

DISS. ETH NO. 28873

RECONSTRUCTING EXPRESSIVE 3D HUMANS
FROM RGB IMAGES

A dissertation submitted to attain the degree of
DOCTOR OF SCIENCES OF ETH ZÜRICH
(Dr. sc. ETH Zürich)

presented by
VASILEIOS CHOUTAS
Dipl. ECE, Aristotle University of Thessaloniki
born on 30 September 1993
citizen of Greece

accepted on the recommendation of
Prof. Dr. Luc Van Gool, examiner
Dr. Michael J. Black, co-examiner
Prof. Dr. Otmar Hilliges, co-examiner
Prof. Dr. Gerard Pons-Moll, co-examiner
Prof. Dr. Dimitrios Tzionas, co-examiner

2022

Στην Ευαγγελία
To Evangelia

ABSTRACT

Humans use their entire body to interact with each other and the environment, e.g. we lean towards an object and grasp it with our hands. During a conversation, facial expressions and hand gestures convey important non-verbal cues about our emotional state and intentions to our fellow speakers. Thus, we need the full 3D surface of the body, hands, and face to understand human behavior. This is a complex task, due to the complexity of body articulation, variance in appearance and body shape, occlusions from objects and the body itself. The community has thus far relied on expensive and cumbersome equipment, such as marker-based motion capture, to capture the 3D human body. While this approach is effective, it is limited to few subjects and indoor scenarios. Monocular RGB cameras are an attractive alternative, thanks to their lower cost and ease of use, but introduce further challenges, namely perspective ambiguities and view occlusions. Researchers adopt a divide-and-conquer strategy to simplify the problem, estimating the body, face, and hands with distinct methods using part-specific datasets. However, the hands and face constrain the body and vice-versa, e.g. the position of the wrist depends on the elbow, shoulder, etc.; the divide-and-conquer approach ignores these constraints.

In this thesis, we aim to reconstruct the full 3D human body, using only readily accessible monocular RGB images. First, we introduce SMPL-X, a parametric 3D body model, that represents full-body shape and pose, hand articulation, and facial expression. Next, we present SMPLify-X, an iterative optimization method, that fits SMPL-X to 2D image keypoints. SMPLify-X produces plausible results, if the 2D observations are not noisy, but it is slow and prone to local minima. To address these issues, we introduce ExPose, a neural network regressor, that predicts SMPL-X parameters from an image using body-driven attention, i.e. zooming in on the hands and face, after predicting the body. From the zoomed-in part images, dedicated part networks predict the part parameters. ExPose combines the independent body, hand, and face estimates by trusting them equally, ignoring the correlation between parts. PIXIE fuses features from the body and part images using neural networks called moderators, before predicting the final part parameters. Overall, the addition of the hands and face leads to noticeably more natural and expressive reconstructions.

Creating high fidelity avatars from RGB images requires accurate estimation of 3D body shape. Although existing methods are effective at predicting body pose, they struggle with body shape, due to the lack of data. To resolve this, we collect internet images from fashion models websites, together with anthropometric measurements. At the same time, we ask human annotators to rate images and meshes according to a pre-defined set of linguistic attributes. We use this information as weak supervision to train SHAPY, a neural network that predicts 3D body pose and shape from an RGB image. Existing 3D shape benchmarks lack subject variety or ground-truth shape. To address this issue, we introduce a new dataset, Human Bodies in the Wild (HBW), which contains images of humans and their corresponding 3D ground-truth body scans. SHAPY outperforms existing 3D body shape regressors, demonstrating that easy-to-obtain measurements and linguistic attributes are sufficient for accurate 3D body shape estimation.

Regressors that estimate 3D model parameters are robust and accurate, but often fail to tightly fit the observations. Optimization approaches tightly fit the data, by minimizing an energy function composed of a data term and hand-crafted task-specific priors. Balancing these terms and implementing a performant version of the solver is a time-consuming task. Machine-learned optimizers combine the benefits of both regression and optimization approaches. They learn the priors directly from data, forgoing hand-crafted priors, and benefit from optimized neural network frameworks for fast inference. Inspired by the classic Levenberg-Marquardt algorithm, we propose an update rule which uses a weighted combination of gradient descent and a network-predicted update. We demonstrate the proposed method’s versatility on three problems: (i) face tracking from dense 2D landmarks, (ii) body estimation from 2D keypoints and (iii) head and hand location from a head-mounted device. Our method offers a competitive alternative to traditional model fitting pipelines, both in terms of accuracy and speed.

To summarize, we present SMPL-X, a richer representation of the human body that jointly models the 3D human body pose and shape, facial expressions, and hand articulation, and three methods, SMPLify-X, ExPose and PIXIE, that estimate SMPL-X parameters from monocular RGB images, progressively improving the accuracy and realism of the predictions. Next, we present two ways for collecting proxy 3D body shape annotations for in-the-wild images, anthropometric measurements and linguistic attributes, and use them to train SHAPY, a model that predicts accurate 3D shape from images without explicit shape supervision. Finally, we propose a versatile and effective learned optimizer for parametric human model fitting tasks.

ZUSAMMENFASSUNG

Menschen benutzen ihren gesamten Körper, um miteinander und mit der Umwelt zu interagieren, z.B. man lehnt sich zu einem Objekt hin und greift es mit den Händen. Während einem Gespräch vermitteln Mimik und Handgesten wichtige nonverbale Hinweise über unseren emotionalen Zustand und unsere Absichten an unsere Gesprächspartner. Wir brauchen also die gesamte 3D-Oberfläche des Körpers, der Hände und des Gesichts, um menschliches Verhalten zu verstehen. Dies ist eine vielschichtige Aufgabe aufgrund der Komplexität der Körperartikulation, der Varianz des Aussehens und der Körperform, der Verdeckung durch von Objekten und des Körpers selbst. Bislang hat sich die Gemeinschaft auf teure und schwerfällige Ausrüstung, wie z.B. markerbasierte Bewegungserfassung, angewiesen, um den menschlichen Körper in 3D zu erfassen. Dieser Ansatz ist zwar effektiv, aber ist auf wenige Objekte und Innenraumszenarien beschränkt. Monokulare RGB-Kameras sind dank ihrer geringeren Kosten und ihrer Benutzerfreundlichkeit eine attraktive Alternative, allerdings bringen sie Herausforderungen mit sich, nämlich perspektivische Mehrdeutigkeiten und Verdeckungen. Forscher wenden das Teile-und-herrsche-Verfahren an, um das Problem zu vereinfachen, indem sie den Körper, das Gesicht und die Hände mit unterschiedlichen Methoden mittels teilespezifischen Datensätze abschätzen. Die Hände und das Gesicht schränken jedoch den Körper ein und vice versa, z.B. die Position des Handgelenks ist vom Ellbogen, von der Schulter usw. abhängig; der Teile-und-herrsche-Ansatz ignoriert diese Beschränkungen.

In dieser Arbeit versuchen wir den ganzen menschlichen 3D-Körper zu rekonstruieren, indem wir nur leicht zugänglichen monokularen RGB-Bildern benutzen. Zunächst stellen wir SMPL-X vor, ein parametrisches 3D-Körpermodell, das die Form und Haltung des gesamten Körpers, die Handgelenke und Gesichtsausdrücke darstellt. Als nächstes präsentieren wir SMPLify-X, eine iterative Optimierungsmethode, die SMPL-X an 2D-Bild-Keypoints anpasst. SMPLify-X liefert plausible Ergebnisse, wenn die 2D-Beobachtungen nicht verrauscht sind. Allerdings ist SMPLify-X langsam und anfällig für lokale Minima. Um diese Probleme zu beheben, führen wir ExPose ein, einen neuronalen Netzwerkregressor, der SMPL-X-Parameter aus einem Bild vorhersagt, indem er die Aufmerksamkeit auf den Körper lenkt, d.h. die Hände und das Gesicht heranzoomt, nachdem es den Körper

vorhergesagt hat. Aus diesen vergrösserten Teilbildern die spezielle Teilnetzwerke vorhersagen die Teilparameter. ExPose kombiniert die unabhängigen Körper-, Hand und Gesichtsschätzungen, indem es ihnen gleichermaßen vertraut und die Korrelation zwischen den Teilen ignoriert. PIXIE fusioniert Merkmale aus den Körper- und Teilbildern mithilfe von neuronalen Netzen, den sogenannten Moderatoren, bevor die endgültigen Teileparameter vorhergesagt werden. Im Grossen und Ganzen führt die Hinzunahme der Hände und des Gesichts zu deutlich natürlicheren und ausdrucksstärkeren Rekonstruktionen. Die Erstellung originalgetreuer Avatare aus RGB-Bildern erfordert eine genaue Schätzung der 3D-Körperform.

Bestehende Methoden sind zwar erfolgreich bei der Vorhersage der Körperhaltung, aber bei der Körperform sind sie aufgrund fehlender Daten mangelhaft. Um dieses Problem zu lösen, sammeln wir Internetbilder von Modelagentur-Webseiten, die auch anthropometrischen Messungen zur Verfügung stellen. Dabei bitten wir menschliche Kommentatoren, Bilder und Netze anhand einer Reihe von sprachlichen Attributen zu bewerten. Wir verwenden diese Informationen als schwache Überwachung, um SHAPY zu trainieren. SHAPY ist ein neuronales Netzwerk, das die 3D-Körperhaltung und -form aus einem RGB-Bild vorhersagt. Bei bestehenden 3D-Form-Benchmarks mangelt es an einer Vielzahl von Testpersonen oder an einer wahrheitsgetreuen Form. Um dieses Problem anzugehen, führen wir einen neuen Datensatz ein, den Human Bodies in the Wild (HBW), der Bilder von Menschen und deren wahrheitsgetreuen 3D-Körperscans enthält. SHAPY übertrifft die bestehenden 3D-Körperform-Regressoren und zeigt, dass einfach zu beschaffende Messungen und sprachliche Attribute für eine genaue Schätzung der 3D-Körperform ausreichend sind.

Regressoren, die 3D-Modellparameter schätzen, sind robust und genau, passen aber oft nicht genau zu den Beobachtungen. Optimierungsansätze passen genau zu den Daten, indem sie eine Energiefunktion minimieren, die aus einem Datenterm und handgefertigten aufgabenspezifischen Vorkenntnissen zusammengesetzt ist. Das Ausbalancieren dieser Terme und die Implementierung einer leistungsfähigen Version des Lösers ist eine zeitaufwendige Aufgabe. Maschinengelernte Optimierer kombinieren die Vorteile von Regressions- und Optimierungsverfahren. Sie lernen die Vorkenntnisse direkt aus den Daten, verzichten auf handgefertigte Vorkenntnisse und profitieren von optimierten neuronalen Netzwerken für schnelle Inferenz. Inspiriert durch den klassischen Levenberg-Marquardt-Algorithmus, stellen wir eine Aktualisierungsregel vor, die eine gewichtete Kombination aus Gradientenabstieg und einer vom Netzwerk vorhergesagten Aktua-

lisierung verwendet. Wir zeigen die Vielseitigkeit der vorgeschlagenen Methode anhand von drei Problemen: (i) Face-Tracking aus dichten 2D-Landmarken, Körperschätzung anhand von 2D-Keypoints und Kopf- und Handpositionierung von einem auf dem Kopf getragenen Gerät. Unsere Methode bietet eine wettbewerbsfähige Alternative zu den traditionellen Modellanpassung-Pipelines, sowohl in Bezug auf die Genauigkeit als auch auf die Geschwindigkeit.

Zusammenfassend führen wir SMPL-X ein, eine umfassendere Darstellung des menschlichen Körpers vor, die die 3D-Körperhaltung und -form, Gesichtsausdrücke und Handgelenke gemeinsam modelliert, und drei Methoden, SMPLify-X, ExPose und PIXIE, die die SMPL-X-Parameter aus monokularen RGB-Bildern schätzen und dabei auch die Genauigkeit und den Realitätssinn der Vorhersagen schrittweise verbessern. Darüber hinaus stellen wir zwei Möglichkeiten vor, um Proxy-Annotationen der 3D-Körperform für In-the-wild-Bilder, anthropometrische Messungen und linguistische Attribute zu sammeln. Diese Annotationen nutzen wir, um SHAPY zu trainieren, ein Modell, das aus Bildern und ohne explizite Formüberwachung die genaue 3D-Form vorhersagt. Schliesslich präsentieren wir einen vielseitigen und erfolgreichen Optimierer für die Anpassung parametrischer Menschmodelle.

ACKNOWLEDGEMENTS

*As you set out for Ithaka
hope your road is a long one,
full of adventure, full of discovery.*

Keep Ithaka always in your mind.

*But don't hurry the journey at all.
Better if it lasts for years,
so you're old by the time you reach the island,
wealthy with all you've gained on the way,
not expecting Ithaka to make you rich.*

Ithaka gave you the marvelous journey.

*Wise as you will have become, so full of experience,
you'll have understood by then what these Ithakas mean.*

— Constantine P. Cavafy

I feel very fortunate to have met an amazing set of people that guided and supported me throughout this long journey. Without all of them, I doubt this adventure would have been as fun and as educational.

First, I want to express my gratitude to my advisors at MPI, Michael J. Black and Dimitris Tzionas, for offering me the opportunity to join Perceiving Systems. Michael is always open to new ideas, listening patiently to every wild proposition. He has taught me valuable lessons about every aspect of research, such as how to select problems, communicate the key idea that differentiates a proposed approach from existing work, and present a project. His vision for building faithful digital human doubles is truly inspiring and a great source of motivation. Seeing part of your work contributing to a larger plan is highly fulfilling. In addition to being a great mentor, Michael has been very supportive on a personal level as well, whenever someone faced a difficult moment. I feel very privileged and grateful to have had the opportunity to be one of his students and part of this wonderful group of people that is Perceiving Systems. Dimitris has been there for me, both as a mentor and as a friend, from day one. His attention to detail and meticulousness are unparalleled. From L^AT_EX tips to high-level intuition on every problem we investigated, his

feedback and insights have been priceless and key to my development as a researcher. Despite severe sleep deprivation, every deadline ended up being an entertaining experience, thanks to his humor and lightheartedness. I am thankful to have had him as a supervisor and I look forward to all our future collaborations. I wish him the best of luck in his new adventure in Amsterdam, and I am certain he will rock.

This section would not be complete without a shout-out to the great people at Perceiving Systems. I will treasure every day, from the small talk during coffee sessions to coming together during difficult times. So, thank you Soubhik (chief scientist of the cool side), Mohamed and Zeena (good luck with your new adventures overseas), Omid, Lea, Partha, Qianli, Yao, Muhammed, Nikos Athanasiou, Nikos Kolotouros, Paul, Joachim (may your game list ever grow), Priyanka, Timo, Sergi, Shashank, Alex, Sai, Nadine, Haiwen, Ahmed, Yinghao, Anurag, Paola, Omri, Sergey, Marilyn, Hongwei, Yuliang, David, Jinlong, Nitin, Eric, Benjamin, Silvia, Aamir, Siyu, Arjun, Nima, Tsvetelina, Markus, Mason, Roy, Roman, Rahul, Igor, Richa and Rajat. Big thanks to Melanie, Nicole, and Johanna for all their help with scheduling and for making our daily lives so easy.

Next, I want to thank my advisor at ETH Zürich Luc Van Gool for supporting me during my stay in Zürich and giving me the freedom to pursue my preferred research directions. I feel very grateful to Gurkirt Singh for the great collaboration and fruitful discussions.

Special thanks to Prof. Hilliges and Prof. Pons-Moll for reviewing this thesis and for agreeing to be part of my examination committee.

I will forever be grateful to Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid for offering me the opportunity to work with them at INRIA in Grenoble. I cannot imagine a better place to start doing research. To Vicky, Daan, Vlad, Thomas, Alberto, Francisco, Mikita, Konstantin, Xavier, Pavel, Dexiong, Maha, Alex, Erwan, Alex, thank you, for being so welcoming and fun. Grenoble would not have been so fun without a little piece of home, so thank you to Ioanna, Nikos, Dionysos, Niki, Aris, Leonidas, Christos, Thanasis.

I want to thank Prof. Dr. Matthias Nießner for hosting me in the Visual Computing Laboratory of TUM in München. This was my first experience with 3D computer vision. A special shout-out here to Andreas for the fun GAN debugging sessions and to the rest of the lab Dejan, Armin, Ji, Aljaž, Manuel for all the fond memories from my time in München.

A big thank you to Federica Bogo and Julien Valentin for hosting me at Microsoft. Federica and Julien are great mentors, offering me constant

support, freedom to explore, and insights when I encountered an obstacle. During these 6 months, I learned a lot about model fitting and got a glimpse at how research is transformed into products. Thank you as well to Pashmina, Tom, Sadegh, Jingjing, Matt, Erroll, Andrew, and Filippos for answering any random questions I had and for all the interesting discussions during the remote coffee sessions.

I want to thank Kaiwen, Javier, Chen, Chenglei, Juan, Ali, and Elias for making the internship at Meta Reality Labs Research an amazing experience.

Research can be frustrating, but great friends make ephemeral troubles go away. Armando and Elena have been wonderful friends during our time in Tübingen. From pizza and Lord of the Rings to life choices and paleoarchaeology, there was no boring evening. I will forever cherish their friendship and our moments together. Despoina and Angelos have been precious friends. From innumerable fun moments to research tips and help in settling in Tübingen and Zürich, I cannot thank them enough. I feel very fortunate to count as friends Georgios Pavlakos, remembering fondly our discussions and his kind advice during my first steps. My childhood friend, Christos, has always stood by me and constantly urged me onwards. Thank you to Alexandros and Atalanti, Alexandros and Enora, for all the wonderful moments in Tübingen.

From the late-night rants about work and student life to sailing in the Aegean, they have always been next to me, supporting me every step of the way. To Antonis, Kostantinos, Georgios, Christos, Georgios, Nikos, Orestis, Ilias, thank you, I am very lucky to have you as my friends.

I owe an immense debt of gratitude to my father, Thanasis, my mother, Sofia, and my sister, Stella-Maria. From my undergraduate years in Thessaloniki to the Ph.D. in Tübingen and Zürich, their support has been unending. They always encouraged me to pursue my dreams and never give up, no matter the challenge. Their belief in me gave me the strength and confidence to overcome all the hurdles on my path. I am fortunate to have had the support of my extended family, Thanasis, Chrysoula, and Nikos, during this endeavor and I thank them for their constant encouragement.

All of this would not have been possible without my life partner, Evangelia. She has been there for me every step of the way, supporting me with all her heart. I lack the words to properly express my gratitude and love.

