face detector, while Judd et al. (2009) included detectors for faces, people, cars and horizon. Nevertheless, current saliency models mostly fail to capture these high-level influences which might be the main reason for the poor overall performance of state-of-the-art models. This analysis raises the question whether there are any computational systems capable of capturing such high-level influences.

Independent of these developments, the last two years have seen the rise of deep neural networks to solve multifarious tasks like object detection, speech recognition or automatic translation. Provided with enough training data, deep neural networks show impressive results, often outperforming all competing methods. It has also been shown that deep convolutional networks that have been optimized for object classification can be used to predict neuron responses in higher brain areas of the visual system (Yamins et al., 2014; Razavian et al., 2014). Deep neural networks have also proven to generalize well over tasks (Donahue et al., 2013): a network trained for some task like object detection can often be easily retrained to achieve state-of-the-art performance in some other only loosely related task like scene recognition.

Motivated by these developments, we here try to use pretrained deep neural networks to model fixation selection. The results of Yamins et al. (2014) connect neural network representations with IT

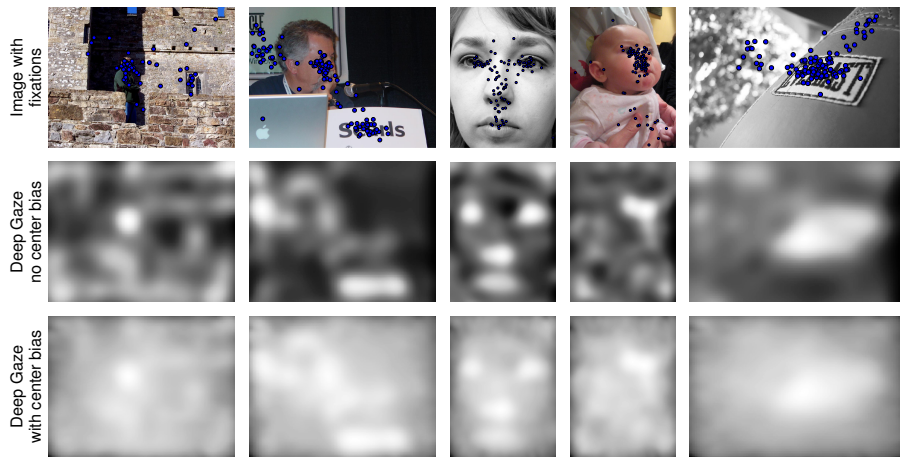


Figure 1: Example saliency maps: The top row shows example images from the dataset by Judd et al. (2009). The fixations of the subjects are indicated by dots. The middle row shows the log-densities produced by Deep Gaze I for these images when assuming a uniform prior distribution instead of a center bias. The bottom row shows the log-densities for the same images when using the center bias of the full dataset. Note that only the first two images were included in the set of images used to train Deep Gaze I.

and similar neural representations. This suggests that we can hope not only to improve prediction performance, but also to improve our understanding of the internal implementation of fixation selection in the brain by formulating new hypotheses that lead to new experimental paradigms. Finally, results from Zeiler & Fergus (2013) show ways to interpret the filters of deeper layers in a way that would allow to formulate predictions that can be tested psychophysically.

A first attempt at modelling saliency with deep convolutional networks has been performed recently by Vig et al. (2014) (eDN), yielding state-of-the-art performance. However, training deep neural networks on fixations suffers from the usually small training sets compared to the training data used in other tasks. To reach their state-of-the-art performance, neural networks trained for object or speech recognition need massive amounts of training data. Most fixation datasets have at most 1000 images with usually not significantly more than 100 fixations per image. Deep neural networks can easily have millions of parameters, which would lead to massive overfitting on these small datasets. Therefore, eDN uses only three convolutional layers, while the Krizhevsky network uses 5 convolutional layers and the most recent networks used in the ImageNet challenge (ILSVRC2014) use around 20 layers.

Here we present a new model of fixation prediction that builds on these results: it uses the well known deep network from Krizhevsky et al. (2012) to generate a high-dimensional feature space, which is then used for the actual fixation prediction. This deep network has been optimized for object recognition using a massive dataset consisting of more than one million images (Deng et al., 2009). Keeping the parameters of the deep network fixed, we train our model on half of the MIT1003 dataset (Judd et al., 2009) and show that it outperforms state-of-the-art models by a large margin, increasing the amount of explained information by 67 %. Furthermore, we analyze how the model exploited the feature space provided by the Krizhevsky network.

1 METHODS

In Figure 2, the model architecture is visualized. After an initial downsampling, the RGB input image is fed into the Krizhevsky network. The Krizhevsky architecture consists of stacked convolutions, each one followed by a rectifiying nonlinearity and optional maxpooling and response normalization. The final three fully connected layers of the Krizhevsky network were removed as we are only interested in spatially located features. Each layer (convolution, rectifier, pooling and normalization) results in a single image of response for each filter in the layer. To predict fixations,

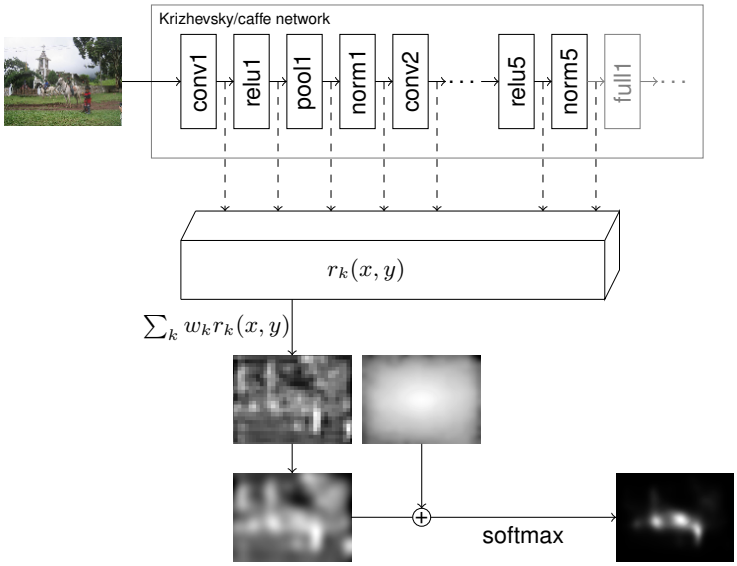


Figure 2: The model structure of Deep Gaze I: The image is first downsampled and preprocessed with the Krizhevsky network. The responses of the layers that are included in the model are then scaled up to the size of the largest network layer and normalized to have unit standard deviation. This list of maps is then linearly combined and blured with a Gaussian kernel. To compensate for the central fixation bias, an estimate of the prior distribution is added. Finally, the model output is fed through a softmax rectification, yielding a two dimensional probability distribution.

we first select one or multiple layers from the network. We rescale all the response images that we want to include in our model to the size of the largest layer of the network, resulting in a list of up to 3712 responses for each location in an image. Each of these responses is then individually normalized to have unit standard deviation on the full dataset. After this preprocessing, the features are fed into the following model.

At leach image location, our saliency model linearly combines the responses $r_k(x, y)$ using weights w_k. The resulting image is then convoled with a Gaussian kernel whose width is controlled by σ, yielding the saliency map

$$s(x, y) = \sum_k w_k r_k(x, y) * G_\sigma.$$

It is well known that fixation locations are strongly biased towards the center of an image (Tatler, 2007). To account for this center bias, the saliency prediction is linearly combined with a fixed center bias prediction $c(x, y)$:

$$o(x, y) = \alpha c(x, y) + s(x, y)$$

To predict fixation probabilities, this output is finally fed into a softmax, yielding a probability distribution over the image:

$$p(x, y) = \frac{\exp\left(o(x, y)\right)}{\sum_{x,y} \exp\left(o(x, y)\right)}$$

For generalization, ℓ_1-regularization on the weights is used to encourage sparsity. For training fixations $(x_1, y_1), \ldots, (x_N, y_N)$ this yields the cost function

$$c(\mu, \alpha, w) = -\frac{1}{N} \sum_i^N \log p(x_i, y_i) + \lambda \frac{\|w\|_1}{\|w\|_2}$$

To quantify which layers help most in predicting the fixations and lead to least overfitting, we trained models on a variety of subsets of layers (see subsection 2.3 and Figure 5). We checked the generalization performance of these models on the remaining 540 images from MIT1003 that have not been used in training. As performance measure we use shuffled area under the curve (shuffled AUC) here (Tatler et al., 2005). In AUC, the saliency map is treated as a classifier score to separate fixations from "nonfixations": presented with two locations in the image, the classifier chooses the location with the higher saliency value as fixation. The AUC measures the classification performance of this classifer. The standard AUC uses a uniform nonfixation distribution, while in the case of shuffled AUC, fixations from other images are used as nonfixations. As shuffled AUC assumes the saliency maps not include the biases of the prior distribution (see Barthelmé et al., 2013) we had to use a uniform center bias for this evaluation.

1.1 Implementation details

For training, we used roughly half of the dataset MIT1003 (Judd et al., 2009). By using only the images of the most common size of 1024×768 pixels (resulting in 463 images), we were able to use the nonparametric estimate of the center bias described in Kümmerer et al. (2014) (mainly a 2d histrogram distribution fitted using the fixations from all other images).

Our implementation of the Krizhevsky network uses the architecture and trained filters as published by Jia et al. (2014) with the following modifications: the original architecture uses a fixed input size of 224×224. As we removed the fully connected layers, we do not need to restrict to a fixed input size but can feed arbitrary images into the network. Furthermore we use convolutions of type *full* (i.e. zero-pad the input) instead of *valid* which would result in convolution outputs that are smaller than the input. This modification is useful, because we need saliency predictions for every point in the image. Note that the caffe implementation of the Krizhevsky network differs slightly from the original architecture in Krizhevsky et al. (2012), as the pooling and the normalization layers have been switched. The subsampling factor for the inital downsampling of the images was set to 2.

The sparsity parameter λ was chosen using grid search and turned out to be 0.001 in the final model. However, even setting it to much smaller values did have very little effect on training and test performance (see subsection 6.1 for more details). All calculations of log-likelihoods, cost functions and gradients were done in theano (Bergstra et al., 2010). To minimize the cost function on the training set of fixations, the mini-batch based BFGS method as described in Sohl-Dickstein et al. (2014) was used. It combines the benefits of batch based methods with the advantage of second order methods, yielding high convergence rates with next to no hyperparameter tuning. To avoid overfitting to the subjects, leave-one-out cross-validation over the 15 subjects contained in the database was used.

The code for our model including training and analysis will be published at `http://www.bethgelab.org/code/deepgaze/`.

2 Results

2.1 Performance Results

We use an information theoretic measure to evaluate our model: log-likelihood. Log-likelihood is a principled measure for probabilistic models and has numerous advantages. See Kümmerer et al. (2014) for an extensive discussion.

Log-likelihoods are much easier to understand when expressed as difference of log-likelihood relative to a baseline model. This *information gain*[1] expresses how much more efficient the model is in describing the fixations than the baseline model: if a model with an information gain of 1 bit/fix is used to encode fixation data, it can save on average one bit per fixation compared to the baseline model.

The information gain is even more intuitive when compared to the explainable information gain, i.e., the information gain of the real distribution compared to the baseline model. This comparison yields a ratio of explained information gain to explainable information gain which will be called

[1] To be more precise, this value is an estimated expected information gain

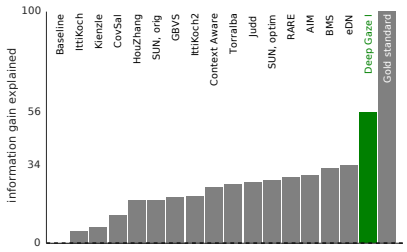


Figure 3: Performance of Deep Gaze I compared to a list of other influential models, expressed as the ratio of explained information (see text for details). All models except for Deep Gaze I have been postprocessed to account for a pointwise nonlinearity, center bias and blurring (see Kümmerer et al. (2014) for details).

"explainable information gain explained" or just "information gain explained" in the following. See Kümmerer et al. (2014) for a more thorough explanation of this notion.

The baseline model is a non-parametric model of the image-independent prior distribution $p(x, y)$, while the explainable information is estimated using a non-parametric model of the fixation distribution $p(x, y \mid I)$ for a given image I (which we call the *gold standard model*). The gold standard model is cross-validated between subjects and thus captures all the structure in the fixations that is purely due to the spatial structure of the image. See Kümmerer et al. (2014) for details on the baseline model and the gold standard model.

By expressing the information gain of a model as a percentage of the possible information gain, we can asses how far we have come in describing the fixations. It is important to note that this interpretation is only possible due to the fact that information gain is on a ratio scale (Michell, 1997): differences and ratios of information gains are meaningful – opposed to other measures like AUC.

In Figure 3, the percentage of information gain explained is plotted for our model in comparison to a range of influential saliency models, including the state-of-the-art models. Of the possible information gain, the best existing model (eDN) is able to explain only $34\,\%$. Deep Gaze I is able to increase this information gain to $56\,\%$.

2.2 Results on MIT Saliency Benchmark

We submitted our model to the MIT Saliency Benchmark (Bylinskii et al.). The benchmark evaluates saliency models on a dataset of 300 images and 40 subjects. The fixations are not available to make training for these fixations impossible.

The MIT Saliency Benchmark evaluates models on a variety of metrics, including AUC with uniform nonfixation distribution and shuffled AUC (i.e. AUC with center bias as nonfixation distribution). The problem with these metrics is that most of them use different definitions of saliency maps. This hold especially for the two most used performance metrics: AUC and shuffled AUC. While AUC expects the saliency maps to model the center bias, shuffled AUC explicitly does not so and penalizes models that do (see Barthelmé et al. (2013) for details). As Deep Gaze I uses an explicit representation of the prior distribution, it is straightforward to produce the saliency maps according to both definitions of AUC: For AUC we use a nonparametric prior estimate, for shuffled AUC we use a uniform prior distribution. As the images of the dataset are of different size, we could not use our non-parametric center bias as is. Instead, we took all fixations from the full MIT-1003 dataset and transformed their position to be relative to a image of size 100×100. Then we trained a Gaussian kernel density estimator on these fixations. This density estimate was then rescaled and renormalized for each image.

Doing so, we beat the state-of-the-art models in the MIT Saliency Benchmark by a large margin in AUC as well as shuffled AUC (see Figure 4): For shuffled AUC, we reach 71.69% compared to 67.90% for the best performing model AWS (center bias is at 50%). For AUC we reach 84.40%

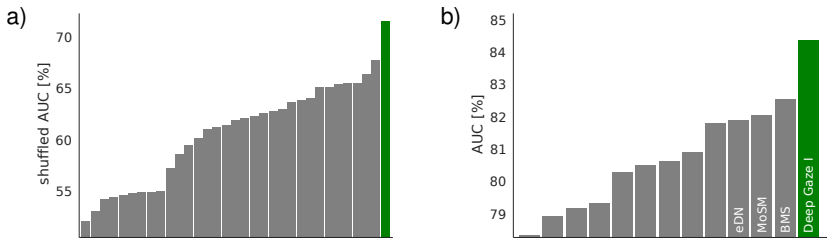


Figure 4: Performance results on the MIT benchmark. **(a)**: Shuffled AUC performance of Deep Gaze I (green bar, 71.69%) compared with all other models in the MIT benchmark. The x-axis is at the level of the center bias model. The three top performing models after Deep Gaze I are in order of decreasing performance: AWS (67.90%, Garcia-Diaz et al. (2012)), RARE2012 (66.54%, Riche et al. (2013)), and AIM (65.64%, Bruce & Tsotsos (2009)). **(b)** AUC performance of Deep Gaze I (green bar, 84.40%) compared with all other models in the MIT benchmark that performed better than the center bias. The x-axis is at the level of the center bias model. The three top performing models after Deep Gaze I are in order of decreasing performance: BMS (82.57%, Zhang & Sclaroff (2013)), Mixture of Saliency Models (82.09%, Han and Satoh, 2014), and eDN (81.92%, Vig et al. (2014)). Notice that AUC and shuffled AUC use different definitions of saliency map: While AUC expects the saliency maps to model the center bias, shuffled AUC explicitly does not and penalizes models that do. Therefore, for the shuffled AUC performances of Deep Gaze I the saliency maps have been calculated with a uniform prior distribution, while for the AUC performances the saliency maps have been calculated with a nonparametric prior (see text for details) [2]. Performances of other models from the MIT benchmark as of September 2014.

compared to 82.57% for the best performing model BMS (center bias is at 78.31%). Relative to the center bias, this is an increase of AUC performance by more than 40%.

2.3 LAYER SELECTION

The final model used only the convolutions of the top-most layer of the Krizhevsky-architecture. This is a principled choice: the top layer can be expected to include most high-level influences and the relu, pool and norm units are often viewed mainly as the nonlinearities needed to provide a new feature space for the next level of convolutions.

But this choice was also backed by a series of comparison models where more or other layers have been included in the model: In Figure 5, performance results are reported for models including layers from a given depth upwards (Figure 5a), layers up to a given depth (Figure 5b), layers of a given depth (Figure 5c) and layers of a given type (Figure 5d). It can be seen that the architecture chosen finally (layer 5 convolutions) generalizes best to the images of the test set in terms of shuffled AUC.

It is also worth noting that models including more layers are substantially better at predicting the test subjects fixations on the images used in training (Figure 5a, left plot): when using all layers, a performance of 83% information gain explained is reached for the test subjects. This suggests that the generalization problems of these models are not due to intersubject variability. They most probably suffer from the fact that the variety of objects in the training images is not rich enough, leading to overfitting to the images (not to the subjects). Therefore we can expect improved performance from using a larger set of images in training.

[2]Note that the MIT Saliency Benchmark webpage reports only performances for the saliency maps with the nonparametric prior. Therefore, there the shuffled AUC performance is lower.

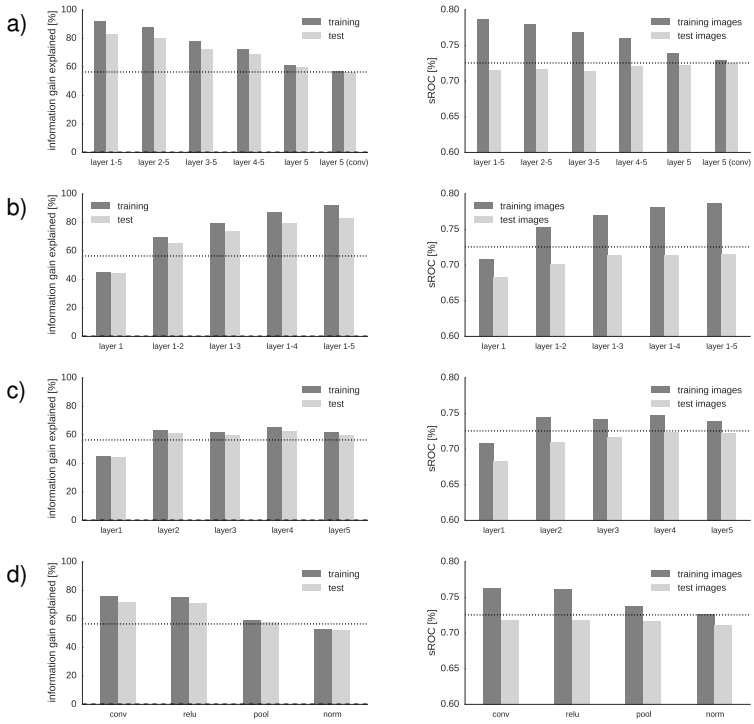


Figure 5: Performance of Deep Gaze I when trained on different subsets of the Krizhevsky layers: **(a)**: Results for models that use layers from a given depth upwards. The left plot shows the percentage of explainable information gain explained on the images used in training for training subjects and test subjects (refer to subsection 2.1 for an explanation of this measure). The dotted line indicates the performance of the model we used in the MIT Saliency Benchmark (which only used the output of the convolutions of layer 5). The right plot shows the shuffled AUC on the images used in training and on the remaining test images. Here, the models have been averaged over all test subjects and the saliency maps assume uniform center bias, as expected by shuffled AUC (see subsection 2.2 for details). The dotted line indicates the performance of the final model on the test images. **(b), (c), (d)**: Results for models that use layers up to a given depth (b), layers of a certain depth (c) and layers of a certain type (d). The plots are as in (a).

