

Self- and Interpersonal Contact in 3D Human Mesh Reconstruction

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Lea Müller
aus Heilbronn

Tübingen
2023

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

Tag der mündlichen Qualifikation:	14.03.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Michael Black
2. Berichterstatter:	Prof. Dr. Kostas Daniilidis
3. Berichterstatter:	Prof. Dr. Alexei Efros

To those who gave me roots and wings.

Abstract

The ability to perceive tactile stimuli is of substantial importance for human beings in establishing a connection with the surrounding world. Humans rely on the sense of touch to navigate their environment and to engage in interactions with both themselves and other people. The field of computer vision has made great progress in estimating a person’s body pose and shape from an image, however, the investigation of self- and interpersonal contact has received little attention despite its considerable significance. Estimating contact from images is a challenging endeavor because it necessitates methodologies capable of predicting the full 3D human body surface, i.e. an individual’s pose *and* shape. The limitations of current methods become evident when considering the two primary datasets and labels employed within the community to supervise the task of human pose and shape estimation. First, the widely used 2D joint locations lack crucial information for representing the entire 3D body surface. Second, in datasets of 3D human bodies, e.g. collected from motion capture systems or body scanners, contact is usually avoided, since it naturally leads to occlusion which complicates data cleaning and can break the data processing pipelines.

In this thesis, we first address the problem of estimating contact that humans make with themselves from RGB images. To do this, we introduce two novel methods that we use to create new datasets tailored for the task of human mesh estimation for poses with self-contact. We create (1) 3DCP, a dataset of 3D body scan and motion capture data of humans in poses with self-contact and (2) MTP, a dataset of images taken in the wild with accurate 3D reference data using pose mimicking. Next, we observe that 2D joint locations can be readily labeled at scale given an image, however, an equivalent label for self-contact does not exist. Consequently, we introduce (3) discrete self-contact (DSC) annotations indicating the pairwise contact of discrete regions on the human body. We annotate three existing image datasets with discrete self-contact and use these labels during mesh optimization to bring body parts supposed to touch into contact. Then we train TOUCH, a human mesh regressor, on our new datasets. When evaluated on the task of human body pose and shape estimation on public benchmarks, our results show that knowing about self-contact not only improves mesh estimates for poses with self-contact, but also for poses without self-contact.

Next, we study contact humans make with other individuals during close social interaction. Reconstructing these interactions in 3D is a significant challenge due to the mutual occlusion. Furthermore, the existing datasets of images taken in the wild with ground-truth contact labels are of insufficient size to facilitate the training of a robust human mesh regressor. In this work, we employ a generative model, BUDDI, to learn

the joint distribution of 3D pose and shape of two individuals during their interaction and use this model as prior during an optimization routine. To construct training data we leverage pre-existing datasets, i.e. motion capture data and Flickr images with discrete contact annotations. Similar to discrete self-contact labels, we utilize discrete human-human contact to jointly fit two meshes to detected 2D joint locations. The majority of methods for generating 3D humans focus on the motion of a single person and operate on 3D joint locations. While these methods can effectively generate motion, their representation of 3D humans is not sufficient for physical contact since they do not model the body surface. Our approach, in contrast, acts on the pose *and* shape parameters of a human body model, which enables us to sample 3D meshes of two people. We further demonstrate how the knowledge of human proxemics, incorporated in our model, can be used to guide an optimization routine. For this, in each optimization iteration, BUDDI takes the current mesh and proposes a refinement that we subsequently consider in the objective function. This procedure enables us to go beyond state of the art by forgoing ground-truth discrete human-human contact labels during optimization.

Self- and interpersonal contact happen on the surface of the human body, however, the majority of existing art tends to predict bodies with similar, “average” body shape. This is due to a lack of training data of paired images taken in the wild and ground-truth 3D body shape and because 2D joint locations are not sufficient to explain body shape. The most apparent solution would be to collect body scans of people together with their photos. This is, however, a time-consuming and cost-intensive process that lacks scalability. Instead, we leverage the vocabulary humans use to describe body shape. First, we ask annotators to label how much a word like “tall” or “long legs” applies to a human body. We gather these ratings for rendered meshes of various body shapes, for which we have ground-truth body model shape parameters, and for images collected from model agency websites. Using this data, we learn a shape-to-attribute (A2S) model that predicts body shape ratings from body shape parameters. Then we train a human mesh regressor, SHAPY, on the model agency images wherein we supervise body shape via attribute annotations using A2S. Since no suitable test set of diverse 3D ground-truth body shape with images taken in natural settings exists, we introduce Human Bodies in the Wild (HBW). This novel dataset contains photographs of individuals together with their body scan. Our model predicts more realistic body shapes from an image and quantitatively improves body shape estimation on this new benchmark.

In summary, we present novel datasets, optimization methods, a generative model, and regressors to advance the field of 3D human pose and shape estimation. Taken together, these methods open up ways to obtain more accurate and realistic 3D mesh estimates from images with multiple people in self- and mutual contact poses and with diverse body shapes. This line of research also enables generative approaches to create more natural, human-like avatars. We believe that knowing about self- and human-human contact through computer vision has wide-ranging implications in other fields as for example robotics, fitness, or behavioral science.

Zusammenfassung

Die Wahrnehmung taktiler Reize ist für den Menschen von großer Bedeutung um eine Verbindung mit unserer Umgebung herzustellen. Dabei verwenden wir unseren Tastsinn um uns in der Umwelt zurechtzufinden und Beziehungen zu uns selbst und zu anderen Menschen aufzubauen. Das maschinelle Sehen hat zwar erhebliche Fortschritte bei der Bestimmung der Pose und Figur einer Person von Bildern gemacht, die Erforschung von körpereigenem- und zwischenmenschlichem Kontakt wurde dabei, trotz seiner Relevanz, jedoch vernachlässigt. Eine Herausforderung beim Schätzen von 3D Kontakt ist, dass die gesamte menschliche Körperoberfläche, also Pose und Figur, berücksichtigt werden muss. Weiterhin werden Grenzen aktueller Methoden erkenntlich, wenn man die vornehmlich für diese Aufgabe verwendeten Datensätze und Annotationen betrachtet. Den häufig verwendeten 2D Gelenkpositionen fehlen wesentliche Informationen über die gesamte 3D Körperoberfläche. Und bei der Aufnahme 3D Körperdaten, z.B. durch Motion-Capture Systeme oder Körperscanner, wird Kontakt oft vermieden, weil er Okklusionen verursacht welche die Datensäuberung und -verarbeitung erschweren.

In unserer Arbeit befassen wir uns zunächst mit der Schätzung von körpereigenem Kontakt (“Eigenkontakt”) aus Farbbildern. Dafür entwickeln wir zwei neue Methoden, zugeschnitten auf die Schätzung menschlicher 3D Meshes mit Eigenkontakt, und erstellen damit neue Datensätze. Wir erstellen 1) 3DCP, einen Datensatz bestehend aus 3D Meshes von Personen in Posen mit Eigenkontakt und 2) MTP, einen Datensatz bestehend aus unter realen Bedingungen aufgenommenen Bildern mit 3D Referenzdaten, zu dessen Erstellung Personen verschiedene Posen nachstellen. Außerdem beobachten wir, dass 2D Gelenkpositionen in einem Foto leicht annotiert werden können, es für Eigenkontakt jedoch kein äquivalentes Label gibt. Daher führen wir ein neues Label, “diskreten Eigenkontakt”, ein um den Kontakt zwischen jeweils zwei verschiedenen Körperregionen anzuzeigen. Wir annotieren drei existierende Datensätze und verwenden diskreten Eigenkontakt während einer Optimierungsroutine um Körperteile die sich berühren sollten in Kontakt zu bringen. Anschließend trainieren TUCH, wir ein neuronales Netz zur Schätzung der 3D Körperoberfläche. Unsere Ergebnisse zeigen dass das Wissen über körpereigenen Kontakt die Schätzungen von Posen sowohl mit als auch ohne Eigenkontakt verbessert.

Als nächstes betrachten wir Kontakt im Kontext enger sozialer Interaktionen. Die Modellierung solchen Kontakts ist schwierig weil sich die beteiligten Personen gegenseitig verdecken. Zudem sind existierende Datensätze nicht groß genug um robuste neuronale Netze zur 3D Rekonstruktion solcher Interaktionen zu trainieren. In dieser Arbeit nutzen wir deshalb ein generatives Modell, BUDDI, um die gemeinsame Verteilung von

sich in sozialer Interaktion befindenden Personen zu erlernen und verwenden dieses Modell während einer Optimizernugsroutine. Als Trainingsdaten verwenden wir existierende Motion Capture Daten sowie Flickr Bilder mit diskreten Kontakt annotationen. Angelehnt an diskreten Eigenkontakt, nutzen wir diskreten Mensch-zu-Mensch Kontakt um zwei menschliche Meshes an detektierte 2D Gelenkpositionen zu fitten. Die meisten Methoden zur Generierung von 3D Meshes von Menschen, sind auf die Bewegung einer einzelnen Person ausgerichtet und agieren auf 3D Gelenkpositionen. Diese Methoden reichen jedoch nicht aus um physischen Kontakt zu modellieren, weil sie nicht die gesamte Körperoberfläche modellieren. Unser Ansatz agiert auf den Pose *und* Figur Parametern eines Körpermodells, wodurch 3D Meshes zweier Personen zu generiert werden können. Außerdem zeigen wir wie das Wissen über menschliche Proxemik unseres Modells während einer Optimierungsroutine verwendet werden kann. Dafür wird in jedem Optimierungsschritt die aktuelle Schätzung durch BUDDI verfeinert. Die verfeinerte Schätzung dient dann als Supervision in einer Zielfunktion. Im Gegensatz zu existierenden Methoden kann unsere Optimierungsroutine auch dann verwendet werden wenn kein annotierter Mensch-zu-Mensch Kontakt verfügbar ist.

Kontakt findet an der Körperoberfläche statt, jedoch schätzen die meisten starte-of-the-art Methoden Körper mit Durchschnittsfigur mangels Trainingsdaten bestehend aus unter natürlichen Bedingungen aufgenommenen Bildern mit entsprechender 3D Grundwahrheit der Figur. Hinzu kommt, dass 2D Gelenkpositionen nicht ausreichen um die Figur einer Person zu erklären. Eine naheliegende Lösung wäre ein neuer Datensatz bestehend aus Körperscans und Bildern; ein zeitintensives und teures Vorhaben, das nicht skalierbar ist. Stattdessen nutzen wir Vokabular das Meschen zur Beschreibung von Figur verwenden. Zuerst lassen wir Annotatoren beurteilen wie gut Worte wie “groß” oder “lange Beine” auf einen Körper zutreffen. Wir verschiedenste Körperformen, deren 3D Grundwahrheit uns vorliegt, sowie Fotos von Modelagenturwebseiten. Wir verwenden diese Daten um ein Figur-zu-Attribut (A2S) Modell zu fitten das die 3D Figur von Bewertungsvektoren vorhersagt. Anschließend trainieren wir einen Mesh Regressor, SHAPY, auf den Modelagenturbildern, wobei wir Figur durch Bewertungsvektoren überwachen. Um unser Modell zu evaluieren erstellen wir einen neuen Testdatensatz names “Human Bodies in the Wild” (HBW), bestehend aus Fotos und Körperscans. Unser Modell verbessert die Figurschätzung auf HBW quantitativ und sagt realistischere Figur vorher.

Zusammenfassend stellen wir in dieser Arbeit neue Datensätze, Optimierungsmethoden, ein generatives Modell, sowie neuronale Netze vor, um das Feld der 3D menschliche Posen- und Figurschätzung voranzubringen. Im gesamten eröffnen die in dieser Arbeit vorgestellten Methoden neue Wege um genauere und realistischere 3D Meshes von Bildern zu schätzen und zwar für Szenarien mit mehreren Personen mit Eigen- und Mensch-zu-Mensch Kontakt. Zudem ermöglicht die Forschung dieser Arbeit die Entwicklung generativer Modelle um natürlicherer Avatare zu kreieren. Wir glauben, dass das Wissen über Kontakt mit Hilfe des maschinellen Sehens weitreichende Auswirkungen in anderen Gebieten haben wird, zum Beispiel in der Robotik, im Fitnessbereich, oder in den Verhaltenswissenschaften.

Acknowledgements

“You must never think of the whole street at once, understand? You must only concentrate on the next step, the next breath, the next stroke of the broom, and the next, and the next. Nothing else.” Again he paused for thought before adding, “That way you enjoy your work, which is important, because then you make a good job of it. And that’s how it ought to be.”

Beppo Roadsweeper in Momo
Michael Ende, 1973

I consider myself very fortunate because I got to work with great inspiring people during my PhD and because I have a caring, loving and supportive family. Having the right people in your life makes life worth living and work just a lot more fun.

I would like to express my gratitude to Michael Black for his extensive and thorough support throughout my PhD. Michael, you built an environment that gave me room to be creative and grow. You took every single idea seriously or at least didn’t let on - even when I suggested building a foot model to walk in shoes of other people. You brought me in touch with amazing people in our field and paved the way for the next step. Especially, I value your inspiring ideas and support when things didn’t work out - “live and learn” - has become an integral part of my life.

I would also like to deeply thank Angjoo Kanazawa for her support during my research visit in Berkeley and during the last year of my PhD. Angjoo, I very much appreciate your technical understanding and enthusiasm to try new methods. The deadlines couldn’t be more fun and exciting together. Angjoo is not only a great advisor but also a role model to me, professionally and personally.

I would also like to thank my amazing PostDoc advisors Chun-Hao Paul Huang and Dimitris Tzionas for their support, ideas, and suggestions. Especially for many iterations in writing to turn our ideas into something other people can understand. I am very grateful to have worked with Georgios Pavlakos, a great mentor and caring friend.

This thesis wouldn’t be the same without the support of all my co-authors. Especially, I want to thank Ahmed for his honest open conversations and for answering all the ques-

tions a new student might have, Vassilis for making the work-from-home period a lot more fun, and Vickie for pushing through deadlines together.

I thank Katherine Kuchenbecker and Gerard Pons-Moll for being on my thesis advisory committee, for providing valuable feedback about my research projects and mentoring me through the academic jungle. I also want to thank Jitendra Malik for inviting me to the human body group meeting in BAIR; these meetings have given me a new perspective on human mesh regression and downstream tasks.

I would also like to thank Kostas Daniilidis, Alexei Efros, Gerard Pons-Moll, and Andreas Geiger for reviewing this thesis and for participating in the oral exam.

I would also like to thank the PS support teams, i.e. our admins, Nicole and Melanie, for keeping our very special department running and the software and data teams for setting up websites and spending many hours in the capture hall collecting data and running user studies. Especially, I would like to thank Tsvetelina for her support and advice from the beginning of my PhD until today.

It's a long way from an idea seed to a ripe publication with many contributors on the way. So I would also like to thank the many amazing people in PS and BAIR for fruitful, sometimes intensive, discussions during group meetings or when having coffee or ice cream. I am grateful to have met both labs, with their faculty and students and want to thank Shashank, Radek, Muhammed, Mohamed, Omid, Qianli, Priyanka, Paola, Sai, Haoran, Yao, Yuliang, Markos, Omri, Vongani, Jathushan, Ethan, Matt, Frederik, Shubham, Aleks, Nikos, and Kartik whose thoughts, support, critique, code, and discussion helped shape this thesis and my life in Tübingen and Berkeley. Without Soubhik, Partha, Nikos, Marilyn, Victoria, and Nadine this PhD experience would have been a different one. Thank you for suffering and laughing together, decorating my desk with chickens, and ensuring there's always a gum supply.

I would like to thank my friends, family, and in-law family, i.e. the supporters who hold back even when I kept missing birthdays because of deadlines, moves, or travel. Doing a PhD takes a lot but because of you there were many fun activities, weekend trips, deep talks, and nights to remember and recharge my battery. Particularly, I want to thank my younger brother Kai for infusing my life with a good sense of humor and for his care and hugs whenever needed.

A warm special thank you deserve my mum and dad for their consistent love and support throughout my life. You have given me everything a daughter can wish for and I couldn't be more lucky. Even if they don't understand much, my parents will probably be among the few people who will actually read the whole thesis.

Finally, I would like to thank my partner Julian for being in my life, riding this exciting journey together. Thank you for (still) being eager for new adventures even when they challenge our relationship, for your care not only during deadlines, and your patience and support with all these new projects and ideas. You have the marvelous gift of putting things in perspective and make me laugh even after the worst of days ♡

Contents

Abstract	v
Zusammenfassung	vii
Acknowledgements	ix
1 Introduction	1
1.1 Motivation	2
1.2 Summary of Content	3
1.3 Background	7
1.3.1 Human Body Models	7
1.3.2 Human Pose and Shape Estimation	10
2 On Self Contact and Human Pose	15
2.1 Introduction	16
2.2 Related Work	18
2.3 Self-Contact	20
2.3.1 Discrete Self-Contact	21
2.3.2 Vertex-based Self-Contact	21
2.3.3 Mesh Surface Points	22
2.4 Self-Contact Datasets	22
2.4.1 3D Contact Pose (3DCP) Meshes	23
2.4.2 Mimic-The-Pose (MTP) Data	25
2.4.3 Discrete Self-Contact (DSC) Data	29
2.4.4 Summary of the Collected Data	30
2.5 TOUCH	30
2.6 Evaluation	32
2.7 Conclusion	37
3 Generative Proxemics: A Prior for 3D Social Interaction from Images	39
3.1 Introduction	40
3.2 Related Work	42
3.3 Human-Human Contact	43
3.3.1 Discrete Human-Human Contact	43

3.4	Method	44
3.4.1	Reconstructing Bodies with Contact Maps	44
3.4.2	Diffusion Model for 3D Proxemics	45
3.4.3	Optimization with the Proxemics Prior	49
3.5	Experiments	51
3.5.1	Unconditional Generation	53
3.5.2	Fitting with BUDDI	54
3.6	Conclusion	61
4	Accurate 3D Body Shape Regression using Metric and Semantic Attributes	63
4.1	Introduction	64
4.2	Related Work	67
4.3	Representations and Data for Body Shape	69
4.3.1	SMPL-X Body Model	69
4.3.2	Model-Agency Images	69
4.3.3	Linguistic Shape Attributes	70
4.4	Mapping Shape Representations	72
4.4.1	Virtual Measurements (VM)	72
4.4.2	Attributes and 3D Shape	72
4.5	3D Shape Regression from an Image	75
4.6	Experiments	76
4.6.1	Evaluation Datasets	76
4.6.2	Evaluation Metrics	77
4.6.3	Shape-Representation Mappings	78
4.6.4	3D Shape from an Image	81
4.7	Conclusion	82
5	Conclusion	87
5.1	Contributions	88
5.2	Considerations for Future Work	89
5.3	Closing Thoughts	93
A	On Self Contact and Human Pose	97
A.1	Self-Contact Datasets	98
A.1.1	3D Contact Pose (3DCP) Meshes	98
A.1.2	Mimic-The-Pose (MTP) Data	99
A.1.3	Discrete Self-Contact (DSC) Data.	101
A.2	TUCH	103
A.3	Evaluation	104

B	Generative Proxemics: A Prior for 3D Social Interaction from Images	105
B.1	Method	106
B.1.1	Preprocessing	106
B.1.2	Optimization	107
B.1.3	Diffusion model	107
B.2	Training and Testing Datasets	108
B.2.1	Flickr Fits	108
B.2.2	Hi4D	110
B.2.3	CHI3D	110
B.3	Evaluation	110
B.3.1	Baseline Methods	110
B.3.2	User study	113
C	Accurate 3D Body Shape Regression using Metric and Semantic Attributes	115
C.1	Data Collection	116
C.1.1	Model-Agency Identity Filtering	116
C.2	Mapping Shape Representations	116
C.2.1	Shape to Anatomical Measurements (S2M)	116
C.2.2	Mapping Attributes to Shape (A2S)	118
C.2.3	Images to Attributes (I2A)	118
C.3	SHAPY - 3D Shape Regression from Images	119
C.4	Experiments	119
C.4.1	Metrics	119
C.4.2	Shape Estimation	121
C.4.3	Pose evaluation	122
	Bibliography	123

List of Tables

2.1	Existing 3D human mesh datasets	20
2.2	Evaluation of TOUCH on 3DPW and MPI-INF-3DHP	33
2.3	Evaluation of TOUCH on 3DCP Scan	33
2.4	Evaluation of TOUCH on 3DPW contact classes	33
2.5	Ablation study of data and algorithm in TOUCH	37
3.1	Evaluation of 3D Pose on FlickrCI3D Signatures	54
3.2	Evaluation of BUDDI on CHI3D	60
3.3	Evaluation of BUDDI on Hi4D	60
4.1	List of linguistic body shape attributes	72
4.2	Evaluation A2S on CMTS	78
4.3	Results of A2S and its variations	79
4.4	Evaluation of SHAPY on HBW	81
4.5	Evaluation of SHAPY on MMTS	81
4.6	Evaluation of SHAPY on SSP-3D	82
A.1	Ablation of MTP data and DSC data	104
B.1	Weights of the different loss terms in optimization	108
B.2	Evaluation of BUDDI on Hi4D	112
B.3	Ablation study of baseline methods on CHI3D	112
B.4	Ablation study of baseline methods on FlickrCI3D Signatures	112
C.1	Model comparison for A2S and AHW2S	118
C.2	Leave-one-out evaluation on MMTS	120
C.3	Leave-one-out evaluation on HBW	120
C.4	Evaluation of S2A model on CMTS	121
C.5	Evaluation of SHAPY on 3DPW	122

List of Figures

1.1	Side View of Pose with Self-Contact	4
2.1	TUCH teaser	16
2.2	Self-contact maps	21
2.3	Mapping mesh vertices to surface points	23
2.4	Registrations from 3DCP Scan	24
2.5	Self-contact optimization	25
2.6	Mimic-The-Pose (MTP) dataset	25
2.7	Presentation format of 3DCP during MTP data collection	26
2.8	Pushing and pulling terms to regulate the self-contact	28
2.9	MTP results. Meshes presented to AMT workers (blue) and the images they submitted with OpenPose keypoints overlaid. In grey, the pseudo ground-truth meshes computed by SMPLify-XMC.	29
2.10	Discrete Self-Contact (DSC) dataset	30
2.11	SMPLify-DC intermediate steps	31
2.12	SMPLify-DC results compared to SMPLify	32
2.13	Evaluation of TUCH vs. SPIN by body region	34
2.14	Qualitative results of TUCH on 3DPW	35
2.15	Qualitative results of TUCH on 3DPW (failures)	36
3.1	BUDDI teaser	40
3.2	Pseudo-ground truth data of people in close proximity	46
3.3	Illustration of the architecture of BUDDI	48
3.4	Unconditional generation of meshes using BUDDI	49
3.5	Optimization with Generative Proxemics	50
3.6	Amazon Mechanical Turk perceptual study layout and instructions	53
3.7	Automatic reconstruction of people in close social interaction on Flickr images	55
3.8	Qualitative examples from optimization with BUDDI	56
3.9	Qualitative examples from optimization with BUDDI	57
3.10	Qualitative examples from optimization with BUDDI	58
3.11	Failure cases optimization with BUDDI	59
4.1	SHAPY teaser	64
4.2	Model-agency data	66
4.3	Measurements and crowd-sourced linguistic body shape attributes	66

List of Figures

4.4	Shape representations and data collection	70
4.5	Histograms of body measurements collected from model-agency websites	71
4.6	Linguistic body shape attribute annotation template on AMT	73
4.7	SHAPY architecture	75
4.8	“Human Bodies in the Wild” (HBW) dataset	77
4.9	Qualitative results of SHAPY on HBW	80
4.10	Qualitative results of SHAPY (female bodies)	83
4.11	Qualitative results of SHAPY (male bodies)	84
4.12	Qualitative results of SHAPY (failure cases)	85
A.1	Hand on body prior weights visualization	100
A.2	Images in MTP per 3DCP subset	101
A.3	Body segments	102
A.4	Ambiguity in discrete self-contact annotation task	102
A.5	Images from 3DCP Scan dataset	103
B.1	Detailed architecture of BUDDI with conditioning	109
C.1	Automatic anatomical measurements on a 3D mesh	117
C.2	20K mesh surface points	119

Listings

B.1 Pseudo code for optimization with BUDDI.	113
--	-----

Chapter 1

Introduction

1.1 Motivation

The human skin is designed to experience touch – the ability to perceive a stimulus that comes into contact with the body surface. The sense of touch is crucial because it allows us to experience physical sensations, create and deepen social relationships, and establish a connection with the world around us. The field of computer vision should be able to model and reconstruct the full human body surface and capture its complex interactions with ourselves, other humans, and the environment. Understanding contact through computer vision will enable advancements in virtual and augmented reality and impact other fields like robotics and behavioral science. Interestingly, many works in our field investigate interaction between humans and scenes, while research on self- and human-human contact remains relatively scarce in comparison.

Human beings engage in self-touch or self-contact multiple times a day, indicating its behavioral significance and prompting extensive research in behavioral- and neuroscience. For instance, studies have shown that facial self-touch is a recognized indicator of stress in adults [34]. Infants frequently and spontaneously touch their own body, which serves as a way of exploration, facilitating the development of body awareness [90] and even fetuses show increased self-contact when maternal stress is present [160], highlighting its relevance across different life stages. The patterns of self-touch are manifold as they vary in speed, trajectory, or movement duration [13, 108] and their exact function is still unknown. Some work argues that self-touch mainly serves for self-stabilization and self-calming [108, 172] or as a mechanism for down-regulation in high arousal states [172]. Self-touch is also associated with emotional processes that interfere with working memory performance [49]. In particular, suppressing self-touch among individuals who frequently touch their own body leads to significantly worse memory performance in haptic working memory tasks [186]. Other research indicates a connection between different patterns of self-touch and neuropsychological state [13, 197] and mental arousal [105, 109, 204]. In dialogues, self-touching colloquists are rated significantly more honest, outgoing, likable, and positively with respect to the working relationship compared to their non self-touching equivalent [55], indicating a relevance not only for self-regulation but also as an outwardly effective mechanism.

The relevance of interpersonal touch or human-human contact has also extensively been investigated in behavioral science, in particular the role of social touch [167]. In fact, the body of research on social touch is extensive, and this paragraph can only offer a brief glimpse to highlight the relevance of interpersonal touch. Beginning from early childhood, physical contact between parent and child establishes bonds and is associated with immediate stress reduction [187, 38], enhanced object exploration [193], and long-term effects on behaviour [9, 150, 24]. Early vocabulary items may consist of words often linked with caregiver touches [173]. The avoidance of interpersonal touch can be a predictor of autism spectrum disorder in older children [12, 131]. Social touch also has many effects in adulthood. Crusco and Wetzel [30] show that a slight touch increased tips in restaurants, i.e. touch causes a more friendly behaviour towards the touch-giver,

also known as the Midas touch. Later studies find similar effects, e.g. that exposure to social touch increases a bus driver’s willingness to transport customers without having enough money for the ticket [51]. In virtual reality, agents with touch are perceived as more human-like [70].

Despite the great relevance of self- and human-human contact to learn about human behavior and conditions, most research on this topic is constrained by small group sizes because contact usually requires manual annotation as only a few rudimentary detection and reconstruction methods exist. This prevents understanding the importance and functionality of touch on human behavior at scale. The field of computer vision could advance the understanding of human social interactions by providing methods for 3D mesh reconstruction with accurate self- and mutual contact from images and video.

Unfortunately, self- and human-human contact has rarely been studied. One reason is that contact is rare in most human scan and motion capture (MoCap) datasets, because contact naturally leads to occlusion, which hampers data capturing. In body scan datasets, most poses avoid self-contact and in MoCap systems usually only a single person is captured. The implications for our field are evident: recent 3D motion generation methods can perfectly synthesize a single static person [58, 235, 60, 238] or human motion [57, 210, 237, 61, 135, 73], but can not generate two people shaking hands. Another problem is that most methods for estimating 3D pose and shape predominantly rely on 2D joint locations for supervision. However, 2D joints are not sufficient to accurately estimate the body surface, because one set of 2D joints can be explained by multiple body shapes and also by multiple poses when no ground-truth camera information is available. Priors, i.e. mathematical functions or models that incorporate prior knowledge about human pose and shape, are usually learned from scan and MoCap datasets that hardly contain contact poses. This leads to 3D mesh estimates that, when projected onto the image, satisfy reprojection constraints and may perfectly overlay with the image evidence. A rotation to the side, however, reveals that the estimated poses are not correct.

Being able to reconstruct and generate meshes with self- and mutual contact will facilitate the creation of avatars aligned with human behaviour, which will let them appear more human-like, natural, and realistic.

1.2 Summary of Content

Self- and human-human contact plays an important role in our everyday lives, but existing art in 3D computer vision fails to accurately estimate meshes with contact. Our goal is to develop methods to accurately predict human pose and shape from an RGB image when poses involve self- and interpersonal contact. This is a challenging task, because (1) existing datasets and labels, commonly used in human pose and shape estimation like 2D joint locations, are not sufficient to reason about contact, (2) annotating contact on an image in 2D is difficult since contact happens on the 3D body surface, and (3) collecting 3D data such as body scans or via MoCap of contact scenarios is expensive and



Figure 1.1: Estimate of a state-of-the-art human mesh regressor [47]. The estimate projected onto the images perfectly overlays with 2D joint locations. A rotation to the side, however, shows that the predicted pose is not correct.

time-consuming since this requires manual data cleaning. During motion capture, for example, self-contact can obscure markers or cause them to detach from the motion capture suit, which disrupts the automated data cleaning process. Addressing these challenges necessitates the development of novel approaches for constructing datasets tailored for the task of estimating poses with contact and, in addition, methodology that is robust to occlusions and limited training data. Following previous work, our goal is to estimate the parameters of a 3D human body model. In particular, we use the SMPL [123] and SMPL-X [144] body models which we describe in Section 1.3 in detail. In a nutshell, body models are functions, learned from e.g. body scans, that take human pose and shape parameters as input and output a 3D mesh. Such models are frequently used to estimate 3D human meshes from images or video, usually achieved through either parameter regression or optimization techniques. Despite the impressive achievements of previous methods that address this task, these method can not accurately estimate 3D contact. A more comprehensive introduction to mesh estimation approaches is given in Section 1.3.

On Self Contact and Human Pose. In Chapter 2, we study the problem of 3D human pose estimation for poses with self-contact. The first insight in this chapter is that humans can easily detect and label self-contact in images, which provides information beyond keypoints that can serve as additional supervision in regressor training and optimization. To this extend, we introduce discrete self-contact (DSC) labels and demonstrate their use in an optimization routine when fitting 3D meshes to 2D joint locations. To collect such labels, we divide the body into 24 regions and ask humans to annotate the pairwise region-to-region contact given an RGB image. Previous art has demonstrated

how 3D meshes can be fit to images by minimizing the error between 3D joints projected into the image and detected 2D joint locations [19]. Our method, SMPLify-DC, is inspired by previous art, but additionally takes discrete contact labels into account to encourages contact between regions annotated to touch. Discrete self-contact is a valuable signal, but it lacks detail and is susceptible to depth ambiguity in single images. For example, imagine a photo of someone holding their hand close to their eyes. Given a frontal photo of this pose, it can be difficult to tell whether the hand is actually touching the eyes or not. Therefore, we introduce Mimic-The-Pose (MTP), a novel data collection setup, reversing the usual annotation processes. Instead of asking humans to annotate images, we start with a 3D mesh in a pose containing self-contact and then gather photos matching the pose. Since self-contact poses are rare in 3D mesh datasets, we first construct 3DCP, a novel dataset of body scans in poses with self-contact as well as meshes in near-contact poses from a MoCap database [130] that we refine to remove self-intersection and encourage contact. Then we present these meshes to trial participants and ask them to mimic the pose. While the participant is pausing in this pose somebody takes a picture. The presented pose and the pose of the person on the photo are usually already very similar, but they do not match perfectly. To address this, we refine the presented mesh with respect to detected keypoints via optimization. Our fitting approach is inspired by previous work that estimates expressive 3D humans given a single image, i.e. meshes with finger articulation and facial expression besides body pose and shape [144]. However, our method considers the ground-truth height and weight of the person on the photo, uses the presented pose and self-contact as guidance, and incorporates novel losses to encourage contact while resolving intersections. We call this optimization SMPLify-XMC and the dataset of images in the wild and refined meshes MTP. Finally, we use the DSC and MTP datasets to train TOUCH, a 3D human pose and shape regressor. TOUCH has the same design as SPIN [101], where a regressor outputs a pose and shape estimate, which is refined by optimization. The optimized meshes are used as supervision for the regressor. We use MTP data as if it was ground-truth and DSC during optimization via SMPLify-DC. The results show, that using self-contact in regressor training improves 3D pose estimation not only for poses with self-contact, but also for poses without self-contact.

Generative Proxemics: A Prior for 3D Social Interaction from Images. In Chapter 3, we exploit the observations from the first chapter to improve 3D mesh estimation of multiple people in close interaction. Existing regressors like BEV [190] can predict the rough pose and spacial positioning of multiple people from images, but fail to capture the subtle detail of human-human contact. To address this problem, we use an existing dataset of Flickr images with discrete human-human contact annotations [41] and design an optimization routine similar to SMPLify-DC but for two people. Our routine refines an initial regressor estimate with respect to detected keypoints, takes discrete human-human contact labels into account, and resolves intersections between people. We use these fits and motion capture data to train BUDDI, a generative model that learns the joint distribution of humans in close interaction. BUDDI is a diffusion model [185, 68]

that takes SMPL-X parameters of two people disturbed with noise as input to transformer network. The networks task is to “denoise”, i.e. remove the noise, from the noisy input parameters. At test time, we can start from random noise to sample novel pairs of people in close proximity from BUDDI. The majority of generative methods for 3D humans operate on 3D joint locations of a single person, i.e. these methods do not model the human body surface, and are therefore not sufficient for generating two meshes with interpersonal contact. Our approach, in contrast, acts on SMPL-X pose and shape parameters which enables us to sample meshes of two people. We further demonstrate how the knowledge of human proxemics, incorporated in our model, can be used to guide an optimization routine. Previous work introduces an approach that uses 2D text-to-image diffusion models during text-to-3D synthesis [153]. We draw inspiration from this work and demonstrate how diffusion models can be used as prior during optimization for multi-person pose estimation. We find that BUDDI knows enough about how people interact to forgo ground-truth discrete human-human contact labels at test time. This is the first demonstration of human mesh optimization for two humans in close proximity that does not rely on ground-truth labels.

Accurate 3D Body Shape Regression using Metric and Semantic Attributes. While discrete self- and human-human contact labels and pose mimicking are useful to advance 3D human pose estimation, they are not sufficient, because human pose with contact can only be estimated accurately in 3D if we also know a person’s body shape. This task is challenging due to the lack of ground-truth training data of images in the wild with paired ground-truth 3D data and because 2D keypoints can not explain the full variety of human shape. Imagine a person gaining weight; their body shape changes, while the skeleton remains the same. Yet, most optimization methods and regressors for human pose and shape estimation focus on body pose and supervise shape only through keypoints and simple body shape priors. For methods that do estimate body shape, the most commonly used signal to estimate body shape is therefore a person’s silhouette which can easily be detected in images using standard computer vision methods. Without ground-truth camera information, however, the estimated shape is only correct up to a scaling factor and usually a person’s true silhouette is covered by clothing. If keypoints and silhouettes are not sufficient, what information can humans provide to label body shape? Previous work has observed that humans have many words to describe body shape, e.g. “tall” or “pear-shaped”, and shown that the relation between SMPL shape parameters and rating vectors indicating how much each word applies to a 3D body shape can be modeled via linear regression [188]. In Chapter 4, we demonstrate that linguistic body shape attributes can be used as supervision signal in end-to-end learning. To do this, we first collect images with a few body measurements from fashion model agency websites. Then we ask workers on Amazon Mechanical Turk to rate how much a word applies to (i) images of the fashion models and (ii) rendered images of CAESAR [162] bodies. We use (ii) to map SMPL-X parameters and body shape attribute ratings (S2A) and employ this mapping in regressor training using the model images and labels from (i). Our regressor, SHAPY, predicts SMPL-X pose and shape parameters from a single image,

obtain predicted attribute ratings from shape parameters via S2A, and use the distance between the predicted and ground-truth attribute rating vector to supervise body shape. We are the first to demonstrate the use of linguistic body shape attributes in network training to supervise body shape.

In summary, this thesis address the problem of reconstructing humans in contact poses in several ways. TUCH investigates 3D human pose estimation for poses involving self-contact, BUDDI studies contact between people in close interaction, and SHAPY addresses the problem of human shape estimation. We introduce multiple novel datasets for this task and present regressors, optimization methods and a generative model tailored to each specific problem. Taken together, this line of research may enable more realistic reconstruction of pose, shape and, contact of multiple people from images, and open up new research directions.

1.3 Background

1.3.1 Human Body Models

A human body model in the context of computer graphics is a digital representation of the human body. The first attempts of creating human body models reach back into the 1970s. These models take bone length and joint angles as input and output a stick figure [212]. Stick figures are extreme simplifications of the body and lack necessary detail hampering the perception of 3D pose and body shape. Surface figures surround the skeleton of a stick figure with planar or curved patches, where removed hidden lines improve pose perception [40]. Volume-based models define a body surface, i.e. “skin” surrounding the skeleton by decomposing the body into volumes, i.e. cylinders [154], ellipsoids [63, 64], or spheres [7]. Badler and Smohar [8] and Magnenat-Thalmann and Thalmann [129] provide a more comprehensive description of these early versions of human body models. Such models are still not very realistic since they do not model local soft-tissue deformation, i.e. the pose-dependent compression of soft tissue e.g. in the knee pit. In 1988, Komatsu [103] defines a human skin model where the skin deformation is driven by an underlying skeleton. Around the same time, Magnenat-Thalmann et al. [128] model local joint-dependent deformations of hands. Such models can generate more realistic human bodies, but they can not portray the complexity of real human anatomy.

Instead of manually defining local soft-tissue deformation, more recent art proposes to learn how the human body deforms with pose from real-world data. To do this, SCAPE [5] represents the human skin through triangles which they fit via least-squares to body scans. Applications of SCAPE are shape completion by fitting the model to noisy body scans to obtain 3D mesh and 3D animation of a moving person by fitting the model to motion capture markers. SCAPE produces realistic shape and pose deformations learned form data, but it is time-consuming since every pose and shape change

requires solving a least-squares problem. Another limitation of SCAPE is its absence of an underlying skeleton, which is a framework typically employed by animators during the animation process. A recent body model that addresses this problem is SMPL [123]. SMPL is a vertex-based model with an artist-designed topology of triangular faces. In contrast to SCAPE, SMPL has an underlying 3D joint structure that emulates the human skeleton. The linear blend skinning is learned from large real datasets of human pose and shape, i.e. 3D body scans of people in different poses and CAESAR [162] with different body shapes. We introduce SMPL and its expressive version, SMPL-X [144], in detail in Section 1.3.1.

Since 2015, the research on human bodies in computer vision has rapidly developed. In 2017, Romero and Tzionas et al. published MANO, a hand model, and SMPL+H, which extends SMPL with regard to finger articulation [165]. Also in 2017, Li and Bolkart et al. introduced FLAME [114], a model of facial shape and expression. This is followed by Hesse et al.’s SMIL [65] in 2018, a body model for children in SMPL topology. The latest human body models, SMPL-X [144], GHUM [220], SUPR [141], and STAR [140], incorporate finger articulation and facial expression besides body pose and shape.