CONTENTS

LIST OF FIGURES	xviii
LIST OF TABLES	xx
NOTATION	xxi
1 INTRODUCTION	1
2 EXPRESSIVE BODY CAPTURE: 3D HANDS, FACE, AND BODY FROM A SINGLE IMAGE	8
2.1 Introduction	8
2.2 Related work	11
2.2.1 Modeling the body	11
2.2.2 Inferring the body	13
2.3 Technical approach	13
2.3.1 Unified model: SMPL-X	14
2.3.2 SMPLify-X: SMPL-X from a single image	15
2.3.3 Variational human body pose prior	16
2.3.4 Collision penalizer	17
2.3.5 Deep gender classifier	18
2.3.6 Optimization	19
2.4 Experiments	20
2.4.1 Evaluation datasets	20
2.4.2 Qualitative & quantitative evaluations	20
2.5 Conclusion	25
3 MONOCULAR EXPRESSIVE BODY REGRESSION THROUGH BODY-DRIVEN ATTENTION	28
3.1 Introduction	28
3.2 Related Work	30
3.3 Method	32
3.3.1 3D Body Representation	32
3.3.2 Body-driven Attention	32
3.3.3 Implementation Details	36
3.4 Experiments	37
3.4.1 Evaluation Datasets	37
3.4.2 Evaluation Metrics	38
3.4.3 Quantitative and Qualitative Experiments	39
3.5 Conclusion	42

4	COLLABORATIVE REGRESSION OF EXPRESSIVE BODIES USING MODERATION	46
4.1	Introduction	46
4.2	Related work	48
4.3	Method	51
4.3.1	Expressive 3D Body Model	51
4.3.2	PIXIE Architecture	51
4.3.3	Training Losses	53
4.3.4	Implementation Details	56
4.4	Experiments	56
4.4.1	Evaluation Datasets	56
4.4.2	Evaluation Metrics	58
4.4.3	Quantitative Evaluation	58
4.4.4	Qualitative Evaluation	61
4.5	Conclusion	63
5	ACCURATE 3D BODY SHAPE REGRESSION USING METRIC AND SEMANTIC ATTRIBUTES	68
5.1	Introduction	68
5.2	Related Work	71
5.3	Representations & Data for Body Shape	74
5.3.1	SMPL-X Body Model	75
5.3.2	Model-Agency Images	75
5.3.3	Linguistic Shape Attributes	76
5.4	Mapping Shape Representations	77
5.4.1	Virtual Measurements (VM)	77
5.4.2	Attributes and 3D Shape	77
5.5	3D Shape Regression from an Image	79
5.6	Experiments	80
5.6.1	Evaluation Datasets	80
5.6.2	Evaluation Metrics	81
5.6.3	Shape-Representation Mappings	82
5.6.4	3D Shape from an Image	83
5.7	Conclusion	86
6	LEARNING TO FIT MORPHABLE MODELS	89
6.1	Introduction	89
6.2	Related Work	91
6.3	Method	93
6.3.1	Neural Fitter	93
6.3.2	Human Body Model and Fitting Tasks	95

6.3.3	Human Face Model and Fitting Task	97
6.3.4	Data Terms	98
6.3.5	Training Details	99
6.4	Experiments	99
6.4.1	Metrics	99
6.4.2	Quantitative Evaluation	100
6.4.3	Discussion	106
6.5	Conclusion	108
7	SUMMARY	110
7.1	Contributions	110
7.2	Future work	112
7.3	Conclusion	114
	APPENDICES	117
A	EXPRESSIVE BODY CAPTURE: 3D HANDS, FACE, AND BODY FROM A SINGLE IMAGE	118
A.1	Qualitative results	118
A.2	Collision Penalizer	119
A.3	Optimization	122
A.4	Quantitative evaluation on “Total Capture”	124
A.5	Quantitative evaluation on Human3.6M	125
A.6	Qualitative evaluation on MPII	128
A.7	Model	128
A.8	VPoser	128
A.8.1	Data preparation	128
A.8.2	Implementation details	130
A.9	Gender classifier	130
A.9.1	Training data	130
A.9.2	Implementation details	131
B	MONOCULAR EXPRESSIVE BODY REGRESSION THROUGH BODY-DRIVEN ATTENTION	137
B.1	Training details	137
B.2	Data augmentation	138
B.3	Converting SMPL to SMPL-X	138
B.4	SMPLify-X qualitative comparison	141
B.5	In-the-wild qualitative results	141
C	COLLABORATIVE REGRESSION OF EXPRESSIVE BODIES USING MODERATION	150
C.1	Implementation Details	150
C.2	Evaluation	151

c.2.1	Body-face correlations discussion	151
c.2.2	Qualitative Evaluation	152
D	ACCURATE 3D BODY SHAPE REGRESSION USING METRIC AND SEMANTIC ATTRIBUTES	159
D.1	Data Collection	159
D.1.1	Model-Agency Identity Filtering	159
D.1.2	Crowd-sourced Linguistic Shape Attributes	159
D.2	Mapping Shape Representations	160
D.2.1	Shape to Anatomical Measurements (S2M)	160
D.2.2	Mapping Attributes to Shape (A2S)	162
D.2.3	Images to Attributes (I2A)	162
D.3	SHAPY- 3D Shape Regression from Images	163
D.4	Experiments	164
D.4.1	Metrics	164
D.4.2	Shape Estimation	164
D.4.3	Pose evaluation	166
D.4.4	Qualitative Results	168
E	LEARNING TO FIT MORPHABLE MODELS	173
E.1	Errors per iteration	173
E.2	Update rule	173
E.3	Additional ablation	174
E.4	Qualitative comparisons	174
E.5	Training details	175
E.5.1	GRU formulation	175
E.5.2	Training losses	175
E.5.3	Datasets	176
E.5.4	Training schedule	177
E.5.5	Edge loss	177
E.5.6	Runtimes	177
E.5.7	Number of iterations	178
	BIBLIOGRAPHY	181

LIST OF FIGURES

Figure 2.1	SMPLify-X teaser.	8
Figure 2.2	SMPL-X: a new expressive human body model.	9
Figure 2.3	SMPLify-X versus the multi-view method of Joo et al.	22
Figure 2.4	SMPLify-X vs. hands-only approach.	23
Figure 2.5	SMPLify-X qualitative results on LSP.	24
Figure 3.1	Body-driven attention.	29
Figure 3.2	ExPose architecture.	35
Figure 3.3	Curated fits examples.	36
Figure 3.4	ExPose vs naive regression.	43
Figure 3.5	ExPose qualitative, multiple viewpoints.	44
Figure 4.1	PIXIE teaser.	46
Figure 4.2	PIXIE confidence example.	48
Figure 4.3	PIXIE architecture.	52
Figure 4.4	PIXIE evaluation on AGORA.	60
Figure 4.5	PIXIE “gendered” shape loss ablation	63
Figure 4.6	PIXIE vs. ExPose and FrankMocap.	64
Figure 4.7	Comparison with Zhou et al.	65
Figure 5.1	SHAPY teaser.	68
Figure 5.2	Model agency images.	70
Figure 5.3	Crowd-sourcing attribute scores.	71
Figure 5.4	Shape representations and data collection.	74
Figure 5.5	Model agencies measurement histograms.	75
Figure 5.6	SHAPY architecture.	80
Figure 5.7	Human Bodies in the Wild.	82
Figure 5.8	SHAPY HBW qualitative results.	84
Figure 6.1	Learned optimizer teaser	90
Figure 6.2	HMD example data.	96
Figure 6.3	Wood et al. face model.	98
Figure 6.4	Learned fitter 3DPW results.	101
Figure 6.5	Errors per iteration for SMPL+H HMD fitting.	103
Figure 6.6	Learned HMD fitter qualitative results.	106
Figure 6.7	Learned face fitter qualitative examples.	107
Figure A.1	SMPL, SMPL+H and SMPL-X comparison.	119
Figure A.2	Hands-only versus SMPLify-X fitting.	120
Figure A.3	Fitting SMPL-X versus FLAME.	121

Figure A.4	SMPLify-X failure cases.	122
Figure A.5	Body regions where collisions are ignored.	123
Figure A.6	Effect of the collision penalizer.	124
Figure A.7	Sensitivity of the optimization weights.	125
Figure A.8	SMPLify-X Total Capture results.	126
Figure A.9	SMPL-X model evaluation	129
Figure A.10	VPoser model modes.	132
Figure A.11	Gender classifier test set results.	133
Figure A.12	VPoser latent space samples.	134
Figure A.13	Qualitative results of SMPLify-X on MPII.	135
Figure A.14	Body pose prior qualitative comparison.	136
Figure B.1	ExPose feature extractor.	137
Figure B.2	Hand global rotation augmentation.	138
Figure B.3	Head global rotation augmentation.	139
Figure B.4	Jaw rotation augmentation.	139
Figure B.5	Head shape augmentation.	140
Figure B.6	Hand shape augmentation.	140
Figure B.7	Head expression augmentation.	140
Figure B.8	ExPose visual comparison with SMPLify-X.	142
Figure B.9	ExPose versus SPIN.	143
Figure B.10	ExPose versus SPIN.	144
Figure B.11	Body-driven attention vs. naive regression	145
Figure B.12	ExPose predictions from multiple viewpoints	146
Figure C.1	Whole-body shape estimation from only the face	152
Figure C.2	PIXIE vs. MTC.	153
Figure C.3	PIXIE, ExPose, FrankMocap qualitative comparison.	154
Figure C.4	PIXIE qualitative results, 1.	155
Figure C.5	PIXIE qualitative results, 2.	156
Figure C.6	PIXIE qualitative results, 3.	157
Figure C.7	PIXIE failure cases.	158
Figure D.1	AMT task.	161
Figure D.2	Female SHAPY qualitative results	170
Figure D.3	Male SHAPY qualitative results.	171
Figure D.4	anthropometric measurements.	172
Figure D.5	Dense point set for shape evaluation.	172
Figure D.6	SHAPY failure cases.	172
Figure E.1	LGD architecture.	174
Figure E.2	Errors per iteration for fully visible HMD.	175
Figure E.3	Learned fitter vs. Levenberg-Marquardt.	178

Figure E.4	Norm plots per step.	178
Figure E.5	Body parts used for metric computation.	179

LIST OF TABLES

Table 2.1	SMPLify-X quantitative comparison on EHF.	21
Table 2.2	Ablative study for SMPLify-X on the EHF dataset.	21
Table 3.1	ExPose 3DPW comparison.	39
Table 3.2	ExPose EHF ablation.	40
Table 3.3	ExPose EHF evaluation.	41
Table 3.4	ExPose FreiHAND evaluation	42
Table 4.1	PIXIE EHF evaluation.	57
Table 4.2	PIXIE EHF ablation.	59
Table 4.3	PIXIE 3DPW evaluation.	61
Table 4.4	PIXIE NoW evaluation.	62
Table 4.5	PIXIE FreiHAND evaluation.	62
Table 5.1	Linguistic shape attributes for human bodies	76
Table 5.2	Results of A2S variants on CMTS for males.	83
Table 5.3	SHAPY evaluation on HBW.	85
Table 5.4	SHAPY measurement evaluation on MMTS.	85
Table 5.5	SHAPY SSP-3D evaluation.	86
Table 6.1	Learned fitter 3DPW evaluation.	100
Table 6.2	Learned fitter HMD evaluation.	102
Table 6.3	Learned fitter shared weights ablation.	104
Table 6.4	Learned fitter network type ablation.	104
Table 6.5	Learned fitter update rule ablation.	104
Table 6.6	Learned fitter learning rate ablation.	104
Table 6.7	Face fitting to 2D landmarks.	105
Table A.1	Quantitative results on a CMU Panoptic Studio subset.	127
Table A.2	Quantitative results on Human3.6M.	127
Table D.1	Comparison of models for A2S and AHW2S regression.	163
Table D.2	A2S variants results on CMTS test set.	165
Table D.3	Leave-one-out evaluation on HBW and MMTS.	166
Table D.4	SHAPY evaluation on 3DPW.	167
Table D.5	S2A evaluation.	168
Table E.1	Scalar vs. vector λ, γ	176

NOTATION

FREQUENTLY USED SYMBOLS

\mathcal{LBS}	standard linear blend skinning function
V	number of vertices in the mesh
J	number of joints in a kinematic skeleton
J	skeleton joint coordinates
J_T	skeleton joint coordinates in T-pose
\mathcal{J}	sparse human joint regressor
\mathcal{W}	linear blend skinning weights
β	body shape blend-shape coefficients
θ	pose
\hat{h}	hand pose PCA coefficients
\mathcal{H}	hand pose PCA components
ψ	expression blend-shape coefficients
\mathcal{S}	shape / identity blend-shapes
\mathcal{E}	expression blend-shapes
\mathcal{P}	pose corrective blend-shapes
\bar{M}_T	T-pose mesh template vertices
M_S	T-pose mesh vertices, with shape blend shapes
M_T	T-pose mesh vertices, with all blend shapes
M	triangle mesh vertices
\mathcal{T}	triangle mesh faces
\mathcal{T}	array of all mesh triangle coordinates
v	triangle mesh vertex
t	list of vertex indices of a single triangle
E	set of edge indices of a triangle mesh
Π_p	perspective projection operator
Π_o	orthographic projection operator
T	transformation matrix
F	neural network features
A	linguistic shape attributes
H	height
C_c	chest circumference
C_w	waist circumference

C_h hips circumference

ACRONYMS

GPU	Graphics Processing Unit
IMU	Inertial Measurement Unit
HPS	Human Pose and Shape
AR	Augmented Reality
VR	Virtual Reality
ML	Machine Learning
NN	Neural Network
RNN	Recurrent Neural Network
GRU	Gated Recurrent Unit
LSTM	Long-Short Term Memory
HMD	Head-Mounted Device
GD	Gradient Descent
LM	Levenberg-Marquardt
GN	Gauss-Newton
PDL	Powell's Dog Leg method
BVH	Bounding Volume Hierarchy
LBS	Linear Blend Skinning
PCA	Principal Components Analysis
GT	Ground-Truth
FOV	Field Of View
V2V	Vertex-to-Vertex
MPJPE	Mean Per-Joint Point Error
IoU	Intersection over Union
AMT	Amazon Mechanical Turk
ST	Spatial Transformer
SMPL model	Skinned Multi-Person Linear model
SMPL+H	SMPL+ Hands
SMPL-X	SMPL eXpressive
EHF	Expressive Hands and Faces
LSP	Leeds Sports Pose Dataset
MTC	Monocular Total Capture
3DPW	3D Poses in the Wild
AMASS	Archive of Motion Capture as Surface Shapes

NoW	Not quite in-the-Wild
AGORA	Avatars in Geography Optimized for Regression Analysis
SSP-3D	Sports Shape and Pose 3D
MMTS	Model Measurements Test Set
CMTS	CAESAR Meshes Test Set
HBW	Human Bodies in the Wild
SSP-3D	Sports Shape and Pose 3D
BMI	Body-Mass Index
A2S	Attributes to Shape
S2A	Shape to Attributes
VM	Virtual Measurements
I2A	Image to Attributes
LGD	Learned Gradient Descent

INTRODUCTION

Future intelligent agents, both virtual and embodied, should be able to reason about human motion and understand the intents and goals of people in their environment in order to collaborate with and assist humans in their tasks. For example, imagine a scenario where a human asks a robot for a specific object. This can be done verbally, but also by pointing a finger or by the person focusing their gaze on the object. Thus, it is necessary to not only accurately estimate the body pose and shape in 3D, but also facial expressions and hand articulation, especially since we use them to convey subtle cues and interact with our environment. New applications of the Metaverse [328] era, such as those in AR/VR, apparel design, virtual try-on, and fitness, will require accurate, robust, fast and easy-to-use digital cloning methods.

While there exist reliable hardware solutions for reconstructing 3D humans, such as 4D scanners [28, 222, 289, 395], motion capture (MoCap) studios [150, 232, 317, 337, 417], multi-camera systems [21, 76, 139, 147, 165] and light stages [68, 168], they come with several disadvantages. First, they are limited to a single environment, the location of their installation, due to the need for calibration, data storage and compute. At best one can swap objects and furniture during data captures [123]. The cost of the necessary hardware, for capture, compute and storage, is high, limiting the use of such facilities to a few institutes and companies. Calibration and setup time, e.g. placing the infrared markers on the subject in the case of MoCap, are a severe obstacle for scaling up data collection to a large number of people. Last but not least, collecting data that cover the full range of human appearance, ethnicity, and shape using these tools is practically impossible, since the majority of capture subjects comes from a single geographical location.

Due to the above limitations, it is necessary to develop accurate and robust 3D human reconstruction methods that utilize a cheap and easy-to-use sensor. Monocular RGB cameras are sensors with these properties. They are easy to use, much cheaper than the capture equipment described above, and widely available, thanks to the widespread use of modern smartphones. However, these advantages come at the cost of increased difficulty of the 3D human capture problem. The scale ambiguity of RGB images makes it hard

to estimate the absolute depth of each person in the image [334]. Predictors need to be able to deal with occlusions from the environment [183], e.g., a table covering the legs of a sitting person, and self-occlusions from the body itself, e.g. interlacing our hands during a conversation [322]. Having data that covers the full variety of human actions, appearance and behavior, in different environments and settings, is both a blessing and a curse. A blessing, since in theory, we can capture all the data we wish, and a curse, because capture methods need to be accurate and robust under all these conditions. If all of these difficulties were not enough, the difference in size between the body, hands, and face is an additional issue that estimation methods have to deal with.

The community has long worked to resolve these issues and create reliable and robust 3D human reconstruction methods. On the one hand, in an attempt to simplify the problem, it adopted a divide-and-conquer approach, with separate methods, benchmarks, and datasets for 2D/ 3D body [7, 23, 34, 109, 115, 152, 163, 169, 262, 272, 291, 321, 324, 403, 420], hands [20, 36, 83, 126, 151, 193, 247, 296, 365, 427] and face estimation [6, 33, 39, 65, 79, 87, 88, 238, 302, 342, 346]. Initial work attempts to estimate the human body, face and hands, with energy minimization methods [23, 34, 102, 138, 266, 403]. Inspired by the success of deep neural networks for image classification [129, 192, 217, 320], object detection [45, 128, 286] and natural language processing (NLP) [361], neural networks are now the de facto tool used to estimate model parameters from observations. On the other hand, there is an ongoing effort to create richer representations of the human body, starting from simple 2D dots [159] to 3D body surfaces [15, 166, 222, 364, 387], 3D voxel grids [359, 421], occupancy functions [50, 240, 241, 428], point clouds [227, 229] and distance fields [8, 49, 264, 265, 299, 300, 348, 366, 385] and more recently volumetric representations [9, 219, 220, 259, 274, 275, 283, 315, 369, 374, 386, 392, 420].

Part I: Expressive 3D reconstruction from images: In Chapter 2, we start by introducing SMPL-X, a holistic 3D body model that jointly models body pose and shape, hand articulation and facial expression. We then propose an iterative optimization method, named SMPLify-X, that fits SMPL-X to 2D image keypoints. Using SMPLify-X, we collect a large dataset of images and corresponding SMPL-X parameters, with the help of human annotators, who accept or reject invalid fits after seeing the projection of the estimated body on the image. In this way, we circumvent the need for outdoor captures with a multi-view setup or extra instrumentation, such as IMUs [142, 235]. The next natural step would be to train a neural network regressor with

this data, to overcome SMPLify-X’s slow runtime and its sensitivity to initialization.

However, we observe that a single network is unable to jointly predict the body, hands, and face from an image in a single step. This is caused by the difference in size between the body, hands and face, with the latter two occupying very few image pixels compared to the body. We propose to use body-driven attention to resolve this issue, i.e. first predict the body and then zoom-in on the hands and face to refine the corresponding parameters with part specific networks. An important advantage of this decomposition is that we can also use separate body, hand and face datasets to train the respective networks. ExPose, described in Chapter 3, is able to accurately reconstruct the whole body, at a fraction of the runtime of SMPLify-X. Nevertheless, ExPose still does not utilize the full image information, since the part networks only “see” the respective part image, and uses a naive parameter integration mechanism, i.e. it simply copies the part predictions irrespective of context. PIXIE, described in Chapter 4, resolves these issues using moderators, which are neural networks that dynamically aggregate features from the body and part images, using a learned confidence score. In addition to SMPL-X parameters, PIXIE predicts lighting, face albedo and wrinkle details, only when the face moderator’s confidence is above a threshold. Furthermore, it uses gender labels during training to select the appropriate gender shape prior, to infer “gendered” 3D body shapes when possible. The use of an expressive body model, like SMPL-X, and the above prediction techniques leads to higher fidelity reconstruction of humans from RGB images.

Part II: 3D Shape estimation from metric and semantic attributes: Although PIXIE’s gender shape prior brings a noticeable qualitative improvement compared to prior methods, progress in 3D body shape estimation lags behind progress in pose. The reason behind this gap is the lack of in-the-wild images with 3D shape training and evaluation data. In the case of 3D pose estimation, 2D keypoints have been proven to be an effective source of supervision, however there is no equivalent for shape. In Chapter 5, we introduce a solution to this problem by collecting images from online fashion model agencies with anthropometric measurements of each model. The measurements, however, do not fully constrain body shape. While humans cannot estimate metric quantities in images, such as arm length, they can reliably rate images of people according to shape attributes, such as “short/tall”, “long legs” or “pear shaped”. BodyTalk [329] shows that it is possible to predict metrically accurate 3D shape from the aver-

age scores of these linguistic shape attributes. Inspired by this, we ask human annotators to rate a dataset of 3D meshes and the collected fashion model images according to the pre-defined set of attributes. Using the 3D meshes, their attributes and the anthropometric measurements we can define mappings from attributes, measurement to shape and vice-versa. We then train SHAPY, a neural network that predicts 3D shape from images. Supervision for the shape prediction branch comes from novel shape-aware losses that employ the above mappings and data: (i) We compute measurements [281] from the estimated 3D shape and penalize their difference from the ground-truth measurements, downloaded from the model agency websites. (ii) We learn a “Shape-to-Attributes” function that converts a 3D shape to attribute scores. When training an image regressor, we use it to convert the estimated shape to scores and penalize their difference from the human annotations. (iii) Next, we learn the inverse mapping, “Attributes-to-shape”, using it to convert the human attribute scores to shape parameters, which act as supervision for the network. To evaluate the accuracy of 3D human shape estimation methods, we introduce a new benchmark, “Human Bodies in the Wild”, that contains images and 3D body shape annotations, and find that SHAPY outperforms existing methods. In this way, we can sidestep the lack of in-the-wild images with 3D shape annotations using easy-to-obtain anthropometric measurements and linguistic shape attributes.