2.4 Analysis of used features

In this section we analyze which features of the Krizhevsky architecture contributed most to the fixation predictions. By getting a solid understanding of the involved features, we can hope to extract predictions from the model that can be tested psychophysically in the future.

In Figure 6, we took the 10 most weighted features from the 256 convolution features in layer 5. For each of these 10 features, we plotted the 9 patches from the dataset that led to the highest response (resp. lowest response for features with negative weight). In Figure 7, the first four patches of the first four features are shown in more detail: The patches are shown in the context of the entire image and also the feature's response to this image is shown.

Clearly, the most important feature is sensitive to faces. The second most important feature seems to respond mainly to text. The third most important feature shows some sort of pop-out response: it seems to respond to whichever feature sticks out from an image: the sign of a bar in the first patch, two persons in a desert in the second patch and, most notably, the target in a visual search image in the fourth patch. Note that the salient feature depends heavily on the image context, so that a simple luminance or color contrast detector would not achieve the same effect.

Figure 6: Analysis of used features I: **(a)** Patches of maximum response: Each square of patches shows for a specific feature of the Krizhevsky architecture the nine patches that led to highest response (resp. smallest response, if the feature has a negative weight in the model). Each patch corresponds to exactly the part of the image that contributes to the response in the location of maximum response. The features used have been choosen by the absolute value of the weight that Deep Gaze I assigned to them. The numbers over the patches show $|w_k| / \max_k |w_k|$.

This shows that Deep Gaze I is not only able to capture the influence of high level objects like faces or text, but also more abstract high-level concepts (like popout).

3 Discussion

Deep Gaze I was able to increase the explained information gain to 56 % compared to 34 % for state of the art models. On the MIT Saliency Benchmark we were also able to beat the state of the art models by a substantial margin. One main reason for this performance is the ability of our model to capture the influence of several high-level features like faces and text but also more abstract ones like popout (2.4).

It is important to note that all reported results from Deep Gaze I are direct model performances, without any fitting of a pointwise nonlinearity as performed in Kümmerer et al. (2014). This indicates that the deep layers provide a sufficiently rich feature space to enable fixation prediction via simple linear combination of the features. The convolution responses turned out to be most informative about the fixations.

While features trained on ImageNet have been shown to generalize to other recognition and detection tasks (e. g. Donahue et al., 2013; Razavian et al., 2014), to our knowledge this is the first work where ImageNet features have been used to predict behaviour.

Extending state-of-the-art neural networks with attention is an exciting new direction of research (Tang et al., 2014; Mnih et al., 2014). Humans use attention for efficient object recognition and we showed that Krizhevsky features work well for predicting human attention. Therefore it is likely that these networks could be brought closer to human performance by extending them with Krizhevsky features. This could be an interesting field for future research.

4 Conclusions

Our contribution in this work is twofold: First, we have shown that deep convolutional networks that have been trained on computer vision tasks like object detection boost saliency prediction.

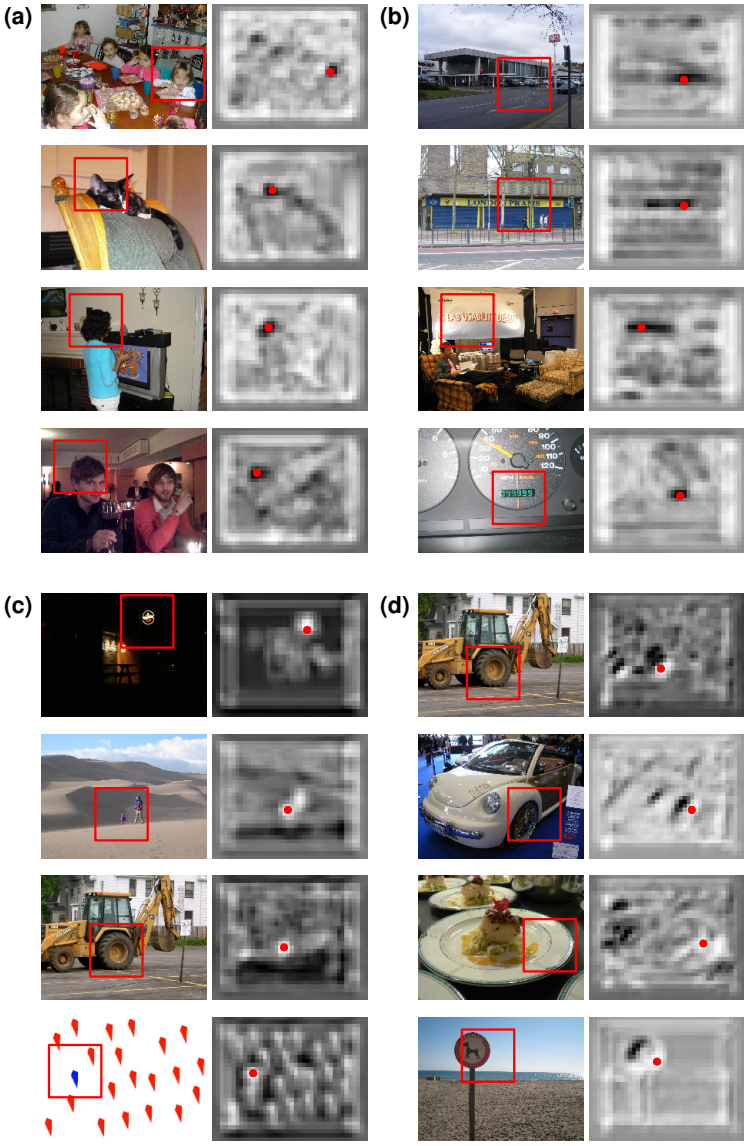


Figure 7: Analysis of used features II: Details for some of the patches from Figure 6 The four double columns (a) to (d) correspond to the first four features shown Figure 6. In each double column, the four rows correspond to the first four patches shown for this feature in Figure 6. The left column of each double column shows the patches in the context of the full image, while the feature's response over the full image is shown in the right column. The position of the maximum is indicated by a dot.

Using the well-known Krizhevsky network (Krizhevsky et al., 2012), we were able to outperform state-of-the-art saliency models by a large margin, increasing the amount of explained information by 67 % compared to state-of-the art. We believe this approach will enable the creation of a new generation of saliency models with high predictive power and deep implications for psychophysics and neuroscience (Yamins et al., 2014; Zeiler & Fergus, 2013). An obvious next step suggested by this approach is to replace the Krizhevsky network by the ImageNet 2014 winning networks such as VGG (Simonyan & Zisserman, 2014) and GoogLeNet (Szegedy et al., 2014).

A second conceptual contribution of this work is to optimize the saliency model by maximizing the log-likelihood of a point process (see Barthelmé et al., 2013; Kümmerer et al., 2014).

We believe that the combination of high performance feature spaces for object recognition as obtained from the ImageNet benchmark with principled maximum likelihood learning opens the door for a "Deep Gaze" program towards explaining all the explainable information in the spatial image-based fixation structure.

5 ACKNOWLEDGEMENTS

This work was mainly supported by the German Research Foundation (DFG; priority program 1527, Sachbeihilfe BE 3848-1) and additionally by the German Ministry of Education, Science, Research and Technology through the Bernstein Center for Computational Neuroscience (FKZ 01GQ1002) and the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307).

REFERENCES

Barthelmé, Simon, Trukenbrod, Hans, Engbert, Ralf, and Wichmann, Felix. Modelling fixation locations using spatial point processes. *Journal of Vision*, 13(12), 2013. doi: 10.1167/13.12.1.

Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

Bruce, Neil DB and Tsotsos, John K. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3), 2009.

Buswell, Guy Thomas. *How people look at pictures*. University of Chicago Press Chicago, 1935.

Bylinskii, Zoya, Judd, Tilke, Durand, Frédo, Oliva, Aude, and Torralba, Antonio. Mit saliency benchmark. http://saliency.mit.edu/.

Cerf, Moran, Harel, Jonathan, Einhaeuser, Wolfgang, and Koch, Christof. Predicting human gaze using low-level saliency combined with face detection. In Platt, J.C., Koller, D., Singer, Y., and Roweis, S.T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 241–248. Curran Associates, Inc., 2008. URL `http://papers.nips.cc/paper/3169-predicting-human-gaze-using-low-level-saliency-combined-with-face-detection.pdf`.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, and Darrell, Trevor. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

Garcia-Diaz, Antón, Leborán, Víctor, Fdez-Vidal, Xosé R, and Pardo, Xosé M. On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of vision*, 12(6):17, 2012.

Itti, Laurent, Koch, Christof, and Niebur, Ernst. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11): 1254–1259, 1998. doi: 10.1109/34.730558.

Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, Jonathan, Girshick, Ross, Guadarrama, Sergio, and Darrell, Trevor. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Judd, Tilke, Ehinger, Krista, Durand, Frédo, and Torralba, Antonio. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pp. 2106–2113. IEEE, 2009.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Kümmerer, M., Wallis, T., and Bethge, M. How close are we to understanding image-based saliency? *arXiv preprint arXiv:1409.7686*, Sep 2014. URL http://arxiv.org/abs/1409.7686.

Michell, Joel. Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3):355–383, 1997.

Mnih, Volodymyr, Heess, Nicolas, Graves, Alex, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pp. 2204–2212, 2014.

Razavian, Ali Sharif, Azizpour, Hossein, Sullivan, Josephine, and Carlsson, Stefan. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pp. 512–519. IEEE, 2014.

Riche, Nicolas, Mancas, Matei, Duvinage, Matthieu, Mibulumukini, Makiese, Gosselin, Bernard, and Dutoit, Thierry. RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642–658, July 2013. ISSN 09235965. doi: 10.1016/j.image.2013.03.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0923596513000489.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

Sohl-Dickstein, Jascha, Poole, Ben, and Ganguli, Surya. An adaptive low dimensional quasi-newton sum of functions optimizer. In *International Conference on Machine Learning*, 2014.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. URL http://arxiv.org/abs/1409.4842.

Tang, Yichuan, Srivastava, Nitish, and Salakhutdinov, Ruslan R. Learning generative models with visual attention. In *Advances in Neural Information Processing Systems*, pp. 1808–1816, 2014.

Tatler, Benjamin W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 2007. doi: 10.1167/7.14.4.

Tatler, Benjamin W., Baddeley, Roland J., and Gilchrist, Iain D. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643 – 659, 2005. ISSN 0042-6989. doi: http://dx.doi.org/10.1016/j.visres.2004.09.017. URL http://www.sciencedirect.com/science/article/pii/S0042698904004626.

Vig, Eleonora, Dorr, Michael, and Cox, David. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Computer Vision and Pattern Recognition, 2014. CVPR'14. IEEE Conference on*. IEEE, 2014.

Yamins, Daniel LK, Hong, Ha, Cadieu, Charles F, Solomon, Ethan A, Seibert, Darren, and DiCarlo, James J. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, pp. 201403112, 2014.

Zeiler, Matthew D and Fergus, Rob. Visualizing and understanding convolutional neural networks. *arXiv preprint arXiv:1311.2901*, 2013.

Zhang, Jianming and Sclaroff, Stan. Saliency detection: a boolean map approach. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 153–160. IEEE, 2013.

Zhang, Lingyun, Tong, Matthew H, Marks, Tim K, Shan, Honghao, and Cottrell, Garrison W. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

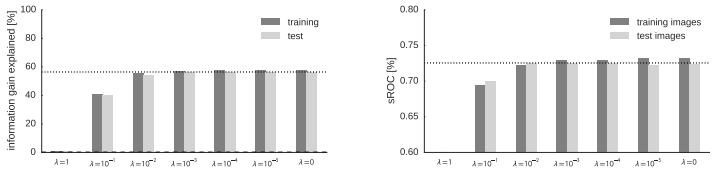


Figure 8: Performance of Deep Gaze I when trained on the conv5-layer with different regularization parameters. The left plot shows the percentage of explainable information gain explained on the images used in training for training subjects and test subjects (refer to subsection 2.1 for an explanation of this measure). The dotted line indicates the performance of the model we used in the MIT Saliency Benchmark ($\lambda = 0.001$). The right plot shows the shuffled AUC on the images used in training and on the remaining test images. Here, the models have been averaged over all test subjects and the saliency maps assume uniform center bias, as expected by shuffled AUC (see subsection 2.2 for details). The dotted line indicates the performance of the final model on the test images.

6 Supplementary Material

6.1 Regularization

The model uses a regularization parameter λ to encourage sparsity in the feature weights (see section 1). This parameter was choosen using grid search. In Figure 8, training and test performances are shown for different choices of λ when fitting the model using only the final convolutional layer (as done in the final model). It can be seen that the choice of the regularization parameter had a visible but only very small effect on the test performance (especially if compared to the influences of the different layers used, see Figure 5).

Understanding Low- and High-Level Contributions to Fixation Prediction

Matthias Kümmerer, Thomas S.A. Wallis, Leon A. Gatys and Matthias Bethge

Published in The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4789-4798

Abstract

Understanding where people look in images is an important problem in computer vision. Despite significant research, it remains unclear to what extent human fixations can be predicted by low-level (contrast) compared to high-level (presence of objects) image features. Here we address this problem by introducing two novel models that use different feature spaces but the same readout architecture. The first model predicts human fixations based on deep neural network features trained on object recognition. This model sets a new state-of-the art in fixation prediction by achieving top performance in area under the curve metrics on the MIT300 hold-out benchmark (AUC = 88%, sAUC = 77%, NSS = 2.34). The second model uses purely low-level (isotropic contrast) features. This model achieves better performance than all models not using features pre-trained on object recognition, making it a strong baseline to assess the utility of high-level features. We then evaluate and visualize which fixations are better explained by low-level compared to high-level image features. Surprisingly we find that a substantial proportion of fixations are better explained by the simple low-level model than the state-of-the-art model. Comparing different features within the same powerful readout architecture allows us to better understand the relevance of low- versus high-level features in predicting fixation locations, while simultaneously achieving state-of-the-art saliency prediction.

Contributions

The idea of improving DeepGaze I with a nonlinear readout network, pretraining on SALICON and replacing AlexNet with VGG was mine. Matthias Bethge suggested to compare to a parsimonious low-level baseline model which developed into what's now the ICF model. I designed, implemented and trained all models and ran all experiments and analyses. The paper was written jointly by Thomas Wallis, Leon Gatys and me. All authors contributed to scientific discussions and paper revisions.

Understanding Low- and High-Level Contributions to Fixation Prediction

Matthias Kümmerer, Thomas S.A. Wallis, Leon A. Gatys, Matthias Bethge
University of Tübingen, Centre for Integrative Neuroscience
{matthias.kuemmerer,thomas.wallis,leon.gatys,matthias}@bethgelab.org

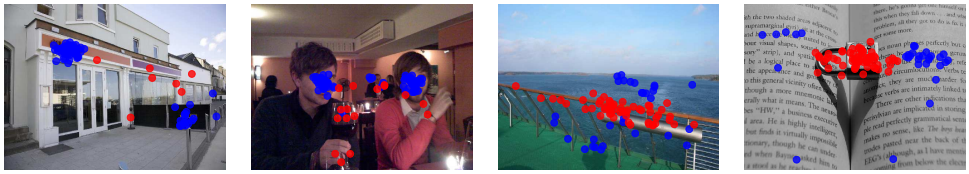


Figure 1: Representative examples for fixation prediction. Fixations are colored depending on whether they are better predicted by the high-level deep object features (DeepGaze II) model (blue) or the low-level intensity contrast features (ICF) model (red). This separates the images into areas where fixations are better predicted by high-level and low-level image features respectively. DeepGaze II is very good at predicting the human tendency to look at text and faces (first and second image), while ICF is better at predicting fixations driven by low-level contrast (third image). In particular, DeepGaze II fails if fixations are primarily driven by low-level features, although high-level features like text are present in the image (fourth image).

Abstract

Understanding where people look in images is an important problem in computer vision. Despite significant research, it remains unclear to what extent human fixations can be predicted by low-level (contrast) compared to high-level (presence of objects) image features. Here we address this problem by introducing two novel models that use different feature spaces but the same readout architecture. The first model predicts human fixations based on deep neural network features trained on object recognition. This model sets a new state-of-the art in fixation prediction by achieving top performance in area under the curve metrics on the MIT300 hold-out benchmark (AUC = 88%, sAUC = 77%, NSS = 2.34). The second model uses purely low-level (isotropic contrast) features. This model achieves better performance than all models not using features pre-trained on object recognition, making it a strong baseline to assess the utility of high-level features. We then evaluate and visualize which fixations are better explained by low-level compared to high-level image features. Surprisingly we find that a substantial proportion of fixations are better explained by the simple low-level model than the state-of-the-art model. Comparing different features within the same powerful readout architecture allows us to better understand the relevance of low- versus high-level features in predicting fixation locations, while simultaneously achieving state-of-the-art saliency prediction.

1. Introduction

Humans make several eye movements per second, *fixating* their high-resolution fovea on things they want to see. Understanding the factors that guide these eye movements is therefore an important component of understanding how humans process visual information and thus has a wide range of applications in image processing. In computer vision this problem is framed as *saliency prediction*[1]: predicting human fixation locations for a given image [21, 26, 25]. Saliency prediction performance has rapidly improved in the last few years, driven by the advent of models based on pre-trained deep neural networks. The models make use of convolutional filters that have been learned on other tasks, most notably object recognition in the ImageNet dataset [10]. The success of these saliency prediction models suggests that the high-level image features encoded by deep networks (e.g. sensitivity to faces, objects and text) are extremely useful to predict human fixation locations.

Despite recent advances, state-of-the-art models remain below the gold standard model of predicting one human's fixations from all others. Given the success of deep learning approaches, it may be tempting to believe that achieving gold standard performance simply requires employing even deeper, more abstracted feature sets. Here, we instead suggest that saliency prediction models may be neglecting low-level image features (local contrast) and overweighting

[1] Note that the term saliency prediction is sometimes also used in different context not related to eye movements.