Body models are used for many tasks beyond the classical applications in graphics and animation, e.g. to predict a person’s pose [87, 189, 190, 102, 18, 39, 115, 85, 201, 198] shape [174, 176] from an image or from video [96, 225], to generate human body motion [149, 196, 191, 241], to learn priors for human pose and motion [199, 161], to learn about contact between human and the world [58, 147, 117, 180], to virtually dress people [126, 28], to reason about camera parameters [99, 225, 228], for action recognition [157, 47] and tracking [158, 159], for reconstructing sign language [104, 44], or to investigate bias between language and body shape [14, 156].

Within this thesis, we employ the SMPL body model [123] and its more expressive variant, SMPL-X [144], to estimate a person’s pose and shape from an images. The subsequent sections of this chapter introduce these two body models and introduce three fundamental publications to the task of human pose and shape estimation, i.e. SMPLify-X [144] and SPIN [101] with HMR [87].

SMPL

In 2015, Loper et al. [124] presented SMPL, a statistical model of the human body learned from data. SMPL is a differentiable function

$$M(\theta, \beta; \Phi) : \mathbb{R}^{|\theta| \times |\beta|} \rightarrow \mathbb{R}^{3N}$$

that maps body pose θ and body shape β via learned parameters Φ to a 3D mesh M . The mesh topology is artist-defined and consists of $N = 6,890$ vertices V , connected through triangular faces F , and $K = 23$ joints connected through a skeletal rig. We can modify the pose of a 3D human mesh by manipulating θ and body shape by manipulating β .

A pose θ is defined by $3 \times K + 3 = 72$ parameters, i.e. 3 rotation angles per joint plus three parameters for rotating the entire body, the global body orientation. The authors use principal component analysis (PCA) to model body shape variations in humans and therefore each scalar in β affects one PCA component. The first component captures for example the variations in body height, the second component in weight etc.

The learnable parameters

$$\Phi = \{T_R, \mathcal{S}, \mathcal{P}, \mathcal{W}, \mathcal{J}\}$$

are determined during SMPL training by minimizing the vertex reconstruction error of the training data, i.e. the Euclidean distance between scan points and the mesh surface.

To place vertices through Φ , SMPL uses linear blend skinning (LBS), a method known from animation in which the surface of a mesh is attached to an underlying skeletal structure. During the skinning process, pose- and shape-dependent blend shapes, i.e. an offset per vertex due to body pose and identity, are additively combined with the template mesh. The pose blend shapes \mathcal{P} , the blend weights \mathcal{W} , and \mathcal{J} , a matrix that transforms the rest vertices (template mesh with shape blend shapes applied) into rest joints, are learned from a dataset of scans of people in different poses. The rest pose template mesh T_R and shape blend shapes \mathcal{S} are learned 3D scans of people with diverse body shape.

The SMPL function is

$$M(\theta, \beta; \Phi) = \sum_{k=1}^K w_k G'_k(\theta, J)(T_R + D_S + D_P).$$

$D_S = B_S(\beta; \mathcal{S})$ and $D_P = B_P(\theta; \mathcal{P})$ denote the shape and pose dependent vertex displacements, respectively, obtained from the blend shape functions B_S and B_P . $w_k \in \mathbb{R}^N$ a vector defining how much each vertex is effected by the rotation of part k , $J = \mathcal{J}(T_R + D_S)$ denotes the 3D joints in rest pose, and $G'_k(\theta, J)$ a function returning the world transformation of joint k after removing transformation due to the rest pose.

The disentanglement of pose and shape-dependent deformation in SMPL is a useful property for the task of human shape estimation; a problem this thesis addresses in Chapter 4. To label and evaluate body shape, we use the rest pose template mesh with only shape blend shapes applied, i.e. $T_S = T_R + D_S$. The authors train a male, female and gender-neutral version of SMPL on corresponding splits of the training data, with the neutral model being trained on data of male *and* female subjects.

SMPL-X

SMPL knows about the effects of major body joints, but misses expressiveness since finger articulation or facial expressions are not modeled. To address this, Pavlakos and Choutas et al. [144] introduce SMPL-X, a body model that extends SMPL with fully articulated finger and an expressive face. The hand pose and facial expression are added

by leveraging blendshapes from existing hand and body models, i.e. MANO [165] and FLAME [114], respectively. In particular, SMPL-X maps pose, $\theta \in \mathbb{R}^{55 \times 3}$, shape, $\beta \in \mathbb{R}^B (B \leq 300)$, and expression, $\psi \in \mathbb{R}^{10}$, parameters to a 3D mesh:

$$M(\theta, \beta, \psi; \Phi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \rightarrow \mathbb{R}^{3N}.$$

The mesh topology consists of $N = 10,475$ vertices, V , connected through triangular faces, F . The SMPL-X template mesh T_R combines artist designed templates of SMPL for the main body and hands and FLAME for the head. The SMPL-X skeleton consists of 55 joints: i.e. 1 joint for body global orientation, 21 for the main body, 3 for yaw and left/right eye, and 15 for each hand. The full SMPL-X function is describes as:

$$M(\theta, \beta; \Phi) = \sum_{k=1}^K w_k G'_k(\theta, J)(T_R + D_S + D_P + D_E).$$

All terms correspond to those from SMPL except the blend shapes for facial expression, $D_E = B_E(\psi; \mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n \mathcal{E}$ with \mathcal{E} being the FLAME blend shapes. To model finger articulation, SMPL-X uses a lower dimensional PCA space of MANO instead of the full 30 finger joint. As for SMPL, the authors train a male, female and gender-neutral version of SMPL-X.

1.3.2 Human Pose and Shape Estimation

A common approach to estimate a person’s body pose and shape in 3D from images is to either optimize [144] over or regress [87, 97] the parameters of a human body model like SMPL. During optimization, the primary objective is to reduce the Euclidean distance between the estimated 3D joints projected onto the image and the ground-truth 2D joints, while pose and shape priors prevent unrealistic estimates. Networks are usually trained on images taken in the wild where 2D joints also serve as supervision signal for pose as well as on datasets of images with paired 3D ground-truth, e.g. from motion capture. In the following paragraphs, we will introduce SMPLify-X [144], an optimization-based approach for fitting SMPL-X to image keypoints, and SPIN [101], a network trained end-to-end for predicting 3D meshes from single images.

SMPLify-X

SMPLify-X [144] is an optimization routine to automatically estimate 3D pose and shape of a person from a single image by fitting SMPL-X meshes to detected 2D joint locations.

To do this, SMPL-X pose θ , shape β , and expression ε parameters are optimized under the objective to minimize the error between detected 2D joints, J^{2D} , and estimated projected 3D body model joints \hat{J}^{3D} . This error is also called “re-projection error/loss” or “2D keypoint error/loss”. To compute this error, we first obtain the 3D body model joints, \hat{J}^{3D} , from the current pose $\hat{\theta}$, shape $\hat{\beta}$, and expression $\hat{\varepsilon}$ estimates via SMPL-X.

Next, a camera model, Π , is employed to depict how objects in the 3D world are transformed into their corresponding 2D representations as they appear on the camera’s sensor or image plane. This camera model encompasses various parameters, K , i.e. the camera’s intrinsic (focal length and optical center) and extrinsic (position/translation and orientation/rotation in the 3D world) properties. The joint error is computed via

$$L_J = \sum_{\text{joint } i} \gamma_i \omega_i \rho(\Pi_K f_i^{3D} - J_i^{2D})$$

with ρ being a robust differentiable Geman-McClure penalty function, ω_i the confidence of the estimate of joint i provided by the detection algorithm, and γ_i a per-joint weight. Additional terms in the optimization objective prevent unrealistic mesh estimates, e.g. extrem bending of knees and elbows and unnatural body shapes or facial expressions. Such terms are called “priors”. SMPLify-X uses multiple priors:

- $L_\alpha = \sum_i \exp(\theta_i)$, where i sums over SMPL pose parameters corresponding to elbow and knee prevents extrem bending of these joints.
- L_{m_h} , L_{θ_f} , L_β , L_ϵ are squared L2 priors for hand and face pose, body shape and facial expression. For example, for the shape parameter β , the squared L2 prior is defined as $L_\beta = \|\beta\|^2$.
- L_{θ_b} is a body pose prior applied to the latent vectors of VPoser, a variational autoencoder.

To further prevent self-interpenetration, SMPLify-X uses a collision term, $L_{\mathcal{T}}$, that pulls vertices that are inside the 3D mesh to the surface. First, a list, \mathcal{T} , of colliding triangles, (t_a, t_b) , is detected, and a local conic 3D distance field, Ψ , computed, defined by the triangles in \mathcal{T} and their normals n_a, n_b . For two colliding triangles, t_a and t_b , the vertices of t_a intrude t_b ’s distance field Ψ_{t_b} and vice versa. The collision term is defined as:

$$L_{\mathcal{T}} = \sum_{(t_a, t_b) \in \mathcal{T}} \left\{ \sum_{v_a \in t_a} \|\Psi_{t_b}(v_a) n_a\|^2 + \sum_{v_b \in t_b} \|\Psi_{t_a}(v_b) n_b\|^2 \right\}$$

Fitting a 3D mesh to keypoints is a challenging task, especially when no good initial estimate, e.g. from a regressor, is available because of which the optimization routine starts from “mean” or T-pose. Therefore, the authors use various tricks that help improve the overall result. First, the SMPLify-X optimization runs in multiple stages to prevent small body parts like fingers to dominate the optimization in the beginning. Second, instead of directly optimizing 3D rotations, the authors train a variational autoencoder, VPoser, that learns a 32-dimensional latent space of human poses. Instead of optimizing axis angle representations of joint rotations, SMPLify-X optimizes θ_{VAE} , i.e. the body pose represented as latent vector of VPoser.

The final method optimizes over θ_{VAE} , β , and ε by minimizing an objective function with a loss weight λ specific to each term:

$$L_{\text{SMPLify-X}} = L_J + \lambda_\alpha L_\alpha + \lambda_{m_h} L_{m_h} + \lambda_{\theta_f} L_{\theta_f} + \lambda_\beta L_\beta + \lambda_\varepsilon L_\varepsilon + \lambda_{\theta_b} L_{\theta_b} + \lambda_{\mathcal{T}} L_{\mathcal{T}} \quad (1.1)$$

SPIN

An important early piece of research addressing the problem of human pose and shape estimating in end-to-end learning is the work of Kolotouros and Pavlakos et al. [101]. Neural networks are usually trained for multiple days on large datasets to directly predict body model parameters from an input image. The network architecture proposed in SPIN consists of two key components: a neural network that directly predicts SMPL pose and shape parameters from an input image and an optimization module, similar to SMPLify-X, that refines the predicted parameters with respect to 2D joint locations. The refined parameters serve as new training data for the regressor.

Regression module. The design of the SPIN regressor was first proposed by Kanazawa et al. [87] in 2018. This regressor was part of the first end-to-end method capable of estimating SMPL pose and shape from images, also known as ‘‘Human Mesh Recovery’’ (HMR). It consists of a ResNet-50 and a 3D regression module. First, the ResNet encodes an input image I into a latent vector ϕ . Then latent is passed to a 3D regression module that predicts SMPL parameters (pose and shape) and camera parameters (rotation, translation, and scale). A key component of HMR is the iterative application of the 3D regression module.

One forward pass of a new image through HMR yields:

$$\text{HMR}(I) = \{\theta_{\text{reg}}, \beta_{\text{reg}}, K_{\text{reg}}\}.$$

In SPIN the authors suggest a small modification to HMR: instead of predicting joint rotations in axis angle format, they use 6D rotation representations as they observe faster convergence during training.

Optimization module. The optimization routine is a simplified version of SMPLify-X. The optimization objective to be minimized is:

$$L_{\text{SPIN-Optimization}} = L_J + \lambda_\theta L_\theta + \lambda_\alpha L_\alpha + \lambda_\beta L_\beta, \text{ where}$$

L_J is the re-projection error between the ground-truth and predicted joints, L_θ a mixture of Gaussians pose prior trained with meshes fit to MoCap data, L_α a prior preventing extreme joint bending of knees and elbow joints, L_β a quadratic L2 shape prior. The optimization routine is initialized from the regressor’s prediction. Since this estimate is usually already close to the ideal pose, a few optimization iterations are usually enough

to converge to a good fit. From the optimization routine we obtain:

$$\text{OPTI}(\theta, \beta, K) = \{\theta_{\text{opt}}, \beta_{\text{opt}}, K_{\text{opt}}\}.$$

SPIN. We now describe how the regression and optimization are combined in SPIN training. First, an image is forwarded through the regression module providing the regressed body model and camera parameters $\{\theta_{\text{reg}}, \beta_{\text{reg}}, K_{\text{reg}}\}$. A common approach would be to apply human pose and shape losses right. In SPIN, however, the parameters are first passed to the optimization module which creates a refinement:

$$\text{OPTI}(\text{HMR}(I)) = \{\theta_{\text{opt}}, \beta_{\text{opt}}, K_{\text{opt}}\}.$$

These new refined parameters serve as training data for the regressor. Note that the refined parameters need to be detached from the gradient. Then the regressed pose and shape can be supervised with the refined pose and shape as follows:

$$L_{\text{SPIN-Regressor}} = \|\theta_{\text{reg}} - \theta_{\text{opt}}\|_2^2 + \|\beta_{\text{reg}} - \beta_{\text{opt}}\|_2^2 + \|J_{\text{reg}} - J_{\text{opt}}\|_2^2.$$

SPIN is trained four *in-the-wild* datasets of images with 2D keypoint annotations, i.e. LSP [79], LSP-extended [81], MPII [4], and MS COCO [120]. Additionally, they use images from datasets captured in the lab with ground truth 3D joint annotations, i.e. Human3.6M and MPI-INF-3DHP.

Chapter 2

On Self Contact and Human Pose

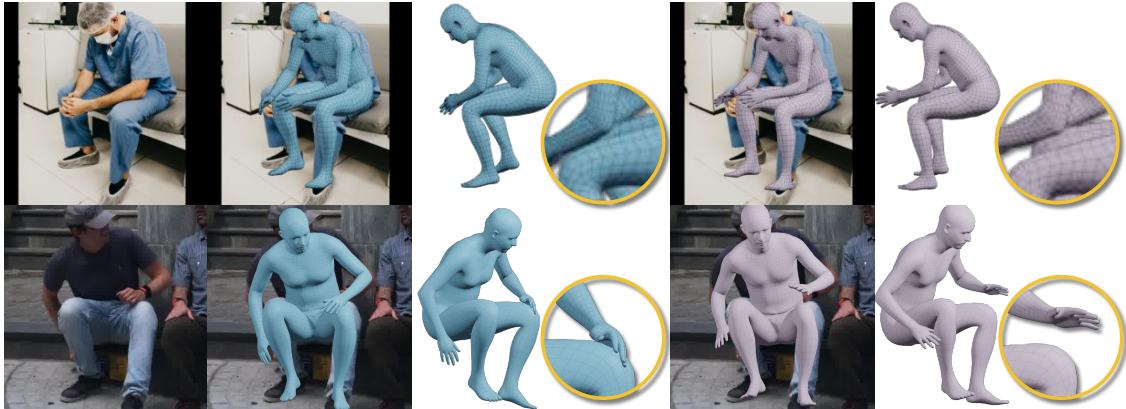


Figure 2.1: The first column shows images containing self-contact. In blue (left), results of our new network, compared to SPIN [101] in violet (right). When rendered from the camera view, the estimated pose may look fine (column two vs. four). However, when rotated, it is clear that training TUCH with self-contact information improves 3D pose estimation (column three vs. five).

In this chapter, we will investigate the problem of estimating human pose from images when a pose involves self-contact. A major challenge in order to solve this problem is the lack of suitable training data of images and self-contact labels and methods to process such information. To address this, we will introduce three new datasets and show how to use them in regressor training.

2.1 Introduction

Self-contact takes many forms. We touch our bodies both consciously and unconsciously [106]. For the major limbs, contact can provide physical support, whereas we touch our faces in ways that convey our emotional state. We perform self-grooming, we have nervous gestures, and we communicate with each other through combined face and hand motions (e.g. “shh”). We may wring our hands when worried, cross our arms when defensive, or put our hands behind our head when confident. A Google search for “sitting person” or “thinking pose” for example, will return images, the majority of which, contain self-contact.

Although self-contact is ubiquitous in human behavior, it is rarely explicitly studied in computer vision. For our purposes, self-contact comprises “self touch” (where the hands touch the body) and contact between other body parts (e.g. crossed legs). We ignore body parts that are frequently in contact (e.g. at the crotch or armpits) and focus on contact that is communicative or functional. Our goal is to estimate 3D human pose and shape (HPS) accurately for any pose. When self-contact is present, the estimated pose should reflect the true 3D contact.

Unfortunately, existing methods that compute 3D bodies from images perform poorly on images with self-contact; see Fig. 2.1. Body parts that should be touching generally are not. Recovering human meshes from images typically involves either learning a regressor from pixels to 3D pose and shape [87, 101], or fitting a 3D model to image features using an optimization method [19, 144, 215, 216]. The learning approaches rely on labeled training data. Unfortunately, current 2D datasets typically contain labeled keypoints or segmentation masks but do not provide any information about 3D contact. Similarly, existing 3D datasets typically avoid capturing scenarios with self-contact because it complicates mesh processing. What is missing is a dataset with in-the-wild images and reliable data about 3D self-contact.

To address this limitation, we introduce three new datasets that focus on self-contact at different levels of detail. Additionally, we introduce two new optimization-based methods that fit 3D bodies to images with contact information. We leverage these to estimate pseudo ground-truth 3D poses with self-contact. To make reasoning about contact between body parts, the hands, and the face possible, we represent pose and shape with the SMPL-X [144] body model, which realistically captures the body surface details, including the hands and face. Our new datasets then let us train neural networks to regress 3D HPS from images of people with self-contact more accurately than state-of-the-art methods.

To begin, we first construct a *3D Contact Pose (3DCP)* dataset of 3D meshes where body parts are in contact. We do so using two methods. First, we use high-quality 3D scans of subjects performing self-contact poses. We extend previous mesh registration methods to cope with self-contact and register the SMPL-X mesh to the scans. To gain more variety of poses, we search the AMASS dataset [130] for poses with self-contact or “near” self-contact. We then optimize these poses to bring nearby parts into full contact while resolving interpenetration. This provides a dataset of valid, realistic, self-contact poses in SMPL-X format.

Second, we use these poses to collect a novel dataset of images with near ground-truth 3D pose. To do so, we show rendered 3DCP meshes to workers on Amazon Mechanical Turk (AMT). Their task is to *Mimic The Pose (MTP)* as accurately as possible, including the contacts, and submit a photograph. We then use the “true” pose as a strong prior and optimize the pose in the image by extending SMPLify-X [144] to enforce contact. A key observation is that, if we know about self-contact (even approximately), this greatly reduces pose ambiguity by removing degrees of freedom. Thus, knowing contact makes the estimation of 3D human pose from 2D images more accurate. The resulting method, SMPLify-XMC (for SMPLify-X with Mimicked Contact), produces high-quality 3D reference poses and body shapes in correspondence with the images.

Third, to gain even more image variety, we take images from three public datasets [80, 82, 122] and have them labeled with discrete body-part contacts. This results in the *Discrete Self-Contact (DSC)* dataset. To enable this, we define a partitioning of the body into regions that can be in contact. Given labeled discrete contacts, we extend SMPLify to optimize body shape using image features and the discrete contact labels. We call this

method SMPLify-DC, for SMPLify with Discrete Self-Contact.

Given the MTP and DSC datasets, we finetune a recent HPS regression network, SPIN [101]. When we have 3D reference poses, i.e. for MTP images, we use these as though they were ground truth and do not optimize them in SPIN. When discrete contact annotations are available, i.e. for DSC images, we use SMPLify-DC to optimize the fit in the SPIN training loop. Fine-tuning SPIN on MTP and DSC significantly improves accuracy of the regressed poses when there is contact (evaluated on 3DPW [207]). Surprisingly, the results on non-self-contact poses also improve, suggesting that (1) gathering accurate 3D poses for in-the-wild images is beneficial, and (2) that self-contact can provide valuable constraints that simplify pose estimation.

We call our regression method *TUCH* (Towards Understanding Contact in Humans). Figure 2.1 illustrates the effect of exploiting self-contact in 3D HPS estimation. By training with self-contact, TUCH significantly improves the physical plausibility.

In summary, the key contributions in Chapter 2 are:

- (1) We introduce TUCH, the first HPS regressor for self-contact poses, trained end-to-end.
- (2) We create a novel dataset of 3D human meshes with realistic contact (3DCP).
- (3) We define a “Mimic The Pose” MTP task and a new optimization method to create a novel dataset of in-the-wild images with accurate 3D reference data.
- (4) We create a large dataset of images with reference poses that use discrete contact labels.
- (5) We show in experiments that taking self-contact information into account improves pose estimation in two ways (data and losses), and in turn achieves state-of-the-art results on 3D pose estimation benchmarks.
- (6) The data and code are available for research purposes at <https://tuch.is.tue.mpg.de>.

2.2 Related Work

3D pose estimation with contact. Despite rapid progress in 3D human pose estimation [86, 87, 101, 134, 144, 171, 215], and despite the role that self-contact plays in our daily lives, only a handful of previous works discuss self-contact. Information about contact can benefit 3D HPS estimation in many ways, usually by providing additional physical constraints to prevent undesirable solutions such as interpenetration between limbs.

Body contact. Lee and Chen [110] approximate the human body as a set of line segments and avoid collisions between the limbs and torso. Similar ideas are adopted

in [15, 45] where line segments are replaced with cylinders. Yin et al. [226] build a pose prior to penalize deep interpenetration detected by the Open Dynamics Engine [184]. While efficient, these stickman-like representations are far from realistic. Using a full 3D body mesh representation, Pavlakos et al. [144] take advantage of physical limits and resolve interpenetration of body parts by adding an interpenetration loss. When estimating multiple people from an image, Zanfir et al. [232] use a volume occupancy exclusion loss to prevent penetration. Still, other work has exploited textual and ordinal descriptions of body pose [145, 151]. This includes constraints like “Right hand above the hips”. These methods, however, do not consider self-contact.

Most similar to our is the work of Fieraru et al. [41], which utilizes discrete contact annotations between people. They introduce contact signatures between people based on coarse body parts. This is similar to how we collect the DSC dataset. Contemporaneous with our work, Fieraru et al. [42] extend this to self-contact with a 2-stage approach. They train a network to predict “self-contact signatures”, which are used for optimization-based 3D pose estimation. In contrast, TOUCH is trained end-to-end to regress body pose with contact information. Recently, Shimada et al. [181] collect hand-face motion and interaction dataset with involves self-contact and a reconstruction method.

World contact. Multiple methods use the 3D scene to help estimate the human pose. Physical constraints can come from the ground plane [208, 232], an object [59, 91, 95, 192, 191, 26, 16, 217, 218], or contextual scene information [54, 223, 200, 72]. Li et al. [116] use a DNN to detect 2D contact points between objects and selected body joints. Narasimhaswamy et al. [137] categorize hand contacts into self, person-person, and object contacts and aim to detect them from in-the-wild images. Their dataset does not provide reference 3D poses or shape. Only recently, after our work on self-contact, Shimada et al. address the problem of 3D face and hand reconstruction for poses with self-contact [181] and Zin et al. contact between two persons [227].

All the above works make a similar observation: human pose estimation is not a stand-alone task; considering additional physical contact constraints improves the results. We go beyond prior work by addressing self-contact and showing how training with self-contact data improves pose estimation overall.

3D body datasets. While there are many datasets of 3D human scans, most of these have people standing in an “A” or “T” pose to explicitly minimize self-contact [163]. Even when the body is scanned in varied poses, these poses are designed to avoid self-contact [5, 21, 22, 152]. For example, the FAUST dataset has a few examples of self-contact and the authors identify these as the major cause of error for scan processing methods [20]. Recently, the AMASS [130] dataset unifies 15 different optical marker-based motion capture (MoCap) datasets within a common 3D body parameterization, offering around 170k meshes with SMPL-H [165] topology. Since MoCap markers are sparse and often do not cover the hands, such datasets typically do not explicitly capture self-contact. As illustrated in Table 2.1, none of these datasets explicitly addresses self-contact.

Name	Meshes	Meshes with self-contact
3DCP Scan (ours)	190	188
3D BodyTex [2]	400	3
SCAPE [5]	70	0
Hasler et al. [56]	520	0
FAUST [20]	100/ 400	20/ 140

Table 2.1: Existing 3D human mesh datasets with the number of poses and the number of contact poses identified by visual inspection. 3DCP Scan is the scan subset of 3DCP (see Section 2.4). FAUST (train/test) includes scans with self-contact, i.e. 20 in the training and 140 in the test set. However, in FAUST the variety is low as each subject is scanned in the same 10/20 poses, whereas in 3DCP Scan each subject does different poses.

Pose mimicking. Our Mimic-The-Pose dataset uses the idea that people can replicate a pose that they are shown. Several previous works have explored this idea in different contexts. Taylor et al. [194] crowd-source images of people in the same pose by imitation. While they do not know the true 3D pose, they are able to train a network to match images of people in similar poses. Marinoiu et al. [132] motion capture subjects reenacting a 3D pose from a 2D image. They found that subjects replicated 3D poses with a mean joint error of around 100mm. This is on par with existing 3D pose regression methods, pointing to people’s ability to approximately recreate viewed poses. Fieraru et al. [42] ask subjects to reproduce contact from an image in a lab setting. They manually annotate the contact, whereas our MTP task is done in people’s homes and SMPLify-XMC is used to automatically optimize the pose and contact.

2.3 Self-Contact

An intuitive definition of contact between two meshes, e.g. a human and an object, is based on intersecting triangles. Self-contact, however, must be formulated to exclude common, but not functional, triangle intersections, e.g. at the crotch or armpits. We can describe self-contact at different levels of granularity: The simplest level is a *binary self-contact* class label, encoding whether a person is touching themselves or not. A more informative description of self-contact can be provided in 3D by indicating which body parts or regions of the body are touching. We call these labels *discrete self-contact*. To consider even more detail, we use Euclidean and geodesic distances on the mesh surface to specify *vertex-based self-contact*. Intuitively, vertices are in self-contact if they are close in Euclidean distance (near zero) but distant in geodesic distance, i.e. far away on the body surface.

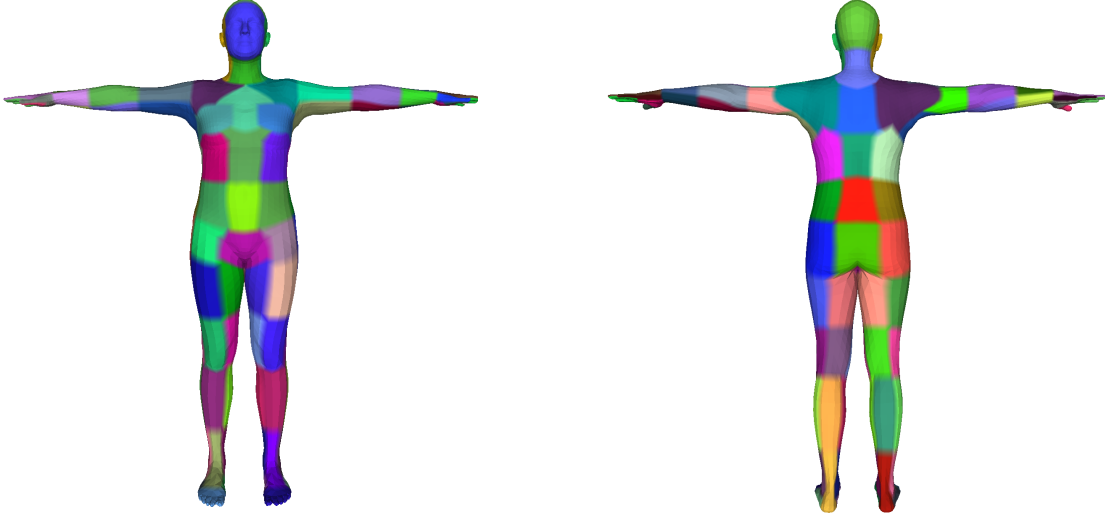


Figure 2.2: To compute self-contact maps, we group vertices into distinct regions, shown here with different colors. This is useful for searching our scan datasets for poses with specific types of contact.

2.3.1 Discrete Self-Contact

To cluster self-contact into distinct types, we define self-contact maps $\mathcal{S}^D \in \{0, 1\}^{R \times R}$; see [42] for a similar definition. To this end, we first segment the faces and vertices of a mesh into R distinct, non-overlapping regions and indicate if two regions are in contact or not via a binary label, i.e.

$$\mathcal{S}_{ij}^D = \begin{cases} 1, & \text{if } r_i \text{ is in contact with } r_j \\ 0, & \text{otherwise.} \end{cases} \quad (2.1)$$

We use fine maps to cluster self-contact meshes from AMASS (see Fig. 2.2) and rough maps (see Fig. 2.10) for human annotation.

2.3.2 Vertex-based Self-Contact

To get a more fine-grained definition of self-contact, we use Euclidean and geodesic distances between vertices.

Definition 2.3.1. Given a mesh M with vertices V , we define two vertices $v, u \in V$ to be in *self-contact*, if (i) $\|v - u\| < t_{eucl}$, and (ii) $geo(v, u) > t_{geo}$, where t_{eucl} and t_{geo} are predefined thresholds and $geo(v, u)$ denotes the geodesic distance between v and u . We use shape-independent geodesic distances precomputed on the neutral, mean-shaped SMPL and SMPL-X models.

Following this definition, we denote the set of vertex pairs in self-contact as $V'_C := \{(v, u) | v, u \in V \text{ and } v, u \text{ satisfy Definition 2.3.1}\}$. We further define V_C as a set of unique vertices in contact. We can obtain V_C from V'_C via an operator $\mathcal{U}(\cdot)$ such that $\mathcal{U}(V'_C) = V_C = \{v_1, v_2, \dots, v_n\}$, where for all $v \in V_C$ exists a vertex $u \in V_C$ such that $(v, u) \in V'_C$. We also define an operator $f_g(\cdot)$ that takes vertex v as input and returns the Euclidean distance to the nearest vertex u that is far enough in the geodesic sense. Formally, $f_g(v) := \min_{u \in V_G(v)} \|v - u\|$, where $V_G(v) := \{u | \text{geo}(v, u) > t_{\text{geo}}\}$. M is a *self-contact mesh* when $|V_C| > 0$.

2.3.3 Mesh Surface Points

To detect self-contact, we need to be able to quickly compute the distance between two points on the body surface. Vertex-to-vertex distance is a poor approximation of this due to the varying density of vertices across the body. Consequently, we introduce HD SMPL-X and HD SMPL to efficiently approximate surface-to-surface distance. For this, we uniformly, and densely, sample mesh surface points $P \in \mathbb{R}^{N_P \times 3}$ with $N_P = 20,000$ on the body. A sparse linear regressor $\mathcal{P} \in \mathbb{R}^{N_P \times N_V}$ regresses P from the mesh vertices V , $P = \mathcal{P}V$. The geodesic distance between two mesh surface points $x, y \in P$ is approximated via $\text{geo}_{HD}(x, y; V) = \text{geo}(\arg \min_{v \in V} \|v - x\|, \arg \min_{u \in V} \|u - y\|)$.

In practice, we use mesh surface points only when contact is present by following a three-step procedure as illustrated in Fig. 2.3. First, we use Definition 2.3.1 to detect vertices in contact, V_C . Then we select all points in P lying on faces that contain vertices in V_C , denoted as P_C . Last, for $x \in P_C$ we find the closest mesh surface point $\min_{y \in P_C} \|x - y\|$ such that $\text{geo}_{HD}(x, y) > t_{\text{geo}}$. With $HD(X) : X \subset V \rightarrow P_C \subset P$ we denote the function that maps from a set of mesh vertices to a set of mesh surface points. As the number of points, P , increases, the point-to-point distance approximates the surface-to-surface distance.

2.4 Self-Contact Datasets

Our goal is to create datasets of in-the-wild images paired with 3D human meshes as pseudo-ground truth. Unlike traditional pipelines that collect images first and then annotate them with pose and shape parameters [83, 207], we take the opposite approach. We first curate meshes with self-contact and then pair them with images through a novel pose mimicking and fitting procedure. We use SMPL-X to create the 3DCP and MTP dataset to better fit contacts between hands and bodies. However, to fine-tune SPIN [101], we convert MTP data to SMPL topology, and use SMPLify-DC when optimizing with discrete contact.

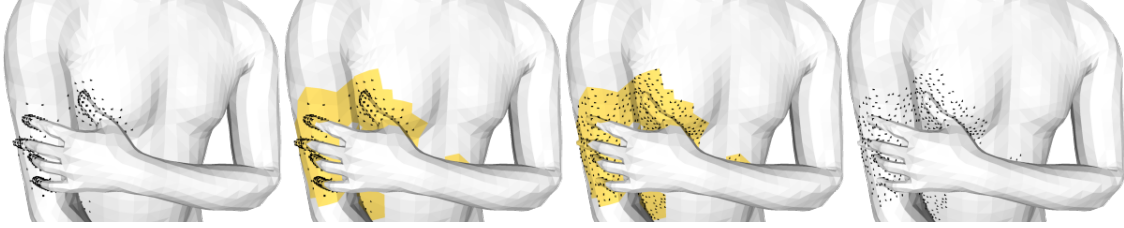


Figure 2.3: Visualization of the function $HD(X)$, that maps from mesh vertices to mesh surface points. The first image shows a SMPL-X mesh with vertices in contact highlighted. Second, in yellow, all faces containing a vertex in contact are selected. In the third image, all points lying on a face containing a vertex in contact are selected from P , denoted as P_C . P is a fixed set of mesh surface points that are regressed from mesh vertices. Note that in the first and second image, the finger vertices are denser than the arm and chest vertices, in contrast to the more uniform density in the third and fourth image.

2.4.1 3D Contact Pose (3DCP) Meshes

We create 3D human meshes with self-contact in two ways: with 3D scans and with motion capture data.

3DCP Scan

We scan 6 subjects (3 males, 3 females) in self-contact poses. Raw scans have varying topology. To bring a corpus of scans to a common topology is the process of “registration”. We register the SMPL-X mesh topology to the raw scans. These registrations are obtained using Co-Registration [67], which iteratively deforms the SMPL-X template mesh M with vertices V to minimize the *point-to-plane* distance between the scan points $S \in \mathbb{R}^{N_S \times 3}$, where N_S is the number of scan points. However, registering poses with self-contact is challenging. When body parts are in close proximity, the standard process can result in interpenetration. To address this, we add a self-contact-preserving energy term to the objective function. If two vertices v and u are in contact according to Definition 2.3.1, we minimize the *point-to-plane* distance between triangles including v and the triangular planes including u . This term ensures that body parts that are in contact remain in contact.

Most traditional registration methods ignore interpenetration and self-contact. Registering our self-contact scans without modeling self-contact would result in self-penetration, particularly where the extremities contact the body. We address this by modifying the registrations objective function to encourage self-contact without penetration.

Specifically, the fitting objective includes a data term E_S evaluating the goodness of fit of the vertices v on the template V to n randomly sampled points, x on the surface of the



Figure 2.4: A representative sample from the registrations. A total of 3 male and 3 female subjects were scanned in a diversity of poses that involve self-contact. The 3D scans are registered to a common mesh topology by fitting the SMPL-X template mesh to them using a self-contact preserving energy term that penalizes body part interpenetration.

scan S

$$E_S(S; V) = \frac{1}{n} \int_{x \in S} \rho(\|x - v\|) \quad (2.2)$$

where ρ is the Geman-McClure robust penalty function.

Additionally, we introduce a self-contact preserving energy term E_C to the objective function. The term E_C helps to minimize and preserve the *point-to-plane* distance between body parts that are in contact. E_C considers the set of contacting vertex pairs M_C defined by Definition 3.1 in the main corpus of this thesis. For each tuple (v_i, v_j) in M_C , we minimize the *point-to-plane* distance between triangles including v_i and the triangular planes including v_j . The contact energy term ensures that body parts that are in contact remain in contact.

The objective function is minimized in two steps: first a model fitting step, where it is minimized with respect to the SMPL-X model pose parameters $\theta \in \mathbb{R}^{55 \times 3}$ and body shape parameters $\beta \in \mathbb{R}^{25}$. Following model fitting, a model-free optimization step minimizes point-to-plane distance between the model vertices v and the scan. A sample of the registrations is shown in Figure 2.4.

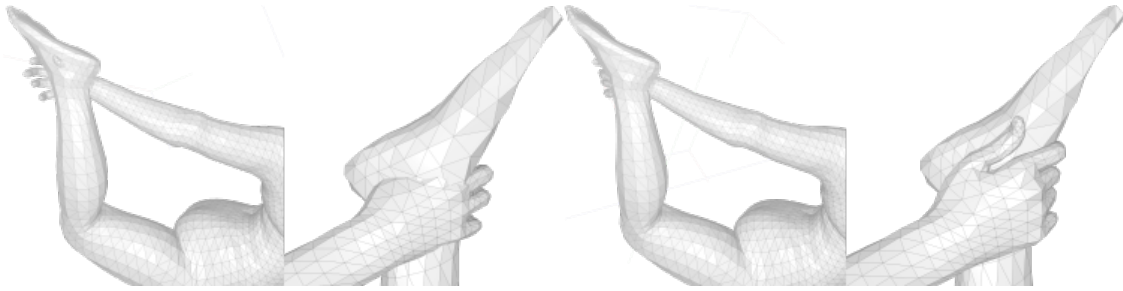


Figure 2.5: Self-contact optimization. Column 1 and 2: a pose selected from AMASS with near self-contact (between the fingertips and the foot) and interpenetration (thumb and foot). Column 3 and 4: after self-contact optimization, all fingers are in contact with the foot and interpenetration is reduced.

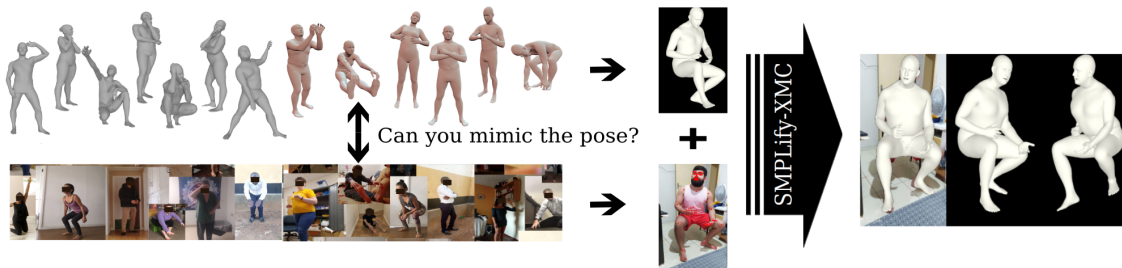


Figure 2.6: Mimic-The-Pose (MTP) dataset. MTP is built via: (1) collecting many 3D meshes that exhibit self-contact. In grey, new 3D scans in self-contact poses, in brown self-contact poses optimized from AMASS mocap data. (2) collecting images in the wild, by asking workers on AMT to mimic poses and contacts. (3) the presented meshes are refined via SMPLify-XMC to match the image features.

3DCP Mocap

While mocap datasets are usually not explicitly designed to capture self-contact, it does occur during motion capture. We therefore search the AMASS dataset for poses that satisfy our self-contact definition. We find that some of the selected meshes from AMASS contain small amounts of self-penetration or near contact. Thus, we perform *self-contact optimization* to fix this while encouraging contact, as shown in Fig. 2.5; see Appendix A.1.1 for details.