Part III: Learned optimization for 3D morphable model fitting: Although neural networks can predict 3D model parameters robustly and accurately, given enough data and proper supervision, they often fail to tightly fit the observations. For example, the predicted body pose parameters might produce a mesh that does not perfectly align with the subject’s limbs [410]. Classic optimization methods on the other hand can tightly fit a parametric model to the input observations through iterative minimization of a hand-crafted energy term. The energy is composed of a data term that measures deviation from the observations and a set of priors that encode knowledge about the problem’s structure. Although optimization-based methods are effective, they require significant effort to formulate and balance the weights of each term, and a good initialization point to achieve convergence to a satisfactory minimum. Furthermore, real-time performance is impossible to achieve without significant time investments in custom implementations from domain experts. Learned continuous optimization offers an attractive solution, combining the advantages of both regression- and optimization-only methods. Learned optimizers learn priors directly from the data, removing the need for hand-crafted heuristics, and

benefit from optimized neural network frameworks for fast inference. We build upon these features and propose a neural fitter for 3D human model fitting, inspired from Levenberg-Marquardt (LM) [199, 236], that (i) keeps information across iterations, (ii) controls the learning rate of each variable independently and (iii) combines gradient descent with a network-predicted update. We apply our method on the problem of fitting a body model to 2D keypoints, face model fitting to 2D dense landmarks [376, 377] and full body fitting to head and hand signals from a head-mounted device [75]. The use of different tasks and settings illustrates the versatility of our method. Quantitative evaluation on all three settings demonstrates the effectiveness of our learned optimizer, which outperforms classic optimization and regression baselines.

In summary, this thesis introduces a new holistic and expressive 3D representation of the human body, SMPL-X, and an iterative optimization method to estimate SMPL-X parameters from monocular RGB images. Using SMPLify-X, we collect a training set of images and SMPL-X parameters. We then use this data to train a regressor that predicts SMPL-X parameters from an RGB image, but observe that it can only estimate a rough configuration of the hands and face, due to their smaller size compared to the body. ExPose overcomes this issue using body-driven attention, i.e. localizing the hands and face from the body, extracting high-resolution crops and improving the rough estimate with dedicated refinement modules. PIXIE employs full-body context to improve the accuracy of hand and face prediction in the presence of ambiguities, such as occlusions. PIXIE uses moderator networks that estimate the confidence of body-part experts and computes a weighted average of their features using this confidence value. By combining SMPLify-X’s training data, ExPose’s body-driven attention and PIXIE’s moderators we can train accurate and fast neural network regressors for expressive 3D body estimation. Next, we show how to use weaker shape annotations, namely anthropometric measurements and linguistic attribute scores, to improve 3D body shape prediction. Finally, we propose a learned optimization method for human model fitting problems that combines the advantages of regression and optimization approaches. This learned optimizer is effective, outperforming the baseline regression and optimization methods, and versatile, as it is easily applicable to different tasks.

PART I
EXPRESSIVE 3D
RECONSTRUCTION FROM IMAGES

EXPRESSIVE BODY CAPTURE: 3D HANDS, FACE, AND BODY FROM A SINGLE IMAGE

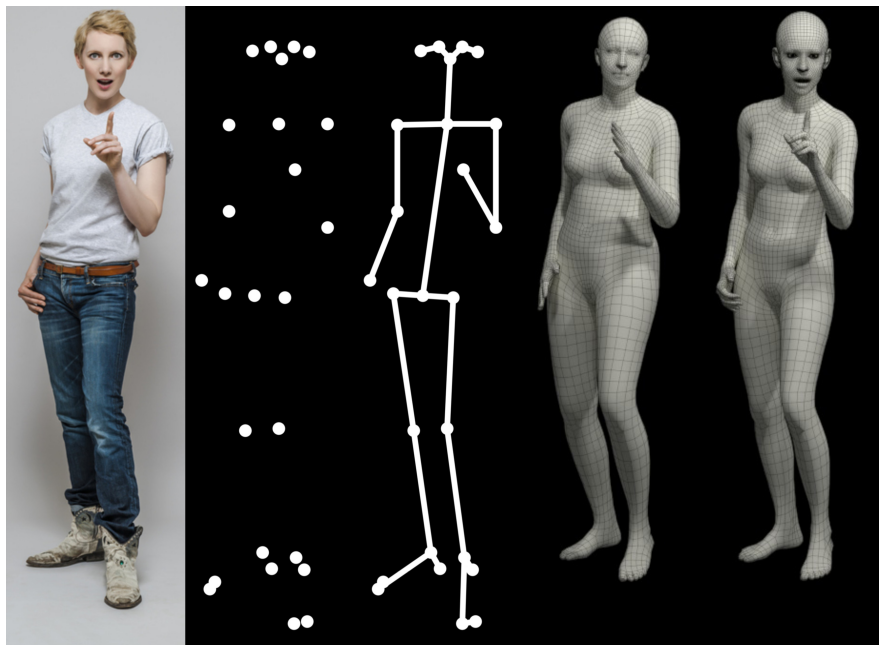


FIGURE 2.1: Communication and gesture rely on the *body* pose, *hand* pose, and *facial* expression, all *together*. The major joints of the body are not sufficient to represent this and current 3D models are not expressive enough. In contrast to prior work, our approach estimates a more detailed and expressive 3D model from a single image. From left to right: RGB image, major joints, skeleton, SMPL (female), SMPL-X (female). The hands and face in SMPL-X enable more *holistic* and *expressive* body capture.

2.1 INTRODUCTION

Humans are often a central element in images and videos. Understanding their posture, the social cues they communicate, and their interactions with the world is critical for holistic scene understanding. Recent methods have



FIGURE 2.2: We learn a new 3D model of the human body called *SMPL-X* that jointly models the human body, face and hands. We fit the female *SMPL-X* model with *SMPLify-X* to single RGB images and show that it captures a rich variety of *natural* and *expressive* 3D human poses, gestures and facial expressions. Image source: <https://www.gettyimages.de/search/stack/546047069>.

shown rapid progress on estimating the major body joints, hand joints and facial features in 2D [44, 148, 319]. Our interactions with the world, however, are fundamentally 3D and recent work has also made progress on the 3D estimation of the major joints and rough 3D pose directly from single images [34, 169, 262, 272].

To understand human behavior, however, we have to capture more than the major joints of the body – we need the full 3D surface of the body, hands and the face. There is no system that can do this today due to several major challenges including the lack of appropriate 3D models and rich 3D training data. Figure 2.1 illustrates the problem. The interpretation of expressive and communicative images is difficult using only sparse 2D information or 3D representations that lack hand and face detail. To address this problem, we need two things. First, we need a 3D model of the body that is able to represent the complexity of human faces, hands, and body pose. Second, we need a method to extract such a model from a single image.

Advances in neural networks and large datasets of manually labeled images have resulted in rapid progress in 2D human “pose” estimation. By “pose”, the field often means the location of the major body joints. This is not sufficient to understand human behavior as illustrated in Fig. 2.1. OpenPose [43, 44, 319] expands this to include the 2D hand joints and 2D

facial features. While this captures much more about the communicative intent, it does not support reasoning about surfaces and human interactions with the 3D world.

Models of the 3D body have focused on capturing the overall shape and pose of the body, excluding the hands and face [10, 11, 15, 121, 222]. There is also an extensive literature on modelling hands [176, 239, 260, 261, 293, 308, 325, 349, 356] and faces [12, 33, 35, 38, 41, 206, 273, 362, 391] in 3D but in isolation from the rest of the body. Only recently has the field begun modeling the body together with hands [293], or together with the hands and face [166]. The Frank model [166], for example, combines a simplified version of the SMPL body model [222], with an artist-designed hand rig, and the FaceWarehouse [41] face model. These disparate models are stitched together, resulting in a model that is not fully realistic.

Here we learn a new, holistic, body model with face and hands from a large corpus of 3D scans. The new *SMPL-X* model (*SMPL eXpressive*) is based on SMPL and retains the benefits of that model: compatibility with graphics software, simple parametrization, small size, efficient, differentiable, etc. We combine SMPL with the FLAME head model [206] and the MANO hand model [293] and then register this combined model to 5586 3D scans that we curate for quality. By learning the model from data, we capture the natural correlations between the shape of bodies, faces and hands and the resulting model is free of the artifacts seen with Frank. The expressivity of the model can be seen in Fig. 2.2 where we fit SMPL-X to expressive RGB images, as well as in Fig. 2.5 where we fit SMPL-X to images of the public LSP dataset [160]. SMPL-X is freely available for research purposes.

Several methods use deep learning to regress the parameters of SMPL from a single image [169, 262, 272]. To estimate a 3D body with the hands and face though, there exists no suitable training dataset. To address this, we follow the approach of SMPLify. First, we estimate 2D image features “bottom up” using OpenPose [44, 319, 371], which detects the joints of the body, hands, feet, and face features. We then fit the SMPL-X model to these 2D features “top down”, with our method called *SMPLify-X*. To do so, we make several significant improvements over SMPLify. Specifically, we learn a new, and better performing, pose prior from a large dataset of motion capture data [221, 232] using a variational auto-encoder. This prior is critical because the mapping from 2D features to 3D pose is ambiguous. We also define a new (self-) interpenetration penalty term that is significantly more accurate and efficient than the approximate method in SMPLify; it

remains differentiable. We train a gender detector and use this to automatically determine what body model to use, either male, female or gender neutral. Finally, one motivation for training direct regression methods to estimate SMPL parameters is that SMPLify is slow. Here we address this with a PyTorch implementation that is at least 8 times faster than the corresponding Chumpy implementation, by leveraging the computing power of modern GPUs. Examples of this SMPLify-X method are shown in Fig. 2.2.

To evaluate the accuracy, we need new data with full-body RGB images and corresponding 3D ground truth bodies. To that end, we curate a new evaluation dataset containing images of a subject performing a wide variety of poses, gestures and expressions. We capture 3D body shape using a scanning system and we fit the SMPL-X model to the scans. This form of pseudo ground-truth is accurate enough to enable quantitative evaluations for models of body, hands and faces together. We find that our model and method perform significantly better than related, and less powerful, models, resulting in natural and expressive results.

We believe that this is a significant step towards *expressive* capture of bodies, hands, and faces *together* from a single RGB image. We make available for research purposes the SMPL-X model, SMPLify-X code, trained networks, model fits, and the evaluation dataset at <https://smpl-x.is.tue.mpg.de/>.

2.2 RELATED WORK

2.2.1 Modeling the body

Bodies, Faces and Hands: The problem of modeling the 3D body has previously been tackled by breaking the body into parts and modeling these parts separately. We focus on methods that learn statistical shape models from 3D scans.

Blanz and Vetter [33] pioneered this direction with their 3D morphable face model. Numerous methods since then have learned 3D face shape and expression from scan data; see [38, 429] for recent reviews. A key feature of such models is that they can represent different face shapes and a wide range of expressions, typically using blend shapes inspired by FACS [80]. Most approaches focus only on the face region and not the whole head. FLAME [206], in contrast, models the whole head, captures 3D head rotations, and also models the neck region; we find this critical

for connecting the head and the body. None of these methods, model correlations in face shape and body shape.

The availability of 3D body scanners enabled learning of body shape from scans. In particular the CAESAR dataset [290] opened up the learning of shape [10]. Most early work focuses on body shape using scans of people in roughly the same pose. Angelov et al. [15] combined shape with scans of one subject in many poses to learn a factored model of body shape and pose based on triangle deformations. Many models followed this, either using triangle deformations [51, 97, 121, 133, 278] or vertex-based displacements [11, 122, 222], however they all focus on modeling body shape and pose without the hands or face. These methods assume that the hand is either in a fist or an open pose and that the face is in a neutral expression.

Similarly, hand modeling approaches typically ignore the body. Additionally, 3D hand models are typically not learned but either are artist designed [325], based on shape primitives [239, 261, 308], reconstructed with multiview stereo and have fixed shape [25, 356], use non-learned per-part scaling parameters [67], or use simple shape spaces [349]. Only recently [176, 293] have learned hand models appeared in the literature. Khamis et al. [176] collect partial depth maps of 50 people to learn a model of shape variation, however they do not capture a pose space. Romero et al. [293] on the other hand learn a parametric hand model (MANO) with both a rich shape and pose space using 3D scans of 31 subjects in up to 51 poses, following the SMPL [222] formulation.

Unified Models: The most similar models to ours are Frank [166], SMPL+H [293] and GHUM/GHUML [387]. Frank stitches together three different models: SMPL (with no pose blend shapes) for the body, an artist-created rig for the hands, and the FaceWarehouse model [41] for the face. The resulting model is not fully realistic. SMPL+H combines the SMPL body with a 3D hand model that is learned from 3D scans. The shape variation of the hand comes from full body scans, while the pose dependent deformations are learned from a dataset of hand scans. SMPL+H does not contain a deformable face. GHUM and the lower resolution GHUML (ite) are unified models of the body, hand and face, trained end-to-end on a dataset of high-resolution body scans and close-ups of the hands and face. In contrast to SMPL, GHUM/GHUML have non-linear shape and expression spaces, implemented as variational auto-encoders (VAE) [179].

We start from the publicly-available SMPL+H [234] and add the publicly-available FLAME head model [94] to it. Unlike Frank, however, we do not

simply graft this onto the body. Instead we take the full model and fit it to 5586 3D scans and learn the shape and pose-dependent blend shapes. This results in a natural looking model with a consistent parameterization. Being based on SMPL, it is differentiable and easy to swap into applications that already use SMPL.

2.2.2 *Inferring the body*

There are many methods that estimate 3D faces from images or RGB-D [429] as well as methods that estimate hands from such data [396]. While there are numerous methods that estimate the location of 3D joints from a single image, here we focus on methods that extract a full 3D body mesh.

Several methods estimate the SMPL model from a single image [169, 195, 262, 272]. This is not trivial due to a paucity of training images with paired 3D model parameters. To address this, SMPLify [34] detects 2D image features “bottom up” and then fits the SMPL model to these “top down” in an optimization framework. In [195] these SMPLify fits are used to iteratively curate a training set of paired data to train a direct regression method. HMR [169] trains a model without paired data by using 2D keypoints and an adversary that knows about 3D bodies. Like SMPLify, NBF [262] uses an intermediate 2D representation (body part segmentation) and infers 3D pose from this intermediate representation. MonoPerfCap [388] infers 3D pose while also refining surface geometry to capture clothing. These methods estimate only the 3D pose of the body without the hands or face.

There are also many multi-camera setups for capturing 3D pose, 3D meshes (performance capture), or parametric 3D models [24, 40, 70, 99, 139, 141, 162, 216, 244, 288, 327, 395]. Most relevant is the Panoptic studio [162] which shares our goal of capturing rich, expressive, human interactions. In [166], the Frank model parameters are estimated from multi-camera data by fitting the model to 3D keypoints and 3D point clouds. The capture environment is complex, using 140 VGA cameras for the body, 480 VGA cameras for the feet, and 31 HD cameras for the face and hand keypoints. We aim for a similar level of expressive detail but from a *single RGB image*.

2.3 TECHNICAL APPROACH

In the following we describe SMPL-X (Sec. 2.3.1), and our approach for fitting SMPL-X to single RGB images (Sec. 2.3.2). Compared to SMPLify [34], SMPLify-X uses a better pose prior (Sec. 2.3.3), a more detailed collision

penalty (Sec. 2.3.4), gender detection (Sec. 2.3.5), and a faster PyTorch implementation (Sec. 2.3.6).

2.3.1 Unified model: SMPL-X

We create a unified model, called *SMPL-X*, for *SMPL expressive*, with shape parameters trained jointly for the face, hands and body. SMPL-X uses standard vertex-based linear blend skinning [200] with learned corrective blend shapes, has $V = 10,475$ vertices and $J = 54$ joints, which includes joints for the neck, jaw, eyeballs and fingers. SMPL-X is defined by a function $M(\theta, \beta, \psi) : \mathbb{R}^{|\theta| \times |\beta| \times |\psi|} \rightarrow \mathbb{R}^{3V}$, parameterized by the pose $\theta \in \mathbb{R}^{3(J+1)}$ where K is the number of body joints in addition to a joint for global rotation. We decompose the pose parameters θ into: θ_f for the jaw joint, θ_h for the finger joints, and θ_b for the remaining body joints. The joint body, face and hands shape parameters are noted by $\beta \in \mathbb{R}^{|\beta|}$ and the facial expression parameters by $\psi \in \mathbb{R}^{|\psi|}$. More formally:

$$M(\beta, \theta, \psi) = \mathcal{LBS}(M_T(\beta, \theta, \psi), J_T(\beta), \theta; \mathcal{W}) \quad (2.1)$$

$$M_T(\beta, \theta, \psi) = \bar{M}_T + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\theta; \mathcal{P}) \quad (2.2)$$

where $B_S(\beta; \mathcal{S}) = \sum_{n=1}^{|\beta|} \beta_n \mathcal{S}_n$ is the shape blend shape function, β are linear shape coefficients, $|\beta|$ is their number, $\mathcal{S}_n \in \mathbb{R}^{3V}$ are orthonormal principal components of vertex displacements capturing shape variations due to different person identity, and $\mathcal{S} = [\mathcal{S}_1, \dots, \mathcal{S}_{|\beta|}] \in \mathbb{R}^{3V \times |\beta|}$ is a matrix of all such displacements. $B_P(\theta; \mathcal{P}) : \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}^{3V}$ is the pose blend shape function, which adds corrective vertex displacements to the template mesh \bar{M}_T as in SMPL [221]:

$$B_P(\theta; \mathcal{P}) = \sum_{n=1}^{9J} (R_n(\theta) - R_n(\theta^*)) \mathcal{P}_n, \quad (2.3)$$

where $R : \mathbb{R}^{|\theta|} \rightarrow \mathbb{R}^{9J}$ is a function mapping the pose vector θ to a vector of concatenated part-relative rotation matrices, computed with the Rodrigues formula [37, 252, 279] and $R_n(\theta)$ is the n^{th} element of $R(\theta)$, θ^* is the pose vector of the rest pose, $\mathcal{P}_n \in \mathbb{R}^{3V}$ are again orthonormal principal components of vertex displacements, and $\mathcal{P} = [\mathcal{P}_1, \dots, \mathcal{P}_{9J}] \in \mathbb{R}^{3V \times 9J}$ is a matrix of all pose blend shapes. $B_E(\psi; \mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n \mathcal{E}$ is the expression blend shape function, where \mathcal{E} are principal components capturing variations due to facial expressions and ψ are PCA coefficients. Since 3D joint locations

J_T vary between bodies of different shapes, they are a function of body shape $J_T(\beta) = \mathcal{J}(\bar{M}_T + B_S(\beta; \mathcal{S}))$, where \mathcal{J} is a sparse linear regressor that regresses 3D joint locations from mesh vertices. A standard linear blend skinning function $\mathcal{LBS}(\cdot)$ [200] rotates the vertices in $M_T(\cdot)$ around the estimated joints $J_T(\beta)$ smoothed by blend weights $\mathcal{W} \in \mathbb{R}^{V \times J}$.

We start with an artist designed 3D template, whose face and hands match the templates of FLAME [206] and MANO [293]. We fit the template to four datasets of 3D human scans to get 3D alignments as training data for SMPL-X. The shape space parameters, $\{\mathcal{S}\}$, are trained on 3800 alignments in an A-pose capturing variations across identities [290]. The body pose space parameters, $\{\mathcal{W}, \mathcal{P}, \mathcal{J}\}$, are trained on 1786 alignments in diverse poses. Since the full body scans have limited resolution for the hands and face, we leverage the parameters of MANO [293] and FLAME [206], learned from 1500 hand and 3800 head high resolution scans respectively. More specifically, we use the pose space and pose corrective blendshapes of MANO for the hands and the expression space \mathcal{E} of FLAME.

The fingers have 30 joints, which correspond to 90 pose parameters (3 DOF per joint as axis-angle rotations). SMPL-X uses a lower dimensional PCA pose space for the hands such that $\theta_h = \sum_{n=1}^{|\mathcal{h}|} h_n \mathcal{H}$, where \mathcal{H} are principal components capturing the finger pose variations and h are the corresponding PCA coefficients. As noted above, we use the PCA pose space of MANO, that is trained on a large dataset of 3D articulated human hands. The total number of model parameters in SMPL-X is 119: 75 for the global body rotation and { body, eyes, jaw } joints, 24 parameters for the lower dimensional hand pose PCA space, 10 for subject shape and 10 for the facial expressions. Additionally there are separate male and female models, which are used when the gender is known, and a shape space constructed from both genders for when gender is unknown. SMPL-X is realistic, expressive, differentiable and easy to fit to data.