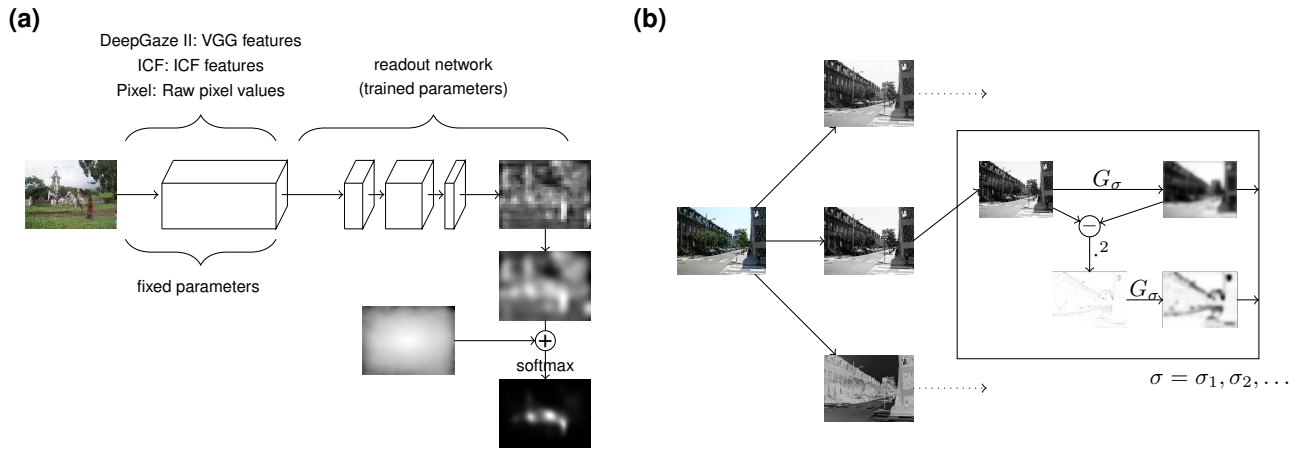


Figure 2: **(a)** The architecture of our models. Each model has a fixed feature space that feeds into the readout network: DeepGaze II uses VGG-19 features, ICF uses simple local intensity and contrast at different scales and the pixel model uses the raw pixel values. These feature activations are passed to a second neural network (the readout network) that is trained for fixation prediction. The readout network consists of four layers of 1×1 convolutions implementing a pixelwise nonlinear function. This results in a saliency map, which is then blurred, combined with a center bias and converted into a probability distribution by means of a softmax. **(b)** The ICF feature space. The network projects an RGB image onto the luminance and two color channels. For each channel we compute local intensities on 5 different scales using Gaussian convolutions. Additionally we square and blur the high-pass residuals from each scale to extract local contrast. The resulting 30 output channels are concatenated and constitute the input to the readout network.

the contribution of high-level features (the presence of objects such as faces or text) in explaining human fixations. We come to this conclusion via three novel contributions:

- A new state-of-the-art model for saliency prediction (the DeepGaze II model) that is based on deep neural network features pre-trained on object recognition [39].The model achieves top performance in area under the curve metrics on the MIT300 hold-out benchmark (AUC = 88%, sAUC = 77%, NSS = 2.34).

- A strong low-level baseline model for saliency prediction (Intensity Contrast Feature or ICF) that is based on local intensity and contrast. The model achieves top performance among all models not using features pre-trained on object recognition.

- Extensive quantitative and qualitative analysis to compare the predictions of these models. While DeepGaze II tends to perform better on images containing faces or text, the ICF model still performs better than DeepGaze II on about 10% of the images in the dataset.

Importantly, because both models use the same well-constrained readout architecture (see below), our comparison only reflects differences in the feature spaces (low- vs high-level).

2. Related Work

Beginning with the seminal image-computable model by Itti and Koch [21], many models have been proposed to predict fixations using local low-level features [52, 27], incorporating global features and statistics [15, 47, 18, 6, 14, 12, 37, 36], using simple heuristics [51] or combinations of low- and high-level features [26] (see [3] for a comprehensive review of saliency models before the advent of pre-trained deep features). In parallel, the effects of biases [42, 43, 45, 7] and tasks [38, 28, 44] on fixation placement have been studied. While these considerations are crucial, in this paper we are concerned not with top-down influences such as task, but rather we seek to understand to what extent fixations in free viewing are driven by low-level features or by high-level features [49, 11, 4, 9, 20, 5].

The state-of-the-art in saliency prediction improved markedly since 2014 with the advent of models using deep neural networks. The first model to use deep features (eDN; [48]) trained them from scratch to predict saliency. Subsequently, the DeepGaze I model showed that DNN features trained on object recognition (AlexNet [29] trained on the ImageNet dataset [10]) could significantly outperform training from scratch [32]. The success of this transfer-learning approach is exciting because it captializes on the presumably tight relationship between high-level tasks such as object recognition and human fixation location selection.

Since the initial success of transfer learning for saliency

prediction, a variety of new models followed this example to further improve saliency prediction performance. The SALICON model [19] fine tunes a mixture of deep features from AlexNet [29], VGG-16 [39] and GoogLeNet [41] for saliency prediction using the SALICON and OSIE [50] datasets. DeepFix [30] and PDP [22] fine-tune features from the VGG-19 network [39] for saliency prediction using the SALICON and the MIT1003 dataset. FUCOS [5] finetunes features trained on PASCAL-Context. SALICON and DeepFix substantially improved performance over DeepGaze I in the MIT benchmark ([8]; see below). The main difference of the new state-of-the art model we introduce here is that rather than fine-tuning the VGG-19 features for saliency prediction, we train a read-out network that uses a point-wise nonlinear combination of deep features. Furthermore we train our model in a probabilistic framework optimising the log-likelihood [31] and model the center bias as an explicit prior (as in Deep Gaze I [32]).

3. Models

We formulate our models as probabilistic models that predict fixation densities. Building on previous work applying probabilistic modelling to fixation prediction [2, 49], Kümmerer *et al*. [31, 33] recently showed that formulating existing models appropriately can remove most of the inconsistencies between existing model evaluation metrics. Furthermore, they argued that using log-likelihood as an evaluation criterion represents a useful and intuitive loss function for model evaluation, with close ties to information theory (though other loss functions may have advantages for some use cases [22]). Therefore we train and evaluate our models using the framework of log-likelihood (specifically reported as information gain explained, see [31]) and additionally report key metrics (AUC, sAUC and NSS) on the MIT1003 dataset and from the MIT Saliency Benchmark.

3.1. Deep Object Features (DeepGaze II) model

Here we describe the architecture of our saliency prediction model that is based on deep features that are trained on object recognition (Fig. 2). A given input image is subsampled by a factor 2 and passed through the normalized VGG-19 network for which all filters have been rescaled to yield feature maps with unit mean over the ImageNet dataset [13]. Next, the feature maps of a selection of high-level convolutional layers (conv5_1, relu5_1, relu5_2, conv5_3, relu5_4; selected via random search, see supplement) are up-sampled by a factor of 8 such that spatial resolution is sufficient for precise prediction. These feature maps are then combined into one 3-dimensional tensor with 2560 (5×512) channels, which is used as input for a second neural network that we term the *readout network*. This readout network consists of four layers of 1×1 convolutions followed by ReLu nonlinearities. The first three layers use 16, 32, and 2 features (see supplement for details). The last layer has only one output channel $O(x, y)$. Crucially, the readout network is only able to represent a *point-wise* nonlinearity in the VGG features. This means that the readout network is only able to learn interactions between existing features across channels but not across pixels—i.e. it cannot learn new spatial features.

The final output from the readout network is convolved with a Gaussian to regularize the predictions:

$$S(x, y) = O(x, y) \star G_\sigma \tag{1}$$

Fixations tend to be near to the center of the image in a way which is strongly task and dataset dependent [42]. Therefore we explicitly model the center bias as a prior distribution that is added to S:

$$S'(x, y) = S(x, y) + \log p_{\text{baseline}}(x, y) \tag{2}$$

We use a Gaussian Kernel density estimate over all fixations from the training dataset for p_{baseline} (for more details see 3.4). Finally, $S'(x, y)$ is converted into a probability distribution over the image by the means of a softmax (as for DeepGaze I and for PDP):

$$p(x, y) = \frac{\exp(S'(x, y))}{\sum_{x,y} \exp(S'(x, y))} \tag{3}$$

3.2. Intensity Contrast Feature (ICF) model

The architecture of our low-level ICF model closely follows that of DeepGaze II (Fig. 2). The main difference is that we replace the VGG features that were trained on object recognition by a feature space that can only extract purely low-level image information (intensity and intensity contrast).

To that end we first subsample the image by a factor of 2 and project the RGB color channels onto their principal components for natural images (computed on the MIT1003 dataset, see supplement), which yields the luminance channel and two color channels. For each of these channels we independently compute local intensity and contrast at different spatial scales. For local intensity we simply compute a Gaussian Pyramid with 5 different scales. The standard deviations the Gaussian kernels are 5,10,20,40,80 px and the window size is 171 px. We use nearest-padding so that the output feature map has the same spatial dimensions as the input feature map. For local contrast we first compute 5 high-pass residuals by subtracting each level of the Gaussian Pyramid from the input channel. Then we square these residuals to compute pixel-wise contrast and finally we blur the squared residuals with the same Gaussian kernel that was used to compute the residual (Fig. 2). This procedure yields 5 intensity and 5 contrast feature maps for each input channel and thus results in 30 feature maps that constitute

the input to the readout network. The readout network and the following stages (blurring and adding of center bias) are the same as for DeepGaze II.

3.3. Pixel model

To compute a baseline that evaluates how powerful the readout network is on its own, we also trained a model that applies the readout network and the following stages directly to the RGB pixel values. This model computes no spatial features and can only learn non-linear combinations of the color channels.

Our models are implemented using Lasagne and Theano [1]. For the computation of the VGG features we used the caffe toolbox [23].

3.4. Model Training

Our models are trained using maximum likelihood learning (see [31] for an extensive discussion of why log-likelihoods are a meaningful metric for saliency modelling). If $p(x, y \mid I)$ denotes the probability distribution over coordinates x and y predicted by our model for an image I, the log-likelihood of a dataset is

$$\frac{1}{N}\sum_{i}^{N}\log p(x_i, y_i|I_i), \tag{4}$$

where i indexes the N fixations in the groundtruth data: The ith fixation occured in the image refered to by I_i, at location $(x_i,\ y_i)$. For both models we minimize this loss function only with respect to the parameters of the readout network and the kernel size of the Gaussian used to regularize the prediction. Since the loss function is differentiable in these parameters, we can use the of-the-shelf *Sum-of-Functions-Optimizer* (SFO, [40]), a mini-batch-based version of L-BFGS.

The feature representations that feed into the readout network (VGG for the high-level and local mean and contrast for the low-level model) are kept fixed during training.

In the pretraining phase, the readout network is initialized with random weights and trained on the SALICON dataset [24]. This dataset consists of 10000 images with pseudofixations from a mouse-contingent task and has proven to be very useful for pretraining saliency models [19, 22, 30]. All images are downsampled by a factor of two. We use 100 images per mini-batch for the SFO. All fixations from the SALICON dataset are used to compute the centerbias.

The MIT1003 dataset is used to determine when to stop the training process. After each iteration over the whole dataset (one epoch) we calculate the performance of the model on the MIT1003 (test) dataset. We wish to stop training when the test performance starts to decrease (due to overfitting). We determine this point by comparing the performance from the last three epochs to the performance five epochs before those. Training runs for at least 20 epochs, and is terminated if all three of the last epochs show decreased performance or if 800 epochs are reached. As it is more expensive to use images of many different sizes, we resized all images from the MIT1003 dataset to either a size of 1024×768 or 768×1024 depending on their aspect ratio, before downsampling by a factor of two. All fixations from the MIT1003 dataset except the ones from the image in question are used to compute the centerbias.

After pre-training, the model is fine-tuned on the MIT1003 dataset and performance is cross-validated over images: the images from the dataset are randomly split into 10 parts of equal size. Then ten models are trained starting from the result of the pre-training, each one using 8 of the 10 parts for training, one part for the stopping criterion (following the stopping criterion as above) and keeping one part for testing. All fixations from the training set are used to compute the centerbias for training, validation and test purposes. We use 10 images per mini-batch in the SFO. For evaluation on the MIT300 benchmark dataset we train on MIT1003 using a ten-fold 9-1 training-validation split and average the predictions from the resulting models, using all fixations from the MIT1003 dataset for the centerbias.

Model	AUC	sAUC	NSS
DeepGaze I [32]	84%	66%	1.22
DSCLRCN [34]	87%	72%	**2.35**
DeepFix [30]	87%	71%	2.26
SALICON [19]	87%	74%	2.12
DeepGaze II	**88%**	**77%**	2.34

Table 1: DeepGaze II performance in the MIT300 Saliency Benchmark. DeepGaze II achieves top performance in both AUC and sAUC, and comes a close second in NSS. Note that we use saliency maps without center bias for the sAUC result (see text for more details).

3.5. Model Evaluation

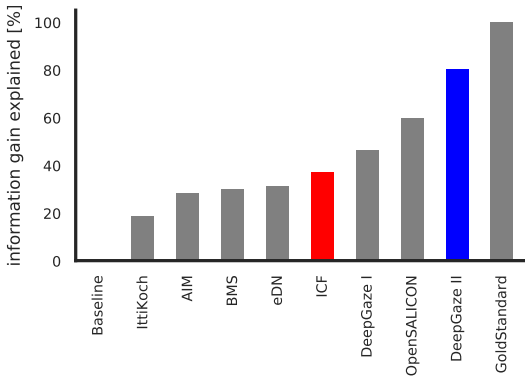
To evaluate model performance we focus on computing *information gain* for its intuitive information-theoretic properties. We additionally report more classic metrics (AUC, sAUC and NSS) to compare to other recent models. Finally, we also report the performance of DeepGaze II on the MIT300 hold-out test set [8].

Information gain tells us what the model knows about the data beyond a given baseline model [31], for which we use the image-independent center bias, expressed in bits / fixation:

$$IG(\hat{p}\|p_{\text{baseline}}) = \frac{1}{N}\sum_{i}\log\hat{p}(x_i, y_i|I_i) - \log p_{\text{baseline}}(x_i, y_i) \tag{5}$$

Here $\hat{p}(x, y|I)$ is the density of the model at location (x, y)

(a)



(b)

Model	IG	IGE	AUC	sAUC	NSS
Centerbias	0.00	0.0	79.6	50.0	1.22
Pixel	0.13	10.7	81.2	60.2	1.38
IttiKoch [16]	0.23	18.6	82.3	64.1	1.41
AIM [6]	0.27	22.6	82.9	65.6	1.50
eDN [48]	0.38	31.1	83.8	68.7	1.61
ICF	0.45	37.2	84.4	70.1	1.74
DeepGaze I [32]	0.56	46.1	85.8	73.0	1.92
OpenSALICON [46]	0.73	59.7	86.4	74.2	2.14
DeepGaze II	**0.98**	**80.3**	**88.3**	**77.7**	**2.48**
Gold Standard	1.22	100.0	89.9	81.2	2.82

Figure 3: Performance on the MIT1003 dataset. **(a)** Ranking of the models according to information gain explained. Our models are marked by the colored bars. All models to the right of ICF use pre-trained deep features. **(b)** Detailed results for a larger set of metrics. IG = information gain (bits / fixation), IGE = information gain explained (%), AUC = area under the ROC curve (%), sAUC = shuffled area under the ROC curve (%), NSS = normalized scanpath saliency.

when viewing image I, and p_{baseline} is the density of the baseline model.

To evaluate the absolute performance of a model we also compute *information gain explained*. This relates the model's performance to the performance of a gold standard model that predicts one subject's fixations for a given image from the fixations of all other subjects using a Gaussian kernel density estimate.

In particular, it is the proportion of the gold standard information gain accounted for by the model:

$$\frac{IG(p\|p_{\text{baseline}})}{IG(p_{\text{gold}}\|p_{\text{baseline}})} \tag{6}$$

where p_{gold} is the density of the gold standard model. Thus it intuitively ranks a model on a scale from 0 to 1, where 0 is a model that does not know the image and 1 is a perfect model that is only limited by inter-subject variablility.

Additionally, we evaluate the traditional area under the ROC curve metrics *AUC* and *sAUC* and the more recent Normalized Scanpath Saliency (NSS, [35]). For AUC and NSS the model's density prediction is the right saliency map to use for evaluation. For sAUC we need to divide the density prediction by the center bias density (which is the non-fixation density in that case) [33].

In all our results we report the test performance of the models. Specifically, for each image in the MIT1003 dataset there is exactly one model from the fine-tuning crossvalidation procedure that did not use that image for training or validation. We use the density prediction from this model to evaluate model performance for that image. For the gold standard model we report leave-one-subject-out crossvalidation performance (which is an image-specific prediction crossvalidated over subjects).

To obtain meaningful results for other models on the information gain metric, we applied the procedure suggested by [31] to convert them to probabilistic models. Specifically, this involves optimizing a pointwise nonlinearity and a center bias (unlike [31], here we do not optimize a blur kernel for the models because all state-of-the-art models produce smooth saliency maps). The conversion usually improves the performance of the models also on the classic metrics. Thus we only report the post-conversion model performances for these models below.

4. Results

4.1. MIT300 Saliency Benchmark

Here we report the performance of our Deep Object Feature model DeepGaze II on the MIT saliency benchmark (the held-out MIT 300 set) (Table 1). DeepGaze II beats the nearest competitors SALICON, DeepFix and DSCLRCN [34] by one percent in AUC. For shuffled AUC, our model beats the nearest competitors by a larger margin. DSCLRCN beats our model by a small margin on NSS (note that this model was optimized for NSS).

Because the MIT Benchmark requires submission of model predictions as JPEG images, one must decide how to store the saliency maps as JPEG images. For AUC, we quantized the density for each image into 256 values such that each value receives the same number of pixels. For sAUC, we divided the density by the density of the MIT1003 center bias and quantized as above. For NSS we quantized the density without histogram normalization. Note that this does not mean we report the results of three different models. The different metrics interpret the saliency maps differently and we translated the predic-

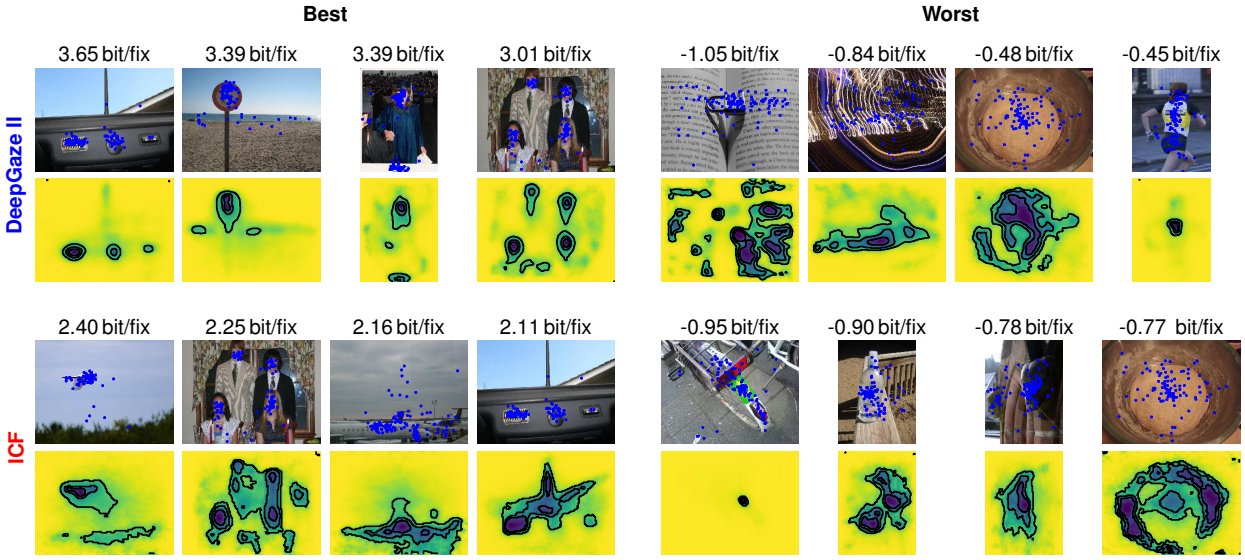


Figure 4: Example images and predictions. For both DeepGaze II and the ICF model we present the best (left) and worst (right) four images with respect to information gain. Ground-truth fixations are plotted in blue over the images. Below each image we show the prediction of the corresponding model. Above each stimulus we report the information gain performance of the model on this image.

tions of our model into the language of the different metrics (without any retraining, see [33] for details). This could partially explain the larger difference between our model and competitor models on sAUC: most state-of-the-art models include a center bias and evaluate sAUC on the saliency maps with a center bias, resulting in a penalty.