2.4.2 Mimic-The-Pose (MTP) Data

Data Collection via AMT

To collect in-the-wild images with near ground-truth 3D human meshes, we propose a novel two-step process (see Fig. 2.6). First, using meshes from 3DCP as examples, workers on AMT are asked to mimic the pose as accurately as possible while someone



Figure 2.7: Presentation format and examples of mimicked poses from the MTP data set. On the left side, the presented pose with contact highlighted in blue. Humans mimicking the poses on the right.

takes their photo showing the full body (the *mimicked pose*). Mimicking poses may be challenging for people when only a single image of the pose is presented [132]. Thus, we render each 3DCP mesh from three different views with the contact regions highlighted (the *presented pose*). We allot 3 hours time for ten poses. Participants also provide their height and weight. All participants gave informed consent for the capture and the use of their imagery. Figure 2.7 shows an example of the rendered 3DCP meshes with images of people mimicking the pose. Please see Appendix A.1.2 for details.

SMPLify-XMC

The second step applies a novel optimization method to estimate the pose in the image, given a strong prior from the presented pose. The presented pose $\hat{\theta}$, shape $\hat{\beta}$, and gender is not mimicked perfectly. To obtain pseudo-ground truth pose and shape, we adapt SMPLify-X [144], a multi-stage optimization method, that fits SMPL-X pose θ , shape

β , and expression ψ to image features starting from the mean pose and shape. We make use of the presented mesh in three ways: first, its used to initialize the optimization and solve for global orientation and camera parameters; second, the presented pose serves as a pose prior; and third its contact is used to keep relevant body parts close to each other. We refer to this new optimization method as SMPLify-XMC.

In the first stage, we optimize body shape β and camera Π (focal length, rotation and translation), and body global orientation ϕ , using the ground-truth height in meters, H , and weight in kg, W . The objective function of the first stage is given as

$$L_{\text{SMPLify-XMC}} = \lambda_J L_J + \lambda_\phi L_\phi + \lambda_M L_M.$$

$L_M = e^{100|\hat{H}-H|} + e^{|\hat{W}-W|}$ is the measurements loss, where \hat{H} and \hat{W} are height and weight of current estimate of mesh M . We compute height and weight from mesh in a zero pose (T-pose). For height, we compute the distance between the top of the head and the mean point between left and right heel. For weight, we compute the mesh volume and multiply it by 985 kg/m^3 , which approximates human body density. L_ϕ is a loss on the body global orientation and L_J denotes the joint re-projection error as specified in SMPLify-X [144].

In the second and third stage, we fix the body global orientation and jointly optimize θ (body and hand pose), β , and Π to minimize

$$L_{\text{SMPLify-XMC}} = \lambda_J L_J + \lambda_{m_h} L_{m_h} + \lambda_{\tilde{\theta}} L_{\tilde{\theta}} + \lambda_M L_M + \lambda_{\tilde{C}} L_{\tilde{C}} + \lambda_S L_S. \quad (2.3)$$

We use the standard SMPLify-X priors for the left and right hand L_{m_h} , where $h \in \{l, r\}$. While the pose prior in SMPLify-X penalizes deviation from the mean pose, here, $L_{\tilde{\theta}} = \|\tilde{\theta} - \tilde{\theta}_{\text{est}}\|_2$ is an L2-Loss that penalizes deviation from the presented pose. The term $L_{\tilde{C}}$ acts on \tilde{V}_C , the vertices in self-contact on the presented mesh. To ensure the desired self-contact, one could seek to minimize the distances between vertices in contact, e.g. $\|v - u\|$ for $(v, u) \in \tilde{V}'_C$. However, with this approach, we observe slight mesh distortions, when presented and mimicked contact are different. Instead, we use a term that encourages every vertex in contact in the presented pose, i.e. vertices in \tilde{V}_C , to be close to a vertex in the current estimate. This loss can formally be described as

$$L_{\tilde{C}} = \frac{1}{|\tilde{V}_C|} \sum_{v \in \tilde{V}_C} \tanh(f_g(v)). \quad (2.4)$$

The third stage activates a new loss, L_S , for fine-grained self-contact optimization, which resolves interpenetration while encouraging contact:

$$L_S = \lambda_C L_C + \lambda_P L_P + \lambda_A L_A.$$

Vertices in contact are pulled together via a contact term L_C , vertices inside the mesh are pushed to the surface via a pushing term L_P , and L_A aligns the surface normals of two

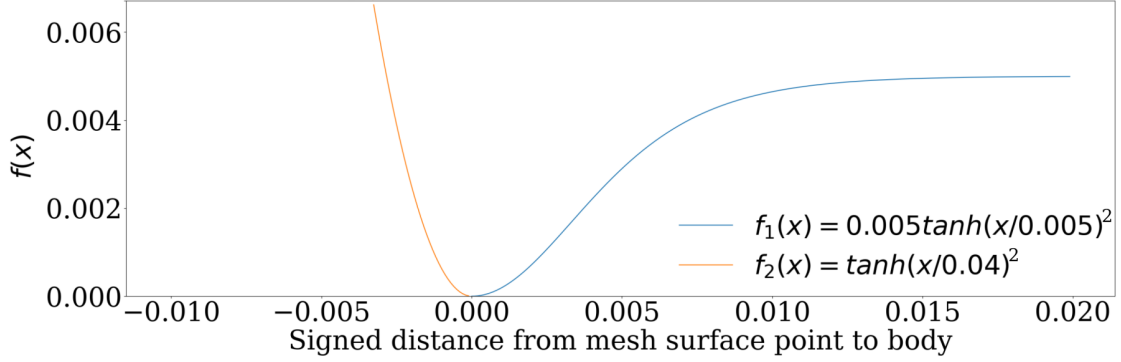


Figure 2.8: Functions to regulate the self-contact pushing and pulling term in SMPLify-XMC. f_1 is used in L_C with $\delta_1 = \delta_2 = 0.005$ and f_2 is used in L_P with $\delta_3 = 1.0$ and $\delta_4 = 0.04$. The parameters ensure that inside vertices are pushed out quickly, while vertices in contact are pulled together as long as they are close enough.

vertices in contact.

To compute these terms, we must first find which vertices are inside, $V_I \subset V$, or in contact, $V_C \subset V$. V_C is computed following Definition 2.3.1 with $t_{geo} = 30\text{cm}$ and $t_{eucl} = 2\text{cm}$. The set of inside vertices V_I is detected by generalized winding numbers [76]. SMPL-X is not a closed mesh and thus complicating the test for penetration. Consequently, we close it by adding a vertex at the back of the mouth. In addition, neighboring parts of SMPL and SMPL-X often intersect, e.g. torso and upper arms. We identify such common self-intersections and filter them out from V_I (see Appendix A.1.2 for details). To capture fine-grained contact, we map the union of inside and contact vertices onto the HD SMPL-X surface, i.e. $S = HD(V_I \cup V_C)$, which is further segmented into an inside S_I and outside S_O subsets by testing for intersections. The self-contact objectives are defined as

$$L_C = \sum_{x \in S_O} \delta_1 \tanh\left(\frac{f_g(x)}{\delta_2}\right)^2,$$

$$L_P = \sum_{x \in S_I} \delta_3 \tanh\left(\frac{f_g(x)}{\delta_4}\right)^2,$$

$$L_A = \sum_{(x,y) \in P_C} 1 + \langle N(x), N(y) \rangle.$$

f_g denotes the function that finds the closest point $y \in P_C$ for x , where P_C is the subset of vertices in contact in P . N denotes the normal of x . We use $\delta_1 = \delta_2 = 0.005$, $\delta_3 = 1.0$, and $\delta_4 = 0.04$. Fig. 2.9 shows examples of our pseudo ground-truth meshes. In Fig. 2.8 we visualize the pushing and pulling terms used in the SMPLify-XMC objective. We use 6 PCA components for the hand pose space [165] and initialize the fitting with a mean

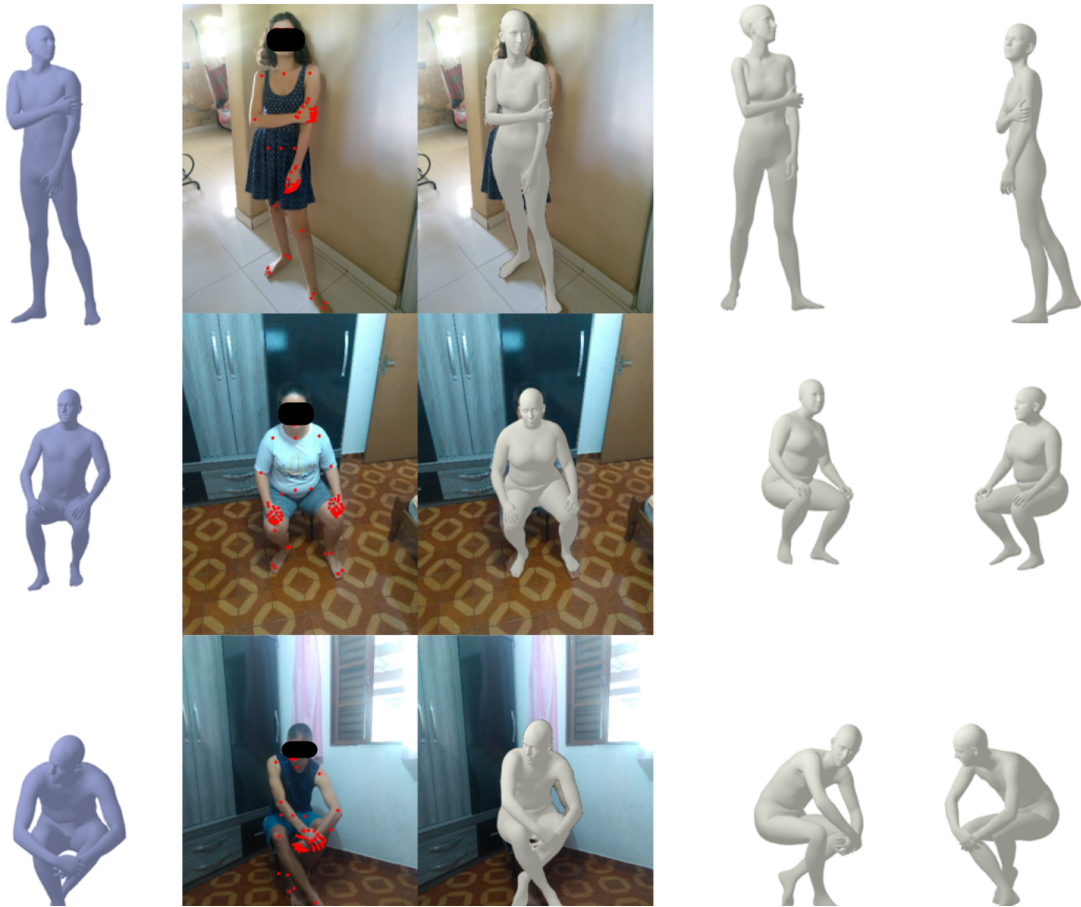


Figure 2.9: MTP results. Meshes presented to AMT workers (blue) and the images they submitted with OpenPose keypoints overlaid. In grey, the pseudo ground-truth meshes computed by SMPLify-XMC.

hand pose. In contrast to SMPLify-X we do not ignore hip joints and double the joint weights for knees and elbows. Before optimization, we resize images and keypoints to a maximum height or width of 500 pixel. Similar to SMPLify-X we use the PyTorch implementation of fast L-BFGS with strong Wolf line search as the optimizer [121]. We do not use the VPoser pose prior for SMPLify-XMC because we have a strong prior from the presented pose.

2.4.3 Discrete Self-Contact (DSC) Data

Images in the wild collected for human pose estimation normally come with 2D keypoint annotations, body segmentation, and/or bounding boxes. Such annotations lack 3D information. Discrete self-contact annotation, however, provides useful 3D information about pose. We use $R = 24$ regions and label their pairwise contact for three publicly available

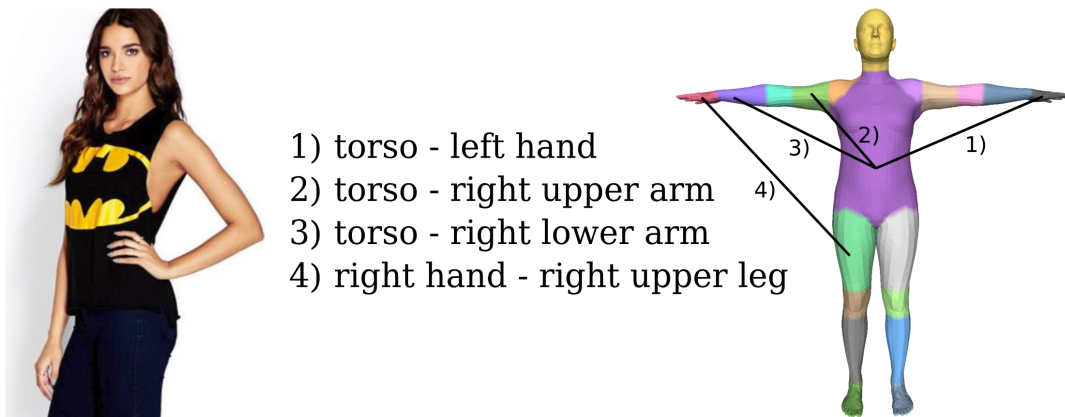


Figure 2.10: DSC dataset. Image with discrete contact annotation on the left. On the right: DSC signature with $R = 24$ regions.

datasets, namely Leeds Sports Pose (LSP), Leeds Sports Pose Extended (LSPet), and DeepFashion (DF). An example annotation is visualized in Fig. 2.10. Of course, such labels are noisy because it can be difficult to accurately determine contact from an image. See Appendix A.1.3 for details.

2.4.4 Summary of the Collected Data

Our 3DCP human mesh dataset consists of 190 meshes containing self-contact from 6 subjects, 159 SMPL-X bodies fit to commercial scans from AGORA [143], and 1304 self-contact optimized meshes from mocap data. From these 1653 poses, we collect 3731 mimicked pose images from 148 unique subjects (52 female; 96 male) for MTP and fit pseudo ground-truth SMPL-X parameters. MTP is diverse in body shapes and ethnicities. Our DSC dataset provides annotations for 30K images.

2.5 TUCH

Finally, we train a regression network that has the same design as SPIN [101]. At each training iteration, the current regressor estimates the pose, shape, and camera parameters of the SMPL model for an input image. Using ground-truth 2D keypoints, an optimizer refines the estimated pose and shape, which are used, in turn, to supervise the regressor. We follow this regression-optimization scheme for DSC data, where we have no 3D ground truth. To this end, we adapt the in-the-loop SMPLify routine to account for discrete self-contact labels, which we term SMPLify-DC. For MTP images, we use the pseudo ground truth from SMPLify-XMC as direct supervision with no optimization involved. We explain the losses of each routine below.

Regressor. Similar to SPIN, the regressor of TUCH predicts pose, shape, and camera,

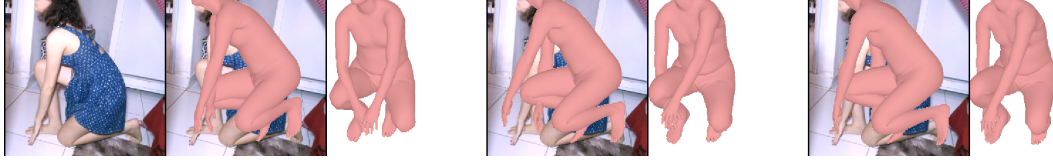


Figure 2.11: Initial wrong contact (left) from the regressor is fixed by SMPLify-DC after 5 (middle) and 10 (right) iterations.

with the loss function:

$$L_{\text{TUCH}} = \lambda_J L_J + \lambda_\theta L_\theta + \lambda_\beta L_\beta + \lambda_C L_C + \lambda_P L_P. \quad (2.5)$$

L_J denotes the joint re-projection loss. L_P and L_C are self-contact loss terms used in L_S in SMPLify-XMC, where L_P penalizes mesh intersections and L_C encourages contact. Further, L_θ and L_β are L2-Losses that penalize deviation from the pseudo ground-truth pose and shape.

Optimizer. We develop SMPLify-DC to fit pose θ_{opt} , shape β_{opt} , and camera Π_{opt} to DSC data, taking ground-truth keypoints and contact as constraints. Typically, in human mesh optimization methods the camera is fit first, then the model parameters follow. However, we find that this can distort body shape when encouraging contact. Therefore, we optimize shape and camera translation first, using the same camera fitting loss as in [101]. After that, body pose and global orientation are optimized under the objective

$$L_{\text{SMPLify-DC}} = \lambda_J L_J + \lambda_\theta L_\theta + \lambda_C L_C + \lambda_P L_P + \lambda_{\mathcal{S}^D} L_{\mathcal{S}^D}. \quad (2.6)$$

The discrete contact loss, $L_{\mathcal{S}^D}$, penalizes the minimum distance between regions in contact. Formally, given a contact signature \mathcal{S}^D where $\mathcal{S}_{ij}^D = \mathcal{S}_{ji}^D = 1$ if two regions r_i and r_j are annotated to be in contact, we define

$$L_{\mathcal{S}^D} = \sum_{i=1}^R \sum_{j=i+1}^R \mathcal{S}_{ij}^D \min_{v \in r_i, u \in r_j} \|v - u\|^2.$$

Given the optimized pose θ_{opt} , shape β_{opt} , and camera Π_{opt} , we compute the re-projection error and the minimum distance between the regions in contact. When the re-projection error improves, and more regions with contact annotations are closer than before, we keep the optimized pose as the current best fit. When no ground truth is available, the current best fits are used to train the regressor.

We make three observations: (1) The optimizer is often able to fix incorrect poses estimated by the regressor because it considers the ground-truth keypoints and contact (see Fig. 2.11). (2) Discrete contact labels bring overall improvement by helping resolve depth ambiguity (see Fig. 2.12). (3) Each batch consists of 50% DSC and 50% MTP data. The direct supervision of MTP data improves the regressor, which benefits SMPLify-DC



Figure 2.12: Impact of discrete self-contact labels in human pose estimation. Body parts labeled in contact are shown in the same color. First row shows an initial SPIN estimate, second row the SMPLify fit, third row the SMPLify-DC fit after 20 iterations.

by providing better initial estimates.

Implementation details. We initialize our regression network with SPIN weights [101]. For SMPLify-DC, we run 10 iterations per stage and do not use the HD operator to speed up the optimization process. For the 2D re-projection loss, we use ground-truth keypoints when available and, for MTP and Deep Fashion images, OpenPose detections weighted by confidence. From DSC data we only use images where the full body is visible and ignore annotated region pairs that are connected in the DSC segmentation (see Appendix A.2).

2.6 Evaluation

We evaluate TUCH on the following three datasets: **3DPW** [207], **MPI-INF-3DHP** [133], and **3DCP Scan**. This last dataset consists of RGB images taken during the 3DCP Scan scanning process. While TUCH has never seen these images or subjects, the contact poses were mimicked in creation of MTP, which is used in training.

We use standard evaluation metrics for 3D pose, namely Mean Per-Joint Position Error (MPJPE) and the Procrustes-aligned version (PA-MPJPE), and Mean Vertex-to-Vertex Error (MV2VE) for shape and contact. Tables 2.2 and 2.3 summarize the results of TUCH on 3DPW and 3DCP Scan. Interestingly, TUCH is more accurate than SPIN on 3DPW. See Fig. 2.14 and Fig. 2.15 for qualitative results of our model.

We further evaluate our results with respect to contact. To this end, we divide the

	MPJPE		PA-MPJPE	
	3DPW	MI	3DPW	MI
SPIN [101]	96.9	105.2	59.2	67.5
EFT [83]	-	-	54.2	68.0
TUCH	84.9	101.2	55.5	68.6

Table 2.2: Evaluation on 3DPW and MPI-INF-3DHP (MI). Bold numbers indicate the best result; units are *mm*. We report the EFT result denoted in their publication when 3DPW was not part of the training data. Please note that SPIN is trained on MI, but we do not include MI in the fine-tuning set. MI contains mostly indoor lab sequences (100% train, 75% test), while DSC and MTP contain only in-the-wild images. This domain gap likely explains the decreased performance in PA-MPJPE.

	MPJPE	PA-MPJPE	MV2VE
SPIN [101]	79.7	50.6	95.7
EFT [83]	71.4	48.3	83.9
TUCH	69.5	42.5	81.5

Table 2.3: Evaluation on 3DCP Scan. Numbers are in *mm*. Note that in contrast to TUCH, this version of SPIN did not see poses in the MTP dataset during training. Please see Table 2.5 and the corresponding text for an ablation study.

3DPW test set into subsets, namely for $t_{geo} = 50\text{cm}$: *self-contact* ($t_{eucl} < 1\text{cm}$), *no self-contact* ($t_{eucl} > 5\text{cm}$), and *unclear* ($1\text{cm} < t_{eucl} < 5\text{cm}$). For 3DPW we obtain 8752 *self-contact*, 16752 *no self-contact*, and 9491 *unclear* poses. Table 2.4 shows a clear improvement on poses with contact and unclear poses compared to a smaller improvement on poses without contact.

To further understand the improvement of TUCH over SPIN, we break down the improved MPJPE in 3DPW *self-contact* into the pairwise body-part contact labels defined in the DSC dataset. Specifically, for each contact pair, we search all poses in 3DPW *self-contact* that have this particular self-contact. We find a clear improvement for a large number of contacts between two body parts, frequently between arms and torso, or

	MPJPE				PA-MPJPE			
	contact	no contact	unclear	total	contact	no contact	unclear	total
SPIN	100.2	95.5	96.7	96.9	59.1	61.7	55.7	59.2
TUCH	85.1	86.6	81.9	84.9	54.1	58.6	51.2	55.5

Table 2.4: Evaluation of TUCH for contact classes in 3DPW. Numbers are in *mm*. See text.

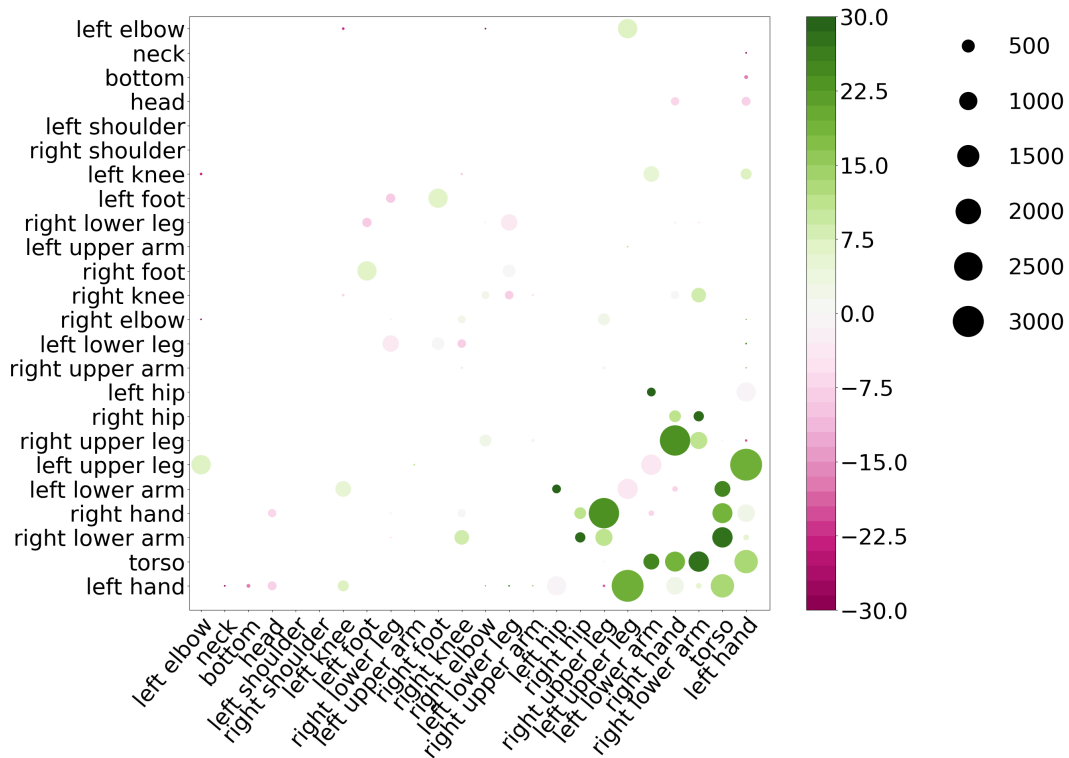


Figure 2.13: Average MPJPE difference (SPIN - TUCH), evaluated on the *self-contact* subset of 3DPW. The axes show labels for the DSC regions. Green indicates that TUCH has a lower error than SPIN on average across all poses with the corresponding regions in contact. The circle size represents the number of images per region. Regions with small circle sizes are less common.

e.g. left hand and right elbow, which is common in arms-crossed poses (see Fig. 2.13).

TUCH incorporates self-contact in various ways: annotations of training data, in-the-loop fitting, and in the regression loss. We evaluate the impact of each in Table 2.5. S+ is SPIN but it sees MTP+DSC images in fine-tuning and runs standard in-the-loop SMPLify with no contact information. S++ is S+ but uses pseudo ground truth computed with SMPLify-XMC on MTP images; thus self-contact is used to generate the data but nowhere else. S+ vs. SPIN suggests that, while poses in 3DCP Scan appear in MTP, just seeing similar poses for training and testing does not yield improvement. S+ vs. TUCH is a fair comparison as both see the same images during training. The improved results of TUCH confirm the benefit of using self-contact.

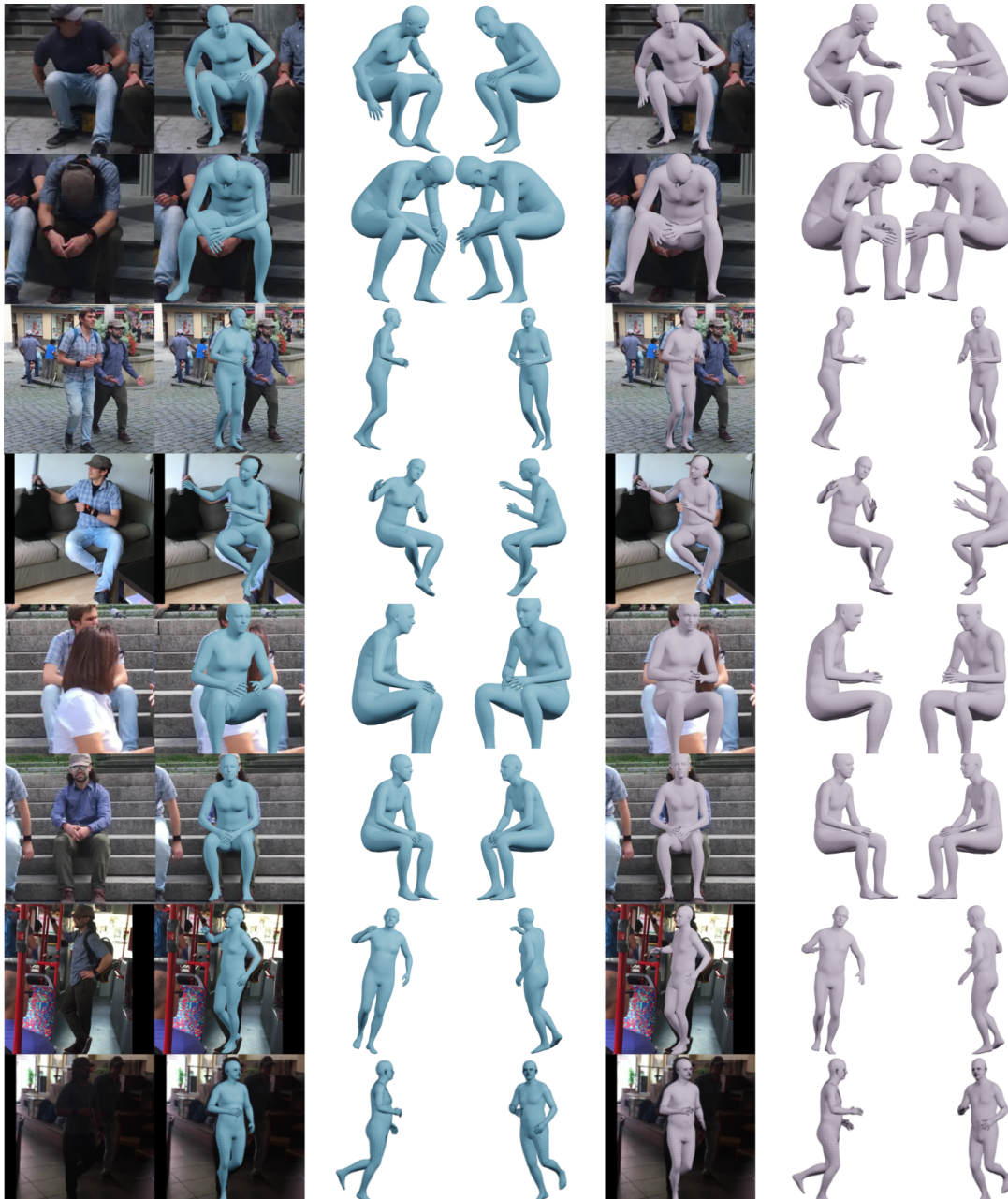


Figure 2.14: Qualitative results on the self-contact subset of 3DPW. We find all images with an improvement on MPJPE and PA-MPJPE ≥ 10 mm. From this subset, we select interesting poses. Left column, RGB image for reference. In blue, TOUCH result and in violet, the SPIN result.

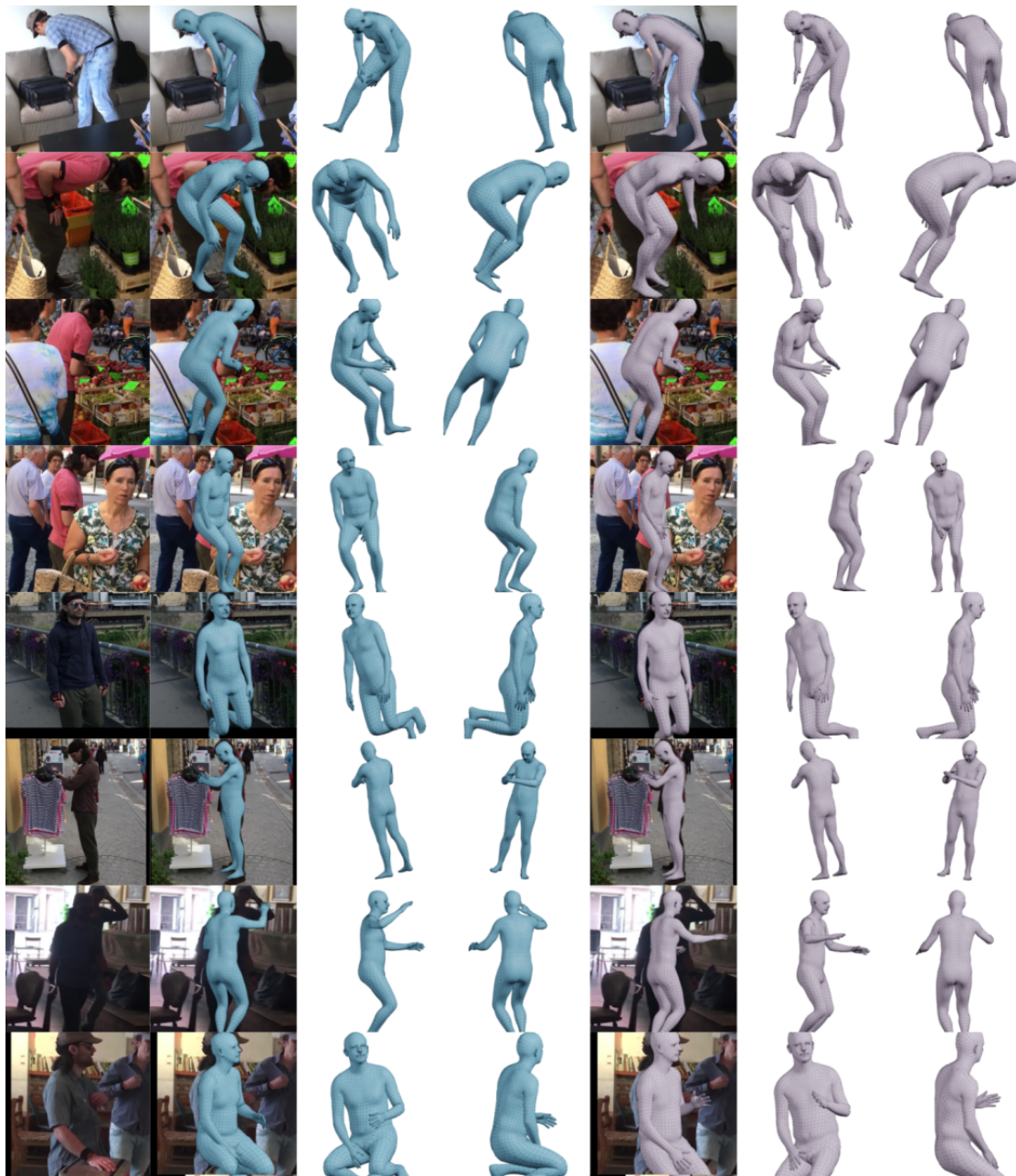


Figure 2.15: Qualitative results on the self-contact subset of 3DPW. We find all images where SPIN is better than TUCH by at least 10 mm for MPJPE and PA-MPJPE. From this subset, we select interesting poses. Left column, RGB image for reference. In blue, TUCH result and in violet, the SPIN result.

	SPIN	S+	S++	TUCH
3DPW	96.9/ 59.2	96.1/ 61.4	85.0/ 56.3	84.9/ 55.5
3DCP Scan	82.2/ 52.1	86.9/ 52.3	74.8/ 45.7	75.2/ 45.4
MI	105.2/ 67.5	105.8/ 69.4	103.1/ 69.0	101.2/ 68.6

Table 2.5: MPJPE/PA-MPJPE (mm) to examine the impact of data and algorithm on 3DPW, 3DCP Scan, and MPI-INF-3DHP (MI).

2.7 Conclusion

In this chapter, we address the problem of HPS estimation when self-contact is present. Self-contact is a natural, common occurrence in everyday life, but state of the art methods fail to estimate it. One reason for this is that no datasets pairing images in the wild and 3D reference poses exist. To address this problem we introduce a new way of collecting data: we ask humans to mimic presented 3D poses. Then we use our new SMPLify-XMC method to fit pseudo ground-truth 3D meshes to the mimicked images, using the presented pose and self-contact to constrain the optimization. We use the new MTP data along with discrete self-contact annotations to train TUCH; the first end-to-end HPS regressor that also handles poses with self-contact. TUCH uses MTP data as if it was ground truth, while the discrete, DSC, data is exploited during SPIN training via SMPLify-DC. Overall, incorporating contact improves accuracy on standard benchmarks like 3DPW, remarkably, not only for poses with self-contact, but also for poses without self-contact.

Chapter 3

Generative Proxemics: A Prior for 3D Social Interaction from Images

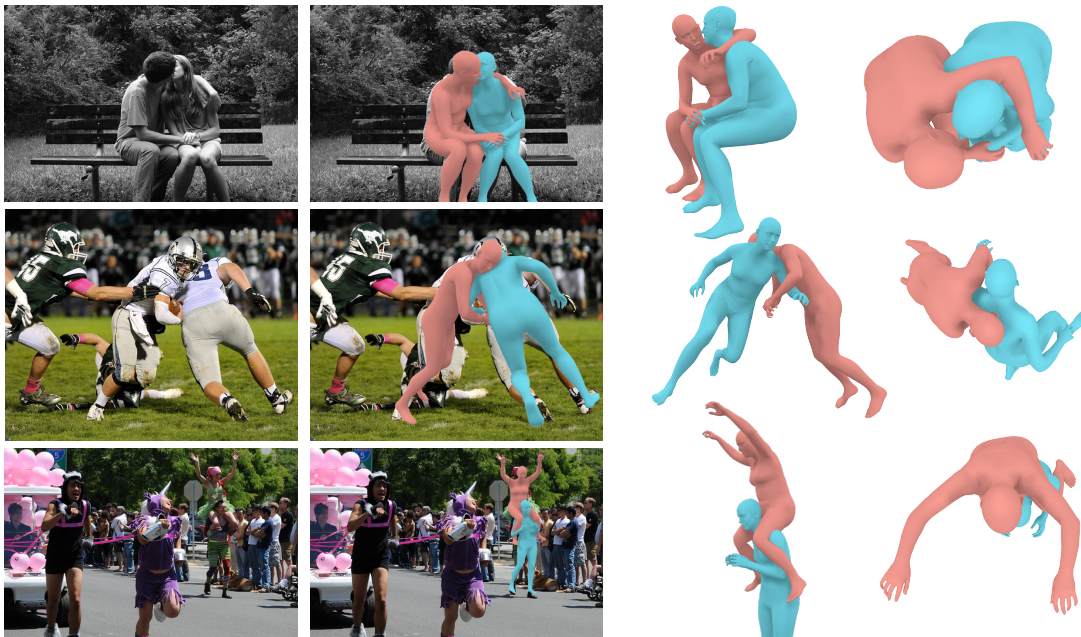


Figure 3.1: **Generative Proxemics.** We propose a diffusion model that learns a 3D generative model of two people in close social interaction. We show how the model can be used to generate samples or as a social prior in the downstream task of reconstructing two people in close proximity from images without any user annotation at test time. Shown here are input test images (left) and our predicted 3D bodies (right).

The previous chapter introduced the problem of human pose estimation for poses with self-contact. We introduced novel datasets and losses to solve this task. In particular, we introduced the concept of discrete self-contact annotations and an optimization method for fitting SMPL bodies to images by taking 2D keypoints and discrete self-contact labels into account. In this chapter, we will introduce discrete human-human contact annotations and use them during optimization in a similar fashion like discrete self-contact labels. We use the optimization method for two people in contact to create pseudo-ground truth fits of SMPL-X bodies to Flickr images and use this data to train a generative model that learns the joint distribution of people in close proximity. We also show how the learned model can be used as prior during optimization to estimate the pose and contact of pairs of people in an image.

3.1 Introduction

Humans are social creatures, and physical interaction plays a crucial role in our daily lives. From a simple handshake to a warm hug, physical touch and other non-verbal communication such as eye contact and body language convey a range of emotions and

meanings, shaping our social relationships. However, capturing and modeling the complexity of physical social interaction in three dimensions is a challenging task. It requires a deep understanding of the intricate interplay between body poses, shape, and proximity. These interactions are hard to model by hand and best learned from data. While 3D data of such social interaction is difficult to obtain at scale, images of people in social interaction are abundant.

In this chapter, we present the first approach that learns a generative model for 3D social proxemics, i.e. the study of interpersonal space in social interactions, from image collections. While diffusion models are widely used for image generation and 3D human motion modeling, here we use them to model the distribution over the 3D body pose and shape parameters of two people interacting. In order to train such a model, we first reconstruct people in close social interaction from images using available ground-truth contact maps [41] via an optimization-based approach. The recovered 3D bodies are used as training data to train our diffusion model. The resulting model is able to generate the realistic 3D social interaction of people depicted in photographs, such as people standing close together, playing sports, hugging, and more, as illustrated in Figure 3.1. Such models can be particularly useful for applications such as augmented reality, 3D content creation, and other scenarios where populating 3D scenes with realistic synthetic people is important.

We further demonstrate the effectiveness of the learned prior by applying it to the downstream task of reconstructing 3D social interactions from a single image. Unlike previous methods that rely on ground truth contact maps, our data-driven prior frees us from the need for explicit 3D training data, making it applicable to real-world scenarios. Our approach is able to obtain plausible and realistic interactions that capture subtle nuances from images, resulting in a significant improvement over the state-of-the-art human pose and shape estimation methods.