2.3.2 SMPLify-X: SMPL-X from a single image

To fit SMPL-X to single RGB images (SMPLify-X), we follow SMPLify [34] but improve every aspect of it. We formulate fitting SMPL-X to the image as an optimization problem, where we seek to minimize the objective function

$$E(\beta, \theta, \psi) = E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_h E_h + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta + \lambda_\psi E_\psi + \lambda_C E_C \quad (2.4)$$

where θ_b , θ_f and \mathbf{h} are the pose vectors for the body, face and the two hands respectively, and θ is the full set of optimizable pose parameters. The body pose parameters are a function $\theta_b(Z)$, where $Z \in \mathbb{R}^{32}$ is a lower-dimensional pose space described in Sec. 2.3.3. $E_J(\beta, \theta; \mathbf{K}, \mathbf{j}_{\text{est}})$ is the data term as described below, while the terms $E_{\mathbf{h}}(\mathbf{h})$, $E_{\theta_f}(\theta_f)$, $E_\beta(\beta)$ and $E_\psi(\psi)$ are simple L_2 priors for the hand pose, facial pose, body shape and facial expressions, penalizing deviation from the neutral state. Since the shape space of SMPL-X is scaled for unit variance, similarly to SMPL+H [293], $E_\beta(\beta) = \|\beta\|^2$ describes the Mahalanobis distance between the shape parameters being optimized and the shape distribution in the training dataset of SMPL-X. $E_\alpha(\theta_b) = \sum_{i \in (\text{elbows}, \text{knees})} \exp(\theta_i)$ follows Bogo et al. [34] and is a simple prior penalizing extreme bending only for elbows and knees. We further employ $E_{\theta_b}(\theta_b)$, which is a VAE-based body pose prior (Sec. 2.3.3), while $E_C(\theta_b, \mathbf{h}, \theta_f, \beta)$ is an interpenetration penalty (Sec. 2.3.4). Finally, λ denotes weights that steer the influence of each term in Eq. (2.4). We empirically find that an annealing scheme for λ helps optimization (Sec. 2.3.6).

For the *data term* we use a re-projection loss to minimize the weighted robust distance between estimated 2D joints \mathbf{j}_{est} and the 2D projection of the corresponding posed 3D joints $J_i(\theta, \beta)$ of SMPL-X for each joint i , where $J_i \theta(\cdot)$ denotes the joints transformed according to the pose θ after traversing the kinematic tree. Similar to Bogo et al. [34], the data term is:

$$E_J(\beta, \theta; \mathbf{K}, \mathbf{j}_{\text{est}}) = \sum_i^J \gamma_i \omega_i \rho(\Pi_p(J_i(\beta, \theta); \mathbf{K}) - \mathbf{j}_{\text{est}, i}) \quad (2.5)$$

where $\Pi_p(\cdot; \mathbf{K})$ denotes the 3D to 2D perspective projection function with intrinsic camera parameters \mathbf{K} . For the 2D detections we rely on the OpenPose library [44, 319, 371], which provides body, hands, face and feet keypoints jointly for each person in an image. To account for noise in the detections, the contribution of each joint in the data term is weighted by the detection confidence score ω_i , while γ_i are per-joint weights for annealed optimization, as described in Sec. 2.3.6. Finally, ρ denotes a robust Geman-McClure error function [107] for down weighting noisy detections.

2.3.3 Variational human body pose prior

We seek a prior over body pose that penalizes impossible poses while allowing possible ones. SMPLify uses an approximation to the negative log of a Gaussian mixture model trained on MoCap data. While effective, we find that the SMPLify prior is not sufficiently strong. Consequently, we

train our body pose prior, VPoser, using a variational autoencoder [179], which learns a latent representation of human pose and regularizes the distribution of the latent code to be a normal distribution. We train our prior using the AMASS [232] dataset. Our training and test data respectively consist of roughly 1M, and 65k poses, in rotation matrix representation. Details on the data preparation procedure is given in Sec. A.8.

The training loss of the VAE is formulated as:

$$\mathcal{L}_{total} = c_1\mathcal{L}_{KL} + c_2\mathcal{L}_{rec} + c_3\mathcal{L}_{orth} + c_4\mathcal{L}_{det1} + c_5\mathcal{L}_{reg} \quad (2.6)$$

$$\mathcal{L}_{KL} = KL(q(Z|R)||\mathcal{N}(0, I)) \quad (2.7)$$

$$\mathcal{L}_{rec} = \|R - \hat{R}\|_2^2 \quad (2.8)$$

$$\mathcal{L}_{orth} = \|\hat{R}\hat{R}' - I\|_2^2 \quad (2.9)$$

$$\mathcal{L}_{det1} = |\det(\hat{R}) - 1| \quad (2.10)$$

$$\mathcal{L}_{reg} = \|\phi\|_2^2, \quad (2.11)$$

where $Z \in \mathbb{R}^{32}$ is the latent space of the autoencoder, $R \in SO(3)$ are 3×3 rotation matrices for each joint as the network input and \hat{R} is a similarly shaped matrix representing the output. The Kullback-Leibler term in Eq. (2.7), and the reconstruction term in Eq. (2.8) follow the VAE formulation in [179], while their role is to encourage a normal distribution on the latent space, and to make an efficient code to reconstruct the input with high fidelity. Equations (2.9) and (2.10) encourage the latent space to encode valid rotation matrices. Finally, Eq. (2.11) helps prevent over-fitting by encouraging smaller network weights ϕ . Implementation details can be found in Sec. A.8.

To employ VPoser in the optimization, rather than to optimize over θ_b directly in Eq. (2.4), we optimize the parameters of a 32 dimensional latent space with a quadratic penalty on Z and transform this back into joint angles θ_b in axis-angle representation. This is analogous to how hands are treated except that the hand pose θ_h is projected into a linear PCA space and the penalty is on the linear coefficients.

2.3.4 Collision penalizer

When fitting a model to observations, there are often self-collisions and penetrations of several body parts that are physically impossible. Our approach is inspired by SMPLify, which penalizes penetrations with an underlying collision model based on shape primitives, i.e. an ensemble

of capsules. Although this model is computationally efficient, it is only a rough approximation of the human body.

For models like SMPL-X, that also model the fingers and facial details, a more accurate collision model is needed. To that end, we employ the detailed collision-based model for meshes from [25, 356]. We first detect a list of colliding triangles \mathcal{C} by employing Bounding Volume Hierarchies (BVH) [341] and compute local conic 3D distance fields Ψ defined by the triangles \mathcal{C} and their normals n . Penetrations are then penalized by the depth of intrusion, efficiently computed by the position in the distance field. For two colliding triangles f_s and f_t , intrusion is bi-directional; the vertices v_t of f_t are the *intruders* in the distance field Ψ_{f_s} of the *receiver* triangle f_s and are penalized by $\Psi_{f_s}(v_t)$, and vice-versa. Thus, the collision term E_C in the objective (Eq. (2.4)) is defined as

$$E_C(\theta) = \sum_{(f_s(\theta), f_t(\theta)) \in \mathcal{C}} \left\{ \sum_{v_s \in f_s} \| -\Psi_{f_t}(v_s)n_s \|^2 + \sum_{v_t \in f_t} \| -\Psi_{f_s}(v_t)n_t \|^2 \right\}. \quad (2.12)$$

For technical details about Ψ , as well as details about handling collisions for parts with permanent or frequent self-contact we redirect the reader to [25, 356] and Sec. A.2. For computational efficiency, we use a highly parallelized implementation of BVH following [172] with a custom CUDA kernel wrapped around a custom PyTorch operator.

2.3.5 Deep gender classifier

Humans of different genders have different proportions and shapes. Consequently, using the appropriate body model to fit 2D data means that we should apply the appropriate shape space. We know of no previous method that automatically takes gender into account in fitting 3D human pose. Here, we train a gender classifier that takes as input an image containing the full body and the OpenPose joints, and assigns a gender label to the detected person. To this end, we first annotate through Amazon Mechanical Turk a large dataset of images from LSP [160], LSP-extended [161], MPII [13], COCO [215], and LIP dataset [212], while following their official splits for train and test sets. The final dataset includes 50216 training examples and 16170 test samples, see Sec. A.9. We use this dataset to fine tune a pretrained ResNet18 [130] for binary gender classification. Moreover, we threshold

the computed class probabilities, by using a class-equalized validation set, to obtain a good trade-off between discarded, correct, and incorrect predictions. We choose a threshold of 0.9 for accepting a predicted class, which yields 62.38% correct predictions, and 7.54% incorrect predictions on the validation set. At test time, we run the detector and fit the appropriate gendered model. When the detected class probability is below the threshold, we fit the gender-neutral body model.

2.3.6 Optimization

SMPLify employs Chumpy and OpenDR [223] which makes the optimization slow. To keep optimization of Eq. (2.4) tractable, we use PyTorch and the Limited-memory BFGS optimizer (L-BFGS) [258] with strong Wolfe line search. Implementation details can be found in Sec. A.3.

We optimize Eq. (2.4) with a multistage approach, similar to [34]. We assume that we know the exact or an approximate value for the focal length of the camera. Then we first estimate the unknown camera translation and global body orientation (see [34]). We then fix the camera parameters and optimize body shape, β , and pose, θ . Empirically, we found that an *annealing scheme* for the weights γ in the data term E_J (Eq. (2.5)) helps optimization of the objective (Eq. (2.4)) to deal with ambiguities and local optima. This is mainly motivated by the fact that small body parts like the hands and face have many keypoints relative to their size, and can dominate in Eq. (2.4), encouraging the optimizer to search first for local optimums for the hands and face. Although these local optima may satisfy the hand and face constraint, they are far from the solution and produce suboptimal body pose estimates.

In the following, we denote by γ_b the weights corresponding to the main body keypoints, γ_h the ones for hands and γ_f the ones for facial keypoints. We then follow three steps, starting with high regularization to mainly refine the global body pose, and gradually increase the influence of hand keypoints to refine the pose of the arms. After converging to a body pose estimate, we increase the influence of both hands and facial keypoints to capture expressivity. Throughout the above steps the weights $\lambda_\alpha, \lambda_\beta, \lambda_\psi$ in Eq. (2.4) start with high regularization that gradually lowers to allow for better fitting, The only exception is λ_C that gradually increases while the influence of hands gets stronger in E_J and more collisions are expected.

2.4 EXPERIMENTS

2.4.1 Evaluation datasets

Despite the recent interest in more expressive models [166, 293] there exists no dataset containing images with ground-truth shape for bodies, hands and faces together. Consequently, we create a dataset for evaluation from currently available data through fitting and careful curation.

Expressive hands and faces dataset (EHF): We begin with the SMPL+H dataset [234], obtaining one full body RGB image per frame. We then align SMPL-X to the 4D scans following [293]. An expert annotator manually curated the dataset to select 100 frames that can be confidently considered pseudo ground-truth, according to alignment quality and interesting hand poses and facial expressions. The pseudo ground-truth meshes enable the use of *vertex-to-vertex* (V2V) error metric [222, 272], in contrast to the common paradigm of reporting 3D joint error, which does not capture surface errors and rotations along the bones.

2.4.2 Qualitative & quantitative evaluations

To test the effectiveness of SMPL-X and SMPLify-X, we perform comparisons to the most related models, namely SMPL [222], SMPL+H [293], and Frank [166]. We fit SMPL-X to the EHF images to evaluate both *qualitatively* and *quantitatively*. Note that we use *only* 1 image and 2D joints as input, while previous methods use *much more* information; i.e. 3D point clouds [166, 293] and joints [166]. Specifically [222, 293] employ 66 cameras and 34 projectors, while Joo et al. [166] employ more than 500 cameras.

We first compare to SMPL, SMPL+H and SMPL-X on the EHF dataset and report results in Tab. 2.1. The table reports *mean vertex-to-vertex* (V2V) error and *mean 3D body joint* error after Procrustes alignment with the ground-truth 3D meshes and body (only) joints respectively. To ease numeric evaluation, for this table only, we “simulate” SMPL and SMPL+H with a SMPL-X variation with locked degrees of freedom, noted as “SMPL” and “SMPL+H” respectively. As expected, the errors show that the standard mean 3D joint error fails to accurately capture the difference in model expressivity. On the other hand, the much stricter V2V metric shows that enriching the body with finger and face modeling results in lower errors. We also fit SMPL with additional features for parts that are not properly modeled, e.g. finger features. The additional features result in an increasing

Model	Keypoints	V2V (mm)	Joint error (mm)
“SMPL”	Body	57.6	63.5
“SMPL”	Body+Hands+Face	64.5	71.7
“SMPL+H”	Body+Hands	54.2	63.9
SMPL-X	Body+Hands+Face	52.9	62.6

TABLE 2.1: Quantitative comparison of “SMPL”, “SMPL+H” and SMPL-X, as described in Sec. 2.4.2, fitted with SMPLify-X on the EHF dataset. We report the mean vertex-to-vertex (V2V) and the standard mean 3D body (only) joint error in mm. The table shows that richer modeling power results in lower errors.

Version	V2V (mm)
SMPLify-X	52.9
gender neutral model	58.0
replace VPoser with GMM	56.4
no collision term	53.5

TABLE 2.2: Ablative study for SMPLify-X on the EHF dataset. The numbers reflect the contribution of each component in overall accuracy.

error, pointing to the importance of richer and more expressive models. We report similar qualitative comparisons in Sec. A.1.

We then perform an ablative study, summarized in Tab. 2.2, where we report the *mean vertex-to-vertex* (V2V) error. SMPLify-X with a gender-specific model achieves 52.9 mm error. The gender neutral model is easier to use, as it does not need gender detection, but comes with a small compromise in terms of accuracy. Replacing VPoser with the GMM of SMPLify [34] increases the error to 56.4 mm, showing the effectiveness of VPoser. Finally, removing the collision term increases the error as well, to 53.5 mm, while also allowing non physically plausible pose estimates.

The closest comparable model to SMPL-X is Frank [166]. Since Frank is not available, nor are the fittings to [66], we show images of results found online. Figure 2.3 shows Frank fittings to 3D joints *and* point clouds, i.e. using more than 500 cameras. Compare this with SMPL-X fitting, which is done with SMPLify-X using *only* 1 RGB image with 2D joints. For a more

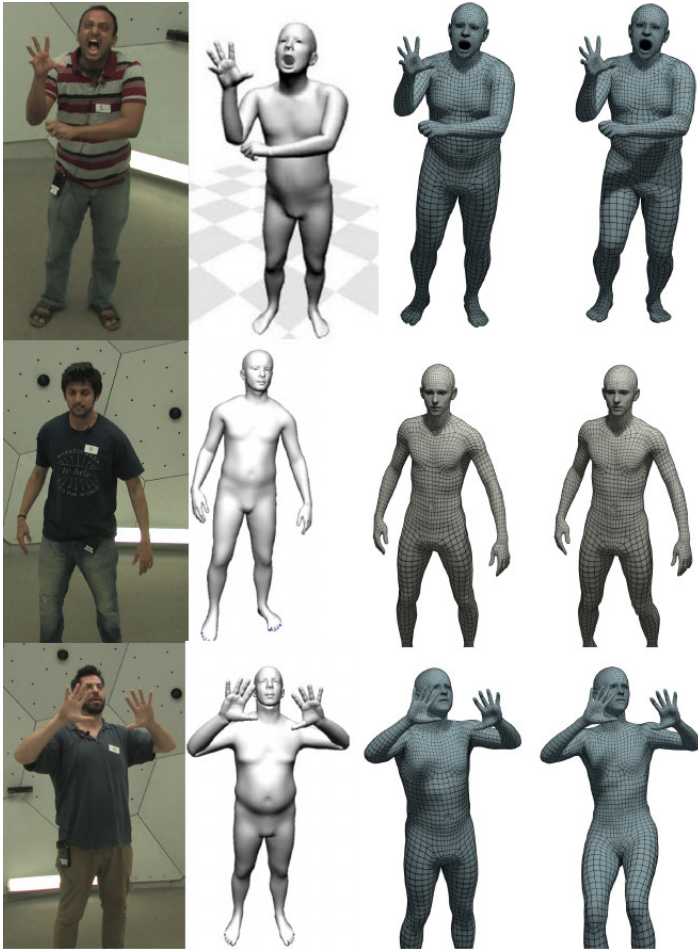


FIGURE 2.3: From left to right: (i) Reference RGB, (ii) [166]: > 500 cameras, (iii) Ours > 500 cameras and (iv) Ours 1 camera. Qualitative comparison of our gender neutral model (top, bottom rows) or gender specific model (middle) against Frank [166] on some of their data. To fit Frank, Joo et al. [166] employ both 3D joints and a point cloud, using more than 500 cameras to obtain the 3D information. In contrast, our method produces a realistic and expressive reconstruction using *only* 2D joints. We show results using the 3D joints of [166] projected in 1 camera view (third column), as well as using joints estimated from only 1 image (last column), to show the influence of noise in 2D joint detection. Compared to Frank, our model does *not* have skinning artifacts around the joints, e.g. elbows.

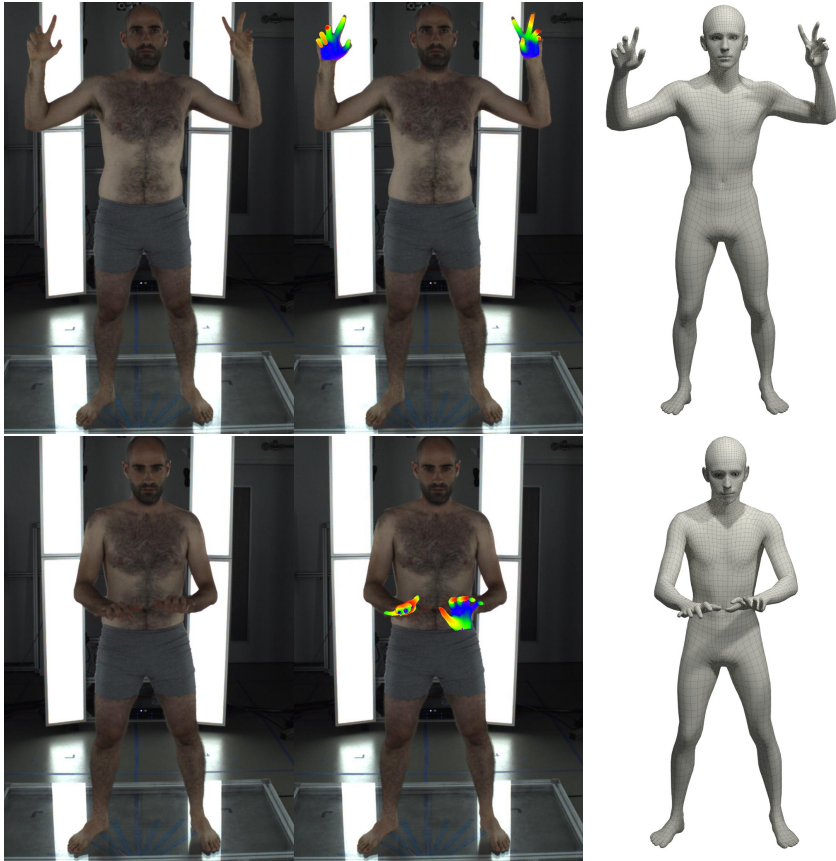


FIGURE 2.4: Comparison of the hands-only approach of Panteleris et al. [266] (middle) against our approach with the male model (right). Both approaches depend on OpenPose. In case of good detections both perform well (top). In case of noisy 2D detections (bottom) our holistic model shows increased robustness.

direct comparison here, we fit SMPL-X to 2D projections of the 3D joints that [166] used for Frank. Although we use *much less* data, SMPL-X shows at least similar expressivity to Frank for both the face and hands. Since Frank does not use pose blend shapes, it suffers from skinning artifacts around the joints, e.g. elbows, as clearly seen in Fig. 2.3. SMPL-X by contrast, is trained to include pose blend shapes and does not suffer from this. As a result it looks more *natural* and *realistic*.



FIGURE 2.5: Qualitative results of SMPL-X for the in-the-wild images of the LSP dataset [160]. A strong holistic model like SMPL-X results in *natural* and *expressive* reconstruction of bodies, hands and faces. Gray color depicts the gender-specific model for confident gender detections. Blue is the gender-neutral model that is used when the gender classifier is uncertain.