4.2. MIT1003 dataset

The MIT300 hold-out set determines the state-of-the-art in saliency prediction. However, precisely because its ground-truth fixations are not publicly available, it is not useful for understanding why models perform the way they do. To develop a deeper understanding of the performance of DeepGaze II and compare it to the ICF model, we therefore evaluate test performance for the MIT1003 dataset. This also allows us to compare models using the intuitive information gain measure described above. Unfortunately we cannot include some recent and competitive models in this analysis (SALICON and DeepFix) because their code is not publicly available. To give at least an approximate result for the previous state-of-the-art, we include results for the OpenSALICON implementation [46].

We evaluate a number of important saliency models using information gain explained (Fig. 3). We display the ranking of the models in Figure 3(a). Our Pixel Model performs the worst, but still remarkably well, accounting for 10% of the information gain of the gold standard over the center bias. Next are models that use hand-crafted low-level features (AIM and BMS) and a convolutional network that is trained from scratch (eDN). Our low-level baseline, the ICF model performs best among all models that do not use pre-trained deep features and accounts for a remarkable 37% of the information gain. Top performance is achieved by models that use pre-trained deep neural network features such as DeepGaze I, OpenSALICON and our new state-of-the-art model DeepGaze II, which can explain 81% of the information gain. Additionally we report the classic measures, AUC, sAUC and NSS to show their consistency with information gain (Fig. 3(b)).

See the supplement for details on how the readout network, the VGG features and pretraining on SALICON contribute to the performance of DeepGaze II.

4.3. What features drive human fixation locations?

Here we compare our low-level ICF and high-level DeepGaze II saliency models to improve our understanding of the features that can explain human fixation locations.

First we look at the images for which each model performs best and worst compared to the center bias and show the respective saliency predictions of the models (Fig. 4). We find that the ICF model performs best on images for which fixations are localized in high contrast regions, for example when there is a single plane in the blue sky (Fig. 4, bottom left panel, first image). At the same time it performs worst when there is a high contrast region that does not attract human fixations or attracts them only in part. For example, it expects people to fixate exclusively on the colored sticker on the bike whereas true fixations are more scattered in the image (Fig. 4, bottom right panel, first image). Note that even though the model only extracts low-level fea-

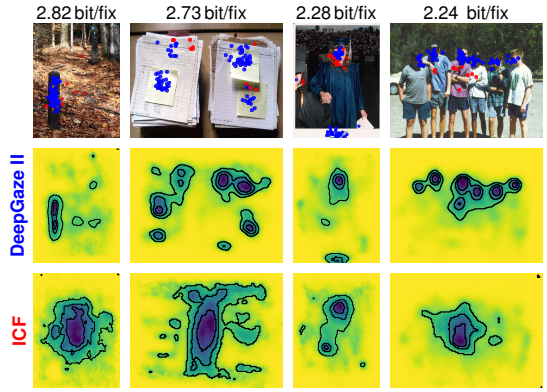


Figure 5: Images for which DeepGaze II has the largest improvement over ICF. Fixations that are better explained by DeepGaze II are colored in blue. Fixations that are better explained by ICF are colored in red. Fixations best explained by the center bias are omitted. Predicted fixation densities for both models are plotted below the images. Above each stimulus we report the difference in information gain between DeepGaze II and ICF for this image.

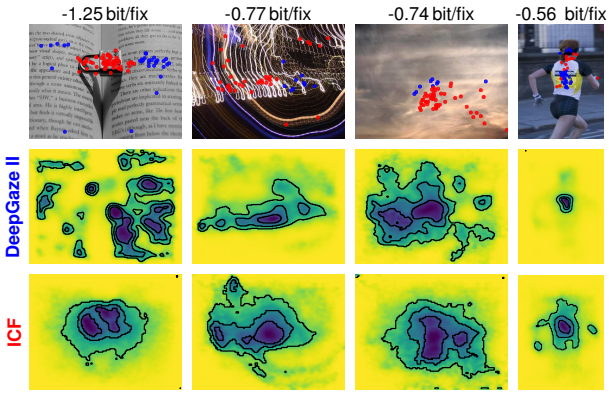


Figure 6: Images for which ICF has the largest improvement over DeepGaze II. Other elements as in Figure 5.

tures, it can still perform well on images where fixations are driven by high-level features such as human faces—if the presence of a face is correlated with the local intensity or contrast of the image (Fig. 4, bottom left panel, second image).

We find that DeepGaze II excels at predicting fixations that are driven by the presence of objects, such as controls in a car, a roadsign or human faces (Fig. 4, top left panel). It fails for images where high-level content is not associated with fixations (e.g. the text in Fig. 4, top right panel, first and last image) or images that are texture-like without any particular objects (Fig. 4, top right panel, second image).

Even though the best images for ICF and DeepGaze II are partly the same, the predicted saliency maps clearly separate the models. While DeepGaze II is extremely accurate

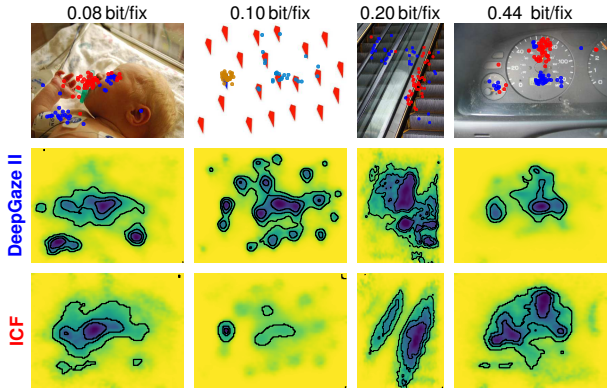


Figure 7: Images for which DeepGaze II and ICF show similar performances but predict the fixations in different locations, separating the image into areas of low-level and high-level fixations. Other elements as in Figure 5, except that in the second image ICF fixations are colored orange and DeepGaze II light blue to better separate them from the blue and red elements in the image.

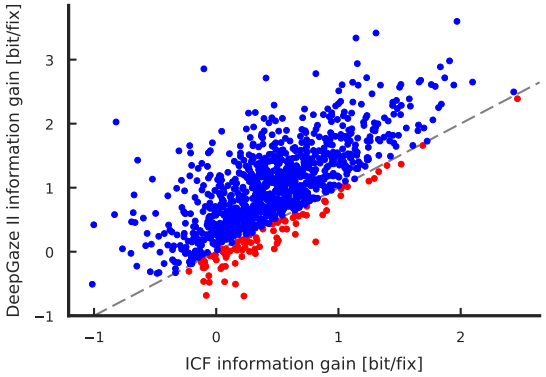


Figure 8: Performances of DeepGaze II and ICF on MIT1003. Each point corresponds to one image, with the performance of DeepGaze II (y-axis) and the ICF model (x-axis) on that image expressed as information gain relative to the center bias. For images above the diagonal (blue dots) DeepGaze II is better than the ICF model, while for images below the diagonal (red dots) the ICF model is better.

at predicting fixations at the location of the important high-level objects (faces, text), the ICF model also predicts fixations at other high-contrast locations in the images.

The difference between the models is made more explicit by looking at the images for which DeepGaze II is maximally better than ICF (Fig. 5). One advantage of training two separate models is that we can easily assess which individual fixations within an image are better explained by each of the models. This allows us to better understand which features drive fixations. In Figure 5, DeepGaze II correctly predicts a concentration of fixations over text (two leftmost images) and faces (two rightmost images) whereas

the ICF model is 'distracted' by high-contrast regions of the image that do not correspond to the presence of objects [49]. For example, ICF strongly predicts fixations in the high-contrast gap between the stacks of papers in the second image. It also predicts high fixation probability for skin on a dark background no matter whether it is the face or the hand of a person (third image). In contrast, DeepGaze II only predicts high fixation probability for the face, in agreement with the ground-truth data.

On the other hand, in Figure 6, we show images where the ICF model performs better than DeepGaze II. In two examples (first and fourth images), DeepGaze II seems to be distracted by high-level features that humans tend not to fixate. For example, in the first image, DeepGaze II predicts that humans will look at the text printed in the book whereas ICF correctly predicts that humans will fixate the padlock lying over the page (forming a high-contrast region). Similarly, in the fourth image, DeepGaze II predicts fixations on the text on the runner's shirt whereas the runner's head and shoulders happen to correspond to higher-local-contrast regions (which are picked up by the ICF model). The middle two images show abstract patterns (motion blur and clouds) for which human fixations appear to be better explained by local contrast in the absence of high-level features.

Finally, we show a sample of images in which DeepGaze II and ICF show similar performance at the image level but predict fixations in different locations (Figure 7). In the first image, DeepGaze II correctly predicts fixations to the baby's eyes and to the text on the arm, but ICF correctly predicts fixations to the pacifier. In the second image, ICF correctly predicts fixations to the color-singleton search item (blue element amongst red) but fails to predict fixations elsewhere. DeepGaze II predicts fixations to the glass window whereas ICF predicts fixations to the high-contrast border of the escalator in the third image, and DeepGaze II predicts fixations to text but not the needle of the speedometer in the fourth image.

The comparison images we have highlighted above show that DeepGaze II can correctly predict fixations to high-level features such as text and faces (see also examples in Figure 1), in accordance with its status as a far more powerful model than ICF (more parameters with pre-trained features). However, there are striking failure cases when comparing against the ICF model, in particular when high-level features are present in the image but are not fixated (e.g. the text and padlock image in Figure 6). On the MIT1003 dataset as a whole, we find that there is a substantial subset of images (94 of 1003) for which the ICF model produces better predictions than DeepGaze II (Figure 8). In terms of individual fixations this proportion is even higher, with around 25% of the fixations in the dataset being better explained by ICF than either DeepGaze II or the center bias. Given the simplicity of the ICF model relative to DeepGaze II, this is remarkable. Because in principle DeepGaze II should also have access to low-level features [17], this result suggests that DeepGaze II may be underweighting the importance of low-level features in guiding fixations.

5. Discussion

In this paper we compare the predictive performance of low- and high-level features for saliency prediction by introducing two new saliency models that use the same readout architecture on top of different feature spaces. DeepGaze II uses transfer learning from the VGG-19 deep neural network to achieve state-of-the-art performance on the MIT300 benchmark. The ICF model uses simple intensity contrast features to achieve better performance than all models that do not use pre-trained deep features.

While the high-level DeepGaze II model significantly outperforms low-level ICF for the dataset as a whole, we find a surprisingly large set of images for which the ICF model is better than DeepGaze II. Thus, while high-level features (the presence of objects, faces and text) are very important for explaining free viewing behaviour in natural scenes [11, 44], our results show that low-level local contrast features do make a small but dissociable contribution over a representative scene database (see also [7, 5]).

The fact that the simple ICF model outperforms all models before transfer learning of deep features shows that the predictive value of low-level features has been historically underestimated. One possible reason for this is that many historical models were not trained on data but rather hand-tuned. On the other hand, the ICF model is isotropic—it does not even have access to orientation filters—which makes its performance improvement relative to earlier models even more remarkable.

Our results suggest that explicitly modelling low-level contributions to saliency could be used to improve the robustness of saliency models. In future work it may prove fruitful to train the DeepGaze II and ICF models jointly, reducing DeepGaze II's tendency to over-emphasize the importance of high-level image structure. Ultimately however, we believe that improvements will come from a better understanding of what features causally drive fixation behaviour, including different task constraints [44, 28].

We provide a webservice to test our models on arbitrary stimuli at `deepgaze.bethgelab.org`.

6. Acknowledgements

Funded by the the German Excellency Initiative (EXC307), the German Science Foundation (DFG; priority program 1527, BE 3848/2-1 and Collaborative Research Centre 1233), the German Academic Foundation and the BCCN Tübingen (BMBF; FKZ: 01GQ1002).

References

[1] R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, A. Belopolsky, Y. Bengio, A. Bergeron, J. Bergstra, V. Bisson, J. Bleecher Snyder, N. Bouchard, N. Boulanger-Lewandowski, X. Bouthillier, A. de Brébisson, O. Breuleux, P.-L. Carrier, K. Cho, J. Chorowski, P. Christiano, T. Cooijmans, M.-A. Côté, M. Côté, A. Courville, Y. N. Dauphin, O. Delalleau, J. Demouth, G. Desjardins, S. Dieleman, L. Dinh, M. Ducoffe, V. Dumoulin, S. Ebrahimi Kahou, D. Erhan, Z. Fan, O. Firat, M. Germain, X. Glorot, I. Goodfellow, M. Graham, C. Gulcehre, P. Hamel, I. Harlouchet, J.-P. Heng, B. Hidasi, S. Honari, A. Jain, S. Jean, K. Jia, M. Korobov, V. Kulkarni, A. Lamb, P. Lamblin, E. Larsen, C. Laurent, S. Lee, S. Lefrancois, S. Lemieux, N. Léonard, Z. Lin, J. A. Livezey, C. Lorenz, J. Lowin, Q. Ma, P.-A. Manzagol, O. Mastropietro, R. T. McGibbon, R. Memisevic, B. van Merriënboer, V. Michalski, M. Mirza, A. Orlandi, C. Pal, R. Pascanu, M. Pezeshki, C. Raffel, D. Renshaw, M. Rocklin, A. Romero, M. Roth, P. Sadowski, J. Salvatier, F. Savard, J. Schlüter, J. Schulman, G. Schwartz, I. V. Serban, D. Serdyuk, S. Shabanian, E. Simon, S. Spieckermann, S. R. Subramanyam, J. Sygnowski, J. Tanguay, G. van Tulder, J. Turian, S. Urban, P. Vincent, F. Visin, H. de Vries, D. Warde-Farley, D. J. Webb, M. Willson, K. Xu, L. Xue, L. Yao, S. Zhang, and Y. Zhang. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[2] S. Barthelmé, H. Trukenbrod, R. Engbert, and F. Wichmann. Modelling fixation locations using spatial point processes. *Journal of Vision*, 13(12), 2013.

[3] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, Jan. 2013.

[4] A. Borji, D. N. Sihite, and L. Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhauser et al.'s data. *Journal of vision*, 13(10):18, Jan. 2013.

[5] N. D. Bruce, C. Catton, and S. Janjic. A deeper look at saliency: feature contrast, semantics, and beyond. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 516–524, 2016.

[6] N. D. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3), 2009.

[7] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95–112, nov 2015.

[8] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[9] M. Cerf, J. Harel, W. Einhaeuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 241–248. Curran Associates, Inc., 2008.

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

[11] W. Einhauser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, Nov. 2008.

[12] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4), 2013.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[14] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012.

[15] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[16] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[17] H. Hong, D. L. K. Yamins, N. J. Majaj, and J. J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622, Feb. 2016.

[18] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[19] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[20] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005.

[21] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[22] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2016.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[24] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical and Electronics Engineers (IEEE), jun 2015.

[25] T. Judd, F. d. Durand, and A. Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations A Benchmark of Computational Models of Saliency to Predict Human Fixations. *CSAIL Technical Reports*, 2012.

[26] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[27] W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann. Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 9(5), 2009.

[28] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein. What do saliency models predict? *Journal of Vision*, 14(3):14–14, mar 2014.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[30] S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *CoRR*, abs/1510.02927, 2015.

[31] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, dec 2015.

[32] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *2015 International Conference on Learning Representations - Workshop Track (ICLR)*, 2015.

[33] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Saliency benchmarking: Separating models, maps and metrics. *arXiv e-prints*, abs/1704.08615, 2017.

[34] N. Liu and J. Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *arXiv preprint arXiv:1610.01708*, 2016.

[35] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, aug 2005.

[36] S. Rahman and N. Bruce. Saliency, scale and information: Towards a unifying theory. In *Advances in Neural Information Processing Systems 28*, pages 2188–2196, 2015.

[37] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit. RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642–658, July 2013.

[38] C. A. Rothkopf, D. H. Ballard, and M. M. Hayhoe. Task and context determine where you look. 7(14):16–16.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[40] J. Sohl-Dickstein, B. Poole, and S. Ganguli. Fast large-scale optimization by unifying stochastic gradient and quasi-newton methods. *CoRR*, abs/1311.2115, 2013.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[42] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, 2007.

[43] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643 – 659, 2005.

[44] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5), 2011.

[45] B. W. Tatler and B. T. Vincent. Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, 2(2):1–18, 2008.

[46] C. Thomas. Opensalicon: An open source implementation of the salicon saliency model. *CoRR*, abs/1606.00110, 2016.

[47] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[48] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Computer Vision and Pattern Recognition, 2014. CVPR'14. IEEE Conference on*. IEEE, 2014.

[49] B. T. Vincent, R. J. Baddeley, A. Correani, T. Troscianko, and U. Leonards. Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, 17(6-7):856–879, 2009.

[50] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014.

[51] J. Zhang and S. Sclaroff. Saliency detection: a boolean map approach. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 153–160. IEEE, 2013.

[52] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

Understanding Low- and High-Level Contributions to Fixation Prediction: Supplementary Material

1 Contributions of architectural components to performance

Our DeepGaze II model uses a similar architecture to DeepGaze I [1], with four primary changes: replacing AlexNet features by VGG features, using a readout network instead of a linear readout, pre-training on the SALICON dataset, and using image-wise crossvalidation over the full MIT1003 dataset rather than subject-wise crossvalidation over only a subset. We quantified the contributions of these changes to achieving our model performance using the full MIT1003 dataset. As seen in Table 1, switching to image-wise crossvalidation on the full dataset (DeepGaze I′) provides a substantial performance boost over the original DeepGaze I model. After considering this change, the largest single improvement over DeepGaze I′ comes from using the pre-trained VGG features in place of AlexNet (though note that we also include more channels from VGG than from AlexNet). Training DeepGaze I′ on the SALICON dataset does not change performance, suggesting that the 258 parameters of this model are already sufficiently constrained by the MIT1003 dataset. Combining SALICON pre-training with the VGG features yields the largest intermediate model performance improvement. Using the readout network without additional pre-training on the SALICON dataset never gives substantially better performance (compare DeepGaze I′ to "readout network", or "VGG" to "readout net + VGG"), suggesting that SALICON pre-training is required for the readout network to avoid overfitting.