Specifically, we propose BUDDI: BUDDies DIffusion Model, a data-driven prior for 3D social proxemics. Unlike prior work on using diffusion to model human motion using 3D joint locations [196], BUDDI is trained to directly operate on the SMPL-X parameters, which represent body shape, pose, global orientation, and translation, through a transformer backbone. After training, BUDDI is able to generate unconditional samples of plausible bodies in social interaction from pure noise. We evaluate the unconditional generation through various qualitative experiments and user studies. The model can also be conditioned on the output of a human pose and shape regressor. In this conditional case, the model takes the noisy output and generates similar poses but with realistic social interaction.

We also introduce a novel optimization-based approach, which uses BUDDI as a data-driven prior while optimizing the 3D poses and shapes of two people in contact from images. We first take the output of BEV [190], a state-of-the-art model that regresses the 3D bodies of multiple people. We then optimize the BEV results to match image evidence with guidance from the diffusion model using a loss inspired by the SDS loss in the diffusion literature [153].

We validate our reconstruction of mesh and contact distance on FlickrCI3D Signatures [41] and show the value of training our generative model from 3D bodies recovered from images. While there has been recent work on generating people in synthetic environments with plausible human-to-object [217] or hand-to-object [213] interactions, our approach is the first that can generate two people with plausible social interaction. This opens a new avenue of research on digital human synthesis. Our data, code, and model are available for research.

3.2 Related Work

Generating 3D humans. There has been a lot of work on the subject of generating 3D humans, in many different contexts. Several methods automatically populate static 3D scenes with 3D humans [60, 236, 234]. More recent methods generate both body and hand poses to interact with 3D objects [191, 213, 195]. Other work generates human motions conditioned on different inputs such as audio [113, 202] or text [148, 149, 196]. Concurrent work proposes text-to-3D diffusion-based approaches to generate motion of two interacting humans [118, 179]. Both methods not predict the full body surface since the method focuses on synthesizing motion by generating either 3D joint locations or SMPL pose parameters, but not SMPL shape.

To model 3D human proxemics probabilistically, we employ diffusion models, which achieve impressive performance on image generation tasks [36, 68, 164, 168]. They have recently been adopted in 3D human motion generation scenarios: MDM [196] generates plausible motions conditioned on text input. PhysDiff [229] incorporates physical constraints in the diffusion process to generate physically plausible motions. EDGE [202] uses a transformer-based diffusion model for dance generation. Related work [25, 31, 125] has investigated different modalities for the conditioning, e.g., audio, text, action classes. EgoEgo [111] generates plausible full-body motions conditioned on the head motion. SceneDiffuser [73] focuses on the scene-conditioned setting. We also rely on techniques from the diffusion literature, but consider the unique setting where two people are in close interaction.

Multi-person 3D human mesh estimation. An extensive line of work focuses on reconstructing the 3D human pose and shape of a single person from images using optimization [19, 50, 107, 144, 161, 199, 220] or regression approaches [6, 52, 84, 87, 101, 136, 139, 222, 231, 233]. Capitalizing on these techniques, recent approaches focus explicitly on reconstructing multiple people jointly from a single image. Zafir et al. [232] propose an optimization solution, while Jiang et al. [78] and Sun et al. [189] rely on deep networks to regress the pose and shape for all people in the image. BEV [190] extends ROMP [189] to reason about the depth of people in a virtual birds-eye-view while taking age/height into account. We use BEV [190] as an initialization for our optimization method, but we demonstrate how we can meaningfully capture the close human-human interactions with the learned 3D social proxemics prior.

The above methods do not address contact between people. To do so, Fieraru et al. [41] introduce the first datasets with ground-truth labels for the body regions in contact between humans. Labels are collected using MoCap (CHI3D) or human annotators (FlickrCI3D Signatures). They propose an optimization approach that uses the ground-truth contact map to reconstruct people in close proximity. They also propose a 2D model that predicts the contact map from images. More recently, REMIPS [43] is a transformer-based method that regresses the 3D pose of multiple people. REMIPS is trained using the above datasets while taking into account contact and interpenetration. In this work, we take a very different approach by learning and exploiting a 3D generative proxemic prior. We use the ground-truth contact maps to generate pseudo-ground truth 3D human fits from which we learn the diffusion model; once this is learned, we show that it can be used as a prior to recover plausible bodies in close proximity from images without explicit knowledge of contact maps.

Data-driven priors in optimization. Optimization-based methods for 3D human pose and shape estimation, like SMPLify [19], are versatile and allow different data-driven prior terms to be incorporated in the objective function. Different methods have been used to learn pose priors including GMMs [19], VAEs [144], neural distance fields [199], and normalizing flows [230]. ProHMR [102] learns a pose prior conditioned on image pixels. HuMoR [161] incorporates a data-driven motion prior in the iterative optimization. POSA [60] learns a prior for human-scene interaction from PROX data [59] and uses it in their optimization. In contrast to these methods, we use a diffusion model to capture the joint distribution over SMPL-X parameters for two people interacting. To our knowledge, this problem has not previously been studied.

3.3 Human-Human Contact

Contact between two humans and their corresponding meshes M^a/M^b , can be annotated at different levels of granularity, similar to our considerations for self-contact; see Section 2.3. The simplest level is a binary class label, encoding whether two humans are touching or not touching; with an optional third class, “uncertain”. A more informative source of contact annotations can be provided directly in 3D. To this extent, Fieraru et al. [41] introduce the concept of discrete contact labels for interacting people.

3.3.1 Discrete Human-Human Contact

To annotate images with a 3D discrete human-human contact label, Fieraru et al. [41] divide the body into $R = 75$ regions and annotate the pairwise contact between both people. We refer to these human-human contact labels as the ground-truth “3D contact map”. Each region, r , roughly covers a similar surface of the body and is associated with SMPL-X faces F_r and, consequently, vertices V_r .

Specifically, in accordance with the definition of discrete self-contact in Chapter 2, discrete 3D human-human contact between M^a and M^b is also represented as a binary contact map $\mathcal{C}^D \in \{0, 1\}^{R \times R}$, where:

$$\mathcal{C}_{ij}^D = \begin{cases} 1, & \text{if } r_i \text{ of } M^a \text{ is in contact with } r_j \text{ of } M^b \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

3.4 Method

First, we describe the optimization process that uses the ground-truth discrete human-human contact maps from the FlickrCI3D Signatures dataset [41]. The output from this process is used as training data to train the diffusion model that learns the 3D proxemics prior between two people. Lastly, we describe how such a prior can be used during optimization for reconstructing two people in close proximity from images without relying on ground-truth contact maps.

For all of the following, we use the SMPL-X [144] body model to represent the human bodies. For the purposes of this chapter, we use the first 10 shape components and keep the facial expression and finger pose fixed. Note that, although we use SMPL-X, we do not optimize hand pose due to the lack of robust hand keypoint detectors for people in interaction. Following [190], we interpolate between the shape space of SMPL-X and SMIL [65] to support producing meshes for infants and children. In practice, we concatenate the interpolation parameter and β such that $\beta \in \mathbb{R}^{11}$. The generated meshes are placed in world coordinates by translating them by $\gamma \in \mathbb{R}^3$ and rotating the body global orientation by $\phi \in \mathbb{R}^3$. Since our goal is to estimate two people a/b , we denote each person’s parameters as $\phi^a, \theta^a, \beta^a, \gamma^a$ and $\phi^b, \theta^b, \beta^b, \gamma^b$. For simplicity, we refer to both people when no index is specified and not stated differently, e.g., ϕ refers to ϕ^a and ϕ^b .

3.4.1 Reconstructing Bodies with Contact Maps

Optimization-based methods for fitting 3D meshes to RGB images usually rely on sparse signals, like 2D keypoints (ground-truth or detected), and priors for human pose and shape [19, 144, 232]. Only a few methods explicitly use self- [136] or human-human [41] contact in their optimization.

Our optimization method takes as input discrete human-human contact annotations, and, for each person, detected 2D keypoints [23, 221], and initial estimates for pose, $\tilde{\theta}$, orientation, $\tilde{\phi}$, shape, $\tilde{\beta}$, and translation, $\tilde{\gamma}$, which are provided from the output of BEV [190]. This is similar to SMPLify-DC in Chapter 2, where we use the regressor output to initialize the optimization. Note that the initial estimates from BEV are in the SMPL format, while our optimization uses SMPL-X. Fitting SMPL-X to SMPL meshes is possible, but slow. Instead, we directly solve for body shape using least-

squares and, knowing that the initial poses are only slightly different, we input the SMPL pose parameters to SMPL-X; see Appendix B.1.1 for more details.

Given these inputs, we take a two-stage approach: In the first stage, we optimize pose, θ , shape, β , and translation, γ , encouraging contact between discretely annotated body regions, while allowing the bodies to intersect. In the second stage, we activate a new loss term to resolve human-human intersection. The output of the first stage is usually close to the final pose with only slight intersections, because of which we optimize only pose and translation and fix the body shape in stage two. The objective function is:

$$L_{\text{fitting}} = \lambda_J L_J + \lambda_{\bar{\theta}} L_{\bar{\theta}} + \lambda_{\theta} L_{\theta} + \lambda_{\beta} L_{\beta} + \lambda_{\mathcal{C}^D} L_{\mathcal{C}^D} + \lambda_P L_P, \quad (3.2)$$

where L_J denotes the 2D re-projection error, $L_{\bar{\theta}}$ a prior on the initial pose, L_{θ} a German McClure pose prior [19], and L_{β} an L2-prior that penalizes deviation from the SMPL-X mean shape. The discrete human-human contact loss, $L_{\mathcal{C}^D}$, minimizes the distance between regions with annotated discrete human-human contact via:

$$L_{\mathcal{C}^D} = \sum_{i,j} \mathcal{C}_{ij}^D \min_{v \in r_i, u \in r_j} \|v - u\|^2. \quad (3.3)$$

L_P denotes an interpenetration loss, active in the second stage only, that pushes inside vertices to the surface. We use winding numbers to find intersecting vertices between M^a and M^b and vice versa. This operation is usually slow and memory intensive, which is why we use low-resolution meshes of SMPL-X with only 1K vertices. With V_I^a we denote vertices of M^a intersecting the low-resolution mesh of M^b ; V_I^b follows the same notation. The intersection loss term is defined as:

$$L_P = \sum_{v \in V_I^a} \min_{u \in V^b} \|v - u\|^2 + \sum_{v \in V_I^b} \min_{u \in V^a} \|v - u\|^2. \quad (3.4)$$

We find functional weights, λ , for each term in the objective function. Note that, in contrast to the definitions in Chapter 2, we do not use mesh surface points to decrease the runtime duration when testing for intersections for multiple people. The result of this fitting approach are illustrated in Figure 3.2, along with the BEV initialization. We use this optimization routine to reconstruct 13K pairs of people in the FlickrCI3D dataset [41]. Since these bodies are obtained with ground-truth contact maps, we refer to them as *pseudo-ground truth*. During training of the diffusion model (see Section 3.4.2), we use the pseudo-ground truth fits as if they were ground-truth.

3.4.2 Diffusion Model for 3D Proxemics

Next we describe how to train a generative model given the pseudo-ground truth 3D body parameters of two people in close social interaction. For the purposes of this section, let

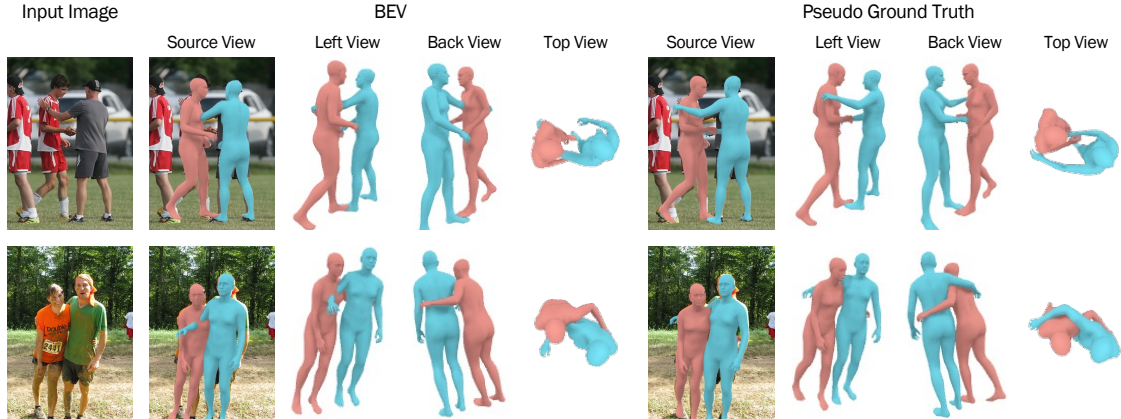


Figure 3.2: **Pseudo-ground truth data.** For each example we show a) the input image, b) the BEV [190] estimates used to initialize our approach (visualized in source and novel views) and c) the output of our optimization process that reconstructs two people in close proximity using ground-truth contact maps (again visualized in source and novel views). We use this 3D pseudo-ground truth as training data for training our diffusion model, BUDDI.

$X^a = [\phi^a, \theta^a, \beta^a, \gamma^a]$ and $X^b = [\phi^b, \theta^b, \beta^b, \gamma^b]$ be the concatenation of the body parameters corresponding to person a/b in a world coordinate frame. Then, our task is to learn the unconditional joint distribution of two people in close social interaction:

$$P(X^a, X^b). \quad (3.5)$$

This is a complex distribution to model, with 176 parameters (3 orientation, $24 * 3$ pose, 10 shape, and 3 translation for each person) affecting the subtle contact relationship present in a realistic interaction. For this task, we use a denoising diffusion model [68], a recent generative model that has shown remarkable capability in modeling the complex joint distribution of image pixels. Note that in this work we do not rely on any textual or image conditioning.

Background. Diffusion models are latent variable generative models that learn to transform random noise into the desired data distribution p_{data} . This is done through a forward Markov process $q(\mathbf{x}_{t+1}|\mathbf{x}_t)$, which gradually adds noise to samples \mathbf{x}_0 from the data distribution over T steps, and a reverse process $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which gradually brings noisy samples back into the data distribution.

The forward process transitions are Gaussian, i.e. in each step $t = 1, \dots, T$ a small amount of Gaussian noise is added to the data sample, such that

$$q(\mathbf{x}_{t+1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t+1}; \sqrt{1 - \sigma_t^2}\mathbf{x}_t, \sigma_t^2\mathbf{I}). \quad (3.6)$$

A convenient property of the forward diffusion process is that instead of adding noise gradually, by T times applying q , we can use the closed form solution:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\sigma}'_t}\mathbf{x}_0, (1 - \bar{\sigma}'_t)\mathbf{I}), \text{ with } \bar{\sigma}'_t = \prod_{i=0}^t (1 - \sigma_i). \quad (3.7)$$

To derive this distribution, let $\varepsilon_0, \dots, \varepsilon_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then $\mathbf{x}_t = \sqrt{1 - \sigma_t}\mathbf{x}_{t-1} + \sqrt{\sigma_t}\varepsilon_{t-1}$. With (1) defining $\sigma'_t = 1 - \sigma_t$, (2) recursively inserting \mathbf{x}_{t-i} , and (3) using the properties when summing over Gaussians, we obtain

$$\mathbf{x}_t \stackrel{\text{use(1)}}{=} \sqrt{\sigma'_t}\mathbf{x}_{t-1} + \sqrt{1 - \sigma'_t}\varepsilon_{t-1} \quad (3.8)$$

$$\stackrel{\text{use(2)}}{=} \sqrt{\sigma'_t\sigma'_{t-1}}\mathbf{x}_{t-2} + \sqrt{\sigma'_t(1 - \sigma'_{t-1})}\varepsilon_{t-2} + \sqrt{1 - \sigma'_t}\varepsilon_{t-1} \quad (3.9)$$

$$\stackrel{\text{use(2)}}{=} \dots \quad (3.10)$$

$$\stackrel{\text{use(3)}}{=} \sqrt{\bar{\sigma}'_t}\mathbf{x}_0 + \sqrt{1 - \bar{\sigma}'_t}\varepsilon_0 \quad (3.11)$$

The variance $\sigma_1^2, \dots, \sigma_T^2 \in (0, 1)$ is defined through a variance schedule a priori, usually small with respect to the data and increases with T . Consequently, for $T \rightarrow \infty$, $\sqrt{(1 - \bar{\sigma}'_t)}$ will converge towards 1 because $\sqrt{\bar{\sigma}'_t}\mathbf{x}_0$ converges to 0. That means the sample \mathbf{x}_0 is modified towards Gaussian noise of zero mean and unit variance.

If we knew the probability density function of the reverse transitions $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ we could draw new samples from p_{data} by gradually removing noise from a sample \mathbf{x}_t . However, estimating $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ would require knowing the entire dataset which is usually not available. With small enough σ_t^2 , the reverse transitions are also Gaussians that can be approximated via a learned model D :

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx \mathcal{N}(\mathbf{x}_{t-1}; \mu_D(\mathbf{x}_t, t), \Sigma_D(\mathbf{x}_t, t)). \quad (3.12)$$

We refer to the process of adding noise as *diffusion* and the process of removing the noise via D as *denoising*.

In contrast to Ho et al. [68], our denoiser directly predicts a sample $\hat{\mathbf{x}}_0 = D(\mathbf{x}_t; t)$ instead of ε_t . During training, the model is trained to minimize

$$\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}} \mathbb{E}_{t \sim \mathcal{U}\{0, T\}, \mathbf{x}_t \sim q(\cdot|\mathbf{x}_0)} \|D(\mathbf{x}_t; t) - \mathbf{x}_0\|. \quad (3.13)$$

Eventually, the model learns to denoise random samples $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into samples from the true data distribution via an iterative sampling process. Diffusion models can also be trained to take conditioning, \mathbf{c} , i.e. side information provided to the model during training and at test time. In this case, $D(\mathbf{x}_t; t)$ becomes $D(\mathbf{x}_t; t, \mathbf{c})$.

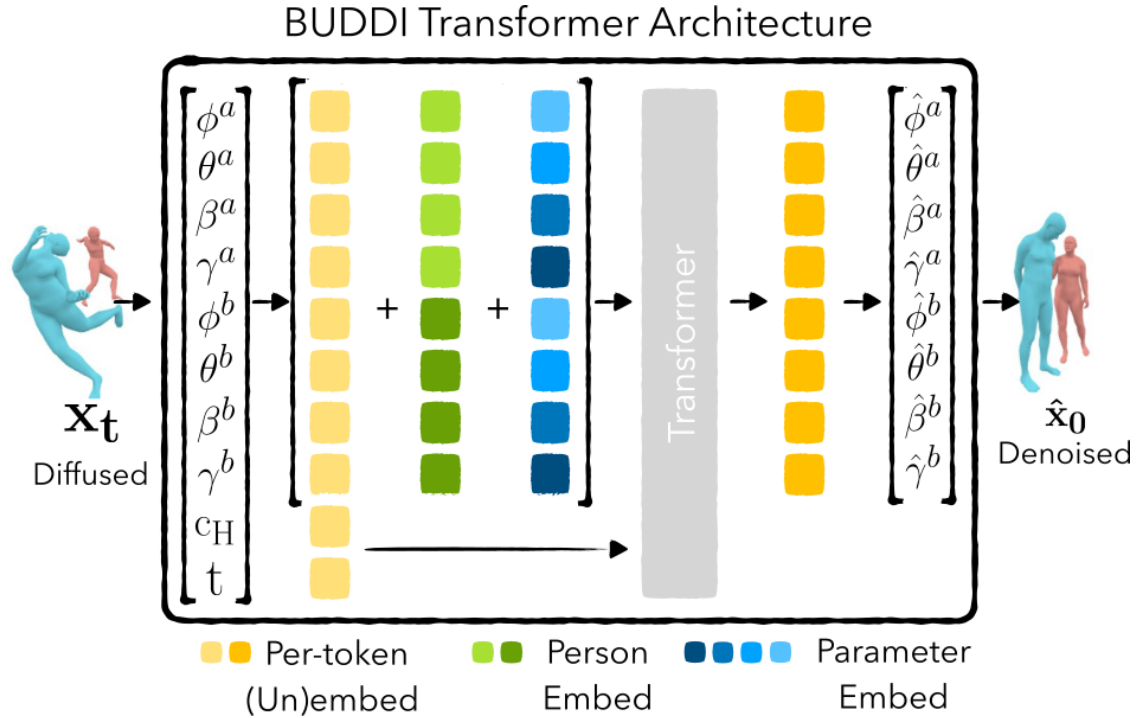


Figure 3.3: **BUDDI: BUDDIES Diffusion model**. We illustrate the architecture of BUDDI, our diffusion model for modeling 3D social proxemics between two people in close interaction. The diffusion process is applied directly on SMPL-X body parameters. To condition BUDDI on estimated body model parameters, c_H , we concatenate the parameters along the token dimension.

Architecture. In our scenario, a data sample $X = [X^a, X^b]$ corresponds to the concatenation of two bodies, i.e. $[\phi^a, \theta^a, \beta^a, \gamma^a, \phi^b, \theta^b, \beta^b, \gamma^b]$. A natural question is in what domain do we diffuse and denoise these parameters? Prior work in modeling human motion used 3D joint locations or pose parameters only [196, 229]. However, since contact requires joint reasoning about the pose as well as the shape of people, in this work we directly operate on the raw parameter space. We treat each type of parameter class (global orientation, pose, shape, and translation per person) as a token input to a transformer encoder.

Specifically, we first diffuse ground-truth $\mathbf{x}_0 = [X^a, X^b]$, by uniformly sampling a noise level t with noise $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$, to then obtain the noisy input signal $\mathbf{x}_t = \sqrt{\sigma_t'} \mathbf{x}_0 + \sqrt{1 - \sigma_t'} \varepsilon_t$. To denoise \mathbf{x}_t , we pass it through our denoiser model D : first, each body model parameter $i \in \{\phi, \theta, \beta, \gamma\}$ of each person $j \in \{a, b\}$ is embedded via linear layers f_{ij} to which we add learnable embeddings, w_i and w_j to encode body model parameters and human identity, respectively. More formally: $g(\mathbf{x}_t; i, j) = f_{ij}(\mathbf{x}_t^{ij}) + w_i + w_j$, where $g(\mathbf{x}_t; i, j) \in \mathbb{R}^{152}$. Noise level t is also encoded and embedded via linear layers g_t such that $g_t(t) \in \mathbb{R}^{152}$. Then the eight model parameter embeddings and the noise level

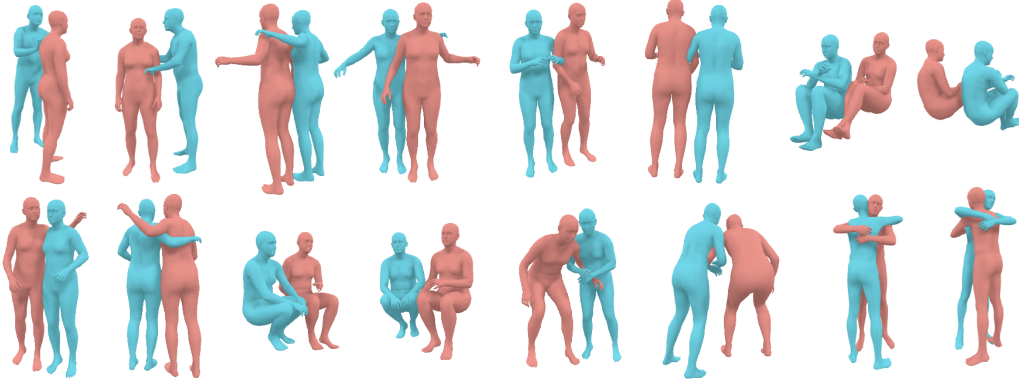


Figure 3.4: **Samples from Generative Proxemics.** All samples are unconditionally generated from pure noise using the trained diffusion model. We select a couple of representative examples and show two views per sample. Note that our model learns to generate the distribution of people in close contact including embracing each other, shaking hands, playing sports, sitting side by side, and taking photographs.

embedding are concatenated, resulting in latent vectors $\mathbf{x}'_t \in \mathbb{R}^{9 \times 152}$, which we pass to the transformer encoder. Finally, we use linear layers h_{ij} to un-embed the transformer encoder output and obtain the estimated denoised parameters $\hat{\mathbf{x}}_0$. Figure 3.3 illustrates the denoiser model D along with the visualization of a diffused set of bodies. For the task of reconstructing humans from images, we condition the denoising network D on \mathbf{c}_H , the SMPL-X parameters of two humans predicted by a regressor. We provide more information about the architecture of the conditional model in Appendix B.1.3

Note that our model directly predicts SMPL-X model parameters, which allows us to employ standard human pose and shape regularization losses. The training objective is:

$$L_D = L_\theta + L_\beta + L_\gamma + L_{v2v}, \quad (3.14)$$

where L_θ , L_β , L_γ denote squared L2-losses on body model parameters and L_{v2v} on model vertices. We use 6D rotation representations [239] for global orientation and pose, and model the relative translation between a and b by setting $\gamma^a = 0$. The trained diffusion model can be used to generate unconditional samples as illustrated in Fig. 3.4.

3.4.3 Optimization with the Proxemics Prior

Given our diffusion generative model D , we now describe how to reconstruct two people in close social interaction from an image, without relying on any ground-truth contact labels. At test time, we propose to use the diffusion model as a prior in optimization, congruent to score distillation sampling, proposed in recent works such as DreamFusion [153] and Score Jacobian Chaining [209], which optimize a 3D scene rep-

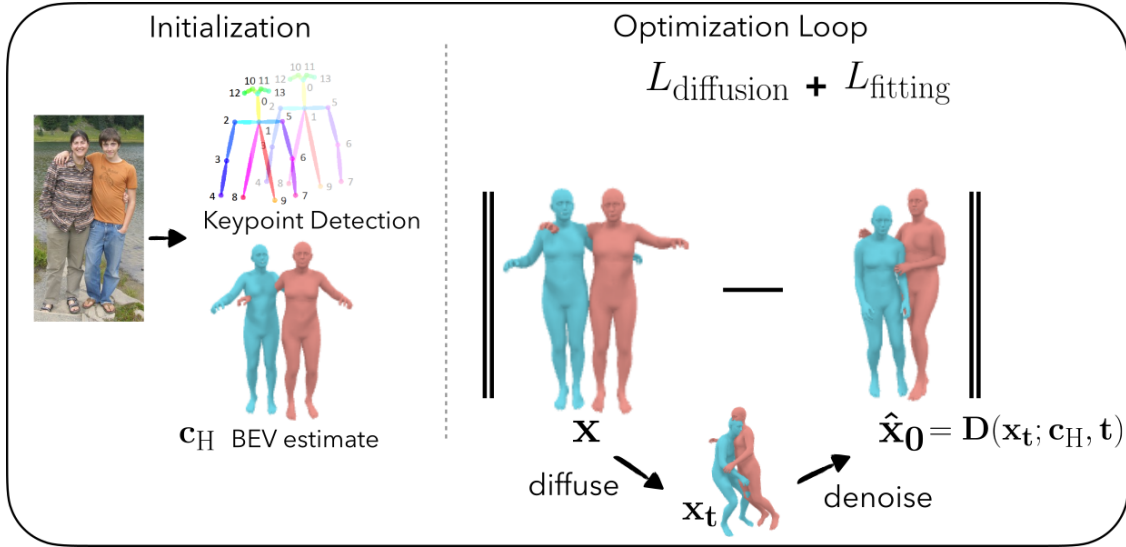


Figure 3.5: **Optimization with Generative Proxemics.** We illustrate the optimization method with BUDDI as prior. Our optimization takes detected keypoints [221, 23] and an initial regressor estimate [190] as input. Given the regressor estimate, we sample from BUDDI to obtain \tilde{x} which we use to initialize the optimization routine. In each optimization iteration, we take a single *diffuse-denoise* step on the current estimate using the learned denoiser model D . Our losses encourage the current estimate to be close to the refined meshes ($L_{\text{diffusion}}$) and to the initial estimate and detected keypoints (L_{fitting}).

representation using a 2D generative model. Specifically, we use the same setup as described in Section 3.4.1 but without the ground-truth contact labels, where we minimize $L_{\text{Optimization w. BUDDI}} = L_{\text{fitting}} + L_{\text{diffusion}}$, with

$$L_{\text{diffusion}} = \|D(\mathbf{x}_t; t, \mathbf{c}_H) - \mathbf{x}\|, \quad (3.15)$$

where $\mathbf{x}_t = \sqrt{\sigma_t'} \mathbf{x}_{\text{no-grad}} + \sqrt{1 - \sigma_t'} \boldsymbol{\varepsilon}_t$ denotes the diffused body model parameters of the current estimate \mathbf{x} . We follow previous work and detach the gradients of \mathbf{x} before sending them to the denoiser D , denoted as $\mathbf{x}_{\text{no-grad}}$. $L_{\text{diffusion}}$ performs a single *diffuse-denoise* step on $\mathbf{x}_{\text{no-grad}}$ to obtain the refined estimate $\hat{\mathbf{x}} = D(\mathbf{x}_t; t, \mathbf{c}_H)$, and encourages \mathbf{x} to be close to $\hat{\mathbf{x}}$; we illustrate this process in Fig. 3.5 (optimization loop). Intuitively, this loss uses the learned denoiser D to take a step from the current estimate towards the true distribution of two people in close proximity.

We also use the current estimate \mathbf{x} to calculate L_{fitting} as described in Section 3.4.1 with $\lambda_{\mathcal{C}D} = 0$. The terms in L_{fitting} ensure that the solution stays close to the image evidence, while $L_{\text{diffusion}}$ is a data-driven prior, similar to those used for 3D pose in previous works such as GMM [19] and V-Poser [144], but for 3D proxemics. In practice, we decode \mathbf{x} and $\hat{\mathbf{x}}$ into model parameters to gain control over their individual contribution, thus, we

can rewrite: $L_{\text{diffusion}} =$

$$\lambda_{\hat{\phi}} \|\hat{\phi} - \phi\| + \lambda_{\hat{\theta}} \|\hat{\theta} - \theta\| + \lambda_{\hat{\beta}} \|\hat{\beta} - \beta\| + \lambda_{\hat{\gamma}} \|\hat{\gamma} - \gamma\|.$$

Note that we only use terms that incorporate parameters we optimize in both stages, i.e. the shape and orientation loss weights are set to zero. Our final optimization routine follows the setup proposed in Section 3.4.1, but does not depend on ground-truth contact maps. We provide more details in Appendix B.1.3.

3.5 Experiments

Datasets. We use FlickrCI3D Signatures [41], a dataset of images showing interacting humans collected from Flickr. This dataset contains various scenes of sports, families, couples, etc. It has discrete 3D contact annotations between pairs of people. The dataset contains 10,631/1,139 images for train/test. One image can have multiple people and therefore multiple contact annotations. For evaluation on FlickrCI3D Signatures Test, we use annotations for which matching BEV, 2D keypoints, and contact labels was possible; i.e., a total 1427 contact pairs.

During training, we also use a small portion of MoCap data: CHI3D [41], which contains 3/2 pairs of training/test subjects performing 127 sequences of two-person interactions, where one frame in the sequence has contact annotated with a contact map. We use sequences from 2 pairs of training subjects to train our diffusion model, which results in 247 mesh pairs for training and the third pair for evaluation. Hi4D [227] contains sequences of 20 pairs of people interacting with each other. The interactions include actions like hugging, dancing, and fighting. We randomly split the data into 14/3/3 pairs for train/val/test and use every fifth frame of the subsequence involving contact as labeled in Hi4D, resulting in about 1K mesh pairs for training. The body representation format in Hi4D is SMPL, which we transfer to SMPL-X using the SMPL-X code repository [144]. Please see the Appendix B for more details about the datasets. Note that while we use SMPL-X model, BUDDI is not trained on hands because none of these datasets contain hand poses.

Baselines. We compare our reconstruction method with BEV [190], which is also used as an input to our conditional model. Since there is no other available work that reasons about people in close social interaction, we experiment with simple but effective baselines. We train the transformer model of BUDDI to directly predict SMPL-X parameters of people in contact from BEV input, essentially a deterministic, single-step ablation of BUDDI. We also evaluate the direct conditional denoised output of BEV by BUDDI without any optimization. As another baseline, we propose an optimization routine that replaces $L_{\text{diffusion}}$ with a simple heuristic that takes the minimal distances between two meshes predicted by BEV and minimizes their distance during optimization along with the other energy terms. Finally, to compare the generation ability we train a VAE which

we also use during the optimization routine in a similar manner to VPoser [144] but for two people by optimizing the VAE latent space instead of SMPL-X parameters. We refer to these models as *Transformer*, *BUDDI (gen.)*, *Contact Heuristic*, and *VAE*, respectively. All baselines are trained on the same datasets as BUDDI with the same sampling strategies. Details about our baselines are provided in the Appendix B.3.1.

Metrics. We use standard evaluation metrics from the human pose and shape estimation literature. Besides MPJPE and PA-MPJPE, we also report the joint PA-MPJPE of both people together. In addition to per-person metrics, this captures the relative orientation and translation of the two people. Previous approaches that predict contact maps evaluate contact metrics based on IoU of the contact map. Please note that previous work [41] only evaluates the contact maps and does not use the predicted contact map to optimize for 3D humans. Since our method directly estimates 3D humans instead of contact maps, we propose a new metric similar to PCK [224] from the 2D pose literature called **PCC**, the percentage of correct contact points with respect to a radius r . Specifically, given two meshes, M^a/M^b and a contact map \mathcal{C}^D we compute the pairwise vertex-to-vertex Euclidean distances $d_{\text{eucl}}(\mathcal{C}^D)$ between annotated contact regions and consider the pair to be correct when $\min(d_{\text{eucl}}(\mathcal{C}^D)) < r$.

Implementation Details. BUDDI is trained with meshes from FlickrCI3D Signatures Fits, CHI3D, and Hi4D. We use 60% Flickr, 20% CHI3D, and 20% Hi4D data distribution per batch with batch size 512. The transformer backbone has six layers and eight heads; we use 10% dropout and randomly shuffle the order of people during training. To train BUDDI, we randomly sample noise levels t up to 1000 using a cosine noise schedule [138]. We use the Adam optimizer [92] with learning rate 10^{-4} . We train two separate networks: an unconditional model for generation and the conditional version for reconstruction. For the conditional model, we use all camera views of the MoCap datasets, i.e. 4/8 cameras for CHI3D/Hi4D and set $\mathbf{c}_H = \emptyset$ with a 20% chance. The unconditional model is trained on 3D MoCap fits in the world coordinate system. To sample new poses, we use DDIM sampling starting at noise levels $t = 1000$ in steps of 10.

During optimization, we experiment with different noise levels, between 10 and 100, and find that $t = 10$ does not disturb the inputs too much, but enough for D to generate new configurations. We use detected 2D keypoints from OpenPose [23] and ViTPose [221] and BEV [190] estimates as conditioning. Unlike single-person mesh regressors, BEV is designed to predict multiple people including their relative depth. Similar to most optimization methods that fix the global orientation [19, 144], we also choose to not update the estimated global orientation from BEV, which is typically reliable. We use Adam optimizer with $lr = .01$ and run each stage for a maximum of 1000 iterations and with early, gradient-dependent stopping. Please see Table B.1 in Appendix B for more details.

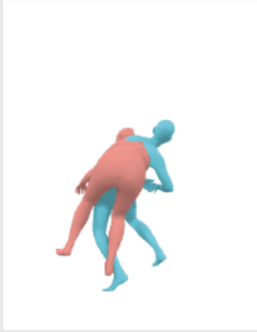
Which video shows a more realistic close social interaction between two people?

In this task you are presented with two videos of rotating characters interacting closely with each other.

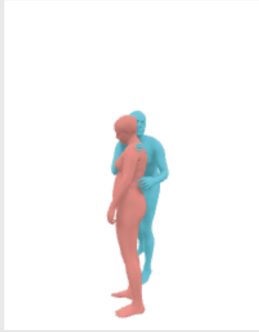
Please decide which video shows a more realistic close social interaction.

Please watch each video long enough to see a full rotation of the characters to see the interaction from all angles.
This usually takes at least 5 to 10 seconds.

Once you have finished all video pairs, the 'SUBMIT' button will be activated and you can submit the HIT.



video A



video B

Which video shows a more realistic close social interaction between two people?

video A

video B

Figure 3.6: **Amazon Mechanical Turk perceptual study layout and instructions.** We show a 360-degree video of the two interacting people. The person working on this task must decide whether video A or B is better.

3.5.1 Unconditional Generation

We qualitatively evaluate BUDDI by showing samples from it in Fig. 3.1 and 3.4. Our approach is able to generate people in close proximity including embraces, handshakes, having a conversation, sitting side by side, and in general plausibly interacting with each other. Since it is trained mostly on Internet image collections, it learns to generate people who are being photographed. It also generates people playing sports.

We further run a perceptual study to evaluate the realism of the generated social interactions against other methods. In a forced choice study, we compare our generated samples with samples from the real data distribution according to the 60/20/20 per-batch ratio for Flickr/CHI3D/Hi4D used during training. We also compare BUDDI against generations from the VAE and a non-parametric random baseline that samples meshes from the pseudo-ground truth after centering the two people. We do a forced choice

	JOINT ↓	PCC at radius ↑				
	PA-MPJPE	5	10	15	20	25
BEV	106	-	-	-	-	-
Transformer	86	14	40	60	73	82
BUDDI (gen.)	92	15	39	58	71	80
Heuristic	68	14	34	49	61	70
VAE	101	11	28	42	55	65
BUDDI	66	19	44	62	73	81

Table 3.1: **3D Pose Evaluation on FlickrCI3D Signatures.** We evaluate methods against the Flickr fits using their joint (two-person) PA-MPJPE expressed in mm. We also evaluate the percentage of correct contact points (PCC) for radius r mm.

comparison between BUDDI and these other methods, asking workers on Amazon Mechanical Turk to choose the sample that shows a more realistic close social interaction. We use 256 samples per method. We collect ratings for 768 pairwise comparisons. In this study, BUDDI was chosen over random in 71.23% of the comparisons, over the VAE in 60.17%, and over the training data in 44.4%. Note that 50% is the upper bound for such forced choice comparisons, in which participants cannot tell the difference between real and generated samples. We show our the design of our user study in Figure 3.6.

3.5.2 Fitting with BUDDI

Lastly, we evaluate our approach that reconstructs people in close proximity from an image.

We show qualitative results in Figure 3.7 comparing against BEV and the Contact Heuristic. Our approach is able to generate various types of human interactions with plausible contact and depth placement. It is also able to capture close interaction between a child and a parent in a plausible manner. Note that all methods take BEV as an input. Although the Contact Heuristic is able to move two people closer together, which helps with image alignment, upon close observation it is not able to capture the subtle interaction between people that happens during intimate interaction. BUDDI’s estimates are more realistic and better capture the subtle details of interaction. We provide additional qualitative examples of optimization with BUDDI and compare them to BEV in Figures 3.8 and 3.9 and the baseline methods in Figure 3.10. Failure cases are provided in Figure 3.11.

We further report the percentage of correct contact (PCC) with respect to the ground truth contact map on the FlickrCI3D Signatures test set in Table 3.1.