To further show the value of a holistic model of the body, face and hands, in Fig. 2.4 we compare SMPL-X and SMPLify-X to the hands-only approach of [266]. Both approaches employ OpenPose for 2D joint detection, while [266] further depends on a hand detector. As seen in Fig. 2.4, in case of good detections both approaches perform nicely, though in case of noisy detections, SMPL-X shows increased robustness due to the context of the body. We further perform a quantitative comparison after aligning the resulting fits to EHF. Due to different mesh topology, for simplicity we use hand joints as pseudo ground-truth, and perform Procrustes analysis of each hand independently, ignoring the body. Panteleris et al. [266] achieve a mean 3D joint error of 26.5 mm, while SMPL-X has 19.8 mm.

Finally, we fit SMPL-X with SMPLify-X to some in-the-wild datasets, namely the LSP [160], LSP-extended [161] and MPII datasets [13]. Figure 2.5 shows some qualitative results for the LSP dataset [160]; see Sec. A.6 for more examples and failure cases. The images show that a strong holistic model like SMPL-X can effectively give *natural* and *expressive* reconstruction from everyday images.

2.5 CONCLUSION

In this chapter, we present SMPL-X, a model that *jointly* captures the body together with face and hands. We additionally present SMPLify-X, an approach to fit SMPL-X to a single RGB image and 2D OpenPose joint detections. We regularize fitting under ambiguities with a new powerful body pose prior and a fast and accurate method for detecting and penalizing penetrations. We present a wide range of qualitative results using images in-the-wild, showing the expressivity of SMPL-X and the effectiveness of SMPLify-X. We introduce a curated dataset with pseudo ground-truth to perform quantitative evaluation, that shows the importance of more expressive models. We believe that this is an important step towards *expressive* capture of bodies, hands, and faces *together* from an RGB image.

MONOCULAR EXPRESSIVE BODY REGRESSION THROUGH BODY-DRIVEN ATTENTION

3.1 INTRODUCTION

A long-term goal of computer vision is to understand humans and their behavior in everyday scenarios using only images. Are they happy or sad? How do they interact with each other and the physical world? What are their intentions? To answer such difficult questions, we first need to *quickly* and *accurately* reconstruct their 3D body, face and hands *together* from a single RGB image. This is very challenging. As a result, the community has broken the problem into pieces with much of the work focused on estimating either the main body [103, 244, 304], the face [429] or the hands [81, 335, 396] separately.

Only recent advances have made the problem tractable in its full complexity. Early methods estimate 2D joints and features [43, 131] for the body, face, and hands. However, this is not enough. It is the skin surface that describes important aspects of humans, e.g. what their precise 3D shape is, whether they are smiling, gesturing, or, holding something. For this reason, strong statistical parametric models for expressive 3D humans were introduced, namely, Adam [166], SMPL-X, described in Chapter 2, and GHUM/GHUML [387]. Such models are attractive because they facilitate reconstruction even from ambiguous data, working as a strong prior.

The first methods that estimate full expressive 3D humans from an RGB image [270, 381, 387], using SMPL-X, Adam and GHUM/GHUML were based on optimization, therefore they are slow, prone to local optima, and rely on heuristics for initialization. These issues significantly limit the applicability of these methods. In contrast, recent body-only methods [169, 187] directly regress 3D SMPL bodies quickly and relatively reliably from an RGB image.

Here we present a *fast* and *accurate* model that reconstructs full *expressive* 3D humans, by estimating SMPL-X parameters directly from an RGB image. This is a hard problem and we show that it is not easily solved by extending SMPL neural-network regressors to SMPL-X for several reasons. First, SMPL-X is a much higher dimensional model than SMPL. Second, there exists no large in-the-wild dataset with SMPL-X annotations for training.

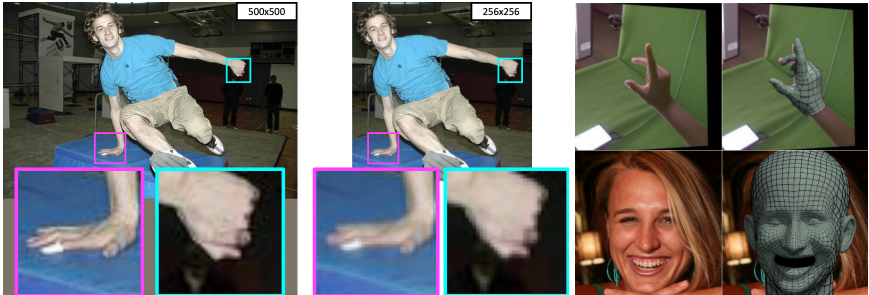


FIGURE 3.1: *Left*: Full-body RGB images of people contain many more pixels on the body than on the face or hands. *Middle*: Images are typically downsized (e.g. to 256×256 px) for use in neural networks. This resolution is fine for the body but the hands and face suffer from low resolution. Our model, see Fig. 3.2, uses *body-driven attention* to restore the lost information for hands and faces from the original image, feeding it to dedicated refinement modules. *Right*: These modules give more expressive hands and faces, by exploiting *part-specific knowledge* learned from higher quality hand-only [427] and face-only [173] datasets; green meshes show example part-specific training data.

Third, the face and hands are often blurry and occluded in images. At any given image resolution, they also occupy many fewer pixels than the body, making them low resolution. Fourth, for technical reasons, full-body images are typically downscaled for input to neural networks [191], e.g. to 256×256 pixels. As shown in Fig. 3.1, this results in even lower resolution for the hands and face, making inference difficult.

Our model and training method, shown in Fig. 3.2, tackles all these challenges. We account for data scarcity by introducing a new dataset with paired in-the-wild images and SMPL-X annotations. To this end, we employ several standard in-the-wild body datasets [13, 160, 161, 215] and fit SMPL-X to them with SMPLify-X, see Sec. 2.3.2. We semi-automatically curate these fits to keep only the good ones as pseudo ground-truth. We then train a model that regresses SMPL-X parameters from an RGB image, similar to [169]. However, this only estimates rough hand and face configurations, due to the problems described above. We observe that the main body is estimated well, on par with [169, 187], providing good rough localization for the face and hands. We use this for *body-driven attention* and focus the network back on the *original* non-downscaled image for the face and hands. We retrieve high-resolution information for these regions and feed this to dedicated *refinement* modules. These modules act as an *expressivity boost*

by distilling *part-specific knowledge* from high-quality hand-only [427] and face-only [218] datasets. Finally, the independent components are fine-tuned jointly end-to-end, so that the part networks can benefit from the full-body initialization.

We call the final model ExPose (EXpressive POse and Shape rEGression). ExPose is at least as accurate as existing optimization-based methods for estimating expressive 3D humans, e.g. SMPLify-X from Sec. 2.3.2, while running two orders of magnitude faster. Our data, model and code are available for research at <https://expose.is.tue.mpg.de>.

3.2 RELATED WORK

Human Modeling: Modeling and capturing the whole human body is a challenging problem. To make it tractable, the community has studied the body, face and hands separately, in a divide-and-conquer fashion. For the human *face*, the seminal work of Blanz and Vetter [33] introduces the first 3D morphable model. Since then, numerous works (see [79]) propose more powerful face models and methods to infer their parameters. For human *hands* the number of models is limited, with Khamis et al. [176] learning a model of hand shape variation from depth images, while Romero et al. [293] learn a parametric hand model with both a rich shape and pose space from 3D hand scans. For the human *body*, the introduction of the CAESAR dataset [290] enables the creation of models that disentangle shape and pose, such as SCAPE [15] and SMPL [222], to name a few. However, these models have a neutral face and the hands are non-articulated. In contrast, Adam [166], SMPL-X, see Sec. 2.3.1, and GHUM [387] are the first models that represent the body, face, and hands jointly. Adam lacks the pose-dependent blendshapes of SMPL and the released version does not include a face model.

Human Pose Estimation: Often pose estimation is treated as the estimation of 2D or 3D keypoints, corresponding to anatomical joints or landmarks [39, 43, 319]. In contrast, recent advances use richer representations of the 3D body surface in the form of parametric [34, 169, 262, 272] or non-parametric [188, 299, 359] models.

To estimate *bodies* from images, many methods break the problem down into stages. First, they estimate some intermediate representation such as 2D joints [34, 113, 114, 140, 169, 237, 272, 318, 352, 419], silhouettes [4, 140, 272], part labels [262, 298] or dense correspondences [116, 294]. Then, they reconstruct the body pose out of this proxy information, by either using it

in the data term of an optimized energy function [34, 140, 403] or “lifting” it using a trained regressor [169, 237, 262, 272, 352]. Due to ambiguities in lifting 2D to 3D, such methods use various priors for regularization, such as known limb lengths [198], a pose prior for joint angle limits [5], or a statistical body model [34, 140, 262, 272] like SMPL [222]. The above 2D proxy representations have the advantage that annotation for them is readily available. Their disadvantage is that the eventual regressor does not get to exploit the original image pixels and errors made by the proxy task cannot be overcome.

Other methods predict 3D pose directly from RGB pixels. Intuitively, they have to learn a harder mapping, but they avoid information bottlenecks and additional sources of error. Most methods infer 3D body joints [205, 271, 331, 332, 340], parametric methods estimate model parameters [169, 170, 187], while non-parametric methods estimate 3D meshes [188], depth maps [98, 323] voxels [359, 421] or distance fields [299, 300]. Datasets of paired indoor images and MoCap data [150, 317] allow supervised training, but may not generalize to in-the-wild data. To account for this, Rogez and Schmid [292] augment these datasets by overlaying synthetic 3D humans, while Kanazawa et al. [169] include in-the-wild datasets [13, 160, 161, 215] and employ a re-projection loss on their 2D joint annotations for weak supervision.

Similar observations can be made in the human hand and face literature. For *hands*, there has been a lot of work on RGB-D data [396], and more recent interest in monocular RGB [20, 36, 120, 126, 151, 194, 247, 339, 426]. Some of the non-parametric methods estimate 3D joints [151, 247, 339, 426], while others estimate 3D meshes [104, 193]. Parametric models [20, 36, 126, 194, 415] estimate configurations of statistical models like MANO [293] or a graph morphable model [194]. For *faces*, 3D reconstruction and tracking has a long history. We refer the reader to a recent comprehensive survey [429].

Attention for Human Pose Estimation: In the context of human pose estimation, attention is often used to improve prediction accuracy. Successful architectures for 2D pose estimation, like Convolutional Pose Machines [371] and Stacked Hourglass [255] include a series of processing stages, where the intermediate pose predictions in the form of heatmaps are used as input to the following stages. This informs the network of early predictions and guides its attention to relevant image pixels. Chu et al. [58] build explicit attention maps, at a global and part-specific level, driving the model to focus on regions of interest. Instead of predicting attention maps, our approach uses the initial body mesh prediction to define the areas of attention for

hands- and face-specific processing networks. A similar practice is used by OpenPose [43], where arm keypoints are used to estimate hand bounding boxes, in a heuristic manner. Additionally, for HoloPose [115], body keypoints are used to pool part-specific features from the image.

A critical difference of ExPose is that, instead of simply pooling already computed features, we also process the region of interest at higher resolution, to capture more subtle face and hand details. In related work, Chandran et al. [46] use a low resolution proxy image to detect facial landmarks and extract high resolution crops that are used to refine facial landmark predictions.

Expressive Human Estimation: Since expressive parametric models of the human body have only recently been introduced [166, 270, 293, 387], there are only a few methods to reconstruct their parameters. Joo et al. [166] present an early approach, but rely on an extended multi-view setup. More recently, Xiang et al. [381], SMPLify-X, described in Sec. 2.3.2, and Xu et al. [387] use a single image to recover Adam, SMPL-X, and GHUM parameters respectively, using optimization-based approaches. This type of inference can be slow and may fail in the presence of noisy feature detections. In contrast, we present the first regression approach for expressive monocular capture and show that it is both more accurate and significantly faster than prior work.

3.3 METHOD

3.3.1 3D Body Representation

To represent the human body, we use SMPL-X, described in Sec. 2.3.1. We denote posed joints with $J(\theta, \beta) \in \mathbb{R}^{J \times 3}$. The final set of predicted SMPL-X parameters is the vector $\Theta = \{\beta, \theta, \psi\} \in \mathbb{R}^{338}$, where $\beta \in \mathbb{R}^{10}$, $\psi \in \mathbb{R}^{10}$ and $\theta \in \mathbb{R}^{J \times D}$, with $J = 53$ and $D = 6$, as we choose to represent the pose parameters θ using the representation of Zhou et al. [422]. Note that we do not predict pose variables for the eyes.

3.3.2 Body-driven Attention

Instead of attempting to regress body, hand, and face parameters from a low-resolution image crop we design an attentive architecture that uses the structure of the body and the full resolution of the image I . Given a bounding box of the body, we extract an image I_b , using an affine trans-

formation $\mathbb{T}_b \in \mathbb{R}^{2 \times 3}$, from the high resolution image I . The body crop I_b is fed to a neural network g , similar to HMR [169], to produce a first set of SMPL-X parameters Θ_b and weak-perspective camera scale $s_b \in \mathbb{R}$ and translation $t_b \in \mathbb{R}^2$. After posing the model and recovering the posed joints \mathbb{J} , we project them on the image:

$$\mathbf{j} = s(\Pi_o(\mathbb{J}) + t) \quad (3.1)$$

where Π_o is the orthographic projection operator. We then compute a bounding box for each hand and the face, from the corresponding subsets of projected 2D joints, \mathbf{j}_h and \mathbf{j}_f . Let (x_{\min}, y_{\min}) and (x_{\max}, y_{\max}) be the top left and bottom right points for a part, computed from the respective joints. The bounding box center is equal to $\mathbf{c} = \left(\frac{x_{\min} + x_{\max}}{2}, \frac{y_{\min} + y_{\max}}{2} \right)$, and its size is $b_s = 2 \cdot \max(x_{\max} - x_{\min}, y_{\max} - y_{\min})$. Using these boxes, we compute affine transformations $\mathbb{T}_h, \mathbb{T}_f \in \mathbb{R}^{2 \times 3}$ to extract higher resolution hand and faces images using spatial transformers (ST) [154]:

$$I_h = \text{ST}(I; \mathbb{T}_h), I_f = \text{ST}(I; \mathbb{T}_f). \quad (3.2)$$

The hand I_h and face I_f images are fed to a hand network h and a face network f , to refine the respective parameter predictions. Hand parameters θ_h include the orientation of the wrist θ^{wrist} and finger articulation θ^{fingers} , while face parameters contain the expression coefficients ψ_f and facial pose θ_f , which is just the rotation of the jaw. We refine the parameters of the body network by predicting offsets for each of the parameters and condition the part specific networks on the corresponding body parameters:

$$\begin{aligned} [\Delta\theta^{\text{wrist}}, \Delta\theta^{\text{fingers}}] &= h(I_h; \theta_b^{\text{wrist}}, \theta_b^{\text{fingers}}), \\ [\Delta\theta_f, \Delta\psi] &= f(I_f; \theta_b^f, \psi_b) \end{aligned} \quad (3.3)$$

where $\theta_b^{\text{wrist}}, \theta_b^{\text{fingers}}, \theta_b^f, \psi_b$ are the wrist pose, finger pose, facial pose and expression predicted by $g(\cdot)$. The hand and head sub-networks also produce a set of weak-perspective camera parameters $\{s_h, t_h\}, \{s_f, t_f\}$ that align the predicted 3D meshes to their respective images I_h and I_f . The final hand and face predictions are then equal to:

$$\theta_h = [\theta^{\text{wrist}}, \theta^{\text{fingers}}] = [\theta_b^{\text{wrist}}, \theta_b^{\text{fingers}}] + [\Delta\theta_{\text{wrist}}, \Delta\theta_{\text{fingers}}] \quad (3.4)$$

$$[\psi, \theta_f] = [\psi_b, \theta_b^f] + [\Delta\psi, \Delta\theta_f]. \quad (3.5)$$

With this approach, we can utilize the full resolution of the original image I to overcome the small pixel resolution of the hands and face in the body image I_b . Another significant advantage is that we are able to leverage hand- and face-only data to supplement the training of the hand and face sub-networks. A detailed visualization of the prediction process can be seen in Fig. 3.2. The loss function used to train the model is a combination of terms for the body, the hands and the face. We train the body network using a combination of a 2D re-projection loss, 3D joint errors, and a loss on the parameters Θ , when available. All variables with a hat denote ground-truth quantities.

$$\mathcal{L} = \mathcal{L}_{\text{body}} + \mathcal{L}_{\text{hand}} + \mathcal{L}_{\text{face}} + \mathcal{L}_h + \mathcal{L}_f \quad (3.6)$$

$$\mathcal{L}_{\text{body}} = \mathcal{L}_{\text{reproj}} + \mathcal{L}_{\text{3D Joints}} + \mathcal{L}_{\text{SMPL-X}} \quad (3.7)$$

$$\mathcal{L}_{\text{reproj}} = \sum_{n=1}^J v_n \|\hat{j}_n - j_n\|_1. \quad (3.8)$$

$$\mathcal{L}_{\text{3D Joints}} = \sum_{n=1}^J \|\hat{J}_n - J_n\|_1 \quad (3.9)$$

$$\mathcal{L}_{\text{SMPL-X}} = \|\{\hat{\beta}, \hat{\theta}, \hat{\psi}\} - \{\beta, \theta, \psi\}\|_2^2 \quad (3.10)$$

We use v_n as a binary variable denoting visibility of each of the J joints. The re-projection losses \mathcal{L}_h and \mathcal{L}_f are applied in the hand and face image coordinate space, using the affine transformations T_h, T_f . The reason for this extra penalty is that alignment errors in the 2D projection of the fingers or the facial landmarks have a much smaller magnitude compared to those of the main body joints when computed on the body image I_b

$$\begin{aligned} \mathcal{L}_h &= \sum_{n \in \text{Hand}} v_n \left\| T_h T_b^{-1} (\hat{j}_n - j_n) \right\|_1, \\ \mathcal{L}_f &= \sum_{n \in \text{Face}} v_n \left\| T_f T_b^{-1} (\hat{j}_n - j_n) \right\|_1. \end{aligned} \quad (3.11)$$

For the hand and head only data we also employ a re-projection loss, using only the subset of joints of each part, and parameter losses:

$$\mathcal{L}_{\text{hand}} = \mathcal{L}_{\text{reproj}} + \|\{\hat{\beta}_h, \hat{\theta}_h\} - \{\beta_h, \theta_h\}\|_2^2 \quad (3.12)$$

$$\mathcal{L}_{\text{face}} = \mathcal{L}_{\text{reproj}} + \|\{\hat{\beta}_f, \hat{\theta}_f, \hat{\psi}_f\} - \{\beta_f, \theta_f, \psi_f\}\|_2^2. \quad (3.13)$$

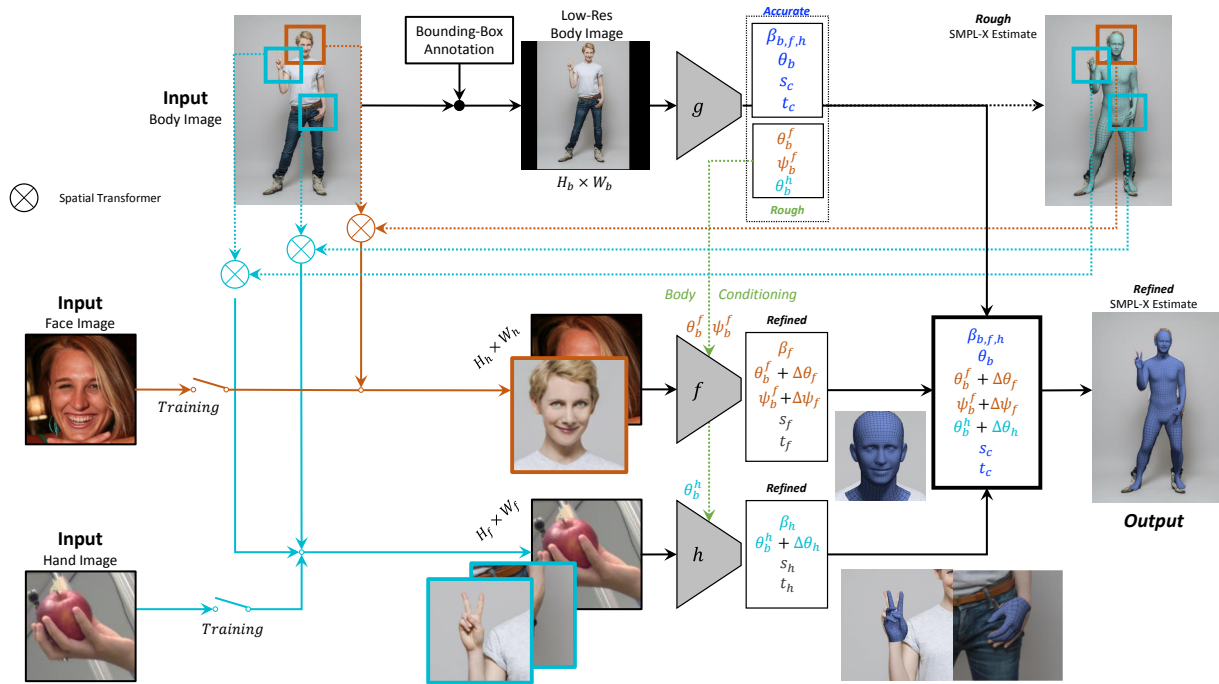


FIGURE 3.2: An image of the body is extracted using a bounding box from the full resolution image and fed to a neural network $g(\cdot)$, that predicts body pose θ_b , hand pose θ_h , facial pose θ_f , shape β , expression ψ , camera scale s and translation t . Face and hand images are extracted from the original resolution image using bilinear interpolation. These are fed to part specific sub-networks $f(\cdot)$ and $h(\cdot)$ respectively to produce the final estimates for the face and hand parameters. During training the part specific networks can also receive hand and face only data for extra supervision.