Model	IG	IGE	AUC	sAUC	NSS
Centerbias	0.00	0.0	79.6	50.0	1.22
DeepGaze I	0.56	46.1	85.8	73.0	1.92
DeepGaze I′	0.76	62.3	86.9	75.0	2.16
readout network	0.75	62.0	87.0	75.0	2.16
SALICON	0.76	62.6	86.9	75.0	2.16
VGG	0.84	69.3	87.7	76.4	2.32
Readout net+SALICON	0.82	67.5	87.3	75.6	2.25
Readout net+VGG	0.85	70.0	87.3	76.2	2.34
SALICON+VGG	0.90	74.3	88.0	76.9	2.42
DeepGaze II	**0.98**	**80.3**	**88.3**	**77.7**	**2.48**
Gold Standard	1.22	100.0	89.9	81.2	2.82

Table 1: Contributions of changes between DeepGaze I and DeepGaze II to performance. DeepGaze I′ is the DeepGaze I model trained with image-wise crossvalidation over the full MIT1003 dataset just like our models. "Readout network" = replacing a linear readout with a nonlinear readout network, "VGG" = replacing AlexNet with VGG features, "SALICON" = pre-training on the SALICON dataset. Metrics as in main paper. The primary improvement in our model compared to DeepGaze I′ comes from using VGG features.

2 Readout network

Our model architecture uses a readout network consisting of multiple layers of 1×1 convolutions on top of a fixed set of features. This allows the models to learn nonlinear combinations of the features and fit the scale of the final log density better while still being comparatively constrained. We estimate how much these two features contribute to the performance when compared to a simple linear readout for ICF and DeepGaze II. In Figure 1, we show models with different readout networks: first, we just use a linear readout as baseline to compare to. Then we use a readout network with layers of 1, 128 and 1 channels ("LN"). Since the first layer has only one feature, this allows the readout network only to learn a nonlinear transformation of a saliency map but keeps it from exploiting interactions between features. Finally we show the performance of the model with the full readout network, which therefore is able to fit the log density scale as well as make use of interactions between features.

We find that the linear DeepGaze II model already accounts for roughly 74% of the explainable information gain. The LN readout network manages to close around two thirds of the performance gap to the full readout network, indicating that DeepGaze II mainly uses the readout network to transform the scale of the saliency prediction and not so much to exploit interactions between features.

For the ICF model on the other hand, the LN readout network increaes the performance only by one third of the difference between the linear readout and the full readout. This shows that the ICF model makes much more use of interactions between features and DeepGaze II.

In Figure 2 we compare the performance of DeepGaze II when using different depths for the readout network. Going from a purely linear readout to one hidden layer gives more than half of the performance gain to the final model with three hidden layers. Two hidden layers yields a performance which is only slightly worse than three hidden layers.

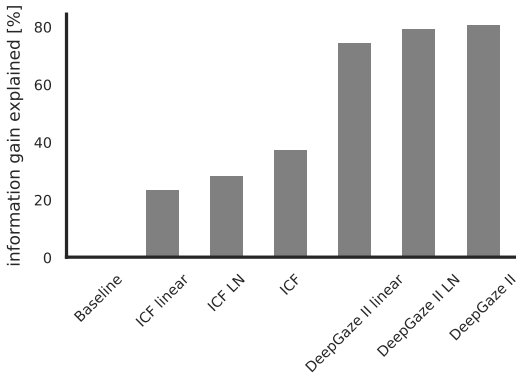


Figure 1: Performances of ICF and DeepGaze II when using either a linear readout, a linear-nonlinear readout network with layers of 1, 128 and 1 channels which cannot exploit feature interactions and the full readout network as described in the main paper.

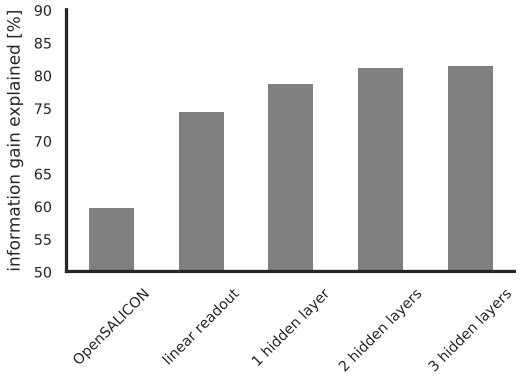


Figure 2: Influence of the depth of the readout network. We show the performance of DeepGaze II when using a linear readout, one hidden layer (16 units), two hidden layers (16 and 32 units) and the final readout network with three hidden layers of 16, 32 and 2 units.

3 VGG features

In DeepGaze II presented in the main paper, we use the conv5_1, relu5_1, relu5_2 conv5_3 and relu5_4 layers from VGG-19 as feature space. These layers have been

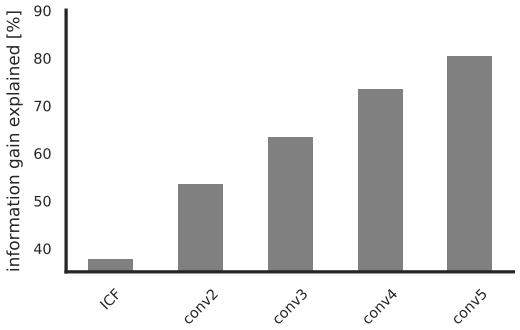


Figure 3: Performance of DeepGaze II when using features from different levels in VGG.

choosen with a random search which trained models using a random selection of layers from the conv4 and conv5 blocks. To compare the predictive power of the different layer blocks in VGG-19, in Figure 3 we show the performance of DeepGaze II when using features from conv2, conv3, conv4 or conv5. For conv3 and conv4 we used conv$n$_1, relu$n$_1, relu$n$_2 conv$n$_3 and relu$n$_4, corresponding to the layers from conv5 used in the final model. For conv2 we used conv2_1, relu2_1, conv2_2, relu2_2. The performances increase steadily from the conv2 model to the conv5 model, but already the conv2 model is significantly better than the ICF model.

4 Principal component analysis for ICF features

The ICF model projects the RGB color channels onto their principal components for natural images. We computed the principal components using all pixels in the MIT1003 dataset. The resulting compontents are up to small deviations: 1) grayscale intensity 2) 50/50 Red/Green 3) 25/50/25 Red/Blue/Green. This color space is not likely to be overfit to the MIT1003 datset, because the SALICON dataset gave almost identical numbers.

References

[1] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In *2015 International Conference on Learning Representations - Workshop Track (ICLR)*, 2015. 1

Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics

Matthias Kümmerer, Thomas S.A. Wallis and Matthias Bethge

Published in The European Conference on Computer Vision (ECCV), 2018, pp. 770-787

Abstract

Dozens of new models on fixation prediction are published every year and compared on open benchmarks such as MIT300 and LSUN. However, progress in the field can be difficult to judge because models are compared using a variety of inconsistent metrics. Here we show that no single saliency map can perform well under all metrics. Instead, we propose a principled approach to solve the benchmarking problem by separating the notions of saliency models, maps and metrics. Inspired by Bayesian decision theory, we define a saliency model to be a probabilistic model of fixation density prediction and a saliency map to be a metric-specific prediction derived from the model density which maximizes the expected performance on that metric given the model density. We derive these optimal saliency maps for the most commonly used saliency metrics (AUC, sAUC, NSS, CC, SIM, KL-Div) and show that they can be computed analytically or approximated with high precision. We show that this leads to consistent rankings in all metrics and avoids the penalties of using one saliency map for all metrics. Our method allows researchers to have their model compete on many different metrics with state-of-the-art in those metrics: "good" models will perform well in all metrics.

Contributions

I developed the idea of deriving metric-specific saliency maps from a fixation density on my own. I also did all the mathematical work of deriving the presented saliency maps. Matthias Bethge suggested that what I had done can be phrased as an application of Bayesian Utility Theory. The way of converting existing saliency map models into probabilistic models of fixation density prediction was taken from Kümmerer et al. 2015b discussed above and also included in this thesis. I implemented all saliency map computations, ran all experiments and did all analyses. The first paper draft was written by me with consulting from Thomas Wallis. This draft was then improved in joint work with Thomas Wallis. All authors contributed to scientific discussions and paper revisions.

Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics

Matthias Kümmerer[1], Thomas S.A. Wallis[1,2], and Matthias Bethge[1]

[1] Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany
[2] Wilhelm-Schickard Institute for Computer Science (Informatik), University of Tübingen, Tübingen, Germany
{matthias.kuemmerer,tom.wallis,matthias}@bethgelab.org

Abstract. Dozens of new models on fixation prediction are published every year and compared on open benchmarks such as MIT300 and LSUN. However, progress in the field can be difficult to judge because models are compared using a variety of inconsistent metrics. Here we show that no single saliency map can perform well under all metrics. Instead, we propose a principled approach to solve the benchmarking problem by separating the notions of saliency models, maps and metrics. Inspired by Bayesian decision theory, we define a saliency model to be a probabilistic model of fixation density prediction and a saliency map to be a metric-specific prediction derived from the model density which maximizes the expected performance on that metric given the model density. We derive these optimal saliency maps for the most commonly used saliency metrics (AUC, sAUC, NSS, CC, SIM, KL-Div) and show that they can be computed analytically or approximated with high precision. We show that this leads to consistent rankings in all metrics and avoids the penalties of using one saliency map for all metrics. Our method allows researchers to have their model compete on many different metrics with state-of-the-art in those metrics: "good" models will perform well in all metrics.

Keywords: saliency, benchmarking, metrics, fixations, Bayesian decision theory, model comparison

1 Introduction

Humans have a foveated visual system: only a small central part of the retina has high receptor density allowing the perception of the details of a scene. Therefore humans make eye movements to place the high resolution fovea on things they want to see. Understanding where they choose to look is therefore an important component of understanding behaviour.

A long-standing account of bottom-up attentional guidance posits the existence of a "saliency map" (or maps) in the human brain [48,26]. Here, a saliency map represents spatial importance, usually defined to be local contrast in low-level features such as luminance, color or orientation. Since Itti and Koch formulated this concept into their seminal image-based model [17], a large number

of models have been proposed for predicting fixations from image features, e.g. [15,56,25,6,24,55,1] and more recently many models based on deep learning, e.g. [49,30,16,28,36,31]; see [4,19] for extensive reviews of the literature. New models are published on a regular basis with contributions coming mainly from the communities of computer vision and psychology. It has been extensively discussed which effects are important for fixation prediction, from low and high-level influences [50,14,3,12,18,7,31] to biases [44,45,46,8], tasks [41,27,43] and semantic effects [11]. Over time, the concept of a saliency map has moved away from its origins in low-level feature integration, and can now refer more generally to "a map that predicts fixations". In practice, saliency maps are now synonymous with saliency models.

The large number of models created the need for quantitative metrics to assess progress in the field and compare models. Many different metrics have been proposed. The AUC-type metrics [45] used to be most common while the last years have seen a shift towards metrics like CC [22], NSS [37] and SIM [23], and recently the information gain metric has been proposed [32]. For an overview of the different metrics in use see e.g. [4,23]. The community uses these metrics in benchmarks to keep track of the progress: the MIT Saliency Benchmark [9,23] and the LSUN Challenge [53,54,52,21].

The most widely accepted MIT benchmark evaluates submissions in eight different metrics. Depending on which metric one chooses, the model rankings and performances change dramatically. This fact has lead to substantial research analyzing the differences between metrics and giving recommendations in which situation to use which metric [33,51,40,10,38,39]. Other authors have instead proposed new approaches to modeling and evaluation: Modeling as point processes [2,42], other loss functions [20] and GLMMs [35].

The general conclusion in the field is that the metrics measure qualitatively different things [51,40,10], and that it is even conceptually impossible to determine a best model independent of the different metrics. Recently, Kümmerer et al. [32] tried to argue for a unique ranking between different models by showing that much of the disagreement between different metrics can be removed via postprocessing of the saliency maps by optimizing the saliency scale and smoothing kernel for information gain (IG, essentially log-likelihood).

However, this does not seem to be a satisfactory solution: For one, this approach requires access to all models one wants to compare to and needs tedious postprocessing for each of them. In addition to this practical barrier the approach also suffers from the major conceptual shortcoming that optimizing for IG cannot be optimal for all metrics. In fact, we show below that the log densities proposed in [32] perform suboptimally on most metrics and can still produce inconsistent rankings. Ideally one would like a model to be able to compete in all metrics on the metric's original scale with other models, even with models that are directly optimized for that metric and where only the metric performances are known. This is not possible when evaluating on log densities as proposed in [32].

In fact, we show in this paper that even with knowledge of the true fixation distribution, no single saliency map can perform well in all metrics. In practice however, researchers must still decide on a particular saliency map to submit to the benchmark. Therefore, their model cannot compete with state-of-the-art models in all metrics – not because the model is intrinsically bad on those metrics, but because different metrics require the saliency maps to look different, independent of the encoded information about fixation placement (see Figure 1). As long as one evaluates all saliency metrics on the *same* saliency maps, it is impossible to solve the benchmarking problem.

Here, we argue that the fundamental problem is that saliency models and saliency maps are considered to be the same. A major insight from Bayesian decision theory is that the derivation of optimal decisions can be decomposed into a task-independent probability distribution over possible outcomes of an experiment and a task-dependent error metric. In the saliency setting, one decides on a saliency map to submit to a certain metric. Correspondingly, saliency *models* should be defined as metric-*independent* probability densities over possible fixations and subsequently many different metric-*dependent* saliency *maps* can be derived from the same density for different error metrics.

We show that saliency maps for the most influential metrics AUC, sAUC, NSS, CC, SIM, and KL-Div can be derived from fixation densities in a principled way. We demonstrate the validity of our approach on real models and real data. By decoupling the notions of saliency models and saliency maps, saliency models can be meaningfully compared on all metrics *in their original scale*, and the MIT saliency benchmark will implement our suggested approach.

2 Theory

Motivated by the line of thoughts presented above we here propose to use the following definitions:

1. a *saliency model* predicts a fixation probability density $p(x, y \mid I)$ given an image I.
2. a *saliency metric* is a performance measure for a saliency map on ground truth data.
3. a *saliency map* $s_{p,\text{metric}}(x, y, I)$ is a metric-specific prediction derived from the model density.

It has been argued before that formulating saliency models as probabilistic models is advantageous (e.g. [2,32]). In this definition, a saliency model predicts a fixation probability density, that is, the probability $p(x, y \mid I)$ of observing a fixation at a given pixel in a given image[3]. The three definitions we propose above follow the rationale of Bayesian decision theory: the saliency model is a posterior density over all possible events and the saliency metric is a utility

[3] Note that we use the fixation probability density for single fixations (as in [32]) whereas [2] define a point process density for a whole scanpath.

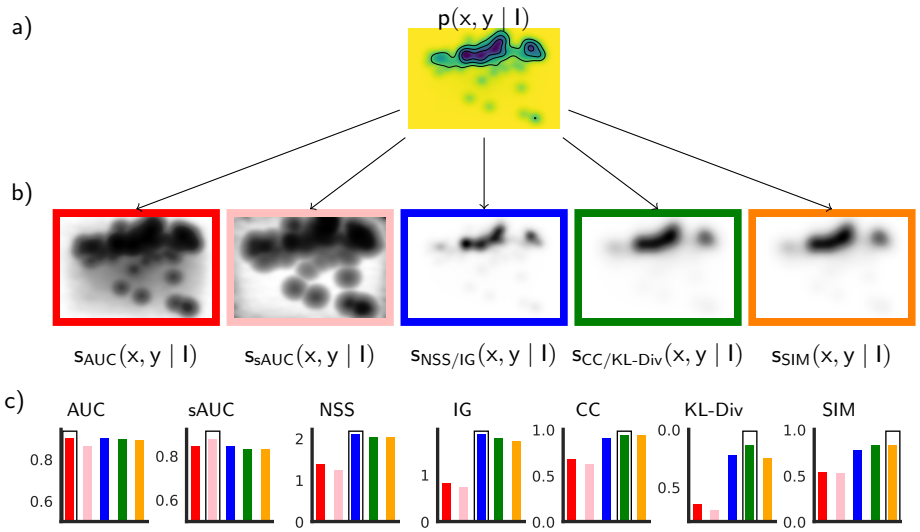


Fig. 1: No single saliency map can perform best in all metrics even when the true fixation distribution is known. This problem can be solved by separating saliency models from saliency maps. **a)** Fixations are distributed according to a ground truth fixation density $p(x, y \mid I)$ for some stimulus I (see supplementary material for details on the visualization). **b)** This ground truth density predicts different saliency maps depending on the intended metric. The saliency maps differ dramatically due to the different properties of the metrics but always reflect the same underlying model. Note that the maps for the NSS and IG metrics are the same, as are those for CC and KL-Div. **c)** Performances of the saliency maps from b) under seven saliency metrics on a large number of fixations sampled from the model distribution in a). Colors of the bars correspond to the frame colors in b). The predicted saliency map for the specific metric (framed bar) yields best performance in all cases.

function. Based on the posterior density and the utility function, a saliency map is then chosen to maximize the expected utility.

2.1 Predicting saliency maps from saliency models

From the predicted fixation density of a model, one can use expected utility maximization to derive the saliency map which the model expects to yield highest performance in some metric[4].

Evaluating a saliency metric involves a saliency map $s(x, y \mid I)$ for a stimulus I and ground truth fixation data (x_i, y_i). Therefore, we can phrase a metric

[4] Note that the term "metric" is a slight abuse of notation: strictly speaking, a metric measures the distance between two objects and is usually desired to be minimal. However, in saliency, the term "metric" denotes the performance that one wants to maximize (with a few exceptions, e. g., KL-Div and earth mover's distance).

as a function $M[s(x, y \mid I); (x_1, y_1), \ldots, (x_n, y_n)]$. Note that some metrics as CC or SIM use an empirical saliency map instead of ground truth fixations (*distribution-based metrics,richeSaliency2013*). However, the empirical saliency map is always constructed from ground truth fixations, usually by convolving them with a Gaussian. This can be taken to be part of the metric evaluation, as we will demonstrate below. Simplifying notation with $D = (x_1, y_1), \ldots, (x_n, y_n)$, the metric evaluation can be written as

$$M[s(x, y \mid I); D].$$

Assuming that the fixations are distributed according to some distribution $(x_i, y_i) \sim p(x, y \mid I)$ and therefore $D \sim \prod_1^n p(x, y)$, the expected performance of the metric on a saliency map is $\mathbb{E}_D M[s(x, y \mid I); D]$. One should choose the saliency map which is expected to yield highest performance for the metric M: that is, the solution of

$$\max_{s(x,y|I)} \mathbb{E}_D M[D, s(x, y \mid I)]$$

Solving this optimization problem for a fixation distribution p given by a model of interest essentially answers the following question: if we assume that the unknown fixations, on which the saliency map later will be evaluated, come from the model density p (and therefore $D = \prod_i^n p$), what would be the best saliency map to use for metric M? For a metric M the solutions to the optimization problem give rise to a transformation $p(x, y \mid I) \mapsto s_M(x, y \mid I)$ from fixation densities to derived metric-specific saliency maps. While the optimization problem might be hard in general, for most commonly-used saliency metrics it can be solved exactly or approximately, as we show below. Importantly, the methods we outline here are deterministic transformations depending only on the model's density prediction. No optimization using ground truth data is necessary.