The table also shows the pose reconstruction accuracy against our Flickr Fits. All met-

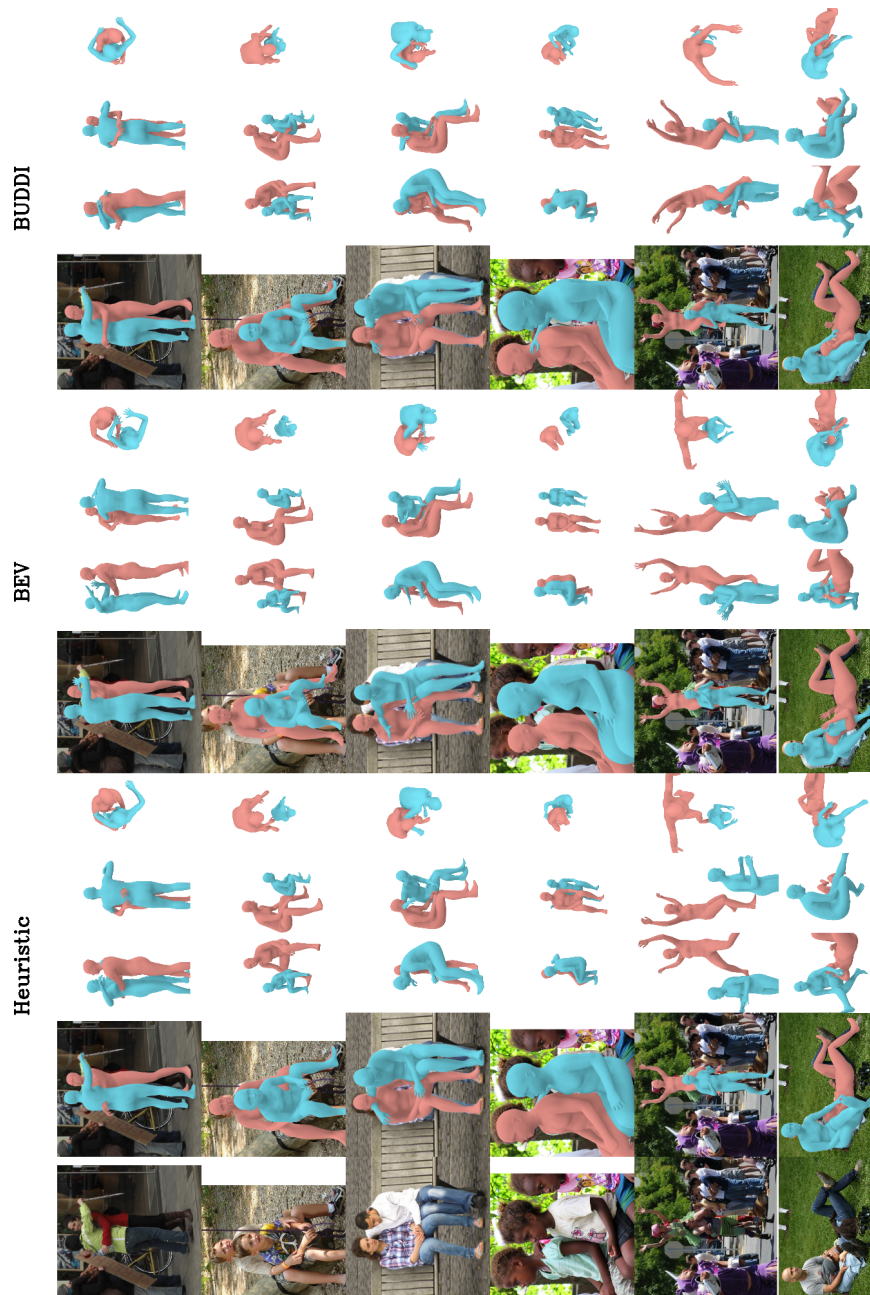


Figure 3.7: **Automatic reconstruction of people in close social interaction on Flickr images.** We show qualitative results from a) BEV, b) contact heuristics, which takes the BEV output and encourages the closest parts to be in contact, and c) our method, which optimizes the BEV estimates against the image evidence with the BUDDI prior. Our approach recovers a plausible reconstruction with subtle details.



Figure 3.8: **Optimization with BUDDI.** Additional qualitative examples from optimization with BUDDI compared to BEV. We provide the overlay and three additional views per method. Optimization with BEV (first method / columns 2-5), optimization with BUDDI (second method / columns 6-9).



Figure 3.9: **Optimization with BUDDI (continuation).** Additional qualitative examples from optimization with BUDDI compared to BEV. We provide the overlay and three additional views per method. Optimization with BEV (first method / columns 2-5), optimization with BUDDI (second method / columns 6-9).

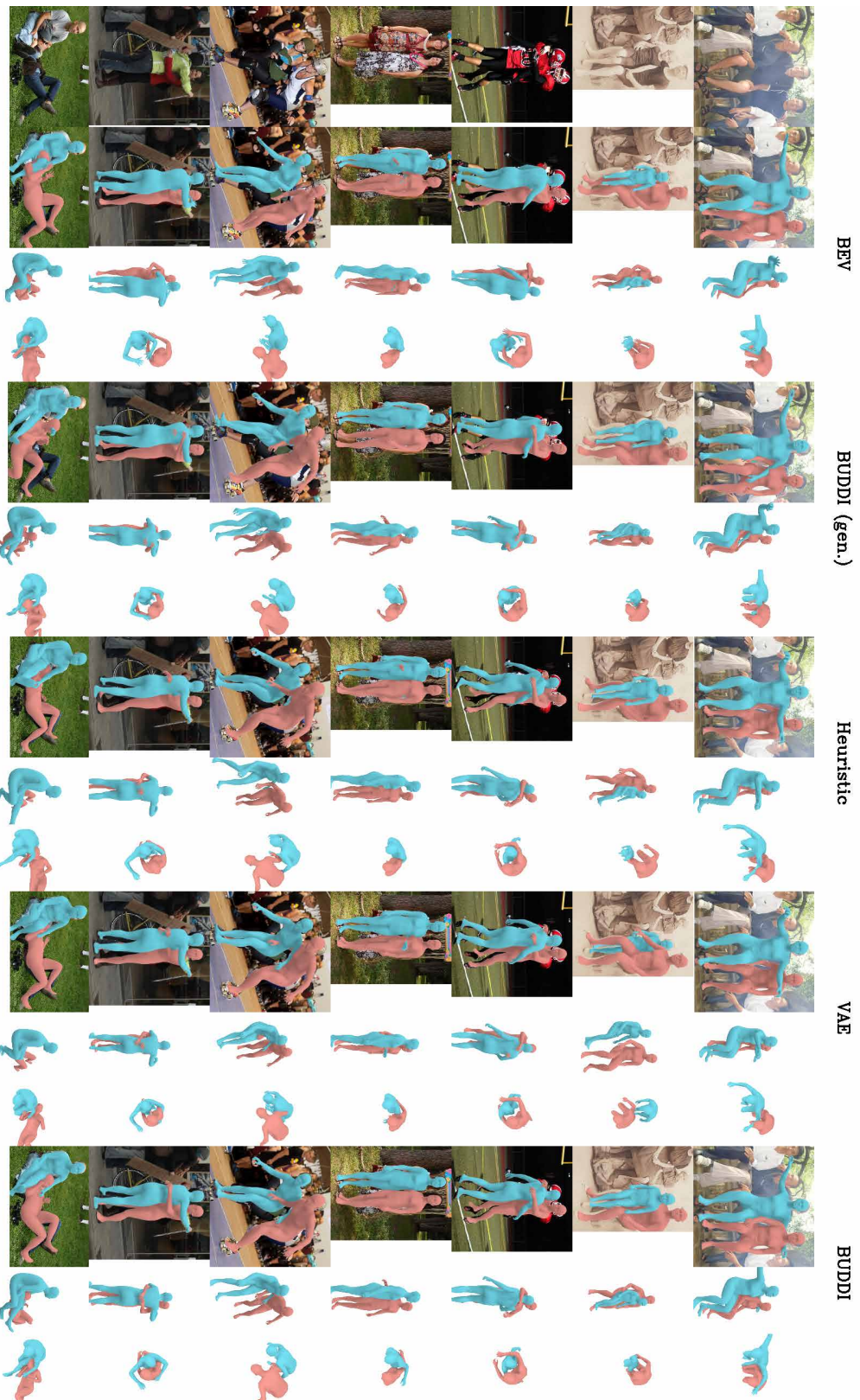


Figure 3.10: **Optimization with BUDDI.** Additional qualitative examples from optimization with BUDDI compared to BEV, BUDDI generations, optimization with heuristic, and optimization with VAE. We provide the overlay and three additional views per method. BEV (first method / columns 2-5), BUDDI (gen.) (second method / columns 6-9), optimization with heuristic (third method / columns 10-13), optimization with VAE (fourth method / columns 14-16), and optimization with BUDDI (fifth method / columns 17-20).

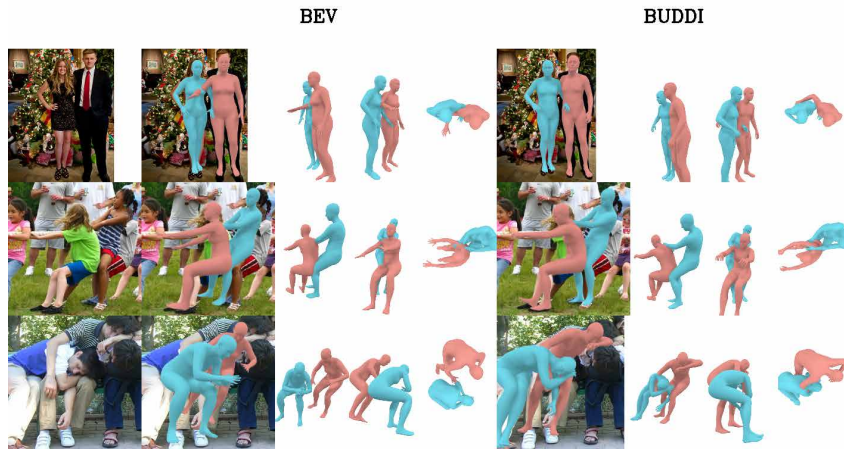


Figure 3.11: **Failure cases optimization with BUDDI.** Failure cases from optimization with BUDDI. In the first row the depth ordering of leg arm is wrong. The image in row 2 contains less common contact so that BUDDI suggests for blue to hold red’s shoulders instead of the rope. The estimated predicted by our method suggests a plausible pose that is not consistent with the image due to single-view ambiguity. The last row shows a failure case due to intersection between arm and torso.

rics show improvement over BEV, in particular the joint PA-MPJPE. Non-optimization methods, i.e. *Transformer* and *BUDDI (gen.)*, are able to predict plausible contacts, with similar PCC accuracy to BUDDI, but struggle to reconstruct the data with a worse joint PA-MPJPE. The *Heuristic*, in contrast, achieves a lower reconstruction error, but worse PCC. Our approach which leverages the learned prior during optimization can recover both the relative positions and contacts between the two people. To provide insights into the performance of single-person mesh regressors when evaluated on the two-person reconstruction task, we run 4D Humans [47] on Flickr Fits. The joint PA-MPJPE is 344 mm which is high, as expected, since these methods are not trained to reason about proximity.

We further evaluate our model against ground truth MoCap data in Table 3.2 and Table 3.3. Optimization with BUDDI consistently improves the two-person reconstruction error over BEV and other baselines. When evaluated per action, the strongest improvements over BEV come from complex close social interactions like hugging or kissing, at 58mm and 54mm absolute improvement over BUDDI respectively.

The *Heuristic* baseline achieves a low PA-MPJPE reconstruction error on all three datasets. Our hypothesis is that the Heuristic baseline is particularly strong for poses with only a few physical contact points, such as a handshake, whereas more complex contact, such as a hug, requires data-driven priors like BUDDI. To quantify this assumption, we compute the percentage of vertices that are in contact (with distance ≤ 10 cm to the other person) for each action in Hi4D. As the amount of contact increases (i.e. becomes more

	PER PERSON ↓ PA-MPJPE	JOINT ↓ PA-MPJPE
BEV	50 52	96
Transformer	54 56	105
BUDDI (gen.)	53 53	80
Heuristic	49 46	105
VAE	54 54	103
BUDDI	48 47	68

Table 3.2: **Quantitative Evaluation on CHI3D.** We compare the output of our model to the baselines on CHI3D (pair s03). All errors reported in mm for 3D Joints.

	PER PERSON ↓ PA-MPJPE	JOINT ↓ PA-MPJPE	JOINT PA-MPJPE ↓										
			backhug	basketball	cheers	dance	fight	highfive	hug	kiss	pose	sidehug	talk
BEV	78 / 84	136	200	126	109	135	121	106	163	139	142	131	118
Heuristic	67 / 71	121	168	83	94	131	94	68	159	159	118	113	109
BUDDI (F, C)	70 / 77	115	200	94	92	128	108	100	133	114	104	107	91
Transformer	79 / 85	120	161	141	103	138	123	128	117	106	120	105	100
BUDDI (gen.)	82 / 90	117	152	139	120	137	130	96	101	97	115	102	101
VAE	80 / 82	138	175	133	114	141	119	87	176	162	135	140	113
BUDDI	70 / 76	98	127	95	92	113	109	72	105	85	88	96	81

Table 3.3: **Evaluation of BUDDI on Hi4D.** We compare the output of BUDDI to the proposed baseline methods on the Hi4D challenge. The first block shows methods that do not use Hi4D data during training or are optimization based without access to priors trained on Hi4D. BUDDI (F,C) in particular, is our model BUDDI trained on Flickr and CHI3D data only. All errors are reported in mm for 3D Joints.

complex), BUDDI significantly outperforms the heuristic. Ordering activities by amount of contact in Hi4D gives: **basketball** (7%), **dance** (9%), **fight** (10%), **highfive** (12%), **talk** (18%), **backhug** (23%), **cheers** (24%), **pose** (29%), **kiss** (46%), **sidehug** (47%), **hug** (53%). Red means the Heuristic is better than BUDDI and green means BUDDI is better than the Heuristic. The contact percentage is indicated in brackets. On average, BUDDI outperforms the heuristic by 23 mm, and particularly improves the reconstruction result for poses with many physical contact points.

Transformer and *BUDDI (gen.)* have lower joint PA-MPJPE errors than BEV and the *Heuristic*, but worse per-person reconstruction errors. The *VAE* results suggest that directly operating in the latent space of a generative model is challenging and not sufficient to accurately recover close social interactions. BUDDI, in contrast, is able to model a wide variety of poses, as supported by the numerical results.

To gain insight into the contribution of each loss term in the optimization method

with BUDDI used a prior, we ablate them starting from L_{J2D} , i.e. the 2D keypoint re-projection loss. The JOINT PA-MPJPE \downarrow on Hi4D is 118/118/111/99/99/98 for $L_{J2D}/+L_P/+L_{\tilde{g}}/+L_{\gamma_{BUDDI}}/+L_{\beta_{BUDDI}}/+L_{\theta_{BUDDI}}$. This result emphasizes the importance of $L_{\gamma_{BUDDI}}$, i.e. the translation prior from BUDDI has the biggest impact on the final result.

3.6 Conclusion

We study 3D human reconstruction in the setting of close human-human interaction. We first leverage a large-scale dataset of images with ground truth annotations for body regions that are in contact for pairs of people; we formulate an optimization method to jointly reconstruct each pair in 3D. We use these human-human reconstructions to learn a data-driven prior of how humans interact in natural images. This prior is based on a denoising diffusion model that enables unconditional sampling of people in close social interaction. More importantly, we demonstrate how this prior can be incorporated in traditional iterative optimization as a novel regularization term that encourages the reconstructed pairs of people to have realistic interactions. Exciting future work is to iteratively apply our method to new images and use the reconstructed examples to further improve the generative prior. Additionally, conditioning modalities can be explored, e.g., conditioning on pixel features, on text, or on action labels. Finally, these insights could be also extended to 3D motion capture and also interactions that involve more than two humans.

Chapter 4

Accurate 3D Body Shape Regression using Metric and Semantic Attributes

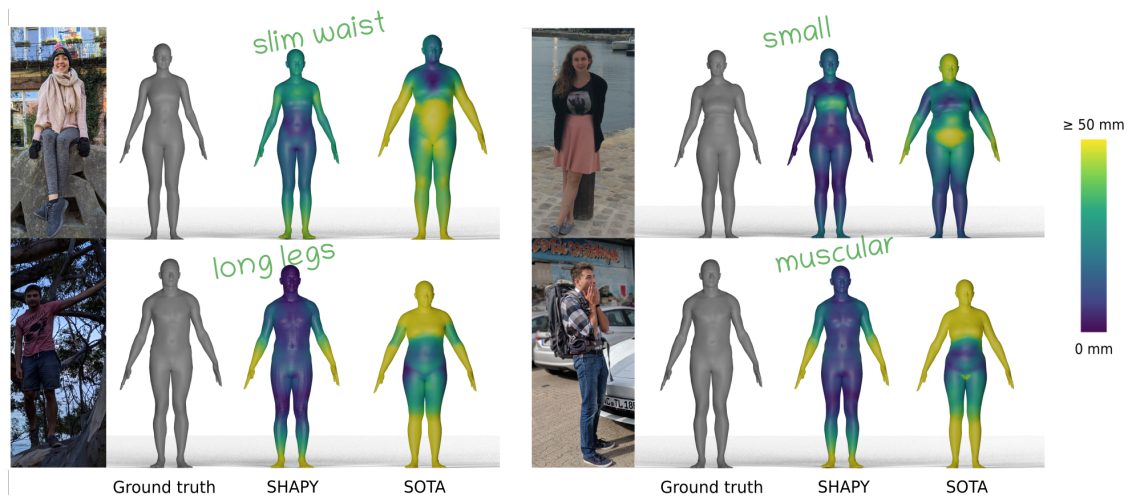


Figure 4.1: Existing work on 3D human reconstruction from a color image focuses mainly on *pose*. We present SHAPY, a model that focuses on body *shape* and learns to predict 3D body shape from a color image, using crowd-sourced *linguistic shape attributes*. Even with this weak supervision, SHAPY outperforms the state of the art (SOTA) [175] on in-the-wild images with varied clothing.

In the previous two chapters, we described methods to reconstruct a single person in a pose with self-contact and two people in close interaction from images. These methods mainly address body pose estimation while body shapes are usually close the SMPL mean shape. This is a problem, because self- and interpersonal contact happens on the body surface. To address this, we introduce SHAPY a network to accurately estimate body shape from images.

4.1 Introduction

The field of 3D human pose and shape (HPS) estimation is progressing rapidly and methods now regress accurate 3D pose from a single image [19, 86, 88, 97, 101, 144, 220, 98, 100, 233]. Unfortunately, less attention has been paid to body shape and many methods produce body shapes that clearly do not represent the person in the image (Fig. 4.1, top right). There are several reasons behind this. Current evaluation datasets focus on pose and not shape. Training datasets of images with 3D ground-truth shape are lacking. Additionally, humans appear in images wearing clothing that obscures the body, making the problem challenging. Finally, the fundamental scale ambiguity in 2D images, makes 3D shape difficult to estimate. For many applications, however, realistic body shape is critical. These include AR/VR, apparel design, virtual try-on, fitness, and not to mention the accurate estimation of self- and human-human contact. To democratize avatars, it is important to represent and estimate all possible 3D body shapes; we make a step in that

direction.

Note that commercial solutions to this problem require users to wear tight fitting clothing and capture multiple images or a video sequence using constrained poses. In contrast, we tackle the unconstrained problem of 3D body shape estimation in the wild from a single RGB image of a person in an arbitrary pose and standard clothing.

Most current approaches to HPS estimation learn to regress a parametric 3D body model like SMPL [123] from images using 2D joint locations as training data. Such joint locations are easy for human annotators to label in images. Supervising the training with joints, however, is not sufficient to learn shape since an infinite number of body shapes can share the same joints. For example, consider someone who puts on weight. Their body shape changes but their bones stay the same. Several recent methods employ additional 2D cues, such as the silhouette, to provide additional shape cues [174, 175]. Silhouettes, however, are influenced by clothing and do not provide explicit 3D supervision. Synthetic approaches [119], on the other hand, drape SMPL 3D bodies in virtual clothing and render them in images. While this provides ground-truth 3D shape, realistic synthesis of clothed humans is challenging, resulting in a domain gap.

To address these issues, we present SHAPY, a new deep neural network that accurately regresses 3D body shape and pose from a single RGB image. To train SHAPY, we first need to address the lack of paired training data with real images and ground-truth shape. Without access to such data, we need alternatives that are easier to acquire, analogous to 2D joints used in pose estimation. To do so, we introduce two novel datasets and corresponding training methods.

First, in lieu of full 3D body scans, we use images of people with diverse body shapes for which we have anthropometric measurements such as height as well as chest, waist, and hip circumference. While many 3D human shapes can share the same measurements, such measurements do constrain the space of possible shapes. Additionally, these are important measurements for applications in clothing and health. Accurate anthropometric measurements like these are difficult for individuals to take themselves but they are often captured for different applications. Specifically, modeling agencies provide such information about their models; accuracy is a requirement for modeling clothing. Thus, we collect a diverse set of such model images (with varied ethnicity, clothing, and body shape) with associated measurements; see Fig. 4.2.

Since sparse anthropometric measurements do not fully constrain body shape, we exploit a novel approach and also use *linguistic shape attributes*. Prior work has shown that people can rate images of others according to shape attributes such as “short/tall”, “long legs” or “pear shaped” [188]; see Fig. 4.3. Using the average scores from several raters, Streuber et al. [188] (BodyTalk) regress metrically accurate 3D body shape. This approach gives us a way to easily label images of people and use these labels to constrain 3D shape. To our knowledge, this sort of linguistic shape attribute data has not previously been exploited to train a neural network to infer 3D body shape from images.

We exploit these new datasets to train SHAPY with three novel *losses*, which can be exploited by any 3D human body reconstruction method:



Figure 4.2: Model-agency websites contain multiple images of models together with anthropometric measurements. A wide range of body shapes are represented; example from [pexels.com](https://www.pexels.com).

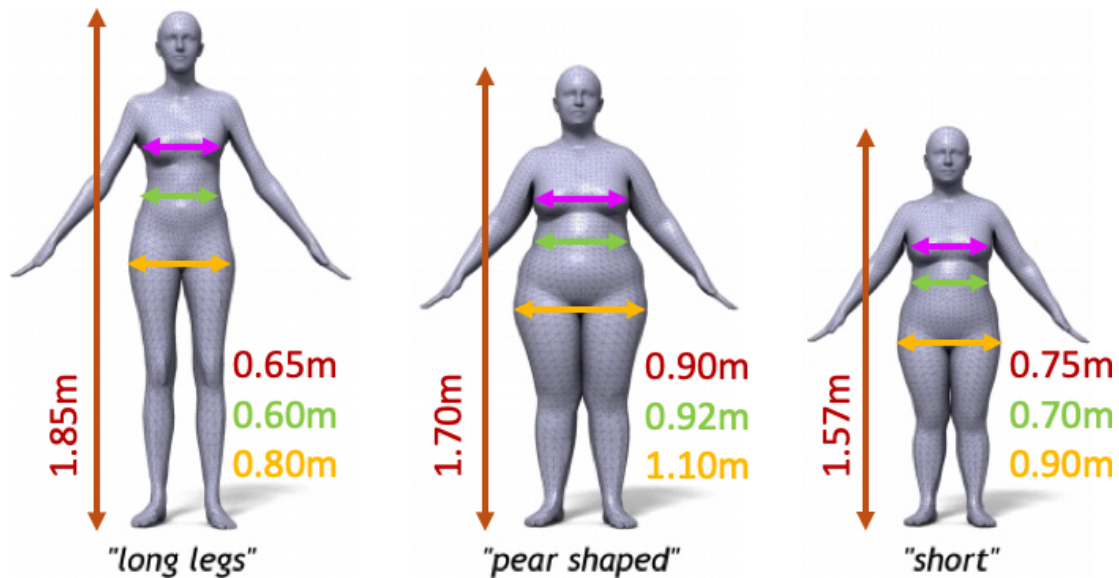


Figure 4.3: We crowd-source scores for linguistic body-shape attributes [188] and compute anthropometric measurements for CAESAR [162] body meshes. We also crowd-source linguistic shape attribute scores for model images, like those in Fig. 4.2

1. We define functions of the SMPL body mesh that return a sparse set of anthropometric measurements. When measurements are available for an image we use a loss that penalizes mesh measurements that differ from the ground-truth (GT).
2. We learn a “Shape to Attribute” (S2A) function that maps 3D bodies to linguistic attribute scores. During training, we map meshes to attribute scores and penalize differences from the ground-truth scores.
3. We similarly learn a function that maps “Attributes to Shape” (A2S). We then penalize body shape parameters that deviate from the prediction.

We study each term in detail to arrive at the final method. Evaluation is challenging because existing benchmarks with ground-truth shape either contain too few subjects [207] or have limited clothing complexity and only pseudo-ground truth shape [174]. We fill this gap with a new dataset, named “Human Bodies in the Wild” (HBW), that contains a ground-truth 3D body scan and several in-the-wild photos of 35 subjects, for a total of 2543 photos. Evaluation on this dataset shows that SHAPY estimates much more accurate 3D shape than existing methods.

We make models, data and code available for research purposes at shapy.is.tue.mpg.de.

4.2 Related Work

3D human pose and shape (HPS). Methods that reconstruct 3D human bodies from one or more RGB images can be split into two broad categories: (1) **parametric methods** that predict parameters of a statistical 3D body model, such as SCAPE [5], SMPL [123], SMPL-X [144], Adam [86], GHUM [220], and (2) **non-parametric methods** that predict a free-form representation of the human body [205, 170, 77, 219]. Parametric approaches lack details with respect to non-parametric ones, e.g., clothing or hair. However, parametric models disentangle the effects of identity and pose on the overall shape. Therefore, their parameters provide control for re-shaping and re-posing. Moreover, pose can be factored out to bring meshes into a canonical pose; this is important for evaluating estimates of an individual’s shape. Finally, since topology is fixed, meshes can be compared easily. For these reasons, we use a SMPL-X body model.

Parametric methods follow two main paradigms, and are based on optimization or regression. **Optimization-based methods** [11, 19, 50, 144] search for model configurations that best explain image evidence, usually 2D landmarks [23], subject to model priors that usually encourage parameters to be close to the mean of the model space. Numerous methods penalize the discrepancy between the projected and ground-truth silhouettes [74, 107] to estimate shape. However, this needs special care to handle clothing [10]; without this, erroneous solutions emerge that “inflate” body shape to explain the “clothed” silhouette. **Regression-based methods** [27, 46, 78, 87, 101, 119, 97, 136, 230] are currently based on deep neural networks that directly regress model parameters from

image pixels. Their training sets are a mixture of data captured in laboratory settings [75, 183], with model parameters estimated from MoCap markers [130], and in-the-wild image collections, such as COCO [120], that contain 2D keypoint annotations. Optimization and regression can be combined, for example via in-the-network model fitting [101, 136].

Estimating 3D body shape. State-of-the-art methods are effective for estimating 3D pose, but *struggle* with estimating *body shape* under clothing. There are several reasons for this. First, 2D keypoints alone are not sufficient to fully constrain 3D body shape. Second, shape priors address the lack of constraints, but bias solutions towards “average” shapes [19, 144, 101, 136]. Third, datasets with in-the-wild images have noisy or biased 3D bodies, recovered by fitting a model to 2D keypoints [19, 144]. Fourth, datasets captured in laboratory settings have a small number of subjects, who do not represent the full spectrum of body shapes. Thus, there is a scarcity of images with known, *accurate*, 3D body shape. Existing methods deal with this in two ways.

First, rendering *synthetic images* is attractive since it gives automatic and precise ground-truth annotation. This involves shaping, posing, dressing and texturing a 3D body model [69, 174, 176, 206, 211], then lighting it and rendering it in a scene. Doing this realistically and with natural clothing is expensive, hence, current datasets suffer from a domain gap. Alternative methods use artist-curated 3D scans [169, 170, 143], which are realistic but limited in variety. Recent work addresses this domain gap via synthetic datasets [143, 18], but methods trained on these datasets are still less accurate than SHAPY on images in the wild.

Second, *2D shape cues* for in-the-wild images, (body-part segmentation masks [139, 166, 37], silhouettes [1, 74, 146]) are attractive, as these can be manually annotated or automatically detected [48, 62]. However, fitting to such cues often gives unrealistic body shapes, by inflating the body to “explain” the clothing “baked” into silhouettes and masks.

Most related to our work is the work of Sengupta et al. [174, 176, 175] who estimate body shape using a probabilistic learning approach, trained on edge-filtered synthetic images. They evaluate on the SSP-3D dataset of real images with pseudo-ground truth 3D bodies, estimated by fitting SMPL to multiple video frames. SSP-3D is biased to people with tight-fitting clothing. Their silhouette-based method works well on SSP-3D but does not generalize to people in normal clothing, tending to over-estimate body shape; see Fig. 4.1.

In contrast to previous work, SHAPY is trained with in-the-wild images paired with linguistic shape attributes, which are annotations that can be easily crowd-sourced for weak shape supervision. We also go beyond SSP-3D to provide HBW, a new dataset with in-the-wild images, varied clothing, and precise GT from 3D scans.

Shape, measurements and attributes. Body shapes can be generated from anthropometric measurements [3, 177, 178]. Tsoli et al. [203] register a body model to multiple high-resolution body scans to extract body measurements. The “Virtual Caliper” [155] allows users to build metrically accurate avatars of themselves using measurements or

VR game controllers. ViBE [71] collects images, measurements (bust, waist, hip circumference, height) and the dress-size of models from clothing websites to train a clothing recommendation network. We draw inspiration from these approaches for data collection and supervision.

Streuber et al. [188] learn BodyTalk, a model that generates 3D body shapes from linguistic attributes. For this, they select attributes that describe human shape and ask annotators to rate how much each attribute applies to a body. They fit a linear model that maps attribute ratings to SMPL shape parameters. Inspired by this, we collect attribute ratings for CAESAR meshes [162] and in-the-wild data as proxy shape supervision to train a HPS regressor. Unlike BodyTalk, SHAPY automatically infers shape from images.

Anthropometry from images. Single-View metrology [29] estimates the height of a person in an image, using horizontal and vertical vanishing points and the height of a reference object. Günel et al. [53] introduce the IMDB-23K dataset by gathering publicly available celebrity images and their height information. Zhu et al. [240] use this dataset to learn to predict the height of people in images. Dey et al. [35] estimate the height of users in a photo collection by computing height differences between people in an image, creating a graph that links people across photos, and solving a maximum likelihood estimation problem. Bieler et al. [17] use gravity as a prior to convert pixel measurements extracted from a video to metric height. These methods do not address body shape.

4.3 Representations and Data for Body Shape

We use linguistic shape attributes and anthropometric measurements as a connecting component between in-the-wild images and ground-truth body shapes; see Fig. 4.4. To that end, we annotate linguistic shape attributes for 3D meshes and in-the-wild images, the latter from fashion-model agencies, labeled via Amazon Mechanical Turk.

4.3.1 SMPL-X Body Model

We use SMPL-X [144] as introduced in Section 1.3 with $B = 100$ shape parameters, i.e. $|\beta| = B$.

4.3.2 Model-Agency Images

Model agencies typically provide multiple color images of each model, in various poses, outfits, hairstyles, scenes, and with a varying camera framing, together with anthropometric measurements and clothing size. We collect training data from multiple model-agency websites, focusing on under-represented body types, namely: curve-models.com, cocainemodels.com, nemesismodels.com, jayjay-models.de, kultmodels.com, modelwerk.de, models1.co.uk, showcast.de,

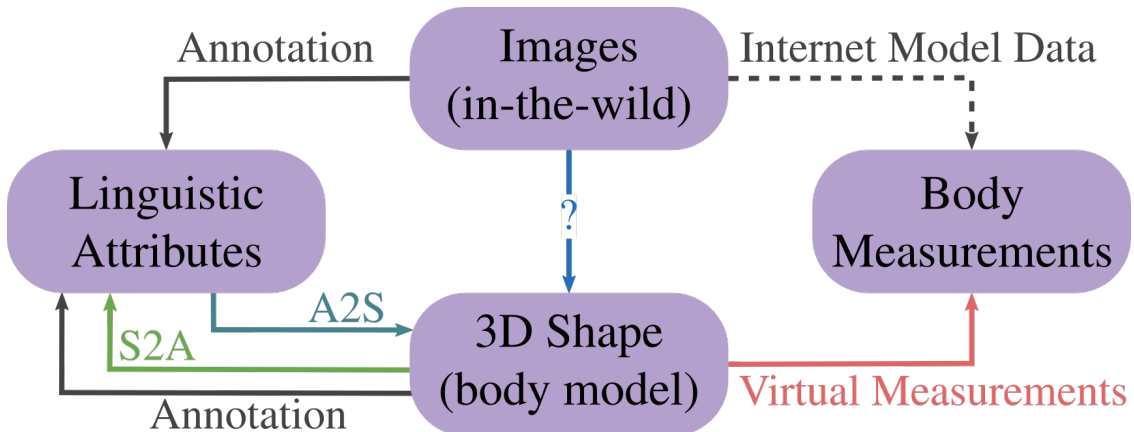


Figure 4.4: Shape representations and data collection. Our goal is 3D body shape estimation from in-the-wild images. Collecting data for direct supervision is difficult and does not scale. We explore two alternatives. **Linguistic Shape Attributes:** We annotate attributes (“A”) for CAESAR meshes, for which we have accurate shape (“S”) parameters, and learn the “A2S” and “S2A” models, to map between these representations. Attribute annotations for images can be easily crowd-sourced, making these scalable. **Anthropometric Measurements:** We collect images with sparse body measurements from model-agency websites. A virtual measurement module [155] computes the measurements from 3D meshes. **Training:** We combine these sources to learn a regressor with weak supervision that infers 3D shape from an image.

the-models.de, and ullamodels.com. In addition to photos, we store gender and four anthropometric measurements, i.e. height, chest, waist and hip circumference, when available. To avoid having the same subject in both the training and test set, we match model identities across websites to identify models that work for several agencies. For details, see Appendix C.1.1.

After identity filtering, we have 94,620 images of 4,419 models along with their anthropometric measurements. However, the distributions of these measurements, shown in Fig. 4.5, reveal a bias for “fashion model” body shapes, while other body types are under-represented in comparison to CAESAR [162]. To enhance diversity in body-shapes and avoid strong biases and log tails, we compute the quantized 2D-distribution for height and weight and sample up to 3 models per bin. This results in $N = 1,185$ models (714 females, 471 males) and 20,635 images.

4.3.3 Linguistic Shape Attributes

Human body shape can be described by linguistic shape attributes [66]. We draw inspiration from Streuber et al. [188] who collect scores for 30 linguistic attributes for 256 3D body meshes, generated by sampling SMPL’s shape space, to train a linear “attribute to

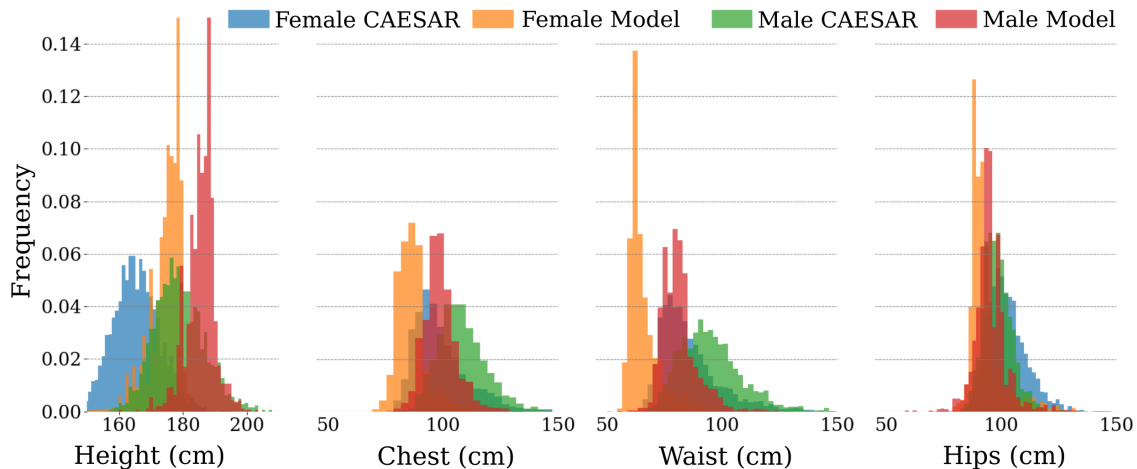


Figure 4.5: Histogram of height and chest/waist/hips circumference for data from model-agency websites (Sec. 4.3.2) and CAESAR. Model-agency data is diverse, yet not as much as CAESAR data.

shape” regressor. In contrast, we train a model that takes as input an image, instead of attributes, and outputs an accurate 3D shape (and pose).

We crowd-source linguistic attribute scores for a variety of body shapes, using images from the following sources: **Rendered CAESAR images.** We use bodies from CAESAR [162] to learn mappings between linguistic shape attributes, anthropometric measurements, and SMPL-X shape parameters, β . Specifically, we register a “gendered” SMPL-X model with 100 shape components to 1,700 male and 2,102 female 3D scans, pose all meshes in an A-pose, and render synthetic images with the same virtual camera. **Model-agency photos.** Each annotator is shown 3 body images per subject, sampled from the image pool of Sec. 4.3.2.

Annotation. To keep annotation tractable, we use $A = 15$ linguistic shape attributes per gender (subset of BodyTalk’s [188] attributes); see Tab. 4.1. Each image is annotated by $K = 15$ annotators on Amazon Mechanical Turk. Their task is to “*indicate how strongly [they] agree or disagree that the [listed] words describe the shape of the [depicted] person’s body*”. Annotations range on a discrete 5-level Likert scale. The rating choices are “strongly disagree” (score 1), “rather disagree” (score 2), “average” (score 3), “rather agree” (score 4), “strongly agree” (score 5). The layout of our CAESAR annotation task is visualized in Fig. 4.6.

We ask multiple persons to rate each body and image, to “average out” the subjectivity of individual ratings [188]. Additionally, we compute the Pearson correlation between averaged attribute ratings and ground-truth measurements. Examples of highly correlated pairs are “Big / Weight”, and “Short / Height”. To ensure good rating quality, we have several qualification requirements per participant: submitting a minimum of 5000 tasks on AMT and an AMT acceptance rate of 95%, as well as having a US residency and passing a language qualification test to ensure similar language skills and cultures

Male & Female		Male only	Female only
short	long neck	skinny arms	pear shaped
big	long legs	average	petite
tall	long torso	rectangular	slim waist
muscular	short arms	delicate build	large breasts
	broad shoulders	soft body	skinny legs
		masculine	feminine

Table 4.1: Linguistic shape attributes for human bodies. Some attributes apply to both genders, but others are gender specific.

across raters.

We use the ratings to create a rating matrix $\mathbf{A} \in \{1, 2, 3, 4, 5\}^{N \times A \times K}$, where N is the number of subjects. In the following, a_{ijk} denotes an element of \mathbf{A} .

4.4 Mapping Shape Representations

In Sec. 4.3 we list three body-shape representations: (1) SMPL-X’s PCA shape space (Sec. 4.3.1), (2) anthropometric measurements (Sec. 4.3.2), and (3) linguistic shape attribute scores (Sec. 4.3.3). Here we learn mappings between these, so that in Sec. 4.5 we can define new losses for training body shape regressors using multiple data sources.

4.4.1 Virtual Measurements (VM)

We obtain anthropometric measurements from a 3D body mesh in a T-pose, namely height, $H(\beta)$, weight, $W(\beta)$, and chest, waist and hip circumferences, $C_c(\beta)$, $C_w(\beta)$, and $C_h(\beta)$, respectively, by following Wuhrer et al. [214] and the “Virtual Caliper” [155]. For details on how we compute these measurements, see Appendix C.2.1.