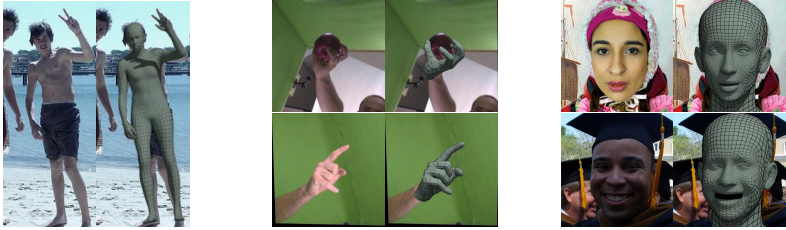


FIGURE 3.3: *Left*: Example curated expressive fit. *Middle*: Hands sampled from the FreiHAND dataset [427]. *Right*: Head training data produced by running RingNet [302] on FFHQ [173] and then fitting to 2D landmarks predicted by [39].

3.3.3 Implementation Details

Training Datasets: We curate a dataset of SMPL-X fits by running vanilla SMPLify-X [270] on the LSP [160], LSP extended [161] and MPII [13] datasets. We then ask human annotators whether the resulting body mesh is plausible and agrees with the image and collect 32,617 pairs of images and SMPL-X parameters. fits of SPIN [187] from SMPL to SMPL-X, see Sec. B.2. Moreover, we use H3.6M [150] for additional 3D supervision for the body. For the hand sub-network we employ the hand-only data of FreiHAND [427]. For the face sub-network we create a pseudo ground-truth face dataset by running RingNet [302] on FFHQ [173]. The regressed FLAME [207] parameters are refined by fitting to facial landmarks [39] for better alignment with the image and more detailed expressions. Figure 3.3 shows samples from all training datasets.

Architecture: For the body network we extract features $\phi \in \mathbb{R}^{2048}$ with HRNet [330]. For the face and hand sub-networks we use a ResNet18 [130] to limit the computational cost. For all networks, rather than directly regressing the parameters Θ from ϕ , we follow the iterative process of [169]. We start from an initial estimate $\Theta_0 = \bar{\Theta}$, where $\bar{\Theta}$ represents the mean, which is concatenated to the features ϕ and fed to an MLP that predicts a residual $\Delta\Theta_1 = \text{MLP}([\phi, \Theta_0])$. The new parameter value is now equal to $\Theta_1 = \Theta_0 + \Delta\Theta_1$ and the whole process is repeated. As in [169], we iterate for $t = 3$ times. The entire pipeline is implemented in PyTorch [268]. For architecture details see Sec. B.1.

Data Pre-processing and Augmentation: We follow the pre-processing and augmentation protocol of [187] for all networks. To make the model robust to partially visible bodies we adopt the cropping augmentation of Joo et al. [163]. In addition, we augment the hand- and face-only images

with random translations, as well as down-sampling by a random factor and then up-sampling back to the original resolution. The former simulates a misaligned body prediction, while the latter bridges the gap in image quality between the full-body and part-specific data. Hand and especially face images usually have a much higher resolution and quality compared to a crop extracted from a full-body image. To simulate body conditioning for the hand- and head-only data we add random noise to the initial point of the iterative regressor. For the hands we replace the default finger pose with a random rotation r_{finger} sampled from the PCA pose space of MANO. For the head we replace the default jaw rotation $\bar{\theta}_f$ with a random rotation of $r_f \sim \mathcal{U}(0, 45)$ degrees around the x-axis. For both parts, we replace their global rotation with a random rotation with angle $r_{\text{global}} \sim \mathcal{U}(r_{\text{min}}, r_{\text{max}})$ and the same axis of rotation as the corresponding ground-truth. We set $(r_{\text{min}}, r_{\text{max}})_{\text{hand}}$ to $(-90, 90)$ and $(r_{\text{min}}, r_{\text{max}})_{\text{face}}$ to $(-45, 45)$ degrees. The default mean shape is replaced with a random vector $\beta \sim \mathcal{N}(\mathbf{0}, I)$, $I \in \mathbb{R}^{10 \times 10}$ and the default neutral expression with a random expression $\psi \sim \mathcal{N}(0, \mathcal{I})$. Some visualizations of the different types of data augmentation can be found in Sec. B.2.

Training: We first pre-train the body, hand and face networks separately, using ADAM [180], on the respective part-only datasets. We then fine-tune all networks jointly on the union of all training data, following Sec. 3.3.2, letting the network make even better use of the conditioning, see Sec. 3.4 and Tab. 3.2. Please note that for this fine-tuning, our new dataset of curated SMPL-X fits plays an instrumental role. Our exact hyper-parameters are included in the released code.

3.4 EXPERIMENTS

3.4.1 Evaluation Datasets

We evaluate on several datasets:

Expressive Hands and Faces (EHF): We use this dataset, described in Sec. 2.4.1, to evaluate our whole-body predictions.

3D Poses in the Wild (3DPW) [235] consists of in-the-wild RGB video sequences annotated with 3D SMPL poses. It contains several actors performing various motions, in both indoor and outdoor environments. It is captured using a single RGB camera and IMUs mounted on the subjects. We use it to evaluate our predictions for the main body area, excluding the head and hands.

FreiHAND [427] is a multi-view RGB hand dataset that contains 3D MANO hand pose and shape annotations. The ground-truth for the test data is held-out and evaluation is performed by submitting the estimated hand meshes to an online server. We use it to evaluate our hand sub-network predictions.

Stirling/ESRC 3D [90] consists of facial RGB images with ground-truth 3D face scans. It contains 2000 neutral face images, namely 656 high-quality (HQ) ones and 1344 low-quality (LQ) ones. We use it to evaluate our face sub-network following the protocol of [90].

3.4.2 Evaluation Metrics

We employ several common metrics below. We report errors with and without rigid alignment to the ground-truth. A “PA” prefix denotes that the metric measures error after solving for rotation, scale, and translation using Procrustes Alignment.

To compare with ground-truth 3D skeletons, we use the **Mean Per-Joint Position Error (MPJPE)**. For this, we first compute the 14 LSP-common joints, by applying a linear joint regressor on the ground-truth and estimated meshes, and then compute their mean Euclidean distance.

For comparing to ground-truth meshes, we use the **Vertex-to-Vertex (V2V)** error, i.e. the mean distance between the ground-truth and predicted mesh vertices. This is appropriate when the predicted and ground-truth meshes have the same topology, e.g. SMPL-X for our overall network, MANO for our hand and FLAME for our face sub-network. For a fair comparison to methods that predict SMPL instead of SMPL-X, like [169, 187], we also report V2V only on the main body, i.e. without the hands and the head, as SMPL and SMPL-X share common topology for this subset of vertices.

For comparing to approaches that output meshes with different topology, like MTC [381] that uses the Adam model and not SMPL-X, we cannot use V2V. Instead, we compute the (mesh-to-mesh) **point-to-surface (P2S)** distance from the ground-truth mesh, as a common reference, to the estimated mesh.

For evaluation on datasets that include ground-truth scans, we compute a **scan-to-mesh** version of the above **point-to-surface** distance, namely from the ground-truth scan points to the estimated mesh surface. We use this for the face dataset of [90] to evaluate the head estimation of our face sub-network.

Method	PA-MPJPE (mm)	MPJPE (mm)	PA-Body V2V (mm)
HMR [169]	81.3	130	65.2
SPIN [187]	59.2	96.9	53.0
ExPose	60.7	93.4	55.6

TABLE 3.1: Comparison on the 3DPW dataset [235] with two state-of-the-art approaches for SMPL regression, HMR [169] and SPIN [187]. The numbers are per-joint and per-vertex errors (in mm) for the body part of SMPL. ExPose outperforms HMR and is on par with SPIN, while also being able to capture details for the hands and the face.

Finally, for the FreiHAND dataset [427] we report all metrics returned by their evaluation server. Apart from PA-MPJPE and PA-V2V described above, we also report the **F-score** [181].

3.4.3 Quantitative and Qualitative Experiments

First, we evaluate our approach on the 3DPW dataset that includes SMPL ground-truth meshes. Although this does not include ground-truth hands and faces, it is ideal for comparing main-body reconstruction against state-of-the-art approaches, namely HMR [169] and SPIN [187]. Table 3.1 presents the results, and shows that ExPose outperforms HMR and is on par with the more recent SPIN. This confirms that ExPose provides a solid foundation upon which to build detailed reconstruction for the hands and face.

We then evaluate on the EHF dataset that includes high-quality SMPL-X ground truth. This allows evaluation for the more challenging task of holistic body reconstruction, including expressive hands and face. Table 3.2 presents an ablation study for our main components. In the first row, we see that the initial body network, which uses a low-resolution body-crop image as input, performs well for body reconstruction but makes mistakes with the hands. The next two rows add *body-driven attention*; they use the body network prediction to locate the hands and face, and then redirect the attention in the original image, crop higher-resolution image patches for them, and feed them to the respective hand and face sub-networks to refine their predictions, while initializing/conditioning their predictions. This conditioning can take place in two ways. The second row shows a naive combination using independently trained sub-networks. This fails

Networks	Attention on high-res. crops	End-to-end fine-tuning	PA-V2V (mm)			
			All	Body	L/R hand	Face
Body only	✗	✗	57.3	55.9	14.3 / 14.8	5.8
Body & Hand & Face	✓	✗	56.4	52.6	14.1 / 13.9	6.0
Body & Hand & Face	✓	✓	54.5	52.6	13.1 / 12.5	5.8

TABLE 3.2: Ablation study on the EHF dataset. The results are vertex-to-vertex errors expressed in mm for the different parts (i.e., all vertices, body vertices, hand vertices and head vertices). We report results for the initial body network applied on the low resolution (first row), for a version that uses the body-driven attention to estimate hands and faces (second row), and for the final regressor that jointly fine-tunes the body, hands and face sub-networks.

to significantly improve the results, since there is a domain gap between images of face- or hand-only [90, 427] training datasets and hand/head image crops from full-body [13, 160, 161] training datasets; the former tend to be of higher resolution and better image quality. Please note that this is similar to [43], but extended for 3D mesh regression. In the third row, the entire pipeline is fine-tuned end-to-end. This results in a boost in quantitative performance, improving mainly hand articulation (best overall performance).

Next, we compare to state-of-the-art approaches again on the EHF dataset. First, we compare against the most relevant baseline, SMPLify-X, which estimates SMPL-X using an optimization approach. Second, we compare against Monocular Total Capture (MTC) [381], which estimates expressive 3D humans using the Adam model. Note that we use their publicly available implementation, which does not include an expressive face model. Third, we compare against HMR [169] and SPIN [187], which estimate SMPL bodies, therefore we perform body-only evaluation, excluding the hand and head regions. We summarize all evaluations in Tab. 3.3. We find that ExPose outperforms the other baselines, both in terms of full expressive human reconstruction and body-only reconstruction. SMPLify-X performs a bit better locally, i.e. for the hands and face, but the full body pose can be inaccurate, mainly due to errors in OpenPose detections. In contrast, our regression-based approach is a bit less accurate locally for the hands and face, but overall it is more robust than SMPLify-X. The two approaches could be combined, with ExPose replacing the heuristic initialization of SMPLify-X with its more robust estimation; we speculate that this

Method	Time (s)	PA-V2V (mm)				PA-MPJPE (mm)		PA-P2S (mm)	
		All	Body	L/R hand	Face	Body Joints	L/R hand	Mean	Median
SMPLify-X'	40-60	52.9	56.37	11.4/12.6	5.3	73.5	11.9/13.2	28.9	18.1
HMR [169]	0.06	N/A	67.2	N/A	N/A	82.0	N/A	34.5	21.5
SPIN [187]	0.01	N/A	60.6	N/A	N/A	102.9	N/A	40.8	28.7
SMPLify-X	40-60	65.3	75.4	11.6/12.9	6.3	87.6	12.2/13.5	36.8	23.0
MTC [381]	20	67.2	N/A	N/A	N/A	107.8	16.3/17.0	41.3	29.0
ExPose (Ours)	0.16	54.5	52.6	13.1/ 12.5	5.8	62.8	13.5/ 12.7	28.9	18.0

TABLE 3.3: Comparison with the state-of-the-art approaches on the EHF dataset. The metrics are defined in Sec. 3.4.2. For SMPLify-X, the results of the first row are generated using ground truth camera parameters, so they are not directly comparable with the other approaches. MTC running time includes calculation of part orientation fields and Adam fitting. The regression-based methods require extra processing to obtain the input human bounding box. For example, if one uses Mask-RCNN [128] with a ResNet50-FPN [214] from Detectron2 [379] the complete running time of these methods increases by 43 milliseconds. All timings were done with an Intel Xeon W-2123 3.60GHz CPU and a Quadro P5000 GPU and are for estimating one person.

would improve both the accuracy and the convergence speed of SMPLify-X. Furthermore, ExPose outperforms MTC across all metrics. Finally, it is approximately two orders of magnitude faster than both SMPLify-X and MTC, which are both optimization-based approaches.

We also evaluate each sub-network on the corresponding part-only datasets. For the hands we evaluate on the FreiHAND dataset [427], and for faces on the Stirling/ESRC 3D dataset [90]. Table 3.4 summarizes all evaluations. The part sub-networks of ExPose match or come close to the performance of state-of-the-art methods. We expect that using a deeper backbone, e.g. a ResNet50, would be beneficial, but at a higher computational cost.

The quantitative findings of Tab. 3.2 are reflected in qualitative results. In Fig. 3.4, we compare our final results with the initial baseline that regresses all SMPL-X parameters directly from a low-resolution image without any attention (first row in Tab. 3.2). We observe that our body-attention mechanism gives a clear improvement for the hand and the face area. Figure 3.5 contains ExPose reconstructions, seen from multiple views, where we again see the higher level of detail offered by our method. For more qualitative results, see Sec. B.5.

FreiHAND	PA-MPJPE (mm)	PA-V2V (mm)	F@5mm	F@15mm
MANO CNN [427]	11.0	10.9	0.516	0.934
ExPose hand sub-network h	12.2	11.8	0.484	0.918
Stirling3D Dataset LQ/HQ	Mean (mm)	Median (mm)	Standard Deviation (mm)	
RingNet [302]	2.08/2.02	1.63/1.58	1.79/1.69	
ExPose face sub-network f	2.27/2.42	1.76/1.91	1.97/2.03	

TABLE 3.4: We evaluate our final hand sub-network on the FreiHAND dataset [427] and the face sub-network on the test dataset of Feng et al. [90]. The final part networks are on par with existing methods, despite using a shallower backbone, i.e. a ResNet-18 vs a Resnet-50.

3.5 CONCLUSION

In this chapter, we present a regression approach for holistic expressive body reconstruction. Considering the different scale of the individual parts and the limited training data, we identify that the naive approach of regressing a holistic reconstruction from a low-resolution body image misses fine details in the hands and face. To improve our regression approach, we investigate a body-driven attention scheme. This results in consistently better reconstructions. Although the pure optimization-based approach [270] recovers the finer details, it is too slow to be practical. ExPose provides competitive results, while being more than two orders of magnitude faster than SMPLify-X. Eventually, the two approaches could be combined effectively, as in [187]. Considering the level of accuracy and speed of our approach, we believe it should be a valuable tool and enable many applications that require expressive human pose information. Future work should extend the inference to multiple humans [156, 403, 404], video sequences [170, 182] and improve the alignment of the body to the image pixels [410]. The rich body representation will also accelerate research on human-scene [124, 305] interaction, human-object [210, 337] interaction, and person-person interaction [91, 204]. In Chapter 5 we describe an approach to improve 3D body shape estimation using partial information, namely anthropometric measurements and linguistic attributes.



FIGURE 3.4: *Left*: The input image. *Middle*: Naive regression from a single body image fails to capture detailed finger articulation and facial expressions. *Right*: ExPose is able to recover these details, thanks to its attention mechanism, and produces results of similar quality as SMPLify-X, while being $200\times$ times faster, see Tab. 3.3.



FIGURE 3.5: Input image, ExPose predictions overlaid on the image and renderings from different viewpoints. ExPose is able to recover detailed hands and faces thanks to its attention mechanism, and produces results of similar quality as SMPLify-X, while being $200\times$ times faster.

COLLABORATIVE REGRESSION OF EXPRESSIVE BODIES USING MODERATION

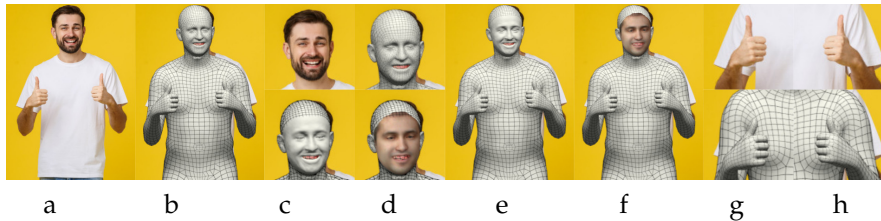


FIGURE 4.1: **PIXIE** estimates expressive 3D humans (b, e, f) from an RGB image (a). For this, it employs experts for the body, face (c, d), and hands (g, h), which are combined (b, e, f) by a novel moderator, according to their confidence (see Fig. 4.2). **PIXIE** estimates appropriate body shapes (b) by implicitly learning to reason about gender from an image. Finally, **PIXIE** estimates fine facial details, i.e. 3D surface displacements (c) and albedo (d), similar to state-of-the-art face-only methods.

4.1 INTRODUCTION

To model human behavior, we need to capture how people look, how they feel, and how they interact with each other. To facilitate this, our goal is to reconstruct whole-body 3D shape and pose, facial expressions, and hand gestures from an RGB image. This is challenging, as humans vary in shape and appearance, they are highly articulated, they wear complex clothing, they are often occluded, and their face and hands are small, yet highly deformable. For these reasons, the community studies the body [34, 169, 187], hands [36, 105, 126, 415] and face [79] mostly separately.

As discussed in the last chapter, recent whole-body statistical models [166, 270, 387] enable approaches to address the problem holistically, by jointly capturing the body, face, and hands. ExPose, see Chapter 3, reconstructs SMPL-X, see Sec. 2.3.1, meshes from an RGB image, using “expert” sub-networks for the body, face, and hands. However, ExPose’s part experts operate completely independently, as they only “see” their respective part

image. Thus, they do not exploit the correlations between parts to overcome challenges like occlusion or motion blur.

Face-only methods [87, 393] are well studied and recover accurate facial shape, albedo, and geometric details, which are important to capture emotions. However, they need a tight crop around the face and struggle with extreme viewing angles and faces that are small, low-resolution or occluded. While whole-body methods [56, 166, 270, 295, 387] handle these challenges well, they estimate average-looking face shapes, without face albedo and fine geometric details.

To get the best of all worlds, we introduce PIXIE (“Pixels to Individuals: eXpressive Image-based Estimation”). PIXIE estimates expressive whole-bodied 3D humans from an RGB image more realistically than existing work. To do so, it pushes the state of the art in three ways.

First, PIXIE learns not only experts for the body, face, and hands, but also a novel moderator that estimates their confidence in each sub-image, and fuses their features weighted by this. The learned fusion helps improve whole-body shape, using SMPL-X’s shared shape space across all body parts. Moreover, it helps to robustly estimate head and hand pose when these are ambiguous (e.g. occlusions or blur) by using full-body context; see Fig. 4.2 for examples.