In the following we give exact or approximate solutions for six of the most widely used metrics, including three metrics which operate directly on ground truth fixations (AUC, sAUC, and NSS) and three distribution-based metrics which first convert the ground truth fixations into a empirical saliency map (CC, SIM, KL-Div). Additionally we include the IG metric introduced in [32] since we use this metric for converting existing saliency map models to probabilistic models.

AUC, sAUC The AUC-type metrics ("Area Under the Curve", [45]) measure the model performance in a 2AFC (2 alternative forced choice) task where the model has to decide which one of two locations has been fixated: in a 2AFC task, a system is presented with one signal and one noise stimulus and chooses which stimulus is the "signal". In the case of the AUC in saliency, signal and noise correspond to fixated and non-fixated image locations respectively (See supplementary material for a proof of the equivalence between the ROC curve and the 2AFC task). Denoting the model's fixation distribution $p_{\text{fix}}(x, y)$, the nonfixation distribution $p_{\text{nonfix}}(x, y)$ (which is uniform for AUC and the image independent center bias for sAUC) and denote the two locations by (x_1, y_1) resp. (x_2, y_2). The 2AFC task reduces to deciding whether these points are sampled from $p_{\text{fix}} \times p_{\text{nonfix}}$ or

from $p_{\text{nonfix}} \times p_{\text{fix}}$. The likelihoods of the two points given these two distributions are $p_{\text{fix}}(x_1, y_1)p_{\text{nonfix}}(x_2, y_2)$ resp. $p_{\text{nonfix}}(x_1, y_1)p_{\text{fix}}(x_2, y_2)$. The model expects optimal performance by choosing the distribution which has higher likelihood, or equivalently, the point for which $p_{\text{fix}}(x, y)/p_{\text{nonfix}}(x, y)$ has the higher value. Therefore the model should expect the saliency map $p_{\text{fix}}(x, y)/p_{\text{nonfix}}(x, y)$ to yield highest performance. In the special case of the standard AUC metric, p_{nonfix} is constant and the saliency map boils down to p_{fix}. An additional practical consideration is that the MIT benchmark currently only accepts submissions as JPEG images. To compensate for this limited precision and possible JPEG-artefacts, one should additionally histogram-equalize the saliency map (see Supplementary Material).

NSS The *Normalized Scanpath Saliency* (NSS, [37]) performance of a saliency map model is defined to be the average saliency value of fixated pixels in the normalized (zero mean, unit variance) saliency maps (i.e., the average z-score of the fixated saliency values).

We can show analytically that one should expect the highest NSS score from the predicted fixation density itself: given an image with N pixels let the probability for a single fixation falling onto pixel i be p_i. Then the expected NSS of a saliency map $q = (q_1, \ldots, q_N)$ with $\frac{1}{N}\sum_i q_i = \bar{q} = 0$, $\|q\|_2^2 = 1$ is $\sum_i^N p_i \cdot q_i = \langle p, q \rangle$. Finding the saliency map with the best possible NSS is equivalent to finding the solution of the problem

$$\max \langle p, q \rangle \quad s.t.\ \bar{q} = 0, \|q\|^2 = 1$$

Since $q \mapsto q' = \bar{p} + \alpha q$ with $\alpha = \sqrt{\|p\|^2 - 1/N}$ induces a maximum-preserving bijection between $\{q \mid \bar{q} = 0, \|q\|^2 = 1\}$ and $\{q' \mid \bar{q}' = \bar{p} = 1/N, \|q\|^2 = \|p\|^2\}$, we can look for the maximum of $\langle p, q' \rangle \quad s.t.\ \bar{q}' = \bar{p}, \|q'\|^2 = \|p\|^2$ instead (and normalize q afterwards to get the normalized saliency map). Because of $\langle x, y \rangle = \frac{1}{2}(\|x\|^2 + \|y\|^2| - \|x - y\|^2)$, the maximum under these conditions is identical with the minimum of $\|p - q\|^2$, which is p.

Therefore, the best possible saliency map with respect to NSS is the density of the fixation distribution.

IG The *information gain* (IG, [32]) metric requires the saliency map to be a probability distribution and compares the average log-probability of fixated pixels to that given by a baseline model (usually the centerbias or a uniform model). The optimal saliency map for IG depends on how the metric interprets saliency maps as probability densities. We normalize the saliency maps to be probability vectors (nonnegative, unit sum) and in this case the predicted density itself yields the highest expected performance: Let $p = (p_1, \ldots, p_N)$ with $p \geq 0$, $\sum_i p_i = 1$ denote the predicted probabilities for each pixel and q with $q \geq 0$, $\sum_i q_i = 1$ a saliency map. Let $p_{bl} = (p_{bl,1}, \ldots, p_{bl,N})$ be the pixel probabilities of the baseline model. Then the expected IG of q is $\mathbb{E}_p IG(q) = \sum_i p_i(\log q_i - \log p_{bl,i})$ and its maximum is $\arg\max_q \mathbb{E}_p IG(q) = \arg\max_q \sum_i p_i(\log q_i - \log p_{bl,i}) = \arg\max_q \sum_i p_i \log q_i = \arg\max_q \sum_i p_i(\log q_i - \log p_i) = \arg\min_q \sum_i p_i(\log p_i - \log q_i) = \arg\min_q KL[p, q] = p$.

CC The *correlation coefficient* (CC, [22]) measures the correlation between model saliency map and empirical saliency map after normalizing both saliency maps to have zero mean and unit variance. This is equivalent to measuring the euclidean distance between the predicted saliency map and the normalized empirical saliency map. The expected euclidean distance to a random variable is minimized by its expectation value. Therefore the optimal saliency map with respect to CC is the expected normalized empirical saliency map.

This shows that predicting the optimal saliency map for CC crucially depends on how the empirical saliency maps are computed. Empirical saliency maps are typically computed by blurring observed fixation positions from eye movement data with a Gaussian kernel of a certain size. In this case the expected empirical saliency map would be $\mathbb{E}_{x_i \sim p} \frac{1}{N} \sum_i^N G_\sigma(x) = \frac{1}{N} \sum_i^N \mathbb{E}_{x \sim p} G_\sigma(x) = \frac{1}{N} \sum_i^N G_\sigma * p = G_\sigma * p$, that is, the density blurred with a Gaussian kernel of size σ.

Unfortunately, the expected empirical saliency map is not the expected normalized empirical saliency map which was earlier shown to be optimal for CC. Normalization involves subtracting the mean and dividing by the standard deviation, and the latter is nonlinear. Effectively, normalizing the variance just changes the weight by which the different empirical saliency maps are averaged in the expectation value. As long as the variances of the different empirical saliency maps don't differ too much, this won't have much of an effect and our simulations suggest that this is the case (Supplementary Material). Therefore, as an approximation to the expected normalized empirical saliency map, we use the expected saliency map in this paper, which is computed by convolving the expected density by a Gaussian.

Obviously, if more involved techniques are used to compute the empirical saliency maps (e.g. cross validation of the kernel size as in [32]), then the expected empirical saliency map is harder or impossible to calculate analytically. However, one can still approximate it numerically by sampling normalized empirical saliency maps from the expected fixation distribution and averaging them.

KL-Div The KL-Div metric computes the *Kullback-Leibler divergence* between the empirical saliency maps and the model saliency maps after converting both of them into probability distributions (by making them nonnegative and normalizing them to have unit sum) Therefore, unlike for most other metrics, in KL-Div lower values are better.

We can show that for the KL-Div metric, the expected empirical saliency map expects the best performance: let $e = (e_1, \ldots, e_N)$ with $e \geq 0$, $\sum_i e_i = 1$ denote the random variable which represents the empirical saliency map and q with $q \geq 0$, $\sum_i q_i = 1$ the model saliency map. Then we are looking for the q which minimizes $\mathbb{E}_p KL[e, q]$. Since $\mathbb{E}_p[KL[e, q]] = \mathbb{E}_p \left[\sum_i e_i \frac{\log e_i}{\log q_i}\right] = \mathbb{E}_p \left[\sum_i e_i \log e_i\right] - \sum_i \mathbb{E}_p[e_i] \log q_i$, this is equivalent to finding the maximum of $\sum_i \mathbb{E}_p[e_i] \log q_i$, which is again equivalent to finding the minimum of $\sum_i \mathbb{E}_p[e_i] \log \mathbb{E}_p[e_i] - \sum_i \mathbb{E}_p[e_i] \log q_i = KL[\mathbb{E}_p[e], q]$. This is obviously minimized by $q = \mathbb{E}_p[e]$, the expected empirical saliency map. As for CC, this is the density blurred by the same kernel size as used for the empirical saliency map.

SIM The *Similarity* (SIM, [23]) metric normalizes the model saliency map and the empirical saliency map to be probability vectors (in the same way as KL-Div) and sums the pixelwise minimum of two saliency maps. As opposed to the CC-metric, which can be interpreted as measuring the l_2-distance between normalized saliency maps, this effectively measures the l_1-distance between saliency maps ($\sum_i \min(p_i, q_i) = \sum_i \frac{1}{2}(p_i + q_i - |p_i - q_i|) = 1 - \frac{1}{2}\|p - q\|_1$,) This optimization problem cannot be solved analytically in general. Instead we solve it numerically: we perform a constrained stochastic gradient descend on sets of fixations sampled from the probability density (see Section 3 for details). Note that the optimal saliency map for SIM, unlike all other saliency maps presented here, depends on the number of fixations per image (see the Supplement for details on this effect).

3 Experiments and Results

We use the pysaliency toolbox [29] to compute saliency metrics (see Supplement for details). From a probability density over an image we compute five types of saliency maps: **AUC saliency maps** are created by equalizing the probability density to yield a uniform histogram over all pixels. **sAUC saliency maps** are created by dividing the probability density by the center bias density and again equalizing the saliency map to yield a uniform histogram over all pixels. The center bias density was estimated using a Gaussian kernel density estimate over all fixations from the MIT1003 dataset and crossvalidated across images. **NSS/IG saliency maps** are simply the probability density. **CC/KL-Div saliency maps** are calculated by convolving the probability density with a Gaussian kernel with $\sigma = 35px$ (corresponding to 1dva, as commonly used on the MIT1003 dataset). **SIM saliency maps**: We divide the CC saliency map by its sum to normalize it. Starting from there, we perform constrained (nonnegative, unit sum) stochastic gradient descend on fixations sampled from the predicted density to maximize the expected SIM performance (see Supplementary Material for implementation details).

3.1 No saliency map to rule them all

Here we illustrate using simulated data that even if the true fixation density is known, no single saliency map can win in all saliency metrics. From a fictional fixation density (Figure 1a) we compute the saliency maps that we predict to be optimal for the seven saliency metrics AUC, sAUC, NSS/IG, CC/KL-Div and SIM (Figure 1b). We sample 1000 sets of 100 fixations from the fixation density and evaluate all five saliency maps using the seven different saliency metrics on this dataset (Figure 1c, raw data in the Supplement).

Although the saliency maps in Figure 1b all are predicted by the same model, they appear visually different: while the AUC saliency map is essentially just the normalized density, the sAUC saliency map removes the center bias contribution

(see above). The NSS/IG saliency map is exactly the density and shows large areas with very low values. The CC/KL-Div saliency map, being a blurred version of the density, is much smoother than the NSS saliency map. The SIM saliency map looks mostly like the CC/KL-Div saliency map but is slightly more sparse.

The ranking of the five saliency maps is highly inconsistent across metrics (Figure 1c): even with knowledge of the real fixation distribution, no saliency map can be optimal for all saliency metrics. However, each saliency map is optimal for exactly those metrics for which it has been predicted to be optimal (framed bars). This illustrates our main result: By deriving metric-specific saliency maps in a principled way from fixation densities, one model can perform optimally in all metrics. Notice that in current practice, the situation faced by an individual research team is rather to pick from one of the maps in Figure 1b and be penalized accordingly on other metrics in Figure 1c.

3.2 MIT1003

In our main experiment, we use our approach to evaluate six saliency models on the popular benchmarking dataset MIT1003 (freeviewing fixations of 15 subjects on 1003 images, [24]). For all evaluated models, the original source code and default parameters have been used. The included models are **AIM** [6], Boolean Map-based Saliency **(BMS)** [55], the Ensemble of Deep Networks **(eDN)** [49], **OpenSALICON** [47], **SalGAN** [36] and **DeepGaze II** [31].

Converting existing models that produce arbitrary saliency maps into probabilistic models is not straightforward [32]. We used the method described in [32] and implemented in the pysaliency toolbox as `optimize_for_information_gain`: we fitted a pixelwise monotone nonlinearity and a center bias for each model to yield maximum information gain for the MIT1003 dataset (see supplementary material for details). Unlike [32] we did not optimize an additional Gaussian convolution to smooth the predictions. Since DeepGaze II is already formulated as a probabilistic model, there was no need to convert this model. For showing the "original saliency map" we use the log density in this case.

Example saliency maps. In Figure 2, we show the probability distribution and the predicted saliency maps (columns) for the saliency models (rows) for one example stimulus. Comparing the saliency maps within and between columns, i.e. metrics, one notices that the process of predicting saliency maps for certain metrics has a strong effect on the shape of the saliency maps that is consistent across models. It influences the visual appearance of the saliency map to a larger degree than the actual model does: the AUC and sAUC maps are very high contrast, while the NSS and CC saliency maps have large areas of very little saliency. The CC and SIM saliency maps are much smoother than all other saliency maps. It is a quite common technique in the field to compare the saliency maps of different models visually (e.g., see [13], Figure 6; [5], Figure 6; [4], Figure 9). Figure 2 shows that this technique can be very misleading unless the saliency maps are of the same type (i.e. intended for the same saliency metric).

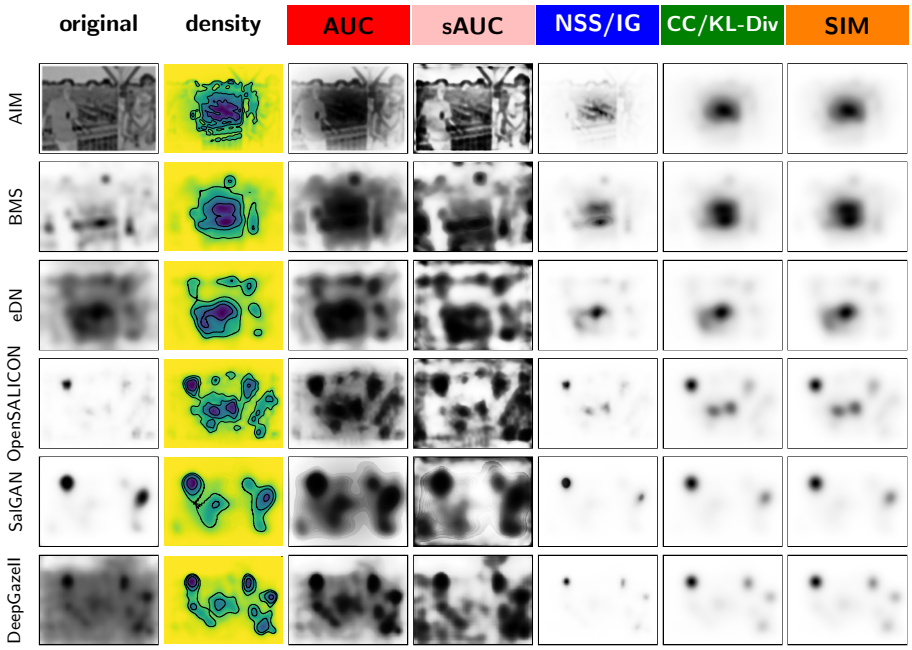


Fig. 2: The predicted saliency map for various metrics according to different models, for the same stimulus. For six models (rows) we show their original saliency map (first column), the probability distribution after converting the model into a probabilistic model (second column) and the saliency maps predicted for seven different metrics (columns three through seven). The predictions of different models for the same metric (column) appear more similar than the predictions of the same model for different metrics (row). In particular, note the inconsistency of the original models (what are typically compared on the benchmark) relative to the per-metric saliency maps. It is therefore difficult to visually compare original model predictions, which have been formulated for different metrics.

Comparing model performance. In Figure 3 we evaluate the saliency maps of the saliency models (AIM, BMS, eDN, OpenSALICON, SalGAN, DeepGaze II; x-axis) on the seven saliency metrics (subplots, raw data in the Supplement). Each line indicates the models' performances in the evaluated metric when using a specific type of saliency map. The dashed lines indicate performance using the models' original saliency maps (i.e. not transformed into true probability densities). The performances are very inconsistent between the different metrics on the original saliency maps. The solid lines indicate the metric performances on the five types of derived saliency maps (red: AUC, pink: sAUC, blue: NSS and IG, green: CC and KL-Div, orange: SIM). Additionally, we included log-density saliency maps as proposed in [32] (purple dotted lines).

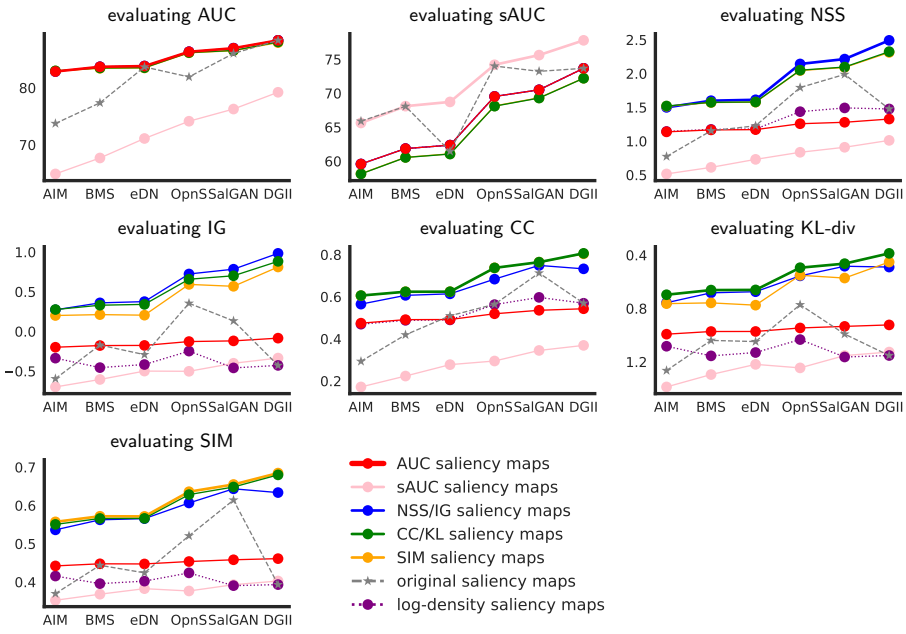


Fig. 3: We reformulated several saliency models in terms of fixation densities and evaluated AUC, sAUC, NSS, IG, CC, KL-Div and SIM on the original saliency maps (dashed line) and the saliency maps derived from the probabilistic model for the different saliency metrics (solid lines) on the MIT1003 dataset. Saliency maps derived for a given metric always yield the highest performance for that metric(thick line), and for each metric the model ranking is consistent when using the correct saliency maps – unlike for the original saliency maps and some other derived saliency maps. Note that AUC metrics yield identical results on AUC saliency maps, NSS saliency maps and log-density saliency maps, therefore the blue and purple lines are hidden by the red line in the AUC and sAUC plots. Also, the CC metric yields only slightly worse results on the SIM saliency map than on the CC saliency map, therefore the orange line is hidden by the green line in the CC plot. OpnS=OpenSALICON, DGII=DeepGaze II.