4.4.2 Attributes and 3D Shape

Attributes to Shape (A2S). We predict SMPL-X shape coefficients from linguistic attribute scores with a second-degree polynomial regression model. For each shape β_i , $i = 1 \dots N$, we create a feature vector, $\mathbf{x}_i^{\text{A2S}}$, by averaging for each of the A attributes the corresponding K scores:

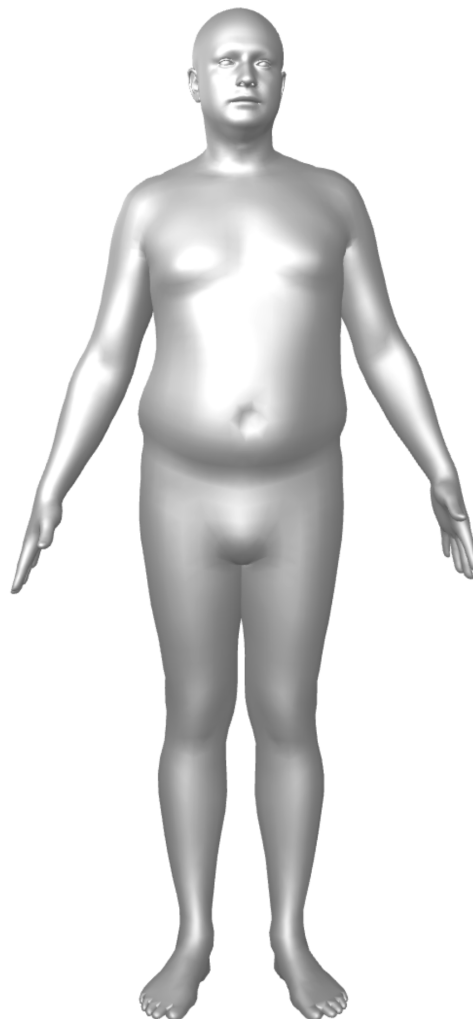
$$\mathbf{x}_i^{\text{A2S}} = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}], \quad \bar{a}_{i,j} = \frac{1}{K} \sum_{k=1}^K a_{ijk}, \quad (4.1)$$

Indicate how strongly you agree or disagree that the words describe the shape of this person's body.

Instructions: Indicate how strongly you agree or disagree that the words describe the shape of this person's body. At the end, enter a weight and age estimate of the person (best guess then hit 'submit').

You must choose one of the following options for each word:

Strongly Disagree (--), **Rather Disagree** (-), **Average** (o), **Rather Agree** (+), **Strongly Agree** (++)



-- - o + ++

Short	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Big	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Tall	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Long Torso	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Long Legs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Short Arms	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Long Neck	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Broad Shoulders	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Skinny Arms	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Average	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rectangular	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Delicate Build	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Soft Body	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Muscular	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Masculine	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please estimate the body weight in pounds:

Please estimate the age:

Figure 4.6: Layout of the AMT task for a male subject. **Left:** the 3D body mesh in A-pose. **Right:** the attributes and ratings buttons.

where i is the shape index (list of “fashion” or CAESAR bodies), j is the attribute index, and k the annotation index. We then define the full feature matrix for all N shapes as:

$$\mathbf{X}^{\text{A2S}} = [\phi(\mathbf{x}_1^{\text{A2S}}), \dots, \phi(\mathbf{x}_N^{\text{A2S}})]^\top, \quad (4.2)$$

where $\phi(\mathbf{x}_i^{\text{A2S}})$ maps \mathbf{x}_i to 2nd order polynomial features.

The target matrix $\mathbf{Y} = [\beta_1, \dots, \beta_N]^\top$ contains the shape parameters $\beta_i = [\beta_{i,1}, \dots, \beta_{i,B}]^\top$. We compute the polynomial model’s coefficients \mathbf{W} via least-squares fitting:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \varepsilon. \quad (4.3)$$

Empirically, the polynomial model performs better than several models that we evaluated; for details, see Appendix C.2.2.

Shape to Attributes (S2A). We predict linguistic attribute scores, A , from SMPL-X shape parameters, β . Again, we fit a second-degree polynomial regression model. S2A has “swapped” inputs and outputs with respect to A2S:

$$\mathbf{x}_i^{\text{S2A}} = [\beta_{i,1}, \dots, \beta_{i,B}], \quad (4.4)$$

$$\mathbf{y}_i = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}]^\top. \quad (4.5)$$

Attributes & Measurements to Shape (AHWC2S). Given a sparse set of anthropometric measurements, we predict SMPL-X shape parameters, β . The input vector is:

$$\mathbf{x}_i^{\text{HWC2S}} = [h_i, w_i, c_c, c_w, c_h], \quad (4.6)$$

where c_c, c_w, c_h are the chest, waist, and hip circumference, respectively, h and w are the height and weight, and **HWC2S** means *Height + Weight + Circumference to Shape*. The regression target is the SMPL-X shape parameters, \mathbf{y}_i .

When both *Attributes* and measurements are available, we combine them for the **AHWC2S** model with input:

$$\mathbf{x}_i^{\text{AHWC2S}} = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}, h_i, w_i, c_c, c_w, c_h]. \quad (4.7)$$

In practice, depending on which measurements are available, we train and use different regressors. Following the naming convention of **AHWC2S**, these models are: **AH2S**, **AHW2S**, **AC2S**, and **AHC2S**, as well as their equivalents without attribute input **H2S**, **HW2S**, **C2S**, and **HC2S**. For an evaluation of the contribution of linguistic shape attributes on top of each anthropometric measurement, see Appendix C.2.2

Training Data. To train the A2S and S2A mappings we use CAESAR data, for which we have SMPL-X shape parameters, anthropometric measurements, and linguistic attribute scores. We train separate gender-specific models.

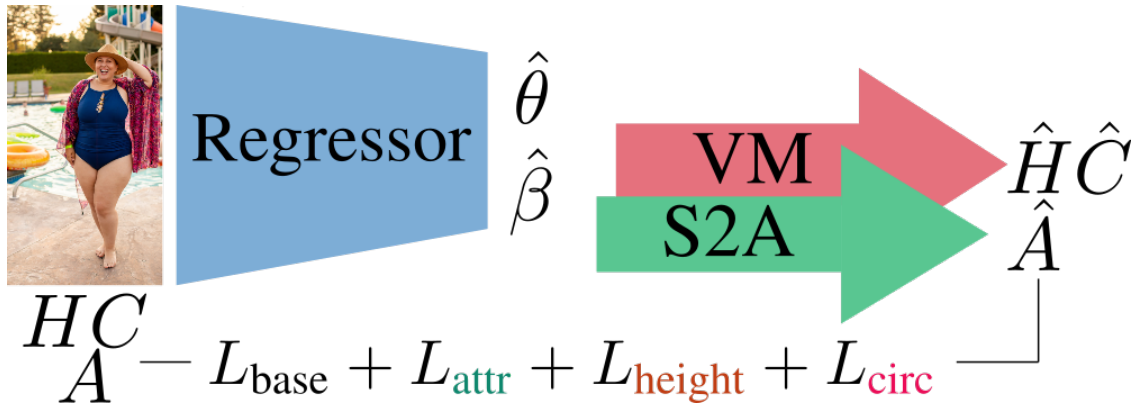


Figure 4.7: SHAPY first estimates shape, $\hat{\beta}$, and pose, $\hat{\theta}$. Shape is used by: (1) our virtual anthropometric measurement (VM) module to compute height, \hat{H} , and circumferences, \hat{C} , and (2) our S2A module to infer linguistic attribute scores, \hat{A} . There are several SHAPY variations, e.g., SHAPY-H uses only VM to infer \hat{H} , while SHAPY-HA uses VM to infer \hat{H} and S2A to infer \hat{A} .

4.5 3D Shape Regression from an Image

We present SHAPY, a network that predicts SMPL-X parameters from an RGB image with more accurate body shape than existing methods. To improve the realism and accuracy of shape, we explore training losses based on all shape representations discussed above, i.e., SMPL-X meshes (Sec. 4.3.1), linguistic attribute scores (Sec. 4.3.3) and anthropometric measurements (Sec. 4.4.1). In the following, symbols with/-out a hat are regressed/ground-truth values. We convert shape $\hat{\beta}$ to height and circumferences values $\{\hat{H}, \hat{C}_c, \hat{C}_w, \hat{C}_h\} = \{H(\hat{\beta}), C_c(\hat{\beta}), C_w(\hat{\beta}), C_h(\hat{\beta})\}$, by applying our virtual measurement tool (Sec. 4.4.1) to the mesh $M(\hat{\beta})$ in the canonical T-pose. We also convert shape $\hat{\beta}$ to linguistic attribute scores, with $\hat{A} = \text{S2A}(\hat{\beta})$.

We train various SHAPY versions with the following ‘‘SHAPY losses’’, using either linguistic shape attributes, or anthropometric measurements, or both:

$$L_{\text{attr}} = \|A - \hat{A}\|_2^2, \quad (4.8)$$

$$L_{\text{height}} = \|H - \hat{H}\|_2^2, \quad (4.9)$$

$$L_{\text{circ}} = \sum_{i \in \{c, w, h\}} \|C_i - \hat{C}_i\|_2^2 \quad (4.10)$$

These are optionally added to a base loss, L_{base} , defined below in ‘‘training details’’. The architecture of SHAPY, with all optional components, is shown in Fig. 4.7. A suffix of color-coded letters describes which of the above losses are used when training a model. For example, SHAPY-AH denotes a model trained with the attribute and height losses, i.e.: $L_{\text{SHAPY-AH2S}} = L_{\text{base}} + L_{\text{attr}} + L_{\text{height}}$.

Training Details. We initialize SHAPY with the ExPose [27] network weights and

use curated fits [27], H3.6M [75], the SPIN [101] training data, and our model-agency dataset (Sec. 4.3.2) for training. In each batch, 50% of the images are sampled from the model-agency images, for which we ensure a gender balance. The “SHAPY losses” of Eqs. (4.8) to (4.10) are applied only on the model-agency images. We use these on top of a standard base loss:

$$L_{\text{base}} = L_{\text{pose}} + L_{\text{shape}}, \quad (4.11)$$

where L_J^{2D} and L_J^{3D} are 2D and 3D joint losses:

$$L_{\text{pose}} = L_J^{2D} + L_J^{3D} + L_{\theta}, \quad (4.12)$$

$$L_{\text{shape}} = L_{\beta} + L_{\beta}^{\text{pixie}}, \quad (4.13)$$

L_{θ} and L_{β} are losses on pose and shape parameters, and L_{β}^{pixie} is PIXIE’s [39] “gendered” shape prior. All losses are L2, unless otherwise explicitly specified. Losses on SMPL-X parameters are applied only on the pose data [75, 27, 101]. For more implementation details, see Appendix C.3.

4.6 Experiments

4.6.1 Evaluation Datasets

3D Poses in the Wild (3DPW) [207]. We use this to evaluate *pose* accuracy. This is widely used, but has only 5 test subjects, i.e., limited shape variation. For results, see Appendix C.4.3.

Sports Shape and Pose 3D (SSP-3D) [174]. We use this to evaluate 3D body *shape* accuracy from images. It has 62 tightly-clothed subjects in 311 in-the-wild images from Sports-1M [89], with *pseudo* ground-truth SMPL meshes that we convert to SMPL-X for evaluation.

Model Measurements Test Set (MMTS). We use this to evaluate anthropometric measurement accuracy, as a proxy for body *shape* accuracy. To create MMTS, we withhold 2699/1514 images of 143/95 female/male identities from our model-agency data, described in Sec. 4.3.2

CAESAR Meshes Test Set (CMTS). We use CAESAR to measure the accuracy of SMPL-X body shapes and linguistic shape attributes for the models of Sec. 4.4. Specifically, we compute: (1) errors for SMPL-X meshes estimated from linguistic shape attributes and/or anthropometric measurements by A2S and its variations, and (2) errors for linguistic shape attributes estimated from SMPL-X meshes by S2A. To create an unseen mesh test set, we withhold 339 male and 410 female CAESAR meshes from the crowd-sourced CAESAR linguistic shape attributes, described in Sec. 4.3.3.

Human Bodies in the Wild (HBW). The field is missing a dataset with varied bod-

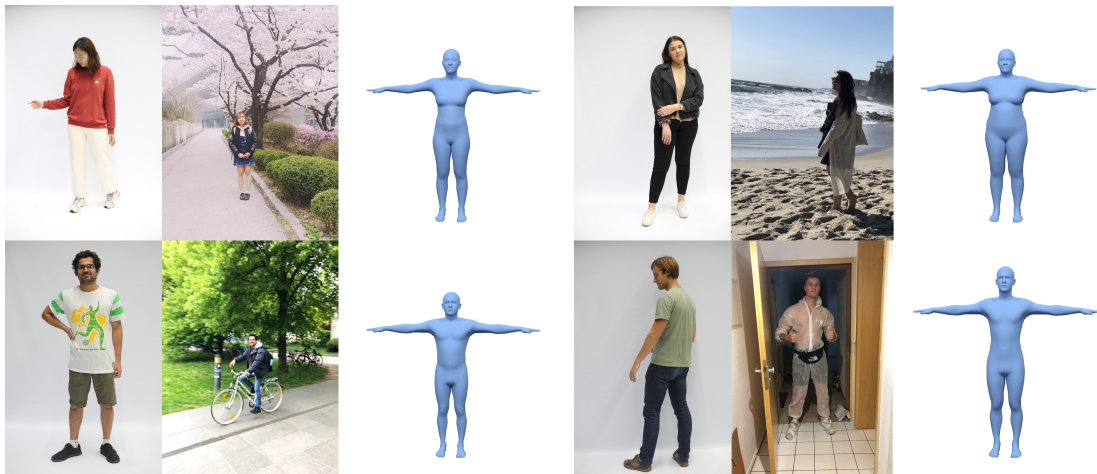


Figure 4.8: “Human Bodies in the Wild” (HBW) color images, taken in the lab and in the wild, and the SMPL-X ground-truth shape.

ies, varied clothing, in-the-wild images, and accurate *3D shape ground truth*. We fill this gap by collecting a novel dataset, called “*Human Bodies in the Wild*” (HBW), with three steps: (1) We collect accurate 3D body scans for 35 subjects (20 female, 15 male), and register a “gendered” SMPL-X model to these to recover 3D SMPL-X ground-truth bodies [152]. (2) We take photos of each subject in “photo-lab” settings, i.e., in front of a white background with controlled lighting, and in various everyday outfits and “fashion” poses. (3) Subjects upload full-body photos of themselves taken in the wild. For each subject we take up to 111 photos in lab settings, and collect up to 126 in-the-wild photos. In total, HBW has 2543 photos, 1,318 in the lab setting and 1,225 in the wild. We split the data into a validation and a test set (val/test) with 10/25 subjects (6/14 female 4/11 male) and 781/1,762 images (432/983 female 349/779 male), respectively. Figure 4.8 shows a few HBW subjects, photos and their SMPL-X ground-truth shapes. All subjects gave prior written informed consent to participate in this study and to release the data. The study was reviewed by the ethics board of the University of Tübingen, without objections.

4.6.2 Evaluation Metrics

We use standard accuracy metrics for 3D body pose, but also introduce metrics specific to 3D body shape.

Anthropometric Measurements. We report the mean absolute error in mm between ground-truth and estimated measurements, computed as described in Sec. 4.4.1. When weight is available, we report the mean absolute error in kg.

MPJPE and V2V metrics. We report in Appendix C.4.3 the mean per-joint point error (MPJPE) and mean vertex-to-vertex error (V2V), when SMPL-X meshes are avail-

	Method	P2P _{20K} (mm)	Height (mm)	Weight (kg)	Chest (mm)	Waist (mm)	Hips (mm)
Male subjects	A2S	11.1 ± 5.2	29 ± 21	5 ± 4	30 ± 22	32 ± 24	28 ± 21
	H2S	12.1 ± 6.1	5 ± 4	11 ± 11	81 ± 66	102 ± 87	40 ± 33
	AH2S	6.8 ± 2.3	4 ± 3	3 ± 3	27 ± 21	29 ± 23	24 ± 18
	HW2S	8.1 ± 2.7	5 ± 4	1 ± 1	24 ± 17	26 ± 20	21 ± 18
	AHW2S	6.3 ± 2.1	4 ± 3	1 ± 1	19 ± 15	19 ± 14	20 ± 16
	C2S	19.7 ± 11.1	59 ± 47	9 ± 8	55 ± 41	63 ± 49	37 ± 28
	AC2S	9.6 ± 4.4	25 ± 19	3 ± 3	23 ± 19	21 ± 17	18 ± 14
	HC2S	7.7 ± 2.6	5 ± 4	2 ± 2	28 ± 23	18 ± 15	13 ± 11
	AHC2S	6.0 ± 2.0	4 ± 3	2 ± 2	21 ± 17	17 ± 14	13 ± 10
	HWC2S	7.3 ± 2.6	5 ± 4	1 ± 1	20 ± 15	14 ± 12	13 ± 11
	AHWC2S	5.8 ± 2.0	4 ± 3	1 ± 1	16 ± 13	13 ± 10	13 ± 10

Table 4.2: Results of A2S variants on CMTS for male subjects, using the male SMPL-X model.

able. The prefix “PA” denotes metrics after Procrustes alignment.

Mean point-to-point error (P2P_{20K}). SMPL-X has a highly non-uniform vertex distribution across the body, which negatively biases the mean vertex-to-vertex (V2V) error, when comparing estimated and ground-truth SMPL-X meshes. To account for this, we use 20K points on SMPL-X’s mesh surface as described in Section 2.3.3 and report the mean point-to-point (P2P_{20K}) error. For details, see Appendix C.4.1.

4.6.3 Shape-Representation Mappings

We evaluate the models A2S and S2A, which map between the various body shape representations (Sec. 4.4).

A2S and its variations. How well can we infer 3D body shape from just linguistic shape attributes, anthropometric measurements, or both of these together? In Tab. 4.2 and Tab. 4.3, we report reconstruction and measurement errors using many combinations of attributes (A), height (H), weight (W), and circumferences (C). Evaluation on CMTS data shows that attributes improve the overall shape prediction across the board. For example, height+attributes (AH2S) has a lower point-to-point error than height alone. The best performing model, AHWC, uses everything, with P2P_{20K}-errors of 5.8 ± 2.0 mm (males) and 6.2 ± 2.4 mm (females). It should be emphasized that even when many measurements are used as input features, i.e. height, weight, and chest/waist/hip circumference, adding attributes still improves the shape estimate, e.g. HWC2S vs. AHWC2S.

S2A. How well can we infer linguistic shape attributes from 3D shape? S2A’s accuracy on inferring the attribute Likert score is 75%/69% for males/females; details in Appendix C.4.2.

	Method	P2P _{20K} (mm)	Height (mm)	Weight (kg)	Chest (mm)	Waist (mm)	Hips (mm)
female	A2S	10.9 ± 5.2	27 ± 21	5 ± 5	30 ± 26	32 ± 31	28 ± 22
	H2S	12.8 ± 7.0	5 ± 5	12 ± 11	93 ± 72	101 ± 88	60 ± 52
	AH2S	7.2 ± 2.8	4 ± 3	3 ± 4	27 ± 23	29 ± 28	23 ± 19
	HW2S	7.9 ± 3.2	5 ± 5	1 ± 1	25 ± 22	22 ± 18	26 ± 25
	AHW2S	6.4 ± 2.5	4 ± 3	1 ± 1	14 ± 12	14 ± 12	17 ± 14
	C2S	19.5 ± 10.8	58 ± 46	8 ± 6	54 ± 36	57 ± 42	47 ± 36
	AC2S	9.6 ± 4.3	24 ± 18	3 ± 2	18 ± 15	19 ± 16	19 ± 14
	HC2S	7.3 ± 2.8	5 ± 5	2 ± 2	19 ± 16	16 ± 14	15 ± 13
	AHC2S	6.3 ± 2.4	4 ± 3	1 ± 1	15 ± 12	14 ± 12	14 ± 12
	HWC2S	7.2 ± 2.9	5 ± 5	1 ± 1	14 ± 12	13 ± 11	14 ± 12
	AHWC2S	6.2 ± 2.4	4 ± 3	1 ± 1	11 ± 9	12 ± 10	13 ± 11
	male	A2S	11.1 ± 5.2	29 ± 21	5 ± 4	30 ± 22	32 ± 24
H2S		12.1 ± 6.1	5 ± 4	11 ± 11	81 ± 66	102 ± 87	40 ± 33
AH2S		6.8 ± 2.3	4 ± 3	3 ± 3	27 ± 21	29 ± 23	24 ± 18
HW2S		8.1 ± 2.7	5 ± 4	1 ± 1	24 ± 17	26 ± 20	21 ± 18
AHW2S		6.3 ± 2.1	4 ± 3	1 ± 1	19 ± 15	19 ± 14	20 ± 16
C2S		19.7 ± 11.1	59 ± 47	9 ± 8	55 ± 41	63 ± 49	37 ± 28
AC2S		9.6 ± 4.4	25 ± 19	3 ± 3	23 ± 19	21 ± 17	18 ± 14
HC2S		7.7 ± 2.6	5 ± 4	2 ± 2	28 ± 23	18 ± 15	13 ± 11
AHC2S		6.0 ± 2.0	4 ± 3	2 ± 2	21 ± 17	17 ± 14	13 ± 10
HWC2S		7.3 ± 2.6	5 ± 4	1 ± 1	20 ± 15	14 ± 12	13 ± 11
AHWC2S		5.8 ± 2.0	4 ± 3	1 ± 1	16 ± 13	13 ± 10	13 ± 10

Table 4.3: Results of A2S and its variations on CMTS test set, in mm or kg. Trained with gender-specific SMPL-X model.



Figure 4.9: Qualitative results from HBW. From left to right: RGB, ground-truth shape, SHAPY and Sengupta et al. [175]. For example, in the upper- and lower- right images, SHAPY is less affected by pose variation and loose clothing.

Method	Model	Height	Chest	Waist	Hips	P2P _{20K}
SMPLR [127]	SMPL	182	267	309	305	69
STRAPS [174]	SMPL	135	167	145	102	47
SPIN [101]	SMPL	59	92	78	101	29
TUCH [136]	SMPL	58	89	75	57	26
Sengupta et al. [175]	SMPL	82	133	107	63	32
ExPose [27]	SMPL-X	85	99	92	94	35
SHAPY (ours)	SMPL-X	51	65	69	57	21

Table 4.4: Evaluation on the HBW test set in mm. We compute the measurement and point-to-point (P2P_{20K}) error between predicted and ground-truth SMPL-X meshes.

Method	Model	Mean absolute error (mm) ↓			
		Height	Chest	Waist	Hips
Sengupta et al. [175]	SMPL	84	186	263	142
TUCH [136]	SMPL	82	92	129	91
SPIN [101]	SMPL	72	91	129	101
STRAPS [174]	SMPL	207	278	326	145
ExPose [27]	SMPL-X	107	107	136	92
SHAPY (ours)	SMPL-X	71	64	98	74

Table 4.5: Evaluation on MMTS. We report the mean absolute error between ground-truth and estimated measurements.

4.6.4 3D Shape from an Image

We evaluate all of our model’s variations (see Sec. 4.5) on the HBW validation set and find, perhaps surprisingly, that SHAPY-A outperforms other variants. We refer to this below (and Fig. 4.1) simply as “SHAPY” and report its performance in Tab. 4.4 for HBW, Tab. 4.5 for MMTS, and Tab. 4.6 for SSP-3D. For images with natural and varied clothing (HBW, MMTS), SHAPY significantly outperforms all other methods (Tabs. 4.4 and 4.5) using only weak 3D shape supervision (Attributes). On these images, Sengupta et al.’s method [175] struggles with the natural clothing.

In contrast, their method is more accurate than SHAPY on SSP-3D (Tab. 4.6), which has tight “sports” clothing, in terms of PVE-T-SC, a scale-normalized metric used on this dataset. These results show that silhouettes are good for tight/minimal clothing and that SHAPY struggles with high BMI shapes due to the lack of such shapes in our training data; see Fig. 4.5. Note that, as HBW has true ground-truth 3D shape, it does not need SSP-3D’s scaling for evaluation.

We show additional qualitative results in Fig. 4.10 and Fig. 4.11. Failure cases are

Method	Model	PVE-T-SC	mIOU
HMR [87]	SMPL	22.9	0.69
SPIN [101]	SMPL	22.2	0.70
STRAPS [174]	SMPL	15.9	0.80
Sengupta et al. [175]	SMPL	13.6	-
SHAPY (ours)	SMPL-X	19.2	-

Table 4.6: Evaluation on the SSP-3D test set [174]. We report the scaled mean vertex-to-vertex error in T-pose [174], and the mean intersection-over-union (mIOU).

shown in Fig. 4.12. To deal with high-BMI bodies, we need to expand the set of training images and add additional shape attributes that are descriptive for high-BMI shapes. Muscle definition on highly muscular bodies is not well represented by SMPL-X, nor do our attributes capture this. The SHAPY approach, however, could be used to capture this with a suitable body model and more appropriate attributes.

A key observation is that training with linguistic shape attributes alone is sufficient, i.e., without anthropometric measurements. Importantly, this opens up the possibility for significantly larger data collections. For a study of how different measurements or attributes impact accuracy, see Appendix C.4.2. Figure 4.9 shows SHAPY’s qualitative results.

4.7 Conclusion

SHAPY is trained to regress more accurate human body shape from images than previous methods, without explicit 3D shape supervision. To achieve this, we present two different ways to collect proxy annotations for 3D body shape for in-the-wild images. First, we collect sparse anthropometric measurements from online model-agency data. Second, we annotate images with linguistic shape attributes using crowd-sourcing. We learn mappings between body shape, measurements, and attributes, enabling us to supervise a regressor using any combination of these. To evaluate SHAPY, we introduce a new shape estimation benchmark, the “Human Bodies in the Wild” (HBW) dataset. HBW has images of people in natural clothing and natural settings together with ground-truth 3D shape from a body scanner. HBW is more challenging than existing shape benchmarks like SSP-3D, and SHAPY significantly outperforms existing methods on this benchmark. We believe this work will open new directions, since the idea of leveraging linguistic annotations to improve 3D shape has many applications.

Limitations. Our model-agency training dataset (Sec. 4.3.2) is not representative of the entire human population and this limits SHAPY’s ability to predict larger body shapes. To address this, we need to find images of more diverse bodies together with anthropometric measurements and linguistic shape attributes describing them.



Figure 4.10: Qualitative results of SHAPY predictions for female bodies.



Figure 4.11: Qualitative results of SHAPY predictions for male bodies.



Figure 4.12: Failure cases. In the first example (upper left) the weight is underestimated. Other failure cases of SHAPY are muscular bodies (upper right) and body shapes with high BMI (second row).

Social impact Knowing the 3D shape of a person has advantages, for example, in the clothing industry to avoid unnecessary returns. If used without consent, 3D shape estimation may invade individuals' privacy. As with all other 3D pose and shape estimation methods, surveillance and deep-fake creation is another important risk. Consequently, SHAPY's license prohibits such uses.

Chapter 5

Conclusion

5.1 Contributions

In this thesis, we address three problem of reconstructing 3D human pose and shape from images for poses with self- and interpersonal contact.

We start by investigating self-contact in Chapter 2, i.e. meaningful physical contact humans make with their own bodies. Such contacts are frequent in human poses, however, prior work has not paid attention to self-contact or even avoided it. To address the lack of datasets, we invent novel labels and data collection methods. In particular, we conceive discrete self-contact, a label that indicates pairwise contact between discrete regions on the human body that can be collected at scale, similar to 2D keypoints. The technique of contact labels on the 3D human mesh surface have been further developed in DECO [200] for annotating contact between human and object or scene. We also demonstrate how these labels can be used during optimization. We also devise “mimic the pose”, a novel way of data collection to create a dataset of 2D images with accurate 3D reference data. For this, we ask humans to mimic a presented 3D poses while somebody takes their photo. Since the presented and mimicked poses do not match exactly, we introduce SMPLify-XMC, a optimization routine that goes beyond existing art since it takes the presented pose *and* its self-contact into account. This is the first time, *in-the-wild* pose mimicking is used to create a dataset for human pose and shape estimation. One advantage of this approach is that complex poses or scenes which existing work fails to reconstruct can be collected at scale. For example, regressors and optimization methods often fail in the 3D reconstruction task when images have strong perspective effects. In SPEC [99], the pose mimicking approach is used to create a novel test set of real images with strong perspective effects and accurate 3D reference meshes. One part of MTP is 3DCP, a collection meshes in poses with self-contact. Previous datasets usually avoid self-contact since it breaks the registration process. We are the first to collect a large number of body scans in various poses with self-contact in minimal clothing. Beyond their use in MTP, future work can use this dataset to learn about soft-tissue deformation due to self-contact. To get more pose variety we also search in existing datasets for poses with self-contact and refine them to resolve slight intersections and encourage contact. In BEDLAM [18], AMASS motion sequences are combined with various body shapes. To resolve intersections due to this process the authors use self-contact optimization. The major insight of Section 2.5 is that self-contact is an excellent signal for the task of human pose estimation. In regressor training, we show that knowing about self-contact not only improves pose estimation for poses with self-contact but also for poses without self-contact.

In Chapter 3, we address the problem of generating and estimating two people in close proximity. To do this, we first fit SMPL-X to a collection of Flickr images with ground-truth discrete human-human contact labels [41] via optimization. Next, we train a generative model that learns the joint distribution of people in close social interaction. Previous art has predominantly focused on generating the motion of a single person. While these methods show how diffusion models can be trained on 3D joint locations,

we are the first to demonstrate how diffusion models can be trained on body model parameters, i.e. pose and shape, of two people; a technique especially relevant for human interactions involving contact. Finally, we demonstrate how diffusion models can serve as prior in human mesh fitting using an SDS loss [153]. Our method is the first optimization routine that can reconstruct two people in close social interaction form images without taking ground-truth contact labels into account.

Contact happens on the surface of the human body. The majority of research in human mesh estimation, however, focuses on body pose and predicts bodies with average or “zero” shape because the training data for body shape misses quality and diversity. The lack of comprehensive training data and labels for body shape estimation is a significant obstacle for accurate mesh regression. In Chapter 4, we address this by leveraging the rich vocabulary humans use to describe body shape. We draw inspiration from prior work that predicts a person’s body shape from linguistic body shape attributes ratings [188], and demonstrate that such information can be used to supervise body shape in end-to-end learning. We create new datasets of ratings for linguistic shape attributes for (1) 3D body shapes and (2) images taken in the wild. We use CAESAR [162] bodies and images with body measurements from model agency websites. Then we learn a model (S2A) that predicts attribute ratings from SMPL-X shape parameters. Our regressor, SHAPY, takes an image as input, predicts SMPL-X shape parameters and obtains attribute annotations using S2A. We formulate new losses for body measurements and attribute ratings to supervise body shape during training. Another obstacle in 3D shape estimation is that existing benchmarks rather focus on body pose and usually contain only a few subjects with little body shape variety or rely on pseudo-ground truth fits. To overcome this, we collect HBW, a new dataset of images taken in the wild with ground truth body shape, i.e. 3D body scans. On our new benchmark, SHAPY predicts more accurate body shape than previous art. More recent methods, published after SHAPY, use synthetic data to estimate body shape [18], but SHAPY is still more accurate on HBW.

5.2 Considerations for Future Work

The recent advances in human pose estimation are impressive, especially when the 3D mesh is projected onto the images, i.e. it is seen from the camera perspective. However, there is still much work to be done to achieve expressive 3D reconstruction of multiple people in poses with self- and interpersonal contact. In particular, because one great potential of SMPL lies beyond pose and shape estimation by using mesh reconstructions for downstream applications in other scientific fields. For example, through precise estimates of human interaction, computer vision methods have the potential to significantly enhance our comprehension of human behaviour. To achieve these long-term goals, the next step is to combine the three topics presented in this theses to enable multi-person mesh regression with accurate estimates of body shape and self-contact and human-human contact. This goal requires an approach similar to BEV [190] or

SLAHMR [225] that also considers the diverse range of body shapes and accounts for both self-contact and human-human interaction.

Level of detail in self- and human-human contact 3D datasets

Future regressors and datasets should use expressive body models like SMPL-X or later versions. Even with recent advancements on data collection for self- and human-human contact, there is still a lack of accurate hand pose estimation and subtle contact, such as touch between hands or between the hand and face. Even everyday actions like rubbing one’s eyes or touching our hair are hardly explored until very recently [182]. While not all human pose estimation tasks may require this level of detail, it is worth considering that in the ongoing pursuit of creating avatars, it might be the nuanced, playful elements humans naturally use in their body language, that make avatars appear truly natural and human-like. To model and reconstruct such subtle contact, we need tools to label contact in images taken in the wild e.g. on vertex level, and high resolution MoCap with hand pose and facial expressions reconstruction. Since some self-contact interactions can be explained by a single term, e.g. “rubbing eyes”, action labels or language descriptions can serve as prior or guidance in reconstruction tasks. Yet, the detailed pose and exact contact type and locations are difficult to describe with text and must be learned from visual input.

From two to multiple people in contact

One strength of BUDDI is that it is trained on model parameters which are of much lower dimension than meshes. Therefore, BUDDI could easily be extended to multiple individuals by increasing the numbers of input tokens to the transformer encoder. However, achieving this requires creating a suitable dataset for training. Optimization methods that use discrete contact labels can also be extended to multiple people. This task requires to resolve self- and pairwise human-human intersections, which necessitates knowing the signed distance of e.g. vertices. TUCH and BUDDI utilize winding numbers to identify interpenetrating vertices; a memory-intensive method despite its relative speed. Therefore, it becomes crucial to develop methods that are not only efficient but also utilize minimal memory resources in order to effectively address the challenge of encouraging contact while resolving intersections.

While individuals may make contact with multiple people in 3-4 person settings, the most common scenarios where multi-person contact can be observed are sports scenes or events with many participants like concerts. Since people usually occlude each other in such scenes, collecting discrete contact labels is challenging and keypoint detectors often fail. Thus, the attempt of expanding the data collection from two to multiple people might fail and may require new ideas. One possible solution are iterative approaches: BUDDI could be conditioned on one person while generating the second person. This can be applied in an iterative manner until a group of interacting people is generated.

Novel human pose and shape regressors are trained on synthetic data only [18] and the generated group of interacting people could be useful for creating more challenging versions of such data.

Level of detail in self- and human-human contact 2D datasets

Discrete contact labels can be scaled and are easy to collect for images in the wild. Yet, the level of detail for contact is constrained to a priori defined regions on the body. To address this, we can collect more detailed self- and human-human contact via novel labeling tools like the one used by Tripathi et al. [200]. Another approach would be to use pose mimicking for human interaction. While pose mimicking is more difficult to scale, it does create additional value because a pair of 2D image and 3D reference pose is aligned with the true human perception of other people’s poses.

Datasets of 3D humans in contact

MoCap and scan data are valuable since they offer high quality 3D information. Yet, there is only few publicly available 3D datasets of closely interacting people. The interactions in these datasets are limited to canonical contacts, like a “hug” or “handshake”, or people dancing and only performed by a few subjects. To go beyond canonical contact, we need to put people with different relation, e.g. couples vs. colleagues, into authentic scenarios from which contact arises in a natural way.

Self-contact, body shape, and soft tissue deformation

SHAPY focuses on addressing the challenge of body shape regression from images captured in various real-world scenarios, while TUCH specializes in estimating people in poses involving self-contact. A method that predicts body shape *and* self-contact poses, will be constrained by the rigidity of SMPL. This limitation is typically negligible for individuals with slender body types, as their bodies tend to have minimal self-intersections even in poses involving self-contact. However, for heavier individuals, self-contact poses often result in strong soft-tissue deformations, leading to self-intersections even in simple standing poses. Utilizing a rigid body model and uniform losses for interpenetration for all body shapes would introduce a bias in pose and motion, particularly for individuals with heavier bodies. This issue could be addressed by either incorporating models that account for soft-tissue deformation or by capturing a broader range of body shapes, coupled with improved motion and pose priors that also take human body shape into account.

Larger and more diverse in-the-wild datasets

While scan data and MoCap data are valuable resources, they are often limited to controlled lab environments and acted scenarios. To achieve accurate pose estimation for

poses with contact, we need *large* image datasets that encompass a truly *diverse* range of internet images, accompanied by discrete labels. Although datasets like DeepFashion (used in TOUCH), Flickr (used in BUDDI), and fashion model images (used in SHAPY) contain useful information, they suffer from biases toward specific poses and body shapes. DeepFashion primarily consists of fashion models, with body shapes common in fashion industry, facing the camera to show and sell clothing items. Although Flickr appears to be a diverse datasource at first glance, we find many images showing individuals smiling towards the camera. This lack of diverse actions, interactions, emotional states etc. hampers the dataset’s ability to capture the full range of touch. The internet model image collection includes curvy and petite models, but it fails to capture the full spectrum of body shape. While addressing variations in height and weight might alleviate the most obvious biases, it is not enough to learn the full variety of shape. For example, changes of body shape due to age or more subtle difference within the group of people of similar height/weight measurements.

Facial expressions and gaze

In this work, we focus on the human body and less on other signal humans use to communicate like facial expressions or eye contact. In many cases, physical contact is accompanied by such signals. Existing methods can effectively predict a persons facial expression and could be combined it with human pose and shape regressors. Additionally, there is an interesting connection between gaze and touch. For example, during a handshake, we look into the other person’s eyes. These are social cues that could be used as conditional or prior knowledge during network training to jointly improve human-human contact and gaze estimation.

Extension in time

In this work, we demonstrate how to estimate poses with contact and body shape from images. A natural extension of our ideas is to apply them in the video domain which already happened for human pose estimation without considering contact. This is not necessarily easy, since the distributions of pose differs between static images and videos, e.g. in photos we often find people statically posing for the picture facing the camera, whereas in video people are moving. This means we find many self-contact poses in static images that people actively take when knowing they need to ‘pose’ for a photo. However, these poses do usually not appear in videos of motion sequences collected in the wild (except the rare cases when a third person records the process of moving into a photo pose). For a proper extension in time domain a larger dataset of at least 1 million images (static and single frames from video) annotated with self-contact labels is necessary to train models similar to 2D keypoint detectors but for self-contact. The human pose and shape estimation community would greatly benefit from such a dataset and model, because even large synthetic datasets like BEDLAM [18] avoid strong self-

contact, because the data creation pipeline breaks when clothes need to be fit between touching surfaces. Discrete labels for images taken in the wild would enable training human mesh regressors like the recent HMR 2.0 [47] on large video datasets with detected keypoints and self-contact supervision.

5.3 Closing Thoughts

In this thesis, we address the problem of estimating poses with self-contact or contact between humans and more accurate body shape, i.e. we reconstruct the full surface of the human body. In general, the field progresses rapidly and recent research demonstrates remarkable results in human mesh regression from images [47, 18]. Nevertheless, the challenge of reconstructing meshes with precise self- and interpersonal contact remains unsolved, particularly when employing neural networks. However, given the speed of progress, we should see regressors to 3D mesh reconstruction with hands, faces, gaze, body shape, and for multiple people within two years. This will open up a whole branch of downstream applications in industry and academia. For example, VR-based medical products to e.g. recover motion after a stroke or accident, let patients practise and train human touch after traumatic experiences, or for medial staff to practise specific hand grips in extreme situations. We can think of applications in the metaverse or of social robots that not only verbally but also physically interact with human beings. In research, we can utilize the scalability of human pose and shape reconstruction to better understand human social behaviour. The manifold contributions in this thesis, i.e. novel labels, losses, ways of capturing data, and models, offer a rich and extensive toolbox and pave the way for accurate human pose and shape reconstruction of multiple people in interactions with self- and interpersonal contact.