Second, PIXIE significantly improves “gendered” body shape realism. While human shape is highly correlated with gender, existing work ignores this and estimates inaccurate body shapes – often with the wrong gender or with a gender-neutral shape. An exception is SMPLify-X, but it uses an offline gender classifier and fits a gender-specific SMPL-X model. Instead, using a single unisex SMPL-X model enables end-to-end training of neural nets. PIXIE adopts this approach and learns to implicitly reason about shape. For this, we define male, female, and non-binary body-shape priors within the SMPL-X shape space. At training time, given automatically created gender labels for input images, we train PIXIE to output plausible shape parameters for the specified gender. At inference time, PIXIE needs no gender labels, is applicable to any in-the-wild image, and supports non-binary genders. Note that this approach is general and is relevant for the broader community (face, body, whole-body). Body shape is also correlated with face shape [100, 117, 185]. Thus, we do the same “gendered” training for our face expert; this allows PIXIE to use face information to inform body shape. This training and network architecture significantly improves body shape both qualitatively and quantitatively.

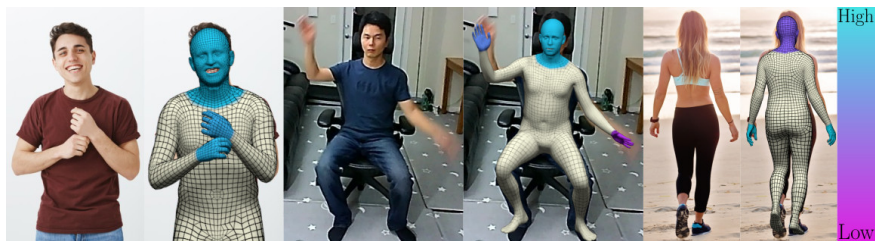


FIGURE 4.2: PIXIE infers the confidence of its body, face and hand experts, and fuses their features accordingly. Challenges, like occlusions, are resolved with full-body context. (L) Input image. (R) Color-coded part-expert confidence.

Third, PIXIE’s face expert additionally infers facial albedo and dense 3D facial-surface displacements. For this, we draw inspiration from Feng et al. [87], and go beyond them in three ways: (1) We use a whole-body shape space, rather than a face-only space, to capture correlations between the body and face shape. (2) We use photometric and identity losses on faces to inform whole-body shape. (3) We use the inferred geometric details only when the face expert is confident, as judged by the moderator. As shown in Fig. 4.1, this results in whole-body 3D humans with detailed faces that can be fully animated.

To summarize, here we make three key contributions: (1) We train a novel moderator, that infers the confidence of body-part experts and fuses their features weighted by this. This improves shape and pose inference under ambiguities. (2) We train the network to implicitly reason about gender, i.e. without gender labels at test time, with a novel “gendered” 3D shape loss that encourages likely body shapes. (3) We extend our face expert with branches that estimate facial albedo and 3D facial-surface displacements, enabling whole-body animation with a realistic face. PIXIE is a step towards automatic, accurate, and realistic 3D avatar creation from a single RGB image. Models and code are available for research purposes: <https://pixie.is.tue.mpg.de>.

4.2 RELATED WORK

Body reconstruction: For years, the community focused on the prediction of 2D or 3D landmarks for the body [43], face [39] and hands [319, 367], with a recent shift towards estimating 3D model parameters [34, 163, 169, 183, 262, 272, 333] or 3D surfaces [188, 213, 299, 300, 359]. One line of work

simplifies the problem by using proxy representations like 2D joints [34, 113, 114, 140, 237, 272, 318, 352, 419], silhouettes [4, 140, 272], part labels [262, 298] or dense correspondences [294, 406]. These are then “lifted” to 3D, either as part of an energy term [34, 140, 403] or using a regressor [237, 262, 272, 352]. To overcome ambiguities, they use priors such as known limb lengths [198], joint angle limits [5], or a statistical body model [34, 140, 262, 270, 272] like SMPL [222] or SMPL-X, see Sec. 2.3.1. While these approaches benefit from 2D annotations, they cannot overcome errors in the proxy features and do not fully exploit image context. The alternative is to directly regress 3D skeletons [205, 271, 331, 332, 340], statistical model parameters [56, 91, 163, 169, 170, 183, 187, 333], 3D meshes [188, 213], depth maps [98, 323], 3D voxels [359, 421] or distance fields [299, 300] from the image pixels.

Face reconstruction: Most modern monocular 3D face estimation methods predict the parameters of a pre-computed statistical face model [79]. Similar to the body literature, this problem is tackled with both optimization [6, 27, 33, 346] and regression methods [89, 153, 302, 343]. Many learning-based approaches follow an analysis-by-synthesis strategy [73, 343, 344], which jointly estimates geometry, albedo, and lighting, to render a synthetic image [223, 284] that is compared with the input. Recent work [73, 87, 108] further employs face-recognition terms [42] during training to reconstruct more accurate facial geometry. Even geometric details, such as wrinkles, can be learned from large collections of in-the-wild images [87, 353]. We refer to Egger et al. [79] for a comprehensive overview. The major downsides of face-specific approaches are their need for tightly cropped face images and their inability to handle non-frontal images. The latter is mainly due to the lack of supervision; 2D landmarks may be missing or the face might not even be detected at all, in which case the photometric term is not applicable. By integrating face and body regression, PIXIE regresses head pose and shape robustly in situations where face-only methods fail and lets the face contribute to whole-body shape estimation.

Hand reconstruction: While hand pose estimation is most often performed from RGB-D data, there has been a recent shift towards the use of monocular RGB images [20, 36, 120, 126, 151, 194, 247, 339, 426]. Similar to the body, we split these into methods that predict 3D joints [151, 247, 339, 426], parameters of a statistical hand model [20, 36, 126, 194, 415], such as MANO [293], or a 3D surface [105, 193].

Whole-body reconstruction: Recent methods approach the problem of human reconstruction holistically. Some of these estimate 3D landmarks

for the body, face and hands [158, 372], but not their 3D surface. This is addressed by whole-body statistical models [166, 270, 387], that jointly capture the 3D surface for the body, face and hands.

SMPLify-X, described in Sec. 2.3.2, fits SMPL-X to 2D body, hand, and face keypoints [43] estimated in an image. Xiang et al. [381] estimate both 2D keypoints and a part orientation field and fit Adam [166] to these. Xu et al. [387] fit GHUM [387] to detected body-part image regions. While these methods work, they are based on optimization, consequently they are slow and do not scale up to large datasets.

Deep-learning methods [56, 295] tackle these limitations, and quickly regress SMPL-X parameters from an image. ExPose, described in Chapter 3, uses “expert” sub-networks for the body, face and hands; the body expert estimates the body and rough part (hand/face) pose from the full-body image, while part experts refine the rough part poses using only local image information (hand/face crop). ExPose merges the output of its experts by always trusting them. Instead, we evaluate the confidence of each expert for each sub-image and fuse body/face and body/hand features weighted by this. To account for different body-part sizes, we use ExPose’s body-driven attention, and multiple data sources for both part-only and whole-body supervision. FrankMocap [295] is similar to ExPose and adds an (optional) optimization step to better align the estimated SMPL-X mesh with the image. Zhou et al. [423] train a network to regress a body-and-hands (SMPL+H) model [293] and the detailed MoFA [344] face model from an RGB image, following a body-part attention mechanism and multi-source training like ExPose. Note that SMPL+H and MoFA are disparate models, which are (offline) manually cut-and-stitched together. Instead, we use the whole-body SMPL-X model [270] that captures the shape of all body parts together, thus no stitching is required. Zhou et al. fuse only hand-body features in a “binary” fashion, while their face model is “disconnected” from the body. Instead, we fuse both face-body and hand-body features in a “fully analog” fusion, and thus our face expert can inform the whole-body shape. Zhou et al. do not predict separate face camera parameters and need PnP-RANSAC [93] and Procrustes to align their face to the image. Instead, we infer face-specific camera parameters and need no extra alignment steps. Zhou et al. use a complicated architecture, with several modules that are trained separately, and is applicable only to whole bodies. Instead, we use no intermediate tasks to avoid possible sources of error and train our model end to end. Our full model is applicable to whole bodies, but the part experts are also (separately) applicable to part-only data.

4.3 METHOD

Here we introduce PIXIE, a novel model for reconstructing SMPL-X humans with a realistic face from a single RGB image. It uses a set of expert sub-networks for body, face/head, and hand regression, and combines them in a bigger network architecture with three main novelties: (1) We use a novel moderator that assesses the confidence of part experts and fuses their features weighted by this, for robust inference under ambiguities, like strong occlusions. (2) We use a novel “gendered” shape loss, to improve body shape realism by learning to implicitly reason about gender. (3) In addition to the albedo predicted by our face expert, we employ the surface details branch of Feng et al. [87].

4.3.1 Expressive 3D Body Model

We use the expressive SMPL-X [270] body model, described in Sec. 2.3.1, to represent the human body. We follow the parameter vector definition of Sec. 3.3.1, but with $\beta \in \mathbb{R}^{200}$ and $\psi \in \mathbb{R}^{50}$. We denote with M_f the face subset of the SMPL-X mesh M .

Camera: To reconstruct SMPL-X from images, we use the weak-perspective camera model with scale $s \in \mathbb{R}$ and translation $t \in \mathbb{R}^2$. We denote the joints J and model vertices M projected on the image with $j \in \mathbb{R}^{J \times 2}$ and $m \in \mathbb{R}^{V \times 2}$.

4.3.2 PIXIE Architecture

PIXIE uses the architecture of Fig. 4.3, and is trained end to end. All model components are described below.

Input images: Given an image I with full resolution, we assume a bounding box around the body. We use this to crop and downsample the body to I_b to feed our network. However, this makes hands and faces too low resolution. We thus use the attention mechanism of ExPose, described in Sec. 3.3.2, to extract from I high-resolution crops for the face/head, I_f , and hand, I_h .

Feature encoding: We feed $\{I_b, I_f, I_h\}$ to separate expert encoders $\{E_b, E_f, E_h\}$ to extract features $\{F_b, F_f, F_h\}$. We use ResNet-50 [130] for the face/head and hand experts to generate $F_f, F_h \in \mathbb{R}^{2048}$. The body expert E_b uses HRNet [330], followed by convolutional layers that aggregate the multi-scale feature maps, to generate $F_b \in \mathbb{R}^{2048}$.

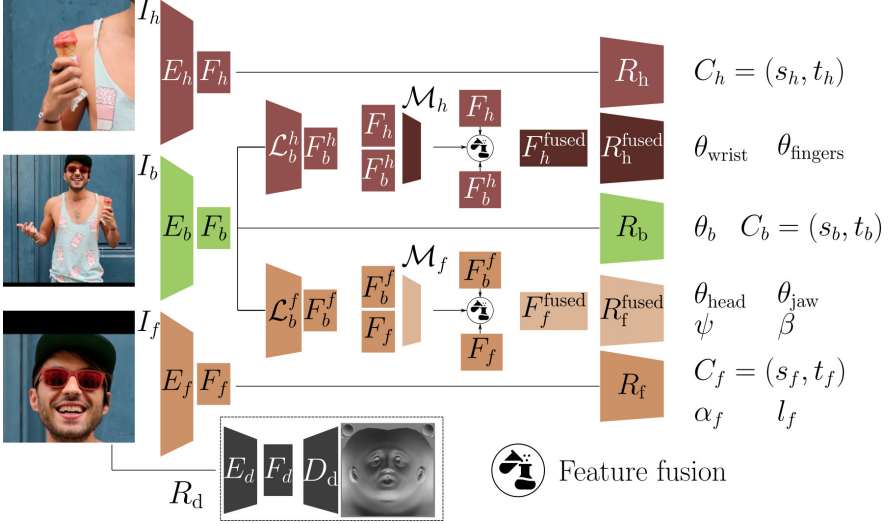


FIGURE 4.3: Body, face/head and hand image crops $\{I_b, I_f, I_h\}$ are fed to the expert encoders $\{E_b, E_f, E_h\}$ to produce part-specific features $\{F_b, F_f, F_h\}$. Our novel moderators $\{\mathcal{M}_f, \mathcal{M}_h\}$ estimate the confidence of experts for these images, and fuse face-body and hand-body features weighted by this, to create $\{F_f^{\text{fused}}, F_h^{\text{fused}}\}$. These are fed to $\{\mathcal{R}_f^{\text{fused}}, \mathcal{R}_h^{\text{fused}}\}$ for robust regression. DECA’s [87] R_d estimates fine geometric details. Icon from [Freepik](#).

Feature fusion (moderator): We identify the expert pairs of {body, head} and {body, hand} as complementary, and learn the novel moderators $\{\mathcal{M}_f, \mathcal{M}_h\}$ that build “fused” features $\{F_f^{\text{fused}}, F_h^{\text{fused}}\}$ and feed them to face/head and hand regressors $\{\mathcal{R}_f^{\text{fused}}, \mathcal{R}_h^{\text{fused}}\}$ (described below) for more informed inference. A moderator is implemented as a multi-layer perceptron (MLP) and gets the body, F_b , and part, F_p (F_f or F_h), features and fuses them with a weighted sum:

$$F_p^{\text{fused}} = w_p F_b^p + (1 - w_p) F_p, \quad (4.1)$$

$$w_p = \frac{1}{1 + \exp\left(-\tau * \mathcal{M}_p(F_b^p, F_p)\right)}, \quad (4.2)$$

where \mathcal{M}_p (\mathcal{M}_f or \mathcal{M}_h) is the part moderator, w_p (w_f or w_h) is the expert’s confidence, and F_b^p (F_b^f or F_b^h) is the body feature F_b transformed by the respective “extractor”, i.e. the linear layer \mathcal{L}^p (\mathcal{L}^h or \mathcal{L}^f) between the body

encoder E_b and part moderator \mathcal{M}_p . Finally, τ is a learned temperature weight, jointly trained with all network weights with the losses of Sec. 4.3.3, with no τ -specific supervision.

Parameter regression: We use two main regressor types: (1) We use the body, face/head, and hand $\{\mathcal{R}_b, \mathcal{R}_f, \mathcal{R}_h\}$ regressors, that get features *only* from the respective expert encoder $\{F_b, F_f, F_h\}$. \mathcal{R}_b infers the camera $c_b = (s_b, t_b)$, and body rotation and pose θ_b up to (excluding) the head and wrist. \mathcal{R}_f infers the camera $c_f = (s_f, t_f)$, face albedo α_f , and lighting l_f . \mathcal{R}_h infers the camera $c_h = (s_h, t_h)$. (2) We use the face/head, $\mathcal{R}_f^{\text{fused}}$, and hand, $\mathcal{R}_h^{\text{fused}}$, regressors that get from moderators the “fused” features, F_f^{fused} and F_h^{fused} . $\mathcal{R}_h^{\text{fused}}$ infers the wrist θ_{wrist} and finger pose θ_{fingers} . $\mathcal{R}_f^{\text{fused}}$ infers expressions ψ , head rotation θ_{head} , and jaw pose θ_{jaw} . Importantly, $\mathcal{R}_f^{\text{fused}}$ also infers body shape β , letting our face expert contribute to whole-body shape.

Detail capture: We use the fine geometric details branch R_d of Feng et al. [87] that, given a face image I_f , estimates dense 3D displacements on top of FLAME’s [207] surface. We convert the displacements from FLAME’s to SMPL-X’s UV map and apply them on PIXIE’s inferred head shape. However, inferring geometric details from full-body images is not trivial; the resolution of faces and image quality is much lower in these compared to face-only images. We account for this with our moderator and use the inferred displacements only when the face/head expert is confident.

4.3.3 Training Losses

To train PIXIE we use body, hand and face losses:

$$\mathcal{L} = \mathcal{L}_{\text{body}} + \mathcal{L}_{\text{hand}} + \mathcal{L}_{\text{face}} + \mathcal{L}_{\text{update}}, \quad (4.3)$$

defined as follows; the hat (e.g. \hat{j}) denotes ground truth.

Body losses: Following [56], we use a combination of a 2D re-projection, a 3D joint, and a SMPL-X parameter loss:

$$\mathcal{L}_{\text{body}} = \mathcal{L}_{\text{2D/3D-Joints}}^{\text{body}} + \mathcal{L}_{\text{params}}^{\text{body}}, \quad (4.4)$$

$$\mathcal{L}_{\text{2D/3D-Joints}}^{\text{body}} = \sum_{n=1}^J \|\hat{j}_n - j_n\|_1 \mathcal{L} + \sum_{n=1}^J \|\hat{j}_n - J_n\|_1, \quad (4.5)$$

$$\mathcal{L}_{\text{params}}^{\text{body}} = \|\hat{\theta} - \theta\|_2^2 + \|\hat{\beta} - \beta\|_2^2. \quad (4.6)$$

Hand losses: We employ a similar set of losses to train the 3D hand pose and shape estimation network:

$$\mathcal{L}_{\text{hand}} = \mathcal{L}_{2\text{D}/3\text{D-Joints}}^{\text{hand}} + \mathcal{L}_{\text{params}}^{\text{hand}}, \quad (4.7)$$

defined similarly to $\mathcal{L}_{2\text{D}/3\text{D-Joints}}^{\text{body}}$ and $\mathcal{L}_{\text{params}}^{\text{body}}$ of the body, but using the hand joints and pose parameters θ_{wrist} and θ_{fingers} .

Face losses: We adopt standard losses used by the 3D face estimation community [73, 87]:

$$\mathcal{L}_{\text{face}} = \mathcal{L}_{\text{lmk}} + \mathcal{L}_{\text{lmk-closure}} + \mathcal{L}_{\text{params}}^{\text{face}} + \mathcal{L}_{\text{pho}} + \mathcal{L}_{\text{id}}. \quad (4.8)$$

The landmark loss penalizes the difference between detected [39] target 2D landmarks \hat{p}_n and respective model landmarks (lying on M_f) projected on the image plane, p_n :

$$\mathcal{L}_{\text{lmk}} = \sum_{n=1}^{N_{\text{lmk}}} \|\hat{p}_n - p_n\|_1. \quad (4.9)$$

Following [87], we also compute a loss for the set UL of landmarks on the upper, lower eyelid and upper, lower lip:

$$\mathcal{L}_{\text{lmk-closure}} = \sum_{(i,j) \in \text{UL}} \|(\hat{p}_i - \hat{p}_j) - (p_i - p_j)\|_1. \quad (4.10)$$

The face parameter loss $\mathcal{L}_{\text{params}}^{\text{face}}$ follows $\mathcal{L}_{\text{params}}^{\text{body}}$, but for face pose θ_{face} only. This loss is only used for face crops from body data, when the target face pose is available.

Given the predicted 3D face mesh M_f as a subset of M , face albedo α_f and lighting l_f , we render a synthetic image I_r for the input subject using the differentiable renderer from Pytorch3D [284]. We then minimize the difference between the input face image I_f and the rendered image I_r :

$$\mathcal{L}_{\text{pho}} = \left\| \mathbb{M} \odot (I_f - I_r) \right\|_{1,1}, \quad (4.11)$$

where \mathbb{M} is a binary face mask with value 1 in the face skin region, and 0 elsewhere, and \odot denotes the Hadamard product. The segmentation mask prevents errors from non-face regions influencing the optimization, and we use the segmentation network of Nirkin et al. [257] to extract \mathbb{M} . The image formation process is the same as in Feng et al. [87].

Following [73, 106], we use a pre-trained face recognition network [42], f_{id} , to compute embeddings for the rendered image I_r and the input I_f . We then maximize the cosine similarity between the two identity embeddings

$$\mathcal{L}_{\text{id}} = 1 - \frac{\langle f_{\text{id}}(I_f), f_{\text{id}}(I_r) \rangle}{\|f_{\text{id}}(I_f)\|_2 \cdot \|f_{\text{id}}(I_r)\|_2}. \quad (4.12)$$

Priors: Due to the difficulty of the problem, we use additional priors to constrain PIXIE to generate plausible solutions. For expression parameters, we use a Gaussian prior:

$$\mathcal{L}_{\text{exp}}(\psi) = \|\psi\|_2^2. \quad (4.13)$$

We also add soft regularization on jaw and face pose:

$$\mathcal{L}_{\text{jaw}}(\theta_{\text{jaw}}) = \left| \theta_{\text{jaw}}^{\text{pitch}} \right|^2 + \left| \theta_{\text{jaw}}^{\text{roll}} \right|^2 + \left| \min(\theta_{\text{jaw}}^{\text{yaw}}, 0) \right|^2, \quad (4.14)$$

$$\mathcal{L}_{\text{face}}(\theta_{\text{face}}) = \left| \max\left(\left| \theta_{\text{face}}^{\text{yaw}} \right|, 90\right) \right|^2. \quad (4.15)$$

All these priors are “standard” regularizers, empirically found to discourage implausible configurations (extreme values, unrealistic shape/pose, interpenetrations, etc.).