For each metric, the saliency map predicted for that metric (thick line in each sub plot) yields highest performance for all models. Conversely, saliency maps derived for other metrics often incur severe penalties (except for very few borderline cases, see below). While the model rankings given by the different metrics on each saliency map type are much more consistent than on the original saliency maps, there is still disagreement between metrics left when evaluating all metrics on the same saliency map type.

Interestingly, the AIM model reaches better NSS performance with the CC saliency map than with the NSS saliency map. This is easy to explain: the AIM model's predicted density improves after blurring. For the better models this effect vanishes. For example, DeepGaze II reaches significantly higher NSS scores

with the NSS saliency map than with the CC saliency map and vice versa for the CC metric. The SIM metric seems to show only slighly better performance on the SIM saliency map than on the CC saliency map, with the average difference being just 0.006. However, the best five models with respect to SIM in the MIT Saliency Benchmark perform within a range of less than 0.02. A difference of 0.006 could easily change a model's ranking by multiple places.

Figure 3 also serves to illustrate a key difference between the metric unification proposed in [32] and our method of predicting saliency maps from fixation densities: the metric results presented in [32] correspond to the purple dotted log-density lines for AUC, sAUC, NSS and to the blue density lines for IG and KL-Div (in our implementation taking the logarithm of the density is part of the metric itself). As reported in [32], the model rankings are more consistent for those lines than for the original saliency maps. However, except for AUC and IG, in all other metrics the models are penalized when evaluated like this and additionally for the best models even the agreement between metric rankings is lost (SalGAN vs DeepGaze II, AUC/sAUC/IG vs NSS/CC/KL-Div). This shows that the method proposed in [32], while managing to remove a significant amount of the disagreement between metrics, is not perfect.

To summarize, Figure 3 illustrates the main result of this paper: No matter what saliency map type you decide for, even state-of-the-art models will perform suboptimally in some metrics and rankings will still be inconsistent. Only by using the right saliency map for each metric given the model density, every model performs as well as it can theoretically and all model rankings agree. Consequently, our evaluation yields a unique winner of the benchmark: from all included models, DeepGaze II performs best in all considered metrics.

4 Discussion

Despite much progress in fixation prediction in recent years, comparing saliency models to each other can be confusing due to the large number of benchmarking metrics, giving inconsistent model rankings. Here we argue that benchmarking can be simplified by considering *saliency models* to be probability density predictors, *saliency metrics* to be performance measures that assess saliency maps against ground truth fixations, and subsequently *saliency maps* to be metric-specific predictions derived from the model's density. We have shown that probabilistic models can predict good saliency maps for the most common saliency metrics: "good models" perform well in many metrics.

Importantly, this metric-specific prediction reflects the same underlying model. It is not the case that the model is being re-trained for each metric. Rather, the saliency maps we show are derived deterministically from the fixation density predicted by a model. In this way it is possible to obtain optimal predictions from a given saliency density for arbitrary metrics without retraining. The saliency model density captures all necessary information in the training data and represents it in a way that it can readily be used in combination with arbitrary error metrics. Information gain (equivalently, log-likelihood) is an ideal optimization

metric because it reflects all information in the structure of the fixation density, independent of any particular metric. Therefore, it should lead to good results in all metrics.

The fact that metrics impose strong constraints on saliency maps means that it is misleading to visually compare saliency maps intended for different metrics (see Figure 2)—but this is commonly done in the field ([13,5,4]) For example, the optimal saliency maps for distribution-based metrics like CC, SIM and KL-Div require blurring unlike those for NSS and IG.

Another consequence of the present work is that the eight metrics available on the MIT benchmark can now be seen as a benefit rather than a possible source of confusion. Since each metric assesses different aspects of the fixation prediction, the benchmark would now allow fair comparison over a number of tasks of interest, which may be more or less relevant for certain applications. For example, sAUC is most relevant when one is interested in a model's predictive performance once the center bias is excluded (e.g., in applying to a setting with a different center bias from the MIT1003 training data).

While the saliency maps we have derived give the optimal metric-specific saliency map for a given fixation density, it is nevertheless still possible that a given model could do better on a metric with a saliency map not intended for that metric, rather than the metric-specific saliency map itself. If the model's density is not the correct one (i.e. does not reflect the data-generating density), then the derived saliency maps can be suboptimal. If the model's density is especially bad, some metrics might even perform better on saliency maps not predicted for this metric than on the one predicted for this metric. For example: if a model's density prediction is too sparse, the AUC metric will perform better on the smoothed CC saliency map than it will perform on the actual AUC saliency map. Therefore, actually optimizing model predictions for each specific metric may yield insights into the differences between the metrics (by comparing the underlying densities). Indeed, this could in practice produce better performance on the training metric than an information gain optimized density. The fact that we don't observe this effect on the original saliency maps (which *were* trained in the case of eDN, OpenSALICON, SalGAN and DeepGaze II: Figure 3, dashed lines) suggests any improvement is likely small, and can come at the price of performing substantially worse in other metrics.

Finally, we would like to note that the distinction between saliency models and saliency maps we draw here does not contradict ideas that a "saliency map" or maps may be instantiated in the human brain, as a corollary of bottom-up attentional guidance or an importance map for (e.g.) choosing the next place to fixate in a scene [34,48,26]. Our nomenclature is rather independent and intended for saliency model benchmarking.

The code for evaluating saliency models as demonstrated in this work has been released as part of the `pysaliency` python library (available at `https://github.com/matthias-k/pysaliency`).

Conclusion Our work solves the problem that one saliency model cannot reach state-of-the-art performance in all relevant saliency metrics. Our key theoretical

contribution is to decouple the notions of saliency models and saliency maps. For benchmarking practice, this means that saliency models can be meaningfully compared on all metrics *in their original scale*. Therefore, our method allows comparing to traditional models that do not use this method; it works even if only metric scores of other models are known (as for example in cases where metric scores are published in a paper). Practically, this means that there is no need to revise an existing benchmark: researchers who submit model densities can have their performance fairly evaluated, but existing models can remain in the table. The MIT saliency benchmark will implement this option.

Acknowledgements This study is part of Matthias Kümmerer's thesis work at the International Max Planck Research School for Intelligent Systems (IMPRS-IS). The research has been funded by the German Science Foundation (DFG; Collaborative Research Centre 1233) and the German Excellency Initiative (EXC307).

References

1. Adeli, H., Vitu, F., Zelinsky, G.J.: A model of the superior colliculus predicts fixation locations during scene viewing and visual search. J. Neurosci. **37**(6), 1453–1467 (Dec 2016). https://doi.org/10.1523/jneurosci.0825-16.2016, `https://doi.org/10.1523/jneurosci.0825-16.2016`
2. Barthelme, S., Trukenbrod, H., Engbert, R., Wichmann, F.: Modeling fixation locations using spatial point processes. Journal of Vision **13**(12), 1–1 (Oct 2013). https://doi.org/10.1167/13.12.1, `https://doi.org/10.1167/13.12.1`
3. Borji, A., Sihite, D.N., Itti, L.: Objects do not predict fixations better than early saliency: A re-analysis of einhauser et al.'s data. Journal of Vision **13**(10), 18–18 (Aug 2013). https://doi.org/10.1167/13.10.18, `https://doi.org/10.1167/13.10.18`
4. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (Jan 2013). https://doi.org/10.1109/tpami.2012.89, `https://doi.org/10.1109/tpami.2012.89`
5. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Trans. on Image Process. **22**(1), 55–69 (Jan 2013). https://doi.org/10.1109/tip.2012.2210727, `https://doi.org/10.1109/tip.2012.2210727`
6. Bruce, N.D.B., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision **9**(3), 5–5 (Mar 2009). https://doi.org/10.1167/9.3.5, `https://doi.org/10.1167/9.3.5`
7. Bruce, N.D.B., Catton, C., Janjic, S.: A deeper look at saliency: Feature contrast, semantics, and beyond. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2016). https://doi.org/10.1109/cvpr.2016.62, `https://doi.org/10.1109/cvpr.2016.62`
8. Bruce, N.D., Wloka, C., Frosst, N., Rahman, S., Tsotsos, J.K.: On computational modeling of visual saliency: Examining what's right, and what's left. Vision Research **116**, 95–112 (Nov 2015). https://doi.org/10.1016/j.visres.2015.01.010, `https://doi.org/10.1016/j.visres.2015.01.010`
9. Bylinskii, Z., Judd, T., Durand, F., Oliva, A., Torralba, A.: MIT saliency benchmark. http://saliency.mit.edu/

10. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv:1604.03605 [cs] (2016), http://arxiv.org/abs/1604.03605
11. Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F.: Where should saliency models look next? In: Computer Vision – ECCV 2016. pp. 809–824. Lecture Notes in Computer Science, Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_49, https://link.springer.com/chapter/10.1007/978-3-319-46454-1_49
12. Cerf, M., Harel, J., Huth, A., Einhäuser, W., Koch, C.: Decoding what people see from where they look: Predicting visual stimuli from scanpaths. In: Attention in Cognitive Systems, pp. 15–26. Springer Berlin Heidelberg (2009). https://doi.org/10.1007/978-3-642-00582-4_2, https://doi.org/10.1007/978-3-642-00582-4_2
13. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting human eye fixations via an LSTM-based saliency attentive model. arXiv:1611.09571 [cs] (2016), http://arxiv.org/abs/1611.09571
14. Einhauser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision **8**(14), 18–18 (Nov 2008). https://doi.org/10.1167/8.14.18, https://doi.org/10.1167/8.14.18
15. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in neural information processing systems. pp. 545–552 (2006)
16. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE (Dec 2015). https://doi.org/10.1109/iccv.2015.38, https://doi.org/10.1109/iccv.2015.38
17. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Machine Intell. **20**(11), 1254–1259 (1998). https://doi.org/10.1109/34.730558, https://doi.org/10.1109/34.730558
18. Itti, L.: Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition **12**(6), 1093–1123 (Aug 2005). https://doi.org/10.1080/13506280444000661, https://doi.org/10.1080/13506280444000661
19. Itti, L., Borji, A.: Computational models: Bottom-up and top-down aspects. In: The Oxford Handbook of Attention. Oxford University Press (2014)
20. Jetley, S., Murray, N., Vig, E.: End-to-end saliency mapping via probability distribution prediction. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2016). https://doi.org/10.1109/cvpr.2016.620, https://doi.org/10.1109/cvpr.2016.620
21. Jiang, M., Huang, S., Duan, J., Zhao, Q.: SALICON: Saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2015). https://doi.org/10.1109/cvpr.2015.7298710, https://doi.org/10.1109/cvpr.2015.7298710
22. Jost, T., Ouerhani, N., Wartburg, R.v., Müri, R., Hügli, H.: Assessing the contribution of color in visual attention. Computer Vision and Image Understanding **100**(1-2), 107–123 (Oct 2005). https://doi.org/10.1016/j.cviu.2004.10.009, https://doi.org/10.1016/j.cviu.2004.10.009
23. Judd, T., Durand, F.d., Torralba, A.: A Benchmark of Computational Models of Saliency to Predict Human Fixations. CSAIL Technical Reports (2012). https://doi.org/1721.1/68590

24. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision. IEEE (Sep 2009). https://doi.org/10.1109/iccv.2009.5459462, https://doi.org/10.1109/iccv.2009.5459462
25. Kienzle, W., Franz, M.O., Scholkopf, B., Wichmann, F.A.: Center-surround patterns emerge as optimal predictors for human saccade targets. Journal of Vision **9**(5), 7–7 (May 2009). https://doi.org/10.1167/9.5.7, https://doi.org/10.1167/9.5.7
26. Koch, C., Ullman, S.: Shifts in selective visual attention: Towards the underlying neural circuitry. Human Neurobiology **4**, 219–227 (1985), https://cseweb.ucsd.edu/classes/fa09/cse258a/papers/koch-ullman-1985.pdf
27. Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? Journal of Vision **14**(3), 14–14 (Mar 2014). https://doi.org/10.1167/14.3.14, https://doi.org/10.1167/14.3.14
28. Kruthiventi, S.S.S., Ayush, K., Babu, R.V.: DeepFix: A fully convolutional neural network for predicting human eye fixations. IEEE Trans. on Image Process. **26**(9), 4446–4456 (Sep 2017). https://doi.org/10.1109/tip.2017.2710620, https://doi.org/10.1109/tip.2017.2710620
29. Kümmerer, M.: pysaliency. https://github.com/matthias-k/pysaliency
30. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze i: Boosting saliency prediction with feature maps trained on ImageNet. In: 2015 International Conference on Learning Representations - Workshop Track (ICLR) (2015), https://arxiv.org/abs/1411.1045
31. Kümmerer, M., Wallis, T.S.A., Gatys, L.A., Bethge, M.: Understanding low- and high-level contributions to fixation prediction. In: The IEEE International Conference on Computer Vision (ICCV). IEEE (Oct 2017)
32. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proc Natl Acad Sci USA **112**(52), 16054–16059 (Dec 2015). https://doi.org/10.1073/pnas.1510393112, https://doi.org/10.1073/pnas.1510393112
33. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. Behav Res **45**(1), 251–266 (Jul 2012). https://doi.org/10.3758/s13428-012-0226-9, https://doi.org/10.3758/s13428-012-0226-9
34. Li, Z.: A saliency map in primary visual cortex. Trends in Cognitive Sciences **6**(1), 9–16 (Jan 2002). https://doi.org/10.1016/s1364-6613(00)01817-9, https://doi.org/10.1016/s1364-6613(00)01817-9
35. Nuthmann, A., Einhäuser, W., Schütz, I.: How well can saliency models predict fixation selection in scenes beyond central bias? a new approach to model evaluation using generalized linear mixed models. Front. Hum. Neurosci. **11** (Oct 2017). https://doi.org/10.3389/fnhum.2017.00491, https://doi.org/10.3389/fnhum.2017.00491
36. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: SalGAN: Visual saliency prediction with generative adversarial networks. arXiv:1701.01081 [cs] (2017), http://arxiv.org/abs/1701.01081
37. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research **45**(18), 2397–2416 (Aug 2005). https://doi.org/10.1016/j.visres.2005.03.019, https://doi.org/10.1016/j.visres.2005.03.019

38. Riche, N.: Metrics for saliency model validation. In: From Human Attention to Computational Attention, pp. 209–225. Springer New York (2016). https://doi.org/10.1007/978-1-4939-3435-5_12, https://doi.org/10.1007/978-1-4939-3435-5_12
39. Riche, N.: Saliency model evaluation. In: From Human Attention to Computational Attention, pp. 245–267. Springer New York (2016). https://doi.org/10.1007/978-1-4939-3435-5_14, https://doi.org/10.1007/978-1-4939-3435-5_14
40. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., Dutoit, T.: Saliency and human fixations: State-of-the-art and study of comparison metrics. In: 2013 IEEE International Conference on Computer Vision. IEEE (Dec 2013). https://doi.org/10.1109/iccv.2013.147, https://doi.org/10.1109/iccv.2013.147
41. Rothkopf, C.A., Ballard, D.H., Hayhoe, M.M.: Task and context determine where you look. Journal of Vision **7**(14), 16 (Jul 2016). https://doi.org/10.1167/7.14.16, https://doi.org/10.1167/7.14.16
42. Schütt, H.H., Rothkegel, L.O.M., Trukenbrod, H.A., Reich, S., Wichmann, F.A., Engbert, R.: Likelihood-based parameter estimation and comparison of dynamical cognitive models. Psychological Review **124**(4), 505–524 (2017). https://doi.org/10.1037/rev0000068, https://doi.org/10.1037/rev0000068
43. Tatler, B.W., Hayhoe, M.M., Land, M.F., Ballard, D.H.: Eye guidance in natural vision: Reinterpreting salience. Journal of Vision **11**(5), 5–5 (May 2011). https://doi.org/10.1167/11.5.5, https://doi.org/10.1167/11.5.5
44. Tatler, B.W.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision **7**(14), 4 (Nov 2007). https://doi.org/10.1167/7.14.4, https://doi.org/10.1167/7.14.4
45. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual correlates of fixation selection: Effects of scale and time. Vision Research **45**(5), 643–659 (Mar 2005). https://doi.org/10.1016/j.visres.2004.09.017, https://doi.org/10.1016/j.visres.2004.09.017
46. Tatler, B.W., Vincent, B.T.: Systematic tendencies in scene viewing. Journal of Eye Movement Research **2**(2), 1–18 (2008), http://csi.ufs.ac.za/resres/files/tatler_2008_jemr.pdf
47. Thomas, C.: OpenSalicon: An open source implementation of the salicon saliency model. CoRR **abs/1606.00110** (2016), http://arxiv.org/abs/1606.00110
48. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology **12**(1), 97–136 (Jan 1980). https://doi.org/10.1016/0010-0285(80)90005-5, https://doi.org/10.1016/0010-0285(80)90005-5
49. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. IEEE (Jun 2014). https://doi.org/10.1109/cvpr.2014.358, https://doi.org/10.1109/cvpr.2014.358
50. Vincent, B.T., Baddeley, R., Correani, A., Troscianko, T., Leonards, U.: Do we look at lights? using mixture modelling to distinguish between low- and high-level factors in natural image viewing. Visual Cognition **17**(6-7), 856–879 (Aug 2009). https://doi.org/10.1080/13506280902916691, https://doi.org/10.1080/13506280902916691
51. Wilming, N., Betz, T., Kietzmann, T.C., König, P.: Measures and limits of models of fixation selection. PLoS ONE **6**(9), e24038 (Sep 2011).

https://doi.org/10.1371/journal.pone.0024038, https://doi.org/10.1371/journal.pone.0024038

52. Xiao, J., Xu, P., Zhang, Y., Ehinger, K., Finkelstein, A., Kulkarni, S.: What can we learn from eye tracking data on 20,000 images? Journal of Vision **15**(12), 790 (Sep 2015). https://doi.org/10.1167/15.12.790, https://doi.org/10.1167/15.12.790
53. Yu, F., Kontschieder, P., Song, S., Jiang, M., Zhang, Y., Zhao, C.Q., Funkhouser, T., Xiao, J.: Large-scale scene understanding challenge. http://http://lsun.cs.princeton.edu/2017/
54. Yu, F., Kontschieder, P., Song, S., Jiang, M., Zhang, Y., Zhao, C.Q., Funkhouser, T., Xiao, J.: SALICON saliency prediction challenge. http://salicon.net/challenge-2017/
55. Zhang, J., Sclaroff, S.: Saliency detection: A Boolean map approach. In: 2013 IEEE International Conference on Computer Vision. IEEE (Dec 2013). https://doi.org/10.1109/iccv.2013.26, https://doi.org/10.1109/iccv.2013.26
56. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision **8**(7), 32 (Dec 2008). https://doi.org/10.1167/8.7.32, https://doi.org/10.1167/8.7.32

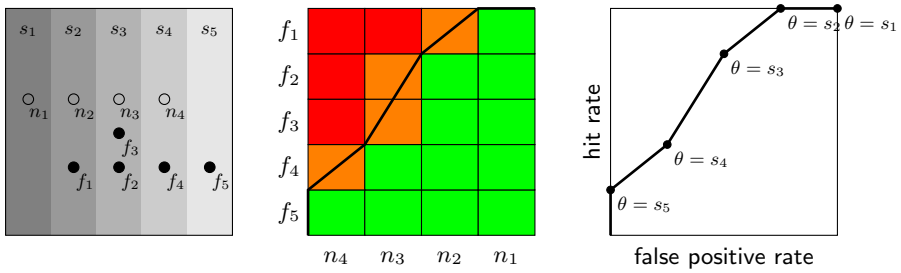


Fig. 4: AUC metrics measure the performance of the saliency map in a 2AFC task where the saliency values of two locations are used to decide which of these two locations is a fixation and which is a nonfixation. **a)** An example saliency map is shown consisting of five saliency values ($s_1 < \cdots < s_5$) and with five fixations ($f_1, \ldots, f_5$) and four nonfixations ($n_1, \ldots, n_4$). **b)** The performance in the 2AFC task can be calculated by going through all fixation-nonfixation pairs (f_i, n_j): The saliency map decides correct if the saliency value of f_i is greater than n_j (green), incorrect if it is smaller (red) and has chance performance if the values are equal (orange). Below the thick line are all correct predictions (green) and half of the chance cases (orange). **c)** The ROC curve of the saliency map with respect to the given fixations and nonfixations. For each threshold θ all values of saliency value greater or equal to θ are classified as fixations. Comparing b) and c) shows that the area under the curve in c) is exactly the performance in the 2AFC task in b).