Appendix

Appendix A

On Self Contact and Human Pose

A.1 Self-Contact Datasets

A.1.1 3D Contact Pose (3DCP) Meshes

3DCP Mocap.

Sampling meshes from AMASS. First, each MoCap sequence is sampled at half of its original frame rate. For each sampled mesh, we compute the contact maps \mathcal{C}^D with $t_{eucl} = 3\text{cm}$, $t_{geo} = 30\text{cm}$ and $K = 98$. The regions are visualized in Fig. 2.2. We select only one pose for each unique signature, while ignoring contact when it occurs in more than 1% of the data. We obtain a subset of 20,114 poses.

Self-Contact Optimization. Here we provide details of the self-contact optimization for body meshes from the AMASS dataset. In this optimization, vertex pairs in M_C are further pulled together via a contact term L_C and vertices inside the mesh are pushed to the surface via a pushing term L_P , while L_O ensures that vertices far away from contact regions stay in place. Note that L_P and L_C are slightly different from the loss terms in the main corpus. L_H is a prior for contact between hand and body and L_A aligns the vertex normals when contact happens.

Given the set of vertices V of mesh M , $V_E \subset V$ denotes the subset of vertices affiliated with extremities, $V_I \subset V$ denotes the subset of vertices inside the mesh, and $V_{EI} = V_E \cap M_I$ denotes the vertices of extremities that are inside the mesh itself and V_{EI}^c its complement. We identify vertices inside the mesh using generalized winding numbers [76]. $V_{V_H} \subset V$ is the subset of hand vertices. Note that we make SMPL-X watertight by closing the back of the mouth. V_C is computed following Definition 3.1 in the main chapters with $t_{geo} = 30\text{cm}$ and $t_{eucl} = 3\text{cm}$ and $V_G(v) = \{u | geo(v, u) > t_{geo}\}$. Given an initial mesh \tilde{I} , we aim to minimize the objective function

$$L(\theta_b, \theta_{h_l}, \theta_{h_r}) = \lambda_C L_C + \lambda_P L_P + \lambda_H L_H + \lambda_O L_O + \lambda_A L_A + \lambda_{\theta_h} L_{\theta_h} + \lambda_{\theta} L_{\theta}, \quad (\text{A.1})$$

where θ_h denote the hand pose vector of the SMPL-X model. Further,

$$L_C = \frac{1}{|V_{EI}^c|} \sum_{v \in M_{EI}^c} a \alpha \tanh\left(\frac{f_g(v)}{\alpha}\right),$$

$$L_P = \frac{1}{|V_{EI}|} \sum_{v \in V_{EI}} \gamma_1 \tanh\left(\frac{f_g(v)}{\gamma_2}\right), \text{ and}$$

$$L_H = \frac{1}{|V_{V_H}|} \sum_{v \in V_{V_H}} \delta_1 h_{v_i} \tanh\left(\frac{f_g(v)}{\delta_2}\right),$$

where f_g denotes a function, that for each vertex v finds the closest vertex in self contact u , or mathematically $f_g(v) = \min_{u \in V_G(v)} \|v - u\|_2$. h_{v_i} denotes the weight per hand vertex from the hand-on-body prior L_H as explained below, if v_i is outside, otherwise $h_{v_i} = 1$.

Further, $a = (\min_{u \in \mathcal{U}(M_C)} \text{geo}(v_i, u) + 1)^{-1}$ is an attraction weight. This weight is higher, for vertices close to vertices in contact of \tilde{I} . L_θ is a L_2 prior that penalizes deviation from the initial pose and L_{θ_h} defines an L_2 prior on the left and right hand pose using the a low-dimensional hand pose space. $\alpha = 0.04$, $\gamma_1 = 0.07$, $\gamma_2 = 0.06$ define slope and offset of the pulling and pushing terms. For the hand-on-body-prior we use $\delta_1 = 0.023$, and $\delta_2 = 0.02$ if v_i is inside and $\delta_1 = \delta_2 = 0.01$ if v_i is outside the mesh.

Self-contact optimization aims to correct interpenetration and encourage near-contact vertices to be in contact by slightly refining the poses around the contact regions. Vertices that are not affected should stay as close to the original positions as possible. In L_O , the displacement of each vertex from its initial position is weighted by its geodesic distance to a vertex in contact. Given \tilde{v} denoting the position of vertex i of \tilde{I} , the outside loss term is

$$L_O = \delta_2 \sum_{v \in V} \min_{u \in \mathcal{U}(M_C)} \text{geo}(v, u)^2 \|v - \tilde{v}\|_2,$$

where $\min_{u \in \mathcal{U}(V_C)} \text{geo}(v, u) = 1$ if $V_C = \emptyset$ and $\delta_2 = 4$. Lastly, we use a term, L_A , that encourages the vertex normals $N(v)$ of vertices in contact to be aligned but in opposite directions:

$$L_A = \frac{1}{|M_C|} \sum_{(v,u) \in V_C} 1 + \langle N(v), N(u) \rangle.$$

Hand-on-Body Prior. Hands and fingers play an important role as they frequently make contact with the body. However, they have many degrees of freedom, which makes their optimization challenging. Therefore, we learn a hand-on-body prior from 1279 self contact registrations. For this, we use only poses where the minimum point-to-mesh distance between hand and body is $< 1\text{mm}$. These are 718 and 701 poses for the right and left hand, respectively. Since left and right hand are symmetric in SMPL-X, we unite left and right hand poses. Across the 1429 poses, the mean distances per hand vertex to the body surface, $d_m(v_i)$ ranges per vertex from 1.79 to 5.52 cm, as visualized in Fig. A.1. To obtain the weights h_{v_i} in L_H , we normalize $d_m(v_i)$ to $[0, 1]$, denoted as $s(d_m(v_i))$, and obtain the vertex weight by $h_{v_i} = -s(d_m(v_i)) + 1$.

A.1.2 Mimic-The-Pose (MTP) Data

AMT task details. It can be challenging to mimic a pose precisely. To simplify the process for workers on AMT, we give detailed instructions, add thumbnails to compare the own image with the presented one and, most importantly, highlight the contact areas. To gain more variety, we also request that participants make small changes in the environment for each image, e.g. by rotating the camera, changing clothes, or turning lights on/off. We also ask participants to mimic the global orientation of the center image. For more variety in global orientation, we vary body roll from -90° to 90° in 30° steps, resulting in seven different presented global orientations. For example, in the first and third row of Fig. 2.7, the center image shows the presented pose from a frontal view.

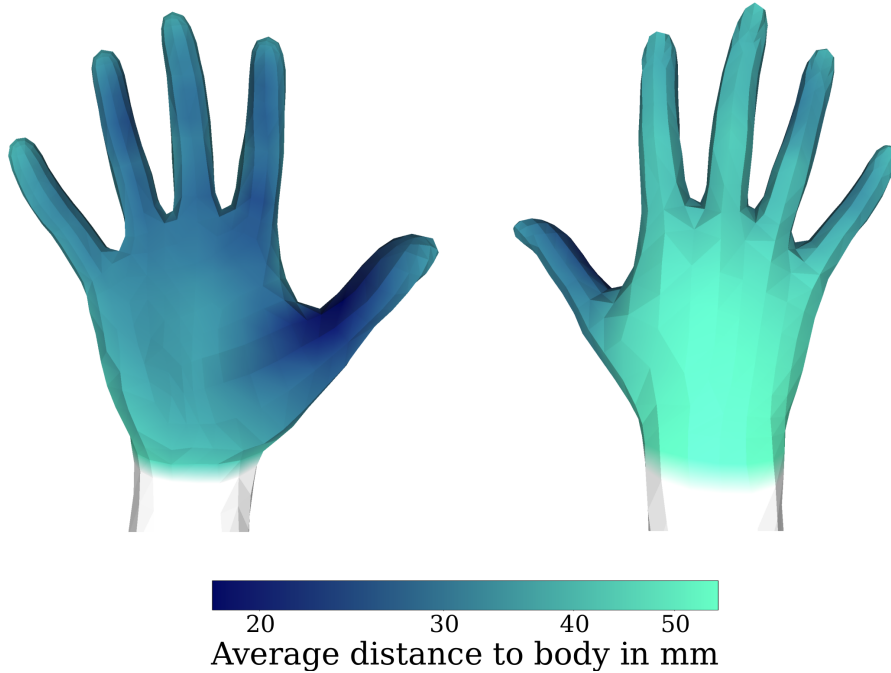


Figure A.1: Hand on body prior. Dark blue indicates small distances to body on average across all registrations where hands are close to the body. The prior is identical for left and right hand.

In the second and fourth row, the center body has different orientations. We also ask participants for their height, weight, and gender (M, F, and Non-Binary).

MTP Dataset Details. We sample meshes from 3DCP Scan, 3DCP Mocap, and AGORA [143] to comprise the presented meshes in MTP dataset. In total, we present 1653 different meshes, from which 1498 (90%) are contact poses following Definition 3.1 in the main document. Of the 1653 meshes, 110 meshes are from 3DCP Scan, 1304 meshes are from 3DCP Mocap, and 159 are from AGORA. We collect at least one image for each mesh. From the 3731 collected images, 3421 (92%) images show a person mimicking a contact pose. Figure A.2 shows how many image we collected per subset.

SMPLify-XMC Details. We notice that the presented global orientation is not always mimicked well. For example, in row 4 of Fig. 2.7 the presented global orientation has a 60 degree rotation, whereas the mimicked image is taken from a frontal view. To better initialize the optimization, we select the best body orientation, ϕ , among the seven presented ones based on their re-projection errors; then we compute the camera translation by again minimizing the re-projection error. We set the initial focal length, f_x and f_y , to 2170, which is the average of available EXIF data. These values, along with mean shape and presented pose are used to initialize the optimization.

In addition, SMPL and SMPL-X have not been trained to avoid self intersection.

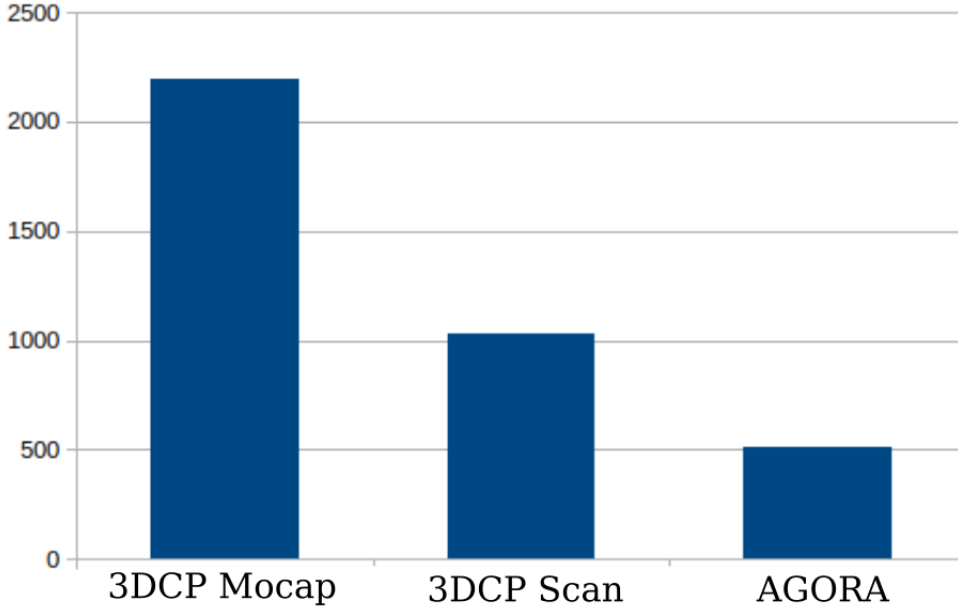


Figure A.2: Image count in MTP Dataset per 3DCP subset.

Therefore, we identify seven body segments that tend to intersect themselves, e.g. torso and upper arms (see Fig. A.3). We test each segment for self intersection and thereby filter irrelevant intersections from V_I .

A.1.3 Discrete Self-Contact (DSC) Data.

Image selection. Discrete self-contact annotation may be ambiguous and we find some annotations that we do not consider to be functional self-contact. For example, in Fig. A.4, some annotators label the left lower arm and left upper arm to be in contact, because of the slight skin touching at the elbow; we do not treat these as in self-contact. Therefore, we leverage the kinematic tree structure provided by SMPL-X and, in order to train TOUCH, ignore the following annotations: left hand - left lower arm, left lower arm - left elbow, left lower arm - left upper arm, left elbow - left upper arm, left upper arm - torso, left foot - left lower leg, left lower leg - left knee, left lower leg - left upper leg, left knee - left upper leg, right hand - right lower arm, right lower arm - right elbow, right lower arm - right upper arm, right elbow - right upper arm, right upper arm - torso, right foot - right lower leg, right lower leg - right knee, right lower leg - right upper leg, right knee - right upper leg.

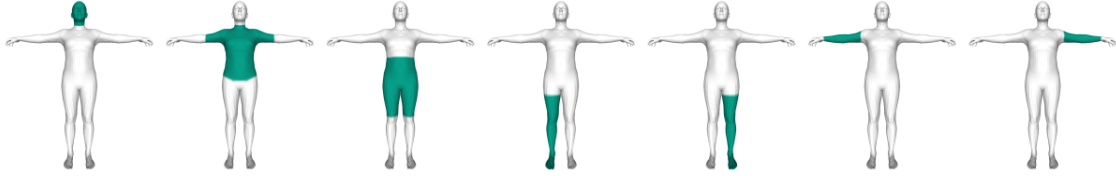


Figure A.3: In addition, SMPL and SMPL-X have not been trained to avoid self intersection. Therefore, we identify seven body segments that tend to intersect themselves, e.g. torso and upper arms (see Fig. A.3). We test each segment for self intersection and thereby filter irrelevant intersections from V_I . regions where intersection can happen, since SMPL and SMPL-X are not trained to avoid self intersection. Per segment, we create closed meshes that allow for individual intersection tests. For self-contact, intersections that happen within a segment are not relevant. The hands are not included in any segment, because self intersections within hands or between hands and lower arm are not plausible and need to be resolved.



Figure A.4: Discrete self-contact can be challenging to annotate. Here we show a few example images that are annotated as having discrete self-contact between the left upper and lower arm (yellow circle). In the last two images, however, the upper and lower arm are barely touching. We do not consider these to be in self-contact. Another ambiguous case, this time due to occlusion, are the two legs in the first image. An annotator can only assume that the shin and calf are touching, based on semantic knowledge about human pose.



Figure A.5: RGB images from 3DCP Scan Scan test set. A subject performing a pose with self-contact in a 3D body scanner.

A.2 TUCH

Here we provide details of the SMPLify-XMC and SMPLify-DC methods and how we apply them on MTP and DSC data respectively.

SMPLify-XMC is applied, before the training, to all MTP images to obtain gender-specific pseudo ground-truth SMPL-X fits. To use these fits for TUCH training, two pre-processing steps are necessary. First, they are converted to neutral SMPL fits. Second, we transform the converted SMPL fits to the camera coordinate frame estimated during SMPLify-XMC. This is necessary since SPIN assumes an identity camera rotation matrix. After that, the data is treated as ground truth during training, which means we apply the regressor loss directly on the converted SMPL pose and shape parameters without in-the-loop fitting. On the contrary, SMPLify-DC is applied during TUCH training to images with discrete self-contact annotations. We run 10 iterations of SMPLify-DC for each image in a mini batch.

MTP and the DeepFashion subset of DSC do not have ground-truth 2D keypoints but we find OpenPose detections good enough in both cases. For the 2D re-projection loss, we use ground-truth keypoints (if available) and OpenPose detections weighted by the detection confidence.

Implementation details. We initialize our regression network with SPIN weights [101]. We use the Adam optimizer [92] and a learning rate of $1e - 5$.

	MPJPE ↓	PA-MPJPE ↓
SPIN	96.9	59.2
TUCH (MTP)	88.7	57.4
TUCH (MTP+DSC)	84.9	55.5

Table A.1: Ablation of MTP data and DSC data.

A.3 Evaluation

3DCP Scan test images. During the scanning process when creating 3DCP Scan, we also take RGB photos of subjects being scanned, as shown in Figure A.5. These images have high-fidelity ground-truth poses and shapes from the registration process, making them a good test set for evaluation purposes. It is worth noting again that TUCH has never seen these images or subjects, but the contact poses were mimicked in creation of MTP, which is used in training TUCH.

TUCH. In Fig. 2.14 we visualize the improvement of TUCH over SPIN qualitatively. One can see that TUCH reconstructs bodies with better self-contact and less interpenetration (row 1 and row 2). Fig. 2.15, on the other hand, shows examples where SPIN is better than TUCH. Four of the images in Fig. 2.15 do not show the full body (rows 3, 4, 5, and 8). A possible reason why SPIN is better than TUCH in these cases is that MTP images always show the full body of a person, thus TUCH could be more sensitive to occlusion than SPIN.

We also evaluate the contribution of MTP data by finetuning SPIN only with it. The results are reported in Table A.1, where TUCH (MTP+DSC) is the same as reported in Chapter 2. This experiment shows that MTP data alone is already sufficient to significantly improve state-of-the-art (SOTA) methods on 3DPW benchmarks. This suggests that the MTP approach is a useful new tool for gathering data to train neural networks.

Appendix B

Generative Proxemics: A Prior for 3D Social Interaction from Images

B.1 Method

Including children. Since SMPL [123] only models adult body shapes, most human pose and shape regressors do not consider child body shapes explicitly. However, we found that FlickrCI3D Signatures does include images of children (roughly 10% of the images). Following the SMPLA [143] convention, BEV also estimates a scale parameter s , which is used to interpolate between SMPL [123] (adult model) and SMIL [65] (infant model) for the template meshes and shape blend shapes. A scale value of $s = 0.0$ is equivalent to SMPL only, a scale value of $s = 1.0$ is equivalent to SMIL only and all the values in between model intermediate stages. To extend this from SMPL to SMPL-X, we use the scale parameter estimated by BEV to interpolate between the SMPL-X and the SMIL template and shape blend shapes in SMPL-X topology. We visually found that this interpolation works well for $s \leq 0.8$, so we exclude pairs where the detected scale is $s > 0.8$ for one of the interaction partners.

B.1.1 Preprocessing

Matching input detections. As input, we have the estimated 3D bodies from BEV [190] and we have a dataset of ground-truth human-human contacts. The bodies in these two data sources are not in correspondence. To generate the pseudo-ground truth, we must first automatically put them in correspondence so that we can optimize the BEV bodies by exploiting the contact information.

In particular, we have (1) detected meshes from BEV, (2) 2D keypoint detections from ViTPose [221], and (3) ground-truth bounding boxes indicating the interacting pair of humans. We observed that the ground-truth bounding boxes typically match with the bounding boxes surrounding OpenPose [23] keypoint detections. As a result, we only need to correspond the OpenPose detections with ViTPose detections and the BEV bodies. Since we can reproject the 3D joints from BEV bodies to 2D keypoints, both correspondence problems require us to solve the assignment between sets of 2D keypoints. To do this, we compute a keypoint-cost matrix taking the detection confidence scores into account. We only consider keypoints with confidence score greater than 0.6 (for BEV all keypoints have by default a score of 1.0 due to the amodal prediction of the human body). We make assignments in a greedy way, while also setting a threshold (0.008) to discard matches with large matching distance.

Merging keypoints. Qualitatively, we found that ViTPose performs better than OpenPose, particularly for people that are heavily occluded. Since ViTPose (unlike OpenPose) does not detect keypoints on the feet, we can extend the ViTPose body detections with feet keypoints detected by OpenPose. We perform this extension only if the L2 distance between ViTPose and OpenPose ankles is less than 5 pixels. Additionally, since many images in FlickrCI3D Signatures include people that are truncated below the waist, we often have missing or wrong keypoint detections for the lower body. Because

of this, we use the projected BEV ankle joints, when the ankle keypoint detection confidence score is less than 0.2. Finally, the original keypoint values k_{orig} are normalized by the keypoint bounding box size via $k = k_{\text{orig}} / (\max(\text{bb}_{\text{height}}, \text{bb}_{\text{length}}) * 512)$. These steps give us a set of 2D keypoints that we use to generate the pseudo-ground truth fits using optimization.

SMPL to SMPL-X body shape conversion. Our method takes BEV estimates as input and optimizes them to fit the image evidence. Since BEV estimates meshes in SMPL topology and the ground-truth contact maps are provided in SMPL-X format, we transfer the BEV estimate to SMPL-X. Ideally, one would fit SMPL-X to SMPL via optimization. This process is time consuming and we found that it is sufficient to initialize the optimization routine with SMPL pose parameters. For body shape, we solve for SMPL-X body shape using a simple least-squares optimization. The shaped vertices, V_{SMPL} and $V_{\text{SMPL-X}}$, are obtained via

$$\begin{aligned} V_{\text{SMPL}} &= T_{\text{SMPL}} + D_{\text{SMPL}}\beta_{\text{SMPL}}, \text{ and} \\ V_{\text{SMPL-X}} &= T_{\text{SMPL-X}} + D_{\text{SMPL-X}}\beta_{\text{SMPL-X}}, \end{aligned} \tag{B.1}$$

where T_{SMPL} and $T_{\text{SMPL-X}}$ are the SMPL and SMPL-X template meshes, D_{SMPL} and $D_{\text{SMPL-X}}$ the shape blend shapes, and β_{SMPL} and $\beta_{\text{SMPL-X}}$ the shape parameters. Only $\beta_{\text{SMPL-X}}$ is unknown. Since the topology between SMPL and SMPL-X is different, we use a SMPL-to-SMPL-X vertex mapping $M \in \mathbb{R}^{10475 \times 6890}$, such that $D_{\text{SMPL-X}} = MD_{\text{SMPL}}$. Then we can directly solve for body shape, $\beta_{\text{SMPL-X}}$, in a least-squares manner:

$$\beta_{\text{SMPL-X}} = (D_{\text{SMPL-X}}^T D_{\text{SMPL-X}})^{-1} D_{\text{SMPL-X}}^T M D_{\text{SMPL}} \beta_{\text{SMPL}}$$

B.1.2 Optimization

In Table B.1 we define the weights of each loss term. Every optimization runs for a maximum of 1000 iterations per stage. For termination, we use early stopping and we keep track of the loss value at the latest 10 iterations. We use these values to fit a line with linear regression $f(x) = ax + b$ and terminate if $a < -1e - 4$. If run for two stages, the second stage’s reference poses, θ_0 , which are used in $L_{\hat{\theta}}$, are taken to be the output / last pose of the first stage. We provide pseudo code below in **Listing 1** showing the BUDDI optimization routine.

B.1.3 Diffusion model

Transformer architecture. To embed each body model parameter x_{ij} of person $j \in \{1, 2\}$ and parameters $i \in \{\phi, \theta, \beta, \gamma\}$ of size d_i in the latent space dimension $d_l = 152$,

	λ_{J2D}	$\lambda_{\bar{\theta}}$	λ_{θ}	λ_{β}	λ_{C^B}	$\lambda_{d_{\min}}$	λ_P	$\lambda_{\theta_{\text{BUDDI}}}$	$\lambda_{\gamma_{\text{BUDDI}}}$	$\lambda_{\beta_{\text{BUDDI}}}$	λ_{VAE}
Flickr Fits	0.04/0.1	200/200	4/4	40/0	10/10	0/0	0/1000	0/0	0/0	0/0	0/0
BUDDI	0.02/0.02	200/200	0/0	0/0	0/0	0/0	0/10	100/100	10/10	1e5/1e5	0/0
VAE	0.02/0.1	200/200	2/2	40/0	0/0	0/0	0/0.1	0/0	0/0	0/0	1/1
Heuristics	0.02/0.1	200/200	2/2	40/0	0/0	1e5/1e5	0/0.1	0/0	0/0	0/0	0/0
Heuristics (a)	0.04/0.1	200/200	4/4	40/0	0/0	1e5/1e5	0/1000	0/0	0/0	0/0	0/0
Heuristics (b)	0.02/0.02	200/200	4/4	40/0	0/0	1e5/1e5	0/10	0/0	0/0	0/0	0/0

Table B.1: **Weights of the different loss term during the optimization.** We consider the case of using pseudo-ground truth contact maps, the heuristics, and BUDDI. Optimizations with BUDDI and pseudo-ground truth are run for two stages. The optimization with heuristics converges quickly so a single stage is enough.

we use linear-SiLU-linear sequences:

$$f_{ij}(x_{ij}) = \text{SiLU}(x_{ij}A_{ij}^T + b_{ij})B_{ij}^T + c_{ij},$$

where $A_{ij} \in \mathbb{R}^{d_l \times d_i}$, $b_{ij} \in \mathbb{R}^{d_l}$, $B_{ij} \in \mathbb{R}^{d_l \times d_l}$, and $c_{ij} \in \mathbb{R}^{d_l}$. After passing these parameters through the transformer, we again use a linear-SiLU-linear sequence to project them back into their original dimension d_i . When BUDDI is trained with BEV [190] conditioning, we embed the conditioning in a similar fashion as the ground truth parameters, concatenate them along the token dimension, and add per-person and per-parameter embedding layers. In Fig. B.1, we show the design of our conditional model.

B.2 Training and Testing Datasets

B.2.1 Flickr Fits

We split the Flickr [41] training images into training and validation sets and use the provided test split for testing. Fits can be noisy for example, when the assignment between contact annotations and keypoints is wrong or when keypoint detectors fail badly. To provide a reliable test set for 3D pose for images taken in the wild, we manually curate the Flickr Fits test set and detect 24 out of 1427 noisy fits. The final curated Flickr Test dataset contains 1403 interactions. We do not curate the training dataset. We further evaluate the optimization method with ground truth contact maps on CHI3D (53/50mm PER-PERSON PA-MPJPE and 80mm JOINT PA-MPJPE) and on FlickrCI3D Signatures (45/87/97/99/100 PCC for radius 5/10/15/20/25).

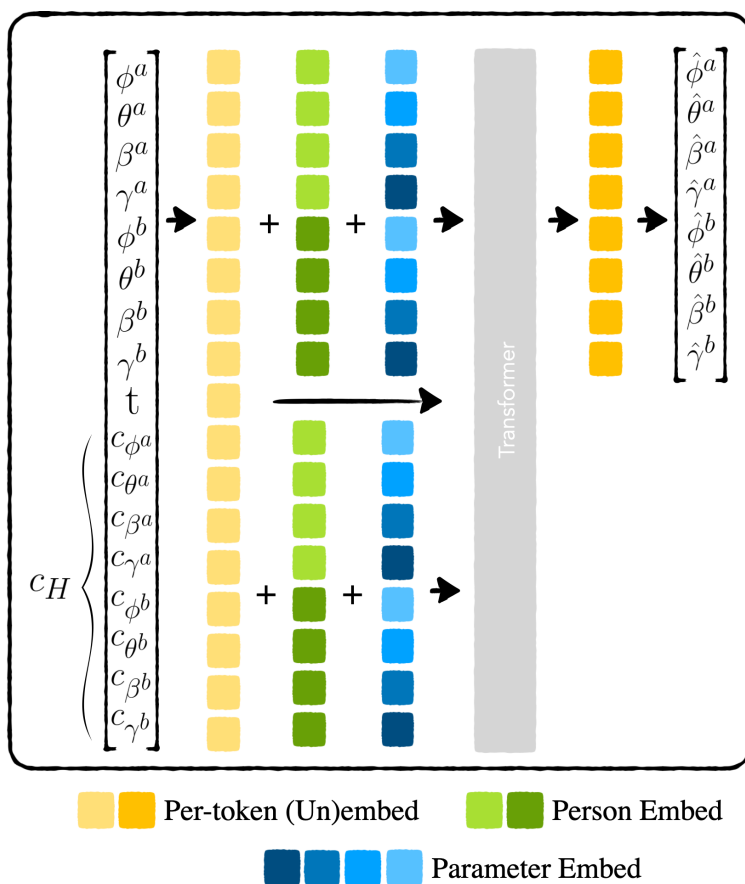


Figure B.1: **Detailed architecture of BUDDI with conditioning.** When BUDDI is conditioned on model parameters, c_H , detected from BEV [190], we concatenate the detected parameters (body global orientation, pose, shape, and translation for person a/b), with the input parameters along the token dimension and add per-person and per-parameter embedding vectors.

B.2.2 Hi4D

Hi4D [227] is a MoCap dataset containing interaction between 20 pairs of people. Each pair performs about five interactions such as dancing, fighting, hugging, doing yoga, talking, etc. We split this dataset by subject pair into 14/3/3 for train/val/test. We use subjects [00, 01, 02, 09, 10, 13, 14, 17, 18, 21, 23, 27, 28, 37] for training, [16, 19, 22] for validation, and [12, 15, 32] for testing. Since Hi4D was originally provided in SMPL format, we fit SMPL-X to the estimates via optimization using the code provided in the SMPL-X repository [144]. The dataset provides a start and end frame from/to which each sequence involves physical contact between two people. We use every 5th frame from the contact sequence for training and testing.

B.2.3 CHI3D

CHI3D [41] is a MoCap dataset containing interactions between 3 pairs of people. Each pair performs eight interactions (grab, handshake, hit, holding hands, hug, kick, posing, and push) in various ways summing up to a total of about 120 sequences per subject pair. We use subjects [02, 04] for training and leave [03] for evaluation. Each sequence has a single frame with contact labels. We use this frame from each sequence for training and evaluation.

B.3 Evaluation

B.3.1 Baseline Methods

Transformer

We use the network design of BUDDI, i.e. embedding, person, and parameter layers, the transformer encoder block and layers to bring the latents back into parameter space. The network takes BEV [190] estimates as input and its task is to predict the correct SMPL-X parameters. We train this network on the same data as the conditional version of BUDDI. This baseline is equivalent to a single-shot (non-iterative) version of our diffusion model.

Contact Heuristic

We design an optimization method which is similar to the routine we use to create Flickr Fits, but replaces the $L_{\mathcal{C}D}$, i.e. the loss that takes ground-truth contact maps into account, with a contact heuristic loss $L_{d_{\min}}$. The contact heuristic loss encourages contact between the two people by minimizing their minimum distance. Given the vertices of each mesh, $v \in V_{X1}$ and $u \in V_{X2}$, we define the contact heuristic loss as

$$L_{d_{\min}} = \min_{v,u} \|v - u\|$$

and the overall objective function to be minimized becomes

$$L_{\text{Heuristic-fitting}} = \lambda_J L_J + \lambda_{\bar{\theta}} L_{\bar{\theta}} + \lambda_{\theta} L_{\theta} + \lambda_{\beta} L_{\beta} + \lambda_P L_P + \lambda_{d_{\min}} L_{d_{\min}}. \quad (\text{B.2})$$

BUDDI (gen.)

The conditional version of BUDDI can generate human meshes in close social interaction from noise given a BEV estimate. We use these generations to initialize the optimization routine and evaluate them against the ground truth.

VAE

We also compare against VAE [94] using the same training data. This model projects the SMPL-X parameters of two people into latent vectors of size 64, modeling a distribution, and from the latent space back into parameter space. Similar to the design of BUDDI, we embed each parameter via an MLP. We use two encoder and two decoder layers. The VAE training loss is

$$L_{\text{VAE-training}} = L_{\theta} + L_{\beta} + L_{\gamma} + L_{v2v} + L_{\text{KL}}.$$

We use the same body model parameter losses as during BUDDI training. L_{KL} is a standard KL-divergence loss between two Gaussians:

$$L_{\text{KL}} = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

During optimization, instead of optimizing body model parameters, we optimize in the VAE’s latent space. The optimization objective is:

$$L_{\text{VAE-fitting}} = \lambda_J L_J + \lambda_{\bar{\theta}} L_{\bar{\theta}} + \lambda_{\theta} L_{\theta} + \lambda_{\beta} L_{\beta} + \lambda_P L_P + \lambda_{\text{VAE}} L_{\text{VAE}}, \quad (\text{B.3})$$

where L_{VAE} denotes a squared L2-loss on the VAE latent vector.

Ablation of baseline methods

We run our baseline methods under different conditions, i.e. we use different weights for the Heuristic for a better comparison against the weights used in Flickr Fits and when optimizing with BUDDI used as a prior. The loss weights of Heuristic (a) are similar to those of Flickr Fits and the weights of Heuristic (b) to those of BUDDI. We report these numbers in Table B.2, Table B.3, and Table B.4.

	PER PERSON ↓	JOINT ↓	JOINT PA-MPJPE ↓										
	PA-MPJPE	PA-MPJPE	backhug	basketball	cheers	dance	fight	highfive	hug	kiss	pose	sidehug	talk
Heuristic	67 / 71	121	168	83	94	131	94	68	159	159	118	113	109
Heuristic (a)	68 / 72	122	166	82	93	126	92	68	161	158	122	122	114
Heuristic (b)	68 / 73	124	164	90	92	130	95	68	161	158	125	124	117

Table B.2: **Evaluation of BUDDI on Hi4D.** We compare the output of BUDDI to the proposed baseline methods on the Hi4D challenge. The first block shows methods that do not use Hi4D data during training or are optimization based without access to priors trained on Hi4D. BUDDI (F,C) in particular, is our model BUDDI trained on Flickr and CHI3D data only. All errors are reported in mm for 3D Joints.

	PER PERSON ↓	JOINT ↓
	PA-MPJPE	PA-MPJPE
Heuristic	49 46	105
Heuristic (a)	49 47	103
Heuristic (b)	47 45	103

Table B.3: **Quantitative Evaluation on CHI3D.** We compare different versions of the baseline optimization with contact heuristic on CHI3D (pair s03). All errors reported in mm for 3D Joints.

	JOINT ↓	PCC at radius ↑				
	PA-MPJPE	5	10	15	20	25
Heuristic	68	14	34	49	61	70
Heuristic (a)	69	11	30	45	57	66
Heuristic (b)	72	12	30	45	57	67

Table B.4: **3D Pose Evaluation on FlickrCI3D Signatures.** We compare different versions of the baseline optimization with contact heuristic on the Flickr fits using their joint (two-person) PA-MPJPE expressed in mm. We also evaluate the percentage of correct contact points (PCC) for radius r mm.

B.3.2 User study

We provide several quantitative evaluations of BUDDI but there are aspects of human interaction that are subtle and best judged by people. In the main part of this thesis we present the results of the perceptual study that evaluates how realistic the generated interactions sampled from BUDDI are compared to meshes sampled from a VAE, the training data, and a random configuration of meshes. For this evaluation, we randomly sample 256 meshes from one training batch of size 512 created with a 60/20/20 ratio of meshes from Flickr/Hi4D/CHI3D. The meshes from the training batch are real samples from MoCap or by fitting SMPL-X to images with ground-truth contact map annotations. We further sample 256 meshes from BUDDI (unconditional model) and the VAE. To create the random baseline, we center all meshes in the training batch, shuffle the people along batch and person dimensions, and sample 256 mesh pairs. This is equivalent to real samples, except that each person are sampled randomly and not as a pair. Each participant was asked to rate 68 video comparisons per human intelligence task (HIT) with each video showing one pair of meshes at 360-degree views. Each HIT starts with 10 training videos (not used in evaluation) and contains 10 catch trials. Catch trials show implausible interaction, e.g. two people with random poses placed on top of each other. The training videos are presented at the beginning of the task, and the method and catch trial videos appear in random order. The remaining 48 comparisons show one sample from BUDDI against either VAE / random baseline / or training data (12 comparisons per method). We randomly shuffle the video order per HIT and left / right. Each HIT is conducted by 6 participants. We exclude HITS where participants fail three or more catch trials. Our final results were computed with the responses from the 83/96 participants who passed.

```

1 import smplx
2 import buddi
3
4 # optimization params
5 num_stages = 2
6 max_iterations = 100
7 t = 10 # noise level
8
9 # create smpl and buddi
10 smpl = smplx.create(model_folder)
11
12 # load buddi denoiser model (D)
13 buddi = buddi.create(checkpoint_path).eval()
14
15 # load detected keypoints and bev
16 kpts = load_keypoint_detections(img_path)
17 bev = load_bev_estimate(img_path)
18

```

```
19 # sample from buddi conditioned on BEV
20 buddi_sample = sample_from_buddi(cond=bev)
21
22 # initialize the optimization
23 smpl.params = buddi_sample
24
25 # run optimization
26 for ss in range(num_stages):
27     optimizer = setup_optimizer(smpl, ss)
28
29     for ii in range(max_iterations):
30         # fitting losses
31         fitting_loss = get_fitting_loss(
32             smpl, buddi_sample, kpts)
33
34         # detach current smpl, then diffuse & denoise
35         with torch.no_grad():
36             diffused_smpl = smpl + sample_noise(t)
37             denoised_smpl = buddi(diffused_smpl, t)
38
39         # compute diffusion losses
40         diffusion_loss = get_diffusion_loss(
41             smpl, denoised_smpl)
42
43         # final loss of iteration ii of stage ss
44         total_loss = fitting_loss + diffusion_loss
45
46         # backprop
47         optimizer.zero_grad()
48         total_loss.backward()
49         optimizer.step()
50
51         # check stopping criterium
52         if converted:
53             break
```

Listing B.1: Pseudo code for optimization with BUDDI.

Appendix C

Accurate 3D Body Shape Regression using Metric and Semantic Attributes

C.1 Data Collection

C.1.1 Model-Agency Identity Filtering

We collect internet data consisting of images and height/chest/waist/hips measurements, from model agency websites. A “fashion model” can work for many agencies and their pictures can appear on multiple websites. To create non-overlapping training, validation and test sets, we match model identities across websites. To that end, we use ArcFace [32] for face detection and RetinaNet [33] to compute identity embeddings $E_i \in \mathbb{R}^{512}$ for each image. For every pair of models (q, t) with the same gender label, let Q, T be the number of query and target model images and $E_Q \in \mathbb{R}^{Q \times 512}$ and $E_T \in \mathbb{R}^{T \times 512}$ the query and target embedding feature matrices. We then compute the pairwise cosine similarity matrix $\mathcal{S} \in \mathbb{R}^{Q \times T}$ between all images in E_Q and E_T , and the aggregate and average similarity:

$$\mathcal{S}_T(t) = \frac{1}{Q} \sum_q \mathcal{S}(q, t), \quad (\text{C.1})$$

$$\mathcal{S}_{TQ} = \frac{1}{QT} \sum_q \sum_t \mathcal{S}(q, t). \quad (\text{C.2})$$

Each pair with \mathcal{S} and \mathcal{S}_T that has no element larger than the similarity threshold $\tau = 0.3$ is ignored, as it contains dissimilar models. Finally, we check if \mathcal{S}_{TQ} is larger than τ , and we keep a list of all pairs for which this holds true.

C.2 Mapping Shape Representations

C.2.1 Shape to Anatomical Measurements (S2M)

An important part of our project is the computation of body measurements. Following “Virtual Caliper” [155], we present a method to compute anatomical measurements from a 3D mesh in the canonical T-pose, i.e. after “undoing” the effect of pose. Specifically, we measure the height, $H(\beta)$, weight, $W(\beta)$, and the chest, waist and hip circumferences, $C_c(\beta)$, $C_w(\beta)$, and $C_h(\beta)$, respectively. Let $v_{\text{head}}(\beta), v_{\text{left heel}}(\beta), v_{\text{chest}}(\beta), v_{\text{waist}}(\beta), v_{\text{hip}}(\beta)$ be the head, left heel, chest, waist and hip vertices. $H(\beta)$ is computed as the difference in the vertical-axis “Y” coordinates between the top of the head and the left heel: $H(\beta) = |v_{\text{head}}^y(\beta) - v_{\text{left heel}}^y(\beta)|$. To obtain $W(\beta)$ we multiply the mesh volume by 985 kg/m^3 , which is the average human body density. We compute circumference measurements using the method of Wuhrer et al. [214].