5 Supplementary Material

5.1 Implementation details on the saliency metrics

We use the pysaliency toolbox [29] to compute metrics. **AUC**: We use all pixels as nonfixations. As thresholds we use the combined saliency values of all fixations and nonfixations. **sAUC**: We use the fixations of all other images of the dataset as nonfixations. As for AUC, we use the combined saliency values of all fixations and nonfixations as thresholds. **NSS** computes the mean saliency of fixation locations after normalizing the saliency map to have zero mean and unit variance. **IG** computes the mean log density of fixation locations for a model's predicted fixation density and substracts the average log density of fixation locations for a baseline model's predicted fixation density. To convert a saliency map to a probability distribution, we check whether any values of the saliency map are negative. If so, we subtract the minimal value from the saliency to make it non-negative. Afterwards we divide the saliency by the sum of all values. For the baseline model we transform the coordinates of all fixations in the MIT1003 dataset to range from 0 to 1. From these points a Gaussian kernel density estimator with a bandwidth of 0.22 is computed (the bandwidth has been tuned with leave-one-image-out crossvalidation). The baseline model scales the density predicted by the estimator to the size of the image in question. For images in

Saliency Map	Model	Binning		Difference
		None	8bit	
density	AIM	0.82883	0.82855	0.00028
	BMS	0.83712	0.83676	0.00035
	eDN	0.83836	0.83810	0.00026
	OpenSALICON	0.86350	0.86200	0.00150
	SalGAN	0.86973	0.86845	0.00128
	DeepGazeII	0.88355	0.87931	0.00424
equalized	AIM	0.82883	0.82882	0.00001
	BMS	0.83712	0.83710	0.00002
	eDN	0.83836	0.83834	0.00001
	OpenSALICON	0.86350	0.86347	0.00003
	SalGAN	0.86973	0.86970	0.00003
	DeepGazeII	0.88355	0.88351	0.00004

Table 1: AUC and low precision: While AUC metrics in theory depend only on the ranking of the saliency values and therefore are invariant to monotone transformations, this does not hold anymore when the saliency map is saved with limited precision (e.g. as 8bit PNG/JPEG as common). In this case, the saliency map should be rescaled to have a uniform histogram before saving.

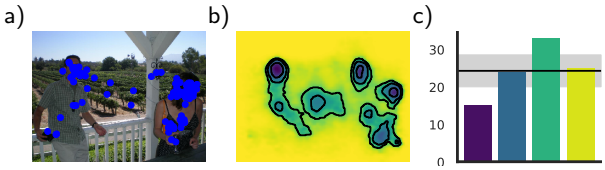


Fig. 5: Visualizing fixation densities: **a)** an example stimulus with $N = 97$ ground truth fixations. **b)** DeepGaze II predicts a fixation density for this stimulus. The contour lines separate the image into four areas of decreasing probability density such that each area has the same total probability mass. **c)** The number of ground truth fixations in each of the four areas. The model expects the same number of fixations for each area (horizontal line: 24.25 fixations for N fixations total). The gray area shows the expected standard deviation from this number. DeepGaze II overestimates the how peaked the density is: there are too few fixations in darkest area. Vice versa, it misses some probability mass in the second to last area. However, the large error margin (gray area) indicates that substantial deviations from the expected number of fixations are to be expected.

the MIT1003 dataset (i.e. for Figure 3), only fixations from all other images in the dataset are used to compute the baseline density for the image. **CC**: As suggested for the MIT1003 dataset used by us [24], we convolve the fixation maps of the ground truth fixations with a gaussian kernel with $\sigma = 35px$ to compute empirical saliency maps. **KL-Div**: We use the same empirical saliency maps as for CC and the same normalization procedure as for IG. **SIM**: We use the same empirical saliency maps as for CC and the same normalization procedure as for IG.

5.2 Converting saliency map models to fixation density models

To convert existing saliency-map based models to probabilistic models, we used the method described in [32] and implemented in the pysaliency toolbox in the method `optimize_for_information_gain`:

It first rescales all saliency maps for the dataset in question jointly to range from 0 for the smallest saliency value (over the full dataset) to 1 for the largest saliency value. The a pixelwise montone nonlinearity is applied to each saliency map. This nonlinearity is implemented as a continous piecewise linear function with 20 equidistant segments from 0 to 1.

The result is multiplied pixelwise with a centerbias which is parametrized with another piecewise linear function applied to

$$\sqrt{(x - \frac{1}{2}x_{max})^2 + \alpha(y - \frac{1}{2}y_{max})^2} / \sqrt{\frac{1}{4}x_{max}^2 + \frac{1}{4}\alpha y_{max}^2}$$

where x_{max} and y_{max} are the maximal x and y coordinates for the image in question. The piecewise linear function for the centerbias is parametrized as a continous piecewise linear function with 12 equidistant segments from 0 to 1. The resulting product is divided by its sum over all pixels to make it a probability distribution.

The parameters for both piecewise linear functions and the eccentricity parameter α are jointly optimized for maximum likelihood on the MIT1003 dataset. Note that unlike [32] we did not optimize an additional Gaussian convolution to smooth the predictions.

5.3 Computing saliency maps for SIM

To compute the saliency map for the SIM metric from a model density, we first divide the CC saliency map (density convolved with a Gaussian of size 35px=1dva) by its sum to normalize it. Starting from there, we perform a constrained stochastic gradient descend on fixations sampled from the predicted density to maximize the expected SIM performance. The (linear) constraints that are enforced in every step of the gradient descend are nonnegativity and unit sum. Each sample consists of 100 fixations (in correspondence to the dataset we are using). We use a batchsize of 50 samples and start with a learning rate of 10^{-7}. We use a fixed set of 1000 samples as validation data. Every 1000 training

samples we compute the validation performance. Whenever it decreases compared to the last epoch, we go back to the point of best validation performance so far and decrease the learning rate by a factor of $\frac{1}{3}$ and continue the gradient descend. We stop when the learning rate is smaller than 10^{-9}.

5.4 The AUC metrics: Digitizing saliency maps

Digitizing the saliency map e.g. by storing them as 8bit images can obviously affect metric performance. The AUC type metrics are sensitive only to the ranking of the saliency values and therefore especially sensitive to mapping similar saliency values to the same value. In Table 1, we evaluate the AUC metric for all included models in four different ways: We use either the model fixation density or we additionally transform it to have a uniform histogram. Also, we optionally bin the saliency values to 256 different values using equidistant bins. Since the AUC metrics are invariant to monotonic transformations, both density and equalized density should have the same AUC performance, as is indeed the case if no binning is applied. In the case of binning, however, the performances change: while for the normalized density binning does not affect performance a lot, for the density it does so. The performance loss after binning the density seems to be the stronger for better models. This is likely the case since better models will map larger areas of the image to very small values that all end up in the lowest bin.

5.5 The CC metric: mean normalized empirical vs normalized mean empirical saliency maps

We use the mean empirical saliency map for the CC metric. As explained in the main text, this is an approximation: the optimal saliency map would be the mean normalized empirical saliency map (i.e. one has to normalize the empirical saliency maps to zero mean and unit variance before taking the mean).

To check the validity of our approximation, we sampled fixations from a distribution (Figure 6a) and used those fixations to compute average empirical saliency maps (Figure 6b) and average normalized empirical saliency maps (Figure 6c) for different numbers of fixations per sample and kernel sizes in the computation of empirical saliency maps.

We evaluated both types of saliency maps on newly sampled fixations and compared the CC performances (Figure 6d). The performances for both saliency maps are very close in all cases, suggesting that the mean empirical saliency map is an adequate approximation for the mean normalized saliency map when computing CC performances.

5.6 The SIM metric depends on the number of fixations per image

Unlike all other metrics presented in this work, the optimal saliency map for the SIM metric depends on how many fixations per image are in the dataset in

question. If ignoring the constraint that the values of the saliency map should sum up to one, this effect is easy to see: The SIM metric effectively measures the l1 distance between empirical saliency map and model saliency map and this distance is minimized by the median empirical saliency map, which will be mostly zero if there are only very little fixations used to compute each empirical saliency map.

This effect is still present when constraining the saliency map to have unit sum, as we demonstrate in Figure 7. For a sample density (Figure 6a), we computed the optimal saliency maps for different numbers of fixations per sample according to our method detailed in Section 3. The resulting saliency maps are shown in Figure 7a. If there are only few fixations per sample, the resulting saliency maps have much larger areas of zeros, effectively being more sparse. For more fixations per sample, the saliency maps visually converge to the CC saliency map (blurred density).

Subsequently, we evaluated those saliency maps (Figure 7b, rows) on newly sampled fixations, again for different numbers of fixations per sample (Figure 7b, columns). Additionally, we included the CC saliency map.

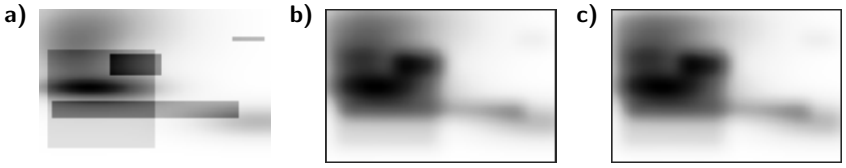
The columns in Figure 7 show that the number of fixations used to compute the saliency map affects the performance: The saliency map computed using the same number of fixations per sample always performs best (bold numbers), and all other saliency maps perform worse – often dramatically. Even in the case of 1000 fixations, there are still measurable differences between the saliency maps computed using 500 fixations, 1000 fixations and the CC saliency map.

5.7 Visualizing probability densities

Visualizing two dimensional densities is harder than it appears to be at the first glance: Although the absolute density values have a very precise meaning, it is hard to read substantially more than the ranking of the values and maybe a very rough idea about the peakyness of the distribution from a color map. When visualizing two dimensional probability densities, we add three contour lines separating the image into four areas of decreasing probability density such that each area has the same total probability mass (i.e. the density predicts each area to receive the same number of fixations, see Figure 5b). If the darkest area is very small, this means the density predicts on fourth of the fixations to be clustered in a very small area. If all areas are roughly of the same size, the density is nearly uniform. Comparing the number of fixations in each area can serve as a simple heuristic to asses a model's quality (see Figure 5c).

5.8 Data

The raw data for Figure 1 can be found in Table 2, the raw data for Figure 3 can be found in Table 3.



d)

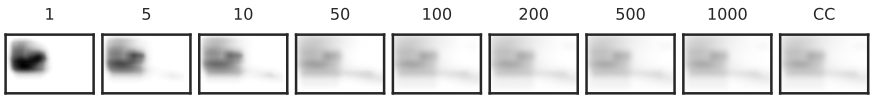
fixations \ kernel size	CC scores (mean empirical/mean normalized empirical)				
	1	5	10	20	30
1	0.0127/0.0127	0.0632/0.0633	0.1210/0.1214	0.2227/0.2242	0.3066/0.3100
10	0.0412/0.0412	0.1928/0.1929	0.3488/0.3491	0.5630/0.5641	0.6855/0.6874
50	0.0920/0.0920	0.3960/0.3960	0.6120/0.6121	0.7728/0.7731	0.8198/0.8206
100	0.1295/0.1295	0.5191/0.5191	0.7307/0.7307	0.8464/0.8465	0.8707/0.8709
200	0.1817/0.1817	0.6509/0.6509	0.8313/0.8314	0.9074/0.9074	0.9203/0.9204

Fig. 6: Predicting optimal saliency maps for the CC metric: Starting from a density (a) we sampled 100000 sets of either 1, 10 or 100 fixations and used them to create empirical saliency maps. Using these empirical saliency maps, we calculated the mean empirical saliency map (shown for 10 fixations per empirical saliency map in (b)). Additionally, we normalized the empirical saliency maps to have zero mean and unit variance to compute the mean normalized empirical saliency map (c) which is optimal with respect to the CC metric. Then we sampled another 100000 empirical saliency maps from the original density and evaluated CC scores of the mean empirical and mean normalized empirical saliency maps (d). The mean normalized saliency map yields slighly higher scores in all cases but the difference to the mean empirical saliency map is tiny, indicating that the expected empirical saliency map is a very good approximation of the optimal saliency map for the CC metric.

Saliency Map	AUC	sAUC	NSS	IG	CC	KL-Div	SIM
AUC	0.897325	0.842109	1.369418	0.826523	0.675036	0.644968	0.541042
sAUC	0.863243	0.875880	1.246049	0.741960	0.618483	0.700238	0.520718
NSS/IG	0.897325	0.842109	2.106131	1.865231	0.907441	0.221233	0.778510
CC/KL	0.892173	0.831612	2.024991	1.765046	0.939149	0.137458	0.824498
SIM	0.891833	0.831888	2.024197	1.700093	0.939007	0.253121	0.827775

Table 2: The raw data plotted in Figure 1

a)



b)

Fixations/sample Saliency Map	1	5	10	50	100	200	500	1000
SIM 1	**0.0936**	0.2662	0.3679	0.5260	0.5526	0.5657	0.5736	0.5761
SIM 5	0.0912	**0.2715**	0.3869	0.6107	0.6598	0.6865	0.7022	0.7070
SIM 10	0.0887	0.2697	**0.3892**	0.6426	0.7057	0.7438	0.7684	0.7760
SIM 50	0.0814	0.2566	0.3779	**0.6713**	0.7583	0.8203	0.8742	0.8994
SIM 100	0.0795	0.2521	0.3725	0.6694	**0.7607**	0.8271	0.8866	0.9159
SIM 200	0.0785	0.2496	0.3692	0.6669	0.7599	**0.8282**	0.8900	0.9211
SIM 500	0.0780	0.2482	0.3674	0.6653	0.7588	0.8278	**0.8906**	0.9223
SIM 1000	0.0778	0.2478	0.3668	0.6647	0.7583	0.8275	0.8905	**0.9224**
CC	0.0776	0.2473	0.3662	0.6640	0.7578	0.8272	0.8903	0.9223

Fig. 7: The optimal SIM saliency map depends on the number of fixations. **(a)** For a sample density (see Figure 6), we calculated the optimal SIM saliency map for different numbers of fixations per sample (numbers on top) and additionally the mean empirical saliency map (CC). **(b)** average performance of those saliency maps (rows) when repeatedly sampling a certain number of fixations (columns) and computing SIM. The best performing saliency map for each sampled dataset (columns) is printed in boldface. It's always the saliency map calculated with the same number of fixations. Note that the CC saliency map – although looking identical – always performs slighly worse

Evaluating AUC

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	0.828831	0.837118	0.838357	0.863505	0.869729	0.883548
sAUC	0.649100	0.677084	0.711299	0.741588	0.762972	0.792404
NSS/IG	0.828831	0.837118	0.838357	0.863505	0.869729	0.883548
CC/KL	0.830038	0.835026	0.835490	0.862059	0.865807	0.880673
SIM	0.829635	0.834474	0.835117	0.861701	0.865361	0.879561

Evaluating sAUC

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	0.595972	0.618699	0.623781	0.695496	0.705236	0.736911
sAUC	0.656440	0.681188	0.687451	0.742167	0.756413	0.778136
NSS/IG	0.595972	0.618699	0.623781	0.695496	0.705236	0.736911
CC/KL	0.581381	0.605622	0.610651	0.681168	0.693121	0.722245
SIM	0.581674	0.605807	0.610982	0.681571	0.693439	0.722389

Evaluating NSS

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	1.139104	1.167812	1.172103	1.259217	1.280780	1.328648
sAUC	0.516497	0.613439	0.731962	0.836885	0.910961	1.012916
NSS/IG	1.501131	1.600230	1.613635	2.143965	2.215962	2.493238
CC/KL	1.521584	1.575962	1.581050	2.051697	2.097997	2.326005
SIM	1.518022	1.572937	1.578647	2.046154	2.092357	2.314424

Evaluating IG

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	-0.197452	-0.176460	-0.174687	-0.125959	-0.117584	-0.083175
sAUC	-0.697486	-0.603408	-0.497874	-0.500561	-0.399373	-0.333968
NSS/IG	0.273559	0.361184	0.376803	0.724795	0.785096	0.984600
CC/KL	0.277253	0.331602	0.342154	0.658744	0.702166	0.884041
SIM	0.202053	0.213307	0.206090	0.593686	0.569699	0.812878

Evaluating CC

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	0.475165	0.491929	0.492461	0.519783	0.536314	0.543618
sAUC	0.172271	0.223996	0.278215	0.295418	0.345583	0.369579
NSS/IG	0.565561	0.607264	0.614484	0.684224	0.749480	0.733241
CC/KL	0.606328	0.624457	0.624698	0.737745	0.764947	0.806748
SIM	0.605789	0.624148	0.624524	0.737826	0.764842	0.804570

Evaluating KL-Div

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	0.991749	0.971731	0.971401	0.945121	0.933347	0.922326
sAUC	1.387180	1.294242	1.218003	1.244400	1.152025	1.125674
NSS/IG	0.756034	0.681768	0.673484	0.554681	0.482559	0.488775
CC/KL	0.696111	0.661160	0.659918	0.493419	0.463078	0.385972
SIM	0.762872	0.757106	0.774335	0.550407	0.571039	0.451175

Evaluating SIM

Saliency Map	AIM	BMS	eDN	OpenSALICON	SalGAN	DeepGaze II
AUC	0.442354	0.447864	0.447482	0.453846	0.458399	0.461371
sAUC	0.352552	0.368417	0.382886	0.376842	0.393249	0.402880
NSS/IG	0.536597	0.562841	0.565940	0.606866	0.643872	0.634031
CC/KL	0.550810	0.566335	0.566567	0.628707	0.648602	0.680294
SIM	0.557526	0.572131	0.571724	0.636311	0.655191	0.684973

Table 3: The raw data plotted in Figure 3