Here, $T \in \mathbb{R}^{F \times 3 \times 3}$, where $F = 20,908$ is the number of triangles in the SMPL-X mesh, denotes “shaped” vertices of all triangles of the mesh $M(\beta, \theta)$; we drop expressions,

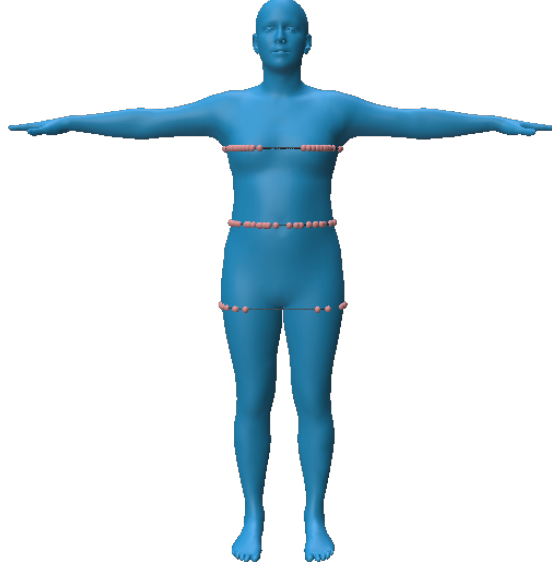


Figure C.1: Automatic anatomical measurements on a 3D mesh. The red points lie on the intersection of planes at chest/waist/hip height with the mesh, while their convex hull is shown with black lines.

ψ , which are not used in this work. Let us explain this using the chest circumference $C_c(\beta)$ as an example. We form a plane P with normal $\mathbf{n} = (0, 1, 0)$ that crosses the point $v_{\text{chest}}(\beta)$. Then, let $\mathcal{S} = \{\mathbf{p}_i\}_{i=1}^N$ be the set of points of P that intersect the body mesh (red points in Fig. C.1). We store their barycentric coordinates (u_i, v_i, w_i) and the corresponding body-triangle index t_i . Let \mathcal{H} be the convex hull of \mathcal{S} (black lines in Fig. C.1), and \mathcal{E} the set of edge indices of \mathcal{H} .

$C_c(\beta)$ is equal to the length of the convex hull:

$$C_c(\beta) = \sum_{(i,j) \in \mathcal{E}} \left\| \begin{pmatrix} u_i \\ v_i \\ w_i \end{pmatrix}^\top T_{t_i} - \begin{pmatrix} u_j \\ v_j \\ w_j \end{pmatrix}^\top T_{t_j} \right\|_2, \quad (\text{C.3})$$

where i, j are point indices for line segments of \mathcal{E} . The process is the same for the waist and hips, but the intersection plane is computed using $v_{\text{waist}}, v_{\text{hip}}$. All of $H(\beta), W(\beta), C_c(\beta), C_w(\beta), C_h(\beta)$ are differentiable functions of body shape parameters, β .

Note that SMPL-X knows the height distribution of humans and acts as a strong prior in shape estimation. Given the ground-truth height of a person (in meter), $H(\beta)$ can be used to directly supervise height and overcome scale ambiguity.

Model	Input	V2V mean \pm std	
		Females	Males
Mean Shape		18.01 \pm 8.73	19.24 \pm 10.36
Linear Regression	A	10.83 \pm 4.77	10.43 \pm 4.63
Polynomial (d=2)	A	10.58 \pm 4.67	10.25 \pm 4.48
MLP	A	10.73 \pm 4.62	10.33 \pm 4.57
Linear Regression	A+H+W	7.00 \pm 2.59	6.56 \pm 2.21
Polynomial (d=2)	A+H+W	7.31 \pm 2.56	6.71 \pm 2.21
MLP	A+H+W	7.03 \pm 2.6	6.68 \pm 2.24
Linear Regression	A+H+ $\sqrt[3]{W}$	6.97 \pm 2.58	6.54 \pm 2.22
Polynomial (d=2)	A+H+ $\sqrt[3]{W}$	6.88 \pm 2.55	6.49 \pm 2.20

Table C.1: Comparison of models for A2S and AHW2S regression.

C.2.2 Mapping Attributes to Shape (A2S)

We introduce A2S, a model that maps the input attribute ratings to shape components β as output. We compare a 2nd degree polynomial model with a linear regression model and a multi-layer perceptron (MLP), using the Vertex-to-Vertex (V2V) error metric between predicted and ground-truth SMPL-X meshes, and report results in Tab. C.1. When using only attributes as input (A2S), the polynomial model of degree $d = 2$ achieves the best performance. Adding height and weight to the input vector requires a small modification, namely using the cubic root of the weight and converting the height from (m) to (cm). With these additions, the 2nd degree polynomial achieves the best performance.

C.2.3 Images to Attributes (I2A)

We briefly experimented with models that learn to predict attribute scores from images (I2A). This attribute predictor is implemented using a ResNet50 for feature extraction from the input images, followed by one MLP per gender for attribute score prediction. To quantify the model’s performance, we use the attribute classification metric described in Chapter 4. I2A achieves 60.7 / 69.3% (fe-/male) of correctly predicted attributes, while our S2A achieves 68.8 / 76% on CAESAR. Our explanation for this result is that it is hard for the I2A model to learn to correctly predict attributes independent of subject pose. Our approach works better, because it decomposes 3D human estimation into predicting pose and shape. Networks are good at estimating pose even without ground-truth shape [112]. “SHAPY ’s losses” affect only the shape branch. To minimize these losses, the network has to learn to correctly predict shape irrespective of pose variations.

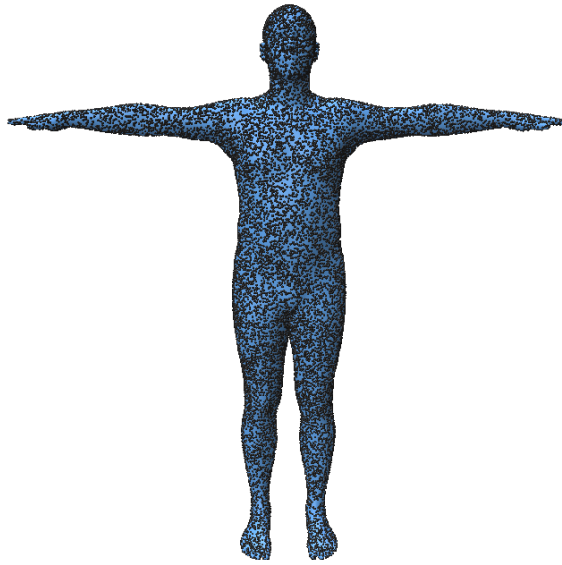


Figure C.2: The 20K body mesh surface points (in black) used to evaluate body shape estimation accuracy.

C.3 SHAPY - 3D Shape Regression from Images

Implementation details. To train SHAPY, each batch of training images contains 50% images collected from model agency websites and 50% images from ExPose’s [27] training set. Note that the overall number of images of males and females in our collected model data differs significantly; images of female models are many more. Therefore, we randomly sample a subset of female images so that, eventually, we get an equal number of male and female images. We also use the BMI of each subject, when available, as a sampling weight for images. In this way, subjects with higher BMI are selected more often, due to their smaller number, to avoid biasing the model towards the average BMI of the dataset.

Our pipeline is implemented in PyTorch [142] and we use the Adam [93] optimizer with a learning rate of $1e-4$. We tune the weights of each loss term with grid search on the MMTS and HBW validation sets. Using a batch size of 48, SHAPY achieves the best performance on the HBW validation set after 80k steps.

C.4 Experiments

C.4.1 Metrics

P2P_{20K}. SMPL-X has more than half of its vertices on the head. Consequently, computing an error based on vertices overemphasizes the importance of the head. To remove this bias, we also report the mean distance between $P = 20k$ mesh surface points; see Fig. C.2

Mean absolute error (mm) ↓				
Method	Height	Chest	Waist	Hips
SHAPY- H	52	113	172	108
SHAPY- HA	60	64	96	77
SHAPY- C	119	66	70	70
SHAPY- CA	74	60	82	69
SHAPY- HC	54	62	72	69
SHAPY- HCA	57	61	85	73

Table C.2: Leave-one-out evaluation on MMTS.

Mean absolute error (mm) ↓					
Method	Height	Chest	Waist	Hips	P2P _{20K}
SHAPY- H	54	90	77	54	22
SHAPY- HA	49	62	71	58	20
SHAPY- C	72	65	77	60	26
SHAPY- CA	54	69	78	58	22
SHAPY- HC	53	61	77	55	23
SHAPY- HCA	47	66	75	52	20

Table C.3: Leave-one-out evaluation on the HBW test set.

for a visualization on the ground-truth and estimated meshes. For this, we uniformly sample the SMPL-X template mesh and compute a sparse matrix $\mathbf{H}_{\text{SMPL-X}} \in \mathbb{R}^{P \times N}$ that regresses the mesh surface points from SMPL-X vertices V , as $\mathbf{P} = \mathbf{H}_{\text{SMPL-X}}V$.

To use this metric in a mesh with different topology, e.g. SMPL, we simply need to compute the corresponding \mathbf{H}_{SMPL} . For this, we align the SMPL model to the SMPL-X template mesh. For each point sampled from the SMPL-X mesh surface, we find the closest point on the aligned SMPL mesh surface. To obtain the SMPL mesh surface points from SMPL vertices, we again compute a sparse matrix, $\mathbf{H}_{\text{SMPL}} \in \mathbb{R}^{P \times 6,890}$. The distance between the SMPL-X and SMPL mesh surface points on the template meshes is 0.073 mm, which is negligible.

Given two meshes M_1 and M_2 of topology T_1 and T_2 we obtain the mesh surface points $P_1 = \mathbf{H}_{T_1}U_1$ and $P_2 = \mathbf{H}_{T_2}U_2$, where U_1 and U_2 denote the vertices of the shaped zero posed (t-pose) meshes. To compute the P2P_{20K} error we correct for translation $t = \bar{P}_2 - \bar{P}_1$ and define

$$\text{P2P}_{20\text{K}}(U_1, U_2) = \|\mathbf{H}_{T_1}U_1 + t - \mathbf{H}_{T_2}U_2\|_2^2.$$

Attribute	Male		Female	
	MAE \pm SD	CCP	MAE \pm SD	CCP
Big	0.25 \pm 0.18	71.68%	0.31 \pm 0.23	70.00%
Broad Shoulders	0.26 \pm 0.20	73.75%	0.33 \pm 0.24	63.90%
Long Legs	0.23 \pm 0.17	81.12%	0.43 \pm 0.33	58.05%
Long Neck	0.27 \pm 0.21	73.75%	0.29 \pm 0.21	69.51%
Long Torso	0.27 \pm 0.20	70.80%	0.36 \pm 0.27	62.68%
Muscular	0.31 \pm 0.24	69.03%	0.26 \pm 0.21	73.17%
Short	0.28 \pm 0.22	72.27%	0.27 \pm 0.21	67.56%
Short Arms	0.20 \pm 0.15	84.07%	0.27 \pm 0.22	72.20%
Tall	0.27 \pm 0.22	70.80%	0.30 \pm 0.23	70.98%
Average	0.27 \pm 0.19	78.76%	n / a	n / a
Delicate Build	0.21 \pm 0.16	78.17%	n / a	n / a
Masculine	0.23 \pm 0.18	78.17%	n / a	n / a
Rectangular	0.27 \pm 0.20	80.24%	n / a	n / a
Skinny Arms	0.25 \pm 0.19	76.40%	n / a	n / a
Soft Body	0.32 \pm 0.23	68.14%	n / a	n / a
Large Breasts	n / a	n / a	0.31 \pm 0.23	72.93%
Pear Shaped	n / a	n / a	0.32 \pm 0.22	64.39%
Petite	n / a	n / a	0.40 \pm 0.30	61.95%
Skinny Legs	n / a	n / a	0.25 \pm 0.18	81.22%
Slim Waist	n / a	n / a	0.30 \pm 0.23	71.71%
Feminine	n / a	n / a	0.26 \pm 0.20	73.41%

Table C.4: S2A evaluation. We report mean, standard deviation and percentage of correctly predicted classes per attribute on CMTS test set.

C.4.2 Shape Estimation

Attribute/Measurement ablation. To investigate the extent to which attributes can replace ground truth measurements in network training, we train SHAPY’s variations in a leave-one-out manner: SHAPY-**H** uses only height and SHAPY-**C** only hip/waist/chest circumference. We compare these models with SHAPY-**AH** and SHAPY-**AC**, which use attributes in addition to height and circumference measurements, respectively. For completeness, we also evaluate SHAPY-**HC** and SHAPY-**AHC**, which use all measurements; the latter also uses attributes. The results are reported in Tab. C.2 (MMTS) and Tab. C.3 (HBW). The tables show that attributes are an adequate replacement for measurements. For example, in Tab. C.2, the height (SHAPY-**C** vs. SHAPY-**CA**) and circumference errors (SHAPY-**H** vs. SHAPY-**AH**) are reduced significantly when attributes are taken into account. On HBW, the P2P_{20K} errors are equal or lower, when attribute information is used, see Tab. C.3. Surprisingly, seeing attributes improves the height error in all

	Model	MPJPE	PA-MPJPE
HMR [87]	SMPL	130	81.3
SPIN [101]	SMPL	96.9	59.2
TUCH [136]	SMPL	84.9	55.5
EFT [83]	SMPL	-	54.2
HybrIK [112]	SMPL	80.0	48.8
STRAPS [174]*	SMPL	-	66.8
Sengupta et al. [176]*	SMPL	-	61.0
Sengupta et al. [175]*	SMPL	84.9	53.6
ExPose [27]	SMPL-X	93.4	60.7
SHAPY (ours)	SMPL-X	95.2	62.6

Table C.5: Evaluation on 3DPW [207]. * uses body poses sampled from the 3DPW training set for training.

three variations. This suggests that training on model images introduces a bias that A2S antagonizes.

S2A. Table C.4 shows the results of S2A in detail. All attributes are classified correctly with an accuracy of at least 58.05% (females) and 68.14% (males). The probability of randomly guessing the correct class is 20%.

AHWC and AHWC2S noise. To evaluate AHWC’s robustness to noise in the input, we fit AHWC using the per-rater scores instead of the average score. The P2P_{20K} ↓ error only increases by 1.0 mm to 6.8 when using the per-rater scores.

C.4.3 Pose evaluation

3D Poses in the Wild (3DPW) [207]: This dataset is mainly useful for evaluating body *pose* accuracy since it contains few subjects and limited body shape variation. The test set contains a limited set of 5 subjects in indoor/outdoor videos with everyday clothing. All subjects were scanned to obtain their ground-truth body shape. The body poses are pseudo ground-truth SMPL fits, recovered from images and IMUs. We convert pose and shape to SMPL-X for evaluation.

We evaluate SHAPY on 3DPW to report pose estimation accuracy (Tab. C.5). SHAPY’s pose accuracy is slightly behind ExPose which also uses SMPL-X. SHAPY’s performance is better than HMR [87] and STRAPS [174]. However, SHAPY is less accurate than recent pose estimation methods, e.g. HybrIK [112]. We assume that SHAPY’s pose estimation accuracy on 3DPW can be improved by (1) adding data from the 3DPW training set (similar to Sengupta et al. [175] who sample poses from 3DPW training set) and (2) creating pseudo ground-truth fits for the model data.

Bibliography

- [1] Ankur Agarwal and Bill Triggs. Recovering 3D human pose from monocular images. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(1):44–58, 2006. 68
- [2] Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Rig Das, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. A survey on deep learning advances on different 3d data representations. *arXiv preprint arXiv:1808.01462*, 2018. 20
- [3] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *Transactions on Graphics (TOG)*, 22(3):587–594, 2003. 68
- [4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 13
- [5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 7, 19, 20, 67
- [6] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 42
- [7] Norman I Badler, Joseph O’Rourke, and Hasida Toltzis. A spherical representation of a human body for visualizing movement. *Proceedings of the IEEE*, 67(10):1397–1403, 1979. 7
- [8] Norman I. Badler and Stephen W. Smoliar. Digital representations of human movement. *ACM Comput. Surv.*, 11(1):19–38, 1979. 7
- [9] Sunhye Bai, Rena L Repetti, and Jacqueline B Sperling. Children’s expressions of positive emotion are sustained by smiling, touching, and playing with parents and siblings: A naturalistic observational study of family life. *Developmental Psychology*, 52(1):88, 2016. 2
- [10] Alexandru Balan and Michael J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision (ECCV)*, volume 5304, pages 15–29, 2008. 67
- [11] Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. Detailed human shape and pose from images. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 67

- [12] Grace T Baranek. Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9–12 months of age. *Journal of Autism and Developmental Disorders*, 29:213–224, 1999. 2
- [13] Felix Barroso, Norbert Freedman, and Stanley Grand. Self-touching, performance, and attentional processes. *Perceptual and Motor Skills*, 50(3_suppl):1083–1089, 1980. 2
- [14] Simone Claire Behrens, Paolo Meneguzzo, Angela Favaro, Martin Teufel, Eva-Maria Skoda, Marion Lindner, Lukas Walder, Alejandra Quiros-Ramirez, Stephan Zipfel, Betty Mohler, Michael Black, and Katrin E. Giel. Weight bias and linguistic body representation in anorexia nervosa: Findings from the bodytalk project. *European Eating Disorders Review*, 29(2):204–215, 2021. 8
- [15] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1929–1942, 2016. 19
- [16] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 19
- [17] Didier Bieler, Semih Gunel, Pascal Fua, and Helge Rhodin. Gravity as a reference for estimating a person’s height from video. In *International Conference on Computer Vision (ICCV)*, pages 8568–8576, 2019. 69
- [18] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 8, 68, 88, 89, 91, 92, 93
- [19] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578, 2016. 5, 17, 42, 43, 44, 45, 50, 52, 64, 67, 68
- [20] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3794–3801, 2014. 19, 20
- [21] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 19
- [22] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008. 19
- [23] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2019. 44, 50, 52, 67, 106
- [24] Carissa J Cascio, David Moore, and Francis McGlone. Social touch and human development. *Developmental Cognitive Neuroscience*, 35:5–11, 2019. 2

- [25] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 42
- [26] Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 17100–17110, 2023. 19
- [27] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, volume 12355, pages 20–40, 2020. 67, 75, 76, 81, 119, 122
- [28] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 8
- [29] Antonio Criminisi, Ian Reid, and Andrew Zisserman. Single view metrology. *International Journal of Computer Vision (IJCV)*, 40(2):123–148, 2000. 69
- [30] April H Crusco and Christopher G Wetzel. The midas touch: The effects of interpersonal touch on restaurant tipping. *Personality and Social Psychology Bulletin*, 10(4):512–517, 1984. 2
- [31] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 42
- [32] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 116
- [33] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 116
- [34] Kyra Densing, Hippokrates Konstantinidis, and Melanie Seiler. Effect of stress level on different forms of self-touch in pre-and postadolescent girls. *Journal of Motor Behavior*, 50(5):475–485, 2018. 2
- [35] Ratan Dey, Madhurya Nangia, Keith W. Ross, and Yong Liu. Estimating heights from photo collections: A data-driven approach. In *Conference on Online Social Networks (COSN)*, page 227–238, 2014. 69
- [36] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Conference on Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 42
- [37] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to regress bodies from images using differentiable semantic rendering. In *International Conference on Computer Vision (ICCV)*, pages 11250–11259, 2021. 68
- [38] Ruth Feldman, Magi Singer, and Orna Zagoory. Touch attenuates infants’ physiological reactivity to stress. *Developmental science*, 13(2):271–278, 2010. 2

- [39] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, pages 792–804, 2021. [8](#), [76](#)
- [40] William A Fetter. A computer graphics human figure system applicable to transportation. *Transportation Research Record*, 657:20–23, 1976. [7](#)
- [41] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7214–7223, 2020. [5](#), [19](#), [41](#), [42](#), [43](#), [44](#), [45](#), [51](#), [52](#), [88](#), [108](#), [110](#)
- [42] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3D human self-contact. In *Conference on Artificial Intelligence (AAAI)*, 2021. [19](#), [20](#), [21](#)
- [43] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. REMIPS: Physically consistent 3D reconstruction of multiple interacting people under weak supervision. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 19385–19397, 2021. [43](#)
- [44] Maria-Paola Forte, Peter Kulits, Chun-Hao Paul Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. Reconstructing signing avatars from video using linguistic priors. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 12791–12801, 2023. [8](#)
- [45] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real-time human pose tracking from range data. In *European Conference on Computer Vision (ECCV)*, pages 738–751, 2012. [19](#)
- [46] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyang Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision (ECCV)*, volume 12362, pages 768–784, 2020. [67](#)
- [47] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. [4](#), [8](#), [59](#), [93](#)
- [48] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7450–7459, 2019. [68](#)
- [49] Martin Grunwald, Thomas Weiss, Stephanie Mueller, and Lysann Rall. Eeg changes caused by spontaneous facial self-touch may represent emotion regulating processes and working memory maintenance. *Brain Research*, 1557:111–126, 2014. [2](#)
- [50] Peng Guan, Alexander Weiss, Alexandru Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009. [42](#), [67](#)
- [51] Nicolas Guéguen and Jacques Fischer-Lokou. Another evaluation of touch and helping behavior. *Psychological Reports*, 92(1):62–64, 2003. [3](#)
- [52] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. [42](#)

- [53] Semih Gunel, Helge Rhodin, and Pascal Fua. What face and body shapes can tell us about height. In *International Conference on Computer Vision Workshops (ICCVw)*, pages 1819–1827, 2019. 69
- [54] Abhinav Gupta, Trista Chen, Francine Chen, Don Kimber, and Larry S Davis. Context and observation driven latent variable model for human pose estimation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 19
- [55] Jinni A Harrigan, John R Kues, John J Steffen, and Robert Rosenthal. Self-touching and impressions of others. *Personality and Social Psychology Bulletin*, 13(4):497–512, 1987. 2
- [56] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346. Wiley Online Library, 2009. 20
- [57] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11354–11364, 2021. 3
- [58] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 3, 8
- [59] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 19, 43
- [60] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, 2021. 3, 42, 43
- [61] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *SIGGRAPH Conf. Track*, 2023. 3
- [62] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(2):386–397, 2020. 68
- [63] Donald Herbison-Evans. *Animated cartoons by computers using ellipsoids*. Basser Department of Computer Science, School of Physics, University of Sydney, 1974. 7
- [64] Don Herbison-Evans. Nudes 2: A numeric utility displaying ellipsoid solids, version 2. *ACM SIGGRAPH Computer Graphics*, 12(3):354–356, 1978. 7
- [65] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 792–800. Springer, 2018. 8, 44, 106
- [66] Matthew Hill, Stephan Streuber, Carina Hahn, Michael Black, and Alice O’Toole. Exploring the relationship between body shapes and descriptions by linking similarity spaces. *Journal of Vision (JOV)*, 15(12):931–931, 2015. 70

- [67] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conf. on Computer Vision (ECCV)*, pages 242–255, 2012. 23
- [68] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 5, 42, 46, 47
- [69] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, pages 609–623, 2019. 68
- [70] Matthias Hoppe, Beat Rossmly, Daniel Peter Neumann, Stephan Streuber, Albrecht Schmidt, and Tonja-Katrin Machulla. A human touch: Social touch increases the perceived human-likeness of agents in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2020. 3
- [71] Wei-Lin Hsiao and Kristen Grauman. ViBE: Dressing for diverse body shapes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11056–11066, 2020. 69
- [72] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexifadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pages 13264–13275, 2022. 19
- [73] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 42
- [74] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, pages 421–430, 2017. 67, 68
- [75] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2013. 68, 76
- [76] Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. Robust inside-outside segmentation using generalized winding numbers. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013. 28, 98
- [77] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12753–12762, 2021. 67
- [78] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5578–5587, 2020. 42, 67
- [79] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMCV*, 2010. 13

-
- [80] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 17
- [81] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, pages 1465–1472, 2011. 13
- [82] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 17
- [83] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, pages 42–52, 2020. 22, 33, 122
- [84] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human model fitting towards in-the-wild 3D human pose estimation. In *International Conference on 3D Vision (3DV)*, 2021. 42
- [85] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *I3DV*, 2022. 8
- [86] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 18, 64, 67
- [87] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 8, 10, 12, 17, 18, 42, 67, 82, 122
- [88] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019. 64
- [89] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 76
- [90] Jason Khoury, Sergiu T. Popescu, Filipe Gama, Valentin Marcel, and Matej Hoffmann. Self-touch and other spontaneous behavior patterns in early infancy. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pages 148–155, 2022. 2
- [91] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Transactions on Graphics (TOG)*, 33(4):120, 2014. 19
- [92] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 52, 103
- [93] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 119
- [94] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 111

- [95] Hedvig Kjellström, Danica Kragić, and Michael J Black. Tracking people interacting with objects. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 747–754. IEEE, 2010. 19
- [96] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 8
- [97] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 10, 64, 67
- [98] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 64
- [99] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11015–11025, 2021. 8, 88
- [100] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *International Conference on Computer Vision (ICCV)*, pages 11035–11045, 2021. 64
- [101] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 5, 8, 10, 12, 16, 17, 18, 22, 30, 31, 32, 33, 42, 64, 67, 68, 76, 81, 82, 103, 122
- [102] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *International Conference on Computer Vision (ICCV)*, 2021. 8, 43
- [103] Koji Komatsu. Human skin model capable of natural shape variation. *The Visual Computer*, 3:265–271, 1988. 7
- [104] Agelos Kratimenos, Georgios Pavlakos, and Petros Maragos. Independent sign language recognition with 3d body, hands, and face reconstruction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4270–4274, 2021. 8
- [105] M Kryger. Bewegungsverhalten von patient und therapeut in als gut und schlecht erlebten therapiesitzungen [movement behavior of patients and therapists in therapy session experienced as good and bad]. *Cologne, Germany: German Sport University*, 2010. 2
- [106] Yen Lee Angela Kwok, Jan Gralton, and Mary-Louise McLaws. Face touching: A frequent habit that has implications for hand hygiene. *Am J Infect Control*, 43(2):112–114, 2015. 16
- [107] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017. 42, 67

-
- [108] Hedda Lausberg. *Understanding body movement: a guide to empirical research on non-verbal behaviour-with an introduction to the NEUROGES coding system*. Peter Lang International Academic Publishers, 2013. 2
- [109] Hedda Lausberg and Monika Kryger. Gestisches verhalten als indikator therapeutischer prozesse in der verbalen psychotherapie: Zur funktion der selbstberührungen und zur repräsentation von objektbeziehungen in gestischen darstellungen. *Psychotherapie-Wissenschaft*, 1(1):41–55, 2011. 2
- [110] Hsi-Jian Lee and Zen Chen. Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing (CGIP)*, 30(2):148–168, 1985. 18
- [111] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023. 42
- [112] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 118, 122
- [113] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *International Conference on Computer Vision (ICCV)*, pages 13401–13412, 2021. 42
- [114] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 8, 10
- [115] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, 2022. 8
- [116] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3d motion and forces of person-object interactions from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8640–8649, 2019. 19
- [117] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)*, 2022. 8
- [118] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *arXiv preprint arXiv:2304.05684*, 2023. 42
- [119] Junbang Liang and Ming C. Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *International Conference on Computer Vision (ICCV)*, pages 4351–4361, 2019. 65, 67
- [120] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755, 2014. 13, 68

- [121] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989. 29
- [122] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 17
- [123] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 4, 8, 65, 67, 106
- [124] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, 2014. 8
- [125] Jianxin Ma, Shuai Bai, and Chang Zhou. Pretrained diffusion models for unified human motion synthesis. *arXiv preprint arXiv:2212.02837*, 2022. 42
- [126] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 8
- [127] Meysam Madadi, Hugo Bertiche, and Sergio Escalera. SMPLR: Deep learning based SMPL reverse for 3D human pose and shape recovery. *Pattern Recognition (PR)*, 106:107472, 2020. 81
- [128] Nadia Magnenat-Thalmann, Richard Laperrière, and Daniel Thalmann. Joint-dependent local deformations for hand animation and object grasping. Technical report, Canadian Inf. Process. Soc, 1988. 7
- [129] Nadia Magnenat-Thalmann, Daniel Thalmann, Nadia Magnenat-Thalmann, and Daniel Thalmann. *Computer animation*. Springer, 1985. 7
- [130] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 5, 17, 19, 68
- [131] Micah A Mammen, Ginger A Moore, Laura V Scaramella, David Reiss, Jody M Ganiban, Daniel S Shaw, Leslie D Leve, and Jenae M Neiderhiser. Infant avoidance during a tactile task predicts autism spectrum behaviors in toddlerhood. *Infant mental health journal*, 36(6):575–587, 2015. 2
- [132] Elisabeta Marinoiu, Dragos Papava, and Cristian Sminchisescu. Pictorial human spaces: How well do humans perceive a 3d articulated pose? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1289–1296, 2013. 20, 26
- [133] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 32
- [134] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian

- Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *ACM Transactions on Graphics*, 39, 2020. 18
- [135] Aymen Mir, Xavier Puig, Angjoo Kanazawa, and Gerard Pons-Moll. Generating continual human motion in diverse 3d scenes. In *arXiv*, 2023. 3
- [136] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, 2021. 42, 44, 67, 68, 81, 122
- [137] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Nguyen. Detecting hands and recognizing physical contact in the wild. In *Advances in neural information processing systems*, 2020. 19
- [138] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171, 2021. 52
- [139] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, pages 484–494, 2018. 42, 68
- [140] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, 2020. 8
- [141] Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. SUPR: A sparse unified part-based human body model. In *European Conference on Computer Vision (ECCV)*, 2022. 8
- [142] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019. 119
- [143] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13468–13478, 2021. 30, 68, 100, 106
- [144] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 4, 5, 8, 9, 10, 17, 18, 19, 26, 27, 42, 43, 44, 50, 51, 52, 64, 67, 68, 69, 110
- [145] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 19

- [146] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 68
- [147] Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023. 8
- [148] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10985–10995, 2021. 42
- [149] Mathis Petrovich, Michael J Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, pages 480–497, 2022. 8, 42
- [150] Andrew Pickles, Helen Sharp, Jennifer Hellier, and Jonathan Hill. Prenatal anxiety, maternal stroking in infancy, and symptoms of emotional and behavioral disorders at 3.5 years. *European child & adolescent psychiatry*, 26(3):325–334, 2017. 2
- [151] Gerard Pons-Moll, David J. Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2345–2352, Columbus, Ohio, USA, 2014. 19
- [152] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *Transactions on Graphics (TOG)*, 34(4):120:1–120:14, 2015. 19, 77
- [153] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 6, 41, 49, 89
- [154] Tom E Potter and Kenneth D Willmert. Three-dimensional human display model. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 102–110, 1975. 7
- [155] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H Bülthoff, and Michael J. Black. The virtual caliper: Rapid creation of metrically accurate avatars from 3D measurements. *Transactions on Visualization and Computer Graphics (TVCG)*, 25(5):1887–1897, 2019. 68, 70, 72, 116
- [156] María Alejandra Quiros-Ramirez, Stephan Streuber, and Michael J. Black. Red shape, blue shape: Political ideology influences the social perception of body shape. *Humanities and Social Sciences Communications*, 8:148, 2021. 8
- [157] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, and Jitendra Malik. On the benefits of 3d pose and tracking for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 8
- [158] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. *Advances in Neural Information Processing Systems*, 34:23703–23713, 2021. 8

- [159] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2740–2749, 2022. 8
- [160] Nadja Reissland, Ezra Aydin, Brian Francis, and Kendra Exley. Laterality of foetal self-touch in relation to maternal stress. *Laterality: Asymmetries of Body, Brain and Cognition*, 20(1):82–94, 2015. 2
- [161] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 8, 42, 43
- [162] Kathleen M. Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeflerlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) final report. Technical Report AFRL-HE-WP-TR-2002-0169, US Air Force Research Laboratory, 2002. 6, 8, 66, 69, 70, 71, 89
- [163] Kathleen M. Robinette and Hein A. M. Daanen. The caesar project: a 3-d surface anthropometry survey. *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*, pages 380–386, 1999. 19
- [164] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 42
- [165] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 8, 10, 19, 28
- [166] Nadine Rueegg, Christoph Lassner, Michael J. Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations. In *Conference on Artificial Intelligence (AAAI)*, pages 5561–5569, 2020. 68
- [167] Aino Saarinen, Ville Harjunen, Inga Jasinskaja-Lahti, Iiro P Jääskeläinen, and Niklas Ravaja. Social touch experience in different contexts: A review. *Neuroscience & Biobehavioral Reviews*, 131:360–372, 2021. 2
- [168] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 42
- [169] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 68
- [170] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 81–90, 2020. 67, 68

- [171] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 152:1–20, 2016. 18
- [172] Klaus R Scherer and Harald G Wallbott. *Nonverbale Kommunikation: Forschungsberichte zum Interaktionsverhalten*. Beltz Weinheim, 1979. 2
- [173] Amanda Seidl, Ruth Tincoff, Christopher Baker, and Alejandrina Cristia. Why the body comes first: Effects of experimenter touch on infants’ word finding. *Developmental Science*, 18(1):155–164, 2015. 2
- [174] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, 2020. 8, 65, 67, 68, 76, 81, 82, 122
- [175] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 64, 65, 68, 80, 81, 82, 122
- [176] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 16094–16104, 2021. 8, 68, 122
- [177] Hyewon Seo, Frederic Cordier, and Nadia Magnenat-Thalmann. Synthesizing animatable body models with parameterized shape modifications. In *Symposium on Computer Animation (SCA)*, pages 120–125, 2003. 68
- [178] Hyewon Seo and Nadia Magnenat-Thalmann. An automatic modeling of human bodies from sizing parameters. In *Symposium on Interactive 3D Graphics (SI3D)*, pages 19–26, 2003. 68
- [179] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 42
- [180] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision (ECCV)*, pages 516–533, 2022. 8
- [181] Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (TOG)*, 2023. 19
- [182] Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. 90
- [183] Leonid Sigal, Alexandru Balan, and Michael J Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1):4–27, 2010. 68
- [184] Russell Smith et al. Open dynamics engine, 2005. 19
- [185] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265, 2015. 5

- [186] Jente L Spille, Martin Grunwald, Sven Martin, and Stephanie M Mueller. The suppression of spontaneous face touch and resulting consequences on memory performance of high and low self-touching individuals. *Scientific Reports*, 12(1):8637, 2022. 2
- [187] Dale M Stack and Darwin W Muir. Tactile stimulation as a component of social interchange: New interpretations for the still-face effect. *British Journal of Developmental Psychology*, 8(2):131–145, 1990. 2
- [188] Stephan Streuber, M. Alejandra Quiros-Ramirez, Matthew Q. Hill, Carina A. Hahn, Silvia Zuffi, Alice O’Toole, and Michael J. Black. Body Talk: Crowdshaping realistic 3D avatars with words. *Transactions on Graphics (TOG)*, 35(4):54:1–54:14, 2016. 6, 65, 66, 69, 70, 71, 89
- [189] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11179–11188, 2021. 8, 42
- [190] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13243–13252, 2022. 5, 8, 41, 42, 44, 46, 50, 51, 52, 89, 106, 108, 109, 110
- [191] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13263–13273, 2022. 8, 19, 42
- [192] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, pages 581–600, 2020. 19
- [193] Yukari Tanaka, Yasuhiro Kanakogi, and Masako Myowa. Social touch in mother–infant interaction affects infants’ subsequent social engagement and object exploration. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021. 2
- [194] Graham W Taylor, Ian Spiro, Christoph Bregler, and Rob Fergus. Learning invariance through imitation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2729–2736. IEEE, 2011. 20
- [195] Purva Tendulkar, Dídac Surís, and Carl Vondrick. FLEX: Full-body grasping without full-body grasps. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 42
- [196] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 8, 41, 42, 48
- [197] Kristin PJ Thompson. Grooming the naked ape: Do perceptions of disease and aggression vulnerability influence grooming behaviour in humans? a comparative ethological perspective. *Current Psychology*, 29:288–296, 2010. 2
- [198] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv:2203.01923*, 2022. 8
- [199] Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 8, 42, 43

- [200] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8001–8013, 2023. [19](#), [88](#), [91](#)
- [201] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. [8](#)
- [202] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. EDGE: Editable dance generation from music. *Computer Vision and Pattern Recognition (CVPR)*, 2023. [42](#)
- [203] Aggeliki Tsoli, Matthew Loper, and Michael J. Black. Model-based anthropometry: Predicting measurements from 3D human scans in multiple poses. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 83–90, 2014. [68](#)
- [204] Gerald Ulrich and K Harms. A video analysis of the non-verbal behaviour of depressed patients before and after treatment. *Journal of affective disorders*, 9(1):63–67, 1985. [2](#)
- [205] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, volume 11211, pages 20–38, 2018. [67](#)
- [206] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. [68](#)
- [207] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume 11214, pages 614–631, 2018. [18](#), [22](#), [32](#), [67](#), [76](#), [122](#)
- [208] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Dynamical simulation priors for human motion tracking. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):52–65, 2012. [19](#)
- [209] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. [49](#)
- [210] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. [3](#)
- [211] Andrew Weitz, Lina Colucci, Sidney Primas, and Brinnae Bent. InfiniteForm: A synthetic, minimal bias dataset for fitness applications. *arXiv:2110.01330*, 2021. [68](#)
- [212] Carol Withrow. *A dynamic model for computer-aided choreography*. PhD thesis, Utah., 1970. [7](#)
- [213] Yan Wu, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang. Saga: Stochastic whole-body grasping with contact. In *European Conference on Computer Vision (ECCV)*, pages 257–274, 2022. [42](#)

- [214] Stefanie Wuhrer and Chang Shu. Estimating 3D human shapes from measurements. *Machine Vision and Applications (MVA)*, 24(6):1133–1147, 2013. 72, 116
- [215] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 17, 18
- [216] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3D Vision (3DV)*, 2020. 17
- [217] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2022. 19, 42
- [218] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 19
- [219] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 67
- [220] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 8, 42, 64, 67
- [221] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 44, 50, 52, 106
- [222] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *International Conference on Computer Vision (ICCV)*, 2019. 42
- [223] Masanobu Yamamoto and Katsutoshi Yagishita. Scene constraints-aided tracking of human body. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 151–156. IEEE, 2000. 19
- [224] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 52
- [225] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 8, 90
- [226] Kangxue Yin, Hui Huang, Edmond SL Ho, Hao Wang, Taku Komura, Daniel Cohen-Or, and Hao Zhang. A sampling approach to generating closely interacting 3d pose-pairs from 2d annotations. *IEEE transactions on visualization and computer graphics*, 25(6):2217–2227, 2018. 19

- [227] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 19, 51, 110
- [228] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 8
- [229] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023. 42, 48
- [230] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, volume 12351, pages 465–481, 2020. 43, 67
- [231] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14484–14493, 2021. 42
- [232] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes—the importance of multiple scene constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 19, 42, 44
- [233] Hongwen Zhang, Yating Tian, Xinchu Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021. 42, 64
- [234] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, pages 642–651, 2020. 42
- [235] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 6193–6203, 2020. 3
- [236] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6194–6204, 2020. 42
- [237] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022. 3
- [238] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022. 3

- [239] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 49
- [240] Rui Zhu, Xingyi Yang, Yannick Hold-Geoffroy, Federico Perazzi, Jonathan Eisenmann, Kalyan Sunkavalli, and Manmohan Chandraker. Single view metrology in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12356, pages 316–333, 2020. 69
- [241] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *arXiv preprint arXiv:2307.10894*, 2023. 8