Gender: As gender strongly affects body shape, we use a gender-specific shape prior during training, when gender labels are available. For this, we register SMPL-X to CAESAR [290] scans, and compute the mean μ and covariance Σ of shape parameters for females and males. Note that CAESAR does not contain non-binary labels. We then use:

$$\mathcal{L}_{\text{shape}}(\beta) = \begin{cases} (\beta - \mu_{\text{F}})^T \Sigma_{\text{F}} (\beta - \mu_{\text{F}}) & \text{if female} \\ (\beta - \mu_{\text{M}})^T \Sigma_{\text{M}} (\beta - \mu_{\text{M}}) & \text{if male} \\ \|\beta\|_2^2 & \text{o/w.} \end{cases} \quad (4.16)$$

When gender is unknown, we use a Gaussian prior computed over all scans/registrations, irrespective of gender. Please note that we do not need gender labels for inference.

Feature update loss: We encourage the transformed body features F_b^p (F_b^f or F_b^h) to match F_p^{fused} with a loss that was empirically found to stabilize network training:

$$\mathcal{L}_{\text{update}} = \left\| F_b^p - F_p^{\text{fused}} \right\|_1. \quad (4.17)$$

4.3.4 Implementation Details

Training data: For whole-body data we use the curated SMPL-X fits of [56], and SMPL-X fits to whole-body COCO data [158]. For hand-only data we use FreiHAND [427] and Total Motion [381]. For face/head data we use VGGFace2 [42] and detect $N_{\text{lmk}} = 68$ 2D landmarks with the method of Bulat et al. [39]. We get gender annotations by running the method of Rothe et al. [297] on many photos per identity and using majority voting to improve robustness. For data augmentation, see Sec. C.1.

Network training: We do multi-step training that empirically aids stability. We pre-train on part-only data, and train on whole-body data end to end; for details see Sec. C.1.

4.4 EXPERIMENTS

4.4.1 Evaluation Datasets

EHF [270]: We evaluate whole-body accuracy on this. It has 100 RGB images of 1 minimally-clothed subject in a lab setting with ground-truth SMPL-X meshes and 3D scans, see Sec. 2.4.1 for more details.

AGORA [269]: We evaluate whole-body and body-only accuracy on this, using its body-face-hands (BFH) subset. It has rendered [357] photo-realistic images of 3D human scans [1, 18, 144, 287] in scenes [127, 358]. It has SMPL-X ground truth recovered from scans, images and semantic labels [407].

3DPW [235]: We evaluate main-body accuracy on this. It captures 5 subjects in indoor/outdoor videos with SMPL pseudo ground truth, recovered from images and IMUs.

NoW [302]: We use it to evaluate face/head-only accuracy. It contains 3D head scans for 100 subjects and 2054 images with various viewing angles and facial expressions.

FreiHAND [427]: We evaluate hand-only accuracy on this. It has 37k hand/hand-object images of 32 subjects, with MANO ground truth, recovered from multi-view images.

Method	Type	Body model	Time (s)	PA-V2V (mm) ↓				TR-V2V (mm) ↓				PA-MPJPE (mm) ↓		PA-P2S (mm) ↓	
				All	Body	L/R hand	Face	All	Body	L/R hand	Face	MPJPE-14	L/R hand	Mean	Median
SMPLify- X'	O	SMPL-X	40-60	52.9	56.37	11.4/12.6	4.4	79.5	92.3	21.3/22.1	10.9	73.5	12.9/13.2	28.9	18.1
SMPLify-X	O	SMPL-X	40-60	65.3	75.4	11.6/12.9	4.9	93.0	116.1	23.8/24.9	11.5	87.6	12.2/13.5	36.8	23.0
MTC [381]	O	Adam	20	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	107.8	16.3/17.0	41.3	29.0
SPIN [187]	R	SMPL	0.01	N/A	60.6	N/A	N/A	N/A	96.8	N/A	N/A	102.9	N/A	40.8	28.7
FrankMocap [295]	R	SMPL-X	0.08	57.5	52.7	12.8/12.4	N/A	76.9	80.1	32.1 / 31.9	N/A	62.3	13.2/12.6	31.6	19.2
ExPose	R	SMPL-X	0.16	54.5	52.6	13.1/12.5	4.8	65.7	76.8	31.2 / 32.4	15.9	62.8	13.5/12.7	28.9	18.0
PIXIE (ours)	R	SMPL-X	0.08-0.10	55.0	53.0	11.2/11.0	4.6	67.6	75.8	25.6/27.0	14.2	61.5	11.7/11.4	29.9	18.4

TABLE 4.1: Evaluation on EHF. PIXIE is on par with the state of the art w.r.t. body and face performance, but predicts better hand poses. SMPLify- X' uses the ground-truth focal length (*excluded from bold*). Run-times were measured on an Intel Xeon W-2123 3.60GHz machine with an NVIDIA Quadro P5000 GPU. “O/R” denotes Optimization/Regression.

4.4.2 Evaluation Metrics

Mesh alignment: Prior to computing a metric, we align estimated meshes to ground-truth ones. The prefix “PA” denotes Procrustes Alignment (solving for scale, rotation and translation), while “TR” denotes translation alignment. “TR” is stricter, as it does not factor out scale and rotation. When reporting hand-/face-only metrics for the full body, we align each part separately.

Mean Per-Joint Position Error (MPJPE): We report the mean Euclidean distance between the estimated and ground-truth joints. For the body-only metric, we compute the 14 LSP-common joints [160] as a common skeleton across different body models, using a linear joint regressor [34, 195] on the estimated and ground-truth vertices. This is a standard metric, but is too sparse; it cannot capture errors in full 3D shape (i.e. surface), or all limb rotation errors.

Vertex-to-Vertex (V2V): For methods that infer meshes with the same topology as the ground-truth ones, e.g. SMPL(-X) estimations and SMPL(-X) ground truth, we compute the mean per-vertex error by taking into account *all* vertices. This is not possible for methods with different topology, e.g. SMPL estimations for SMPL-X ground truth, and vice versa. For such cases, we compute a *main-body* variant of V2V, i.e. without the hands and head, as SMPL and SMPL-X share the same topology for the main body. FB-V2V is the weighted sum of body (B), hand (LH, RH) and face (F) errors: $FB = B + \frac{LH+RH+F}{3}$. V2V is stricter than MPJPE; it also captures 3D shape errors and unnatural limb rotations (for the same joint positions).

Point-to-Surface (P2S): To compare PIXIE with methods that use a different mesh topology to SMPL(-X), e.g. MTC [381], we measure the mean distance from ground-truth vertices to the *surface* of the estimated mesh. P2S is stricter than MPJPE; it captures errors in 3D shape, but not unnatural limb rotations (for the same joint positions).

4.4.3 Quantitative Evaluation

Whole-body: In Tabs. 4.1 and 4.2 we report whole-body metrics (“All”), by taking into account the body, face and hands jointly. We add body-only (“Body”), hand-only (“L/R hand”), and face-only (“Face”) variants for completeness.

EHF [270]: Table 4.1 compares PIXIE to three baseline sets: (1) the optimization-based SMPLify-X, see Chapter 2, and MTC [381] that infer

Method	PA-V2V (mm) ↓		TR-V2V (mm) ↓	
	All	Body	All	Body
Naive Body	59.7	54.3	70.5	83.4
“Copy-paste”	60.3	55.5	72.9	82.4
PIXIE (ours)	55.0	53.0	67.6	75.8

TABLE 4.2: Ablation for our moderator on EHF [270]. “Naive body” denotes a single regressor for the whole body, and “Copy-Paste” denotes a naive integration of the independent expert estimations on the inferred body.

SMPL-X and Adam, (2) the regression-based SPIN [187] that infers SMPL, and (3) the regression-based ExPose, see Chapter 3, and FrankMocap [295] that infer SMPL-X. Note that MTC does not estimate the face. PIXIE outperforms optimization methods on most metrics, while being significantly faster. Moreover, it is on par with regression methods, both in terms of error metrics and runtime, which drops to 0.08 sec for known body-part crops.

AGORA [269]: Figure 4.4 compares PIXIE to whole-body [56, 270, 295] and body-only [163, 169, 183, 187, 213, 333] regressors, for a varying occlusion degree. PIXIE outperforms all methods, and is competitive on body-only metrics even when compared with the occlusion-aware PARE [183]. Note that AGORA is much more complex and natural than EHF, making the results more representative of real-world scenarios.

Ablation for moderators: Table 4.2 compares PIXIE to naive whole-body regression (no body-part experts) and the “copy-paste” fusion strategy. The latter copies pose parameters from the part experts (see [56, 295]), as well as shape parameters from the face expert, to the whole body.

The naive version does not benefit from the expertise of the part experts. “Copy-paste” fusion can lead to erroneous hand/face orientation inference, since the respective experts lack global context. Moreover, estimating whole-body shape from a face image is not always reliable, e.g. when a person faces away from the camera (Fig. 4.2). PIXIE fuses “global” body and “local” part features with its moderators. In this way, it estimates more accurate 3D bodies and is more robust to challenging ambiguities (blur, occlusion) than existing whole-body regressors, especially on stricter metrics without Procrustes alignment.

Ablation for “gendered” shape loss on 3DPW [235]: By removing our “gendered” shape loss, the PA-V2V error increases from 50.9 to 51.7 mm. A

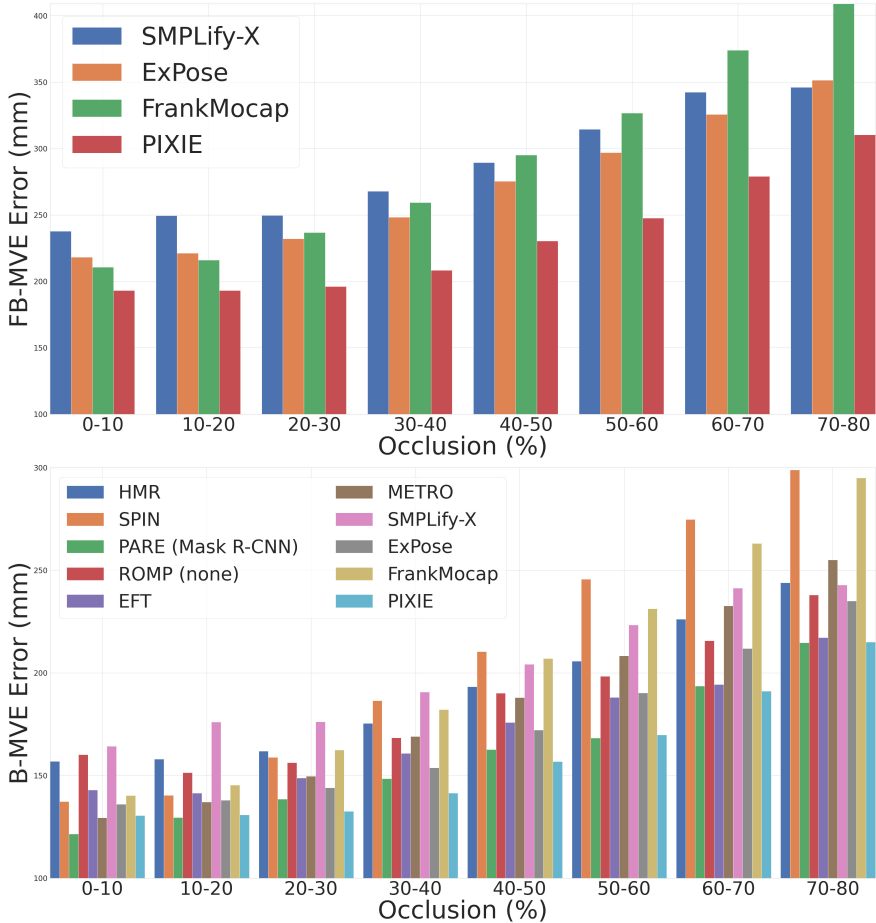


FIGURE 4.4: Comparison against state-of-the-art full-body (top) and body-only (bottom) methods on AGORA [269], using the vertex-to-vertex (V2V) metric (mm) for varying percentages of occlusion. Unless otherwise noted (in parens), we use OpenPose to extract person bounding boxes. PIXIE outperforms existing methods, including the occlusion-aware PARE [183].

qualitative ablation is shown in Fig. 4.5; learned implicit reasoning about gender gives more realistic body shapes. SMPL-X’s shared shape space for the whole body lets parts contribute to the whole.

Parts-only: For completeness, we use standard benchmarks for body-only, face-only, and hand-only evaluation.

Method	Body model	PA-MPJPE (mm) ↓	TR-MPJPE (mm) ↓	Body PA-V2V (mm) ↓
HMR [169]	SMPL	81.3	130.0	65.2
SPIN [187]	SMPL	59.2	96.9	53.0
FrankMocap [295]	SMPL-X	61.9	96.7	55.1
ExPose	SMPL-X	60.7	93.4	55.6
PIXIE (ours)	SMPL-X	61.3	91.0	50.9

TABLE 4.3: Evaluation on 3DPW [235]. PIXIE is the best for the stricter TR-MPJPE (joints) and V2V (surface) metrics.

Body-only on 3DPW [235]: Table 4.3 shows that PIXIE performs on par FrankMocap [295] and ExPose and is worse than SPIN [187], for the PA-MPJPE metric, but outperforms them all in the stricter TR-MPJPE (joints) and V2V (surface) metrics.

Face-only on NoW [302]: Table 4.4 shows that PIXIE’s face expert network outperforms not only the expressive whole-body method ExPose, but also strong and dedicated face-only methods, except for the recent work of Feng et al. [87].

Hand-only on FreiHAND [427]: Table 4.5 shows that our hand expert performs on par with the whole-body ExPose, is a bit worse than the hand-specific “MANO CNN” [427], but outperforms the hand expert of Zhou et al. [423].

4.4.4 Qualitative Evaluation

Figure 4.6 compares PIXIE with FrankMocap [295] and ExPose [56], which also regresses SMPL-X. Both baselines fail when the hand expert faces ambiguities (row 2); PIXIE gains robustness by using the full-body context. Both baselines give body shapes that look average (rows 1, 4) or have the wrong gender (rows 2, 3); PIXIE gives the most realistic shapes due to its “gendered” shape loss. FrankMocap fails for strong occlusions (rows 1, 3). Lastly, ExPose struggles with accurate facial expressions, and FrankMocap with head rotations (rows 1, 3); PIXIE outperforms both with its strong face/head expert and predicts a more realistic face.

Method	PA-P2S for face/head (mm) ↓		
	Median ↓	Mean ↓	Std ↓
3DMM-CNN [355]	1.84	2.33	2.05
PRNet [89]	1.50	1.98	1.88
Deng et al. [73]	1.23	1.54	1.29
RingNet [302]	1.21	1.54	1.31
3DDFA-V2 [118]	1.23	1.57	1.39
DECA [87]	1.09	1.38	1.18
ExPose	1.26	1.57	1.32
PIXIE (ours)	1.18	1.49	1.25

TABLE 4.4: Evaluation on NoW [302]. PIXIE is better than the whole-body ExPose, it outperforms many strong face-specific methods, and is a bit worse than DECA [87].

Method	PA-MPJPE	PA-V2V	PA-F@	PA-F@
	(mm) ↓	(mm) ↓	5mm ↑	15mm ↑
“MANO CNN” [427]	11.0	10.9	0.516	0.934
ExPose hand expert	12.2	11.8	0.484	0.918
Zhou et al. [423]	15.7	-	-	-
PIXIE hand expert	12.0	12.1	0.468	0.919

TABLE 4.5: Evaluation on FreiHAND [427]. PIXIE’s hand expert is on par with the hand expert of ExPose, but clearly outperforms the more related Zhou et al. [423] that also uses hand-body feature fusion.

Figure 4.7 compares PIXIE with Zhou et al. [423], recent work that also estimates a textured face. PIXIE gives more accurate poses (see how hands and faces align to the image), as it fuses both face-body and hand-body expert features, weighted by their confidence. PIXIE also gives more realistic body shapes, both due to its gendered shape loss and due to part experts contributing to whole-body shape, using SMPL-X’s shared body, hand, and face shape space.



FIGURE 4.5: Ablation for the “gendered” shape loss and the shared shape space (body/head). From left to right: (i) RGB Image, (ii) shape prediction only from the body image, and PIXIE without (iii) and with (iv) the “gendered” shape loss. We always use the gender-neutral SMPL-X model.

Future work: Mesh-to-image misalignment is a common limitation of regressors that pool “global” features from the image, losing local information. This could be tackled with “pixel-aligned” features [115, 183, 299, 408]. Moreover, SMPL-X models bodies without clothing; adding clothing models [63, 228] is a challenging but promising avenue. Furthermore, due to the formulation of the photometric term the model prefers to explain image evidence using lighting, rather than albedo, which leads to wrong skin tone predictions. Future work could further improve cases with self-contact [92, 249], or other extreme ambiguities.

4.5 CONCLUSION

We present PIXIE, a novel expressive whole-body reconstruction method that recovers an animatable 3D avatar with a detailed face from a single RGB image. PIXIE uses body-driven attention to leverage dedicated body, head and face experts. It learns a novel moderator that reasons about the



FIGURE 4.6: Qualitative comparison. From left to right: (i) RGB Image, (ii) ExPose [270], (iii) FrankMocap [295], (iv) PIXIE, (v) PIXIE with predicted face albedo and lighting.

confidence of each expert, to fuse their features according to confidence, and exploit their complementary strengths. It uses the best practices from the face community for accurate faces with realistic albedo and geometric details. The face expert can contribute to more realistic whole-body shapes, by using a shared face-body shape space. To further improve shape, PIXIE uses implicit reasoning about gender, to encourage likely “gendered” body shapes. Qualitative results show natural and expressive humans, with im-



FIGURE 4.7: Comparison with Zhou et al. [423]. From left to right: (i) RGB image, (ii) Zhou et al., (iii) PIXIE with inferred facial details and (iv) inferred albedo and lighting. Note that Zhou et al. use tight face crops through Dlib [178] to improve performance; PIXIE needs no tight face crops.

proved body shape, well articulated hands, and realistic faces, comparable to the best face-only methods. We believe that PIXIE will be useful for many applications that need expressive human understanding from images.

PART II

3D SHAPE ESTIMATION FROM METRIC AND SEMANTIC ATTRIBUTES

ACCURATE 3D BODY SHAPE REGRESSION USING METRIC AND SEMANTIC ATTRIBUTES

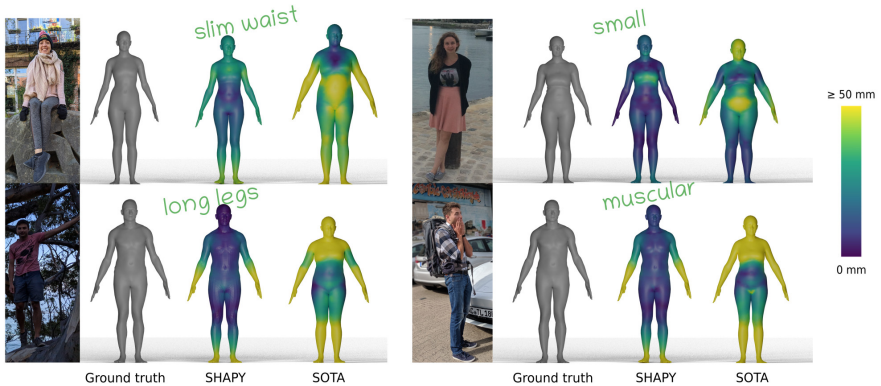


FIGURE 5.1: Existing work on 3D human reconstruction from a color image focuses mainly on *pose*. We present SHAPY, a model that focuses on body *shape* and learns to predict dense 3D shape from a color image, using crowd-sourced *linguistic shape attributes*. Even with this weak supervision, SHAPY outperforms the state of the art (SOTA) [311] on in-the-wild images with varied clothing.

5.1 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 [34, 166, 170, 182, 183, 184, 187, 270, 387, 410]. 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. 5.1, 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, and fitness. 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 [222] 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 joints stay the same. Several recent methods employ additional 2D cues, such as the silhouette, to provide additional shape cues [310, 311]. Silhouettes, however, are influenced by clothing and do not provide explicit 3D supervision. Synthetic approaches [211], 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, they 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



FIGURE 5.2: Model-agency websites contain multiple images of models together with anthropometric measurements. A wide range of body shapes are represented; example from pexels.com.

varied ethnicity, clothing, and body shape) with associated measurements; see Fig. 5.2.

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” [329]; see Fig. 5.3. Using the average scores from several raters, Streuber et al. [329] (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: (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 GT scores. (3) We similarly learn a function that maps “Attributes to Shape” (A2S). We then penalize body shape parameters that deviate from the prediction.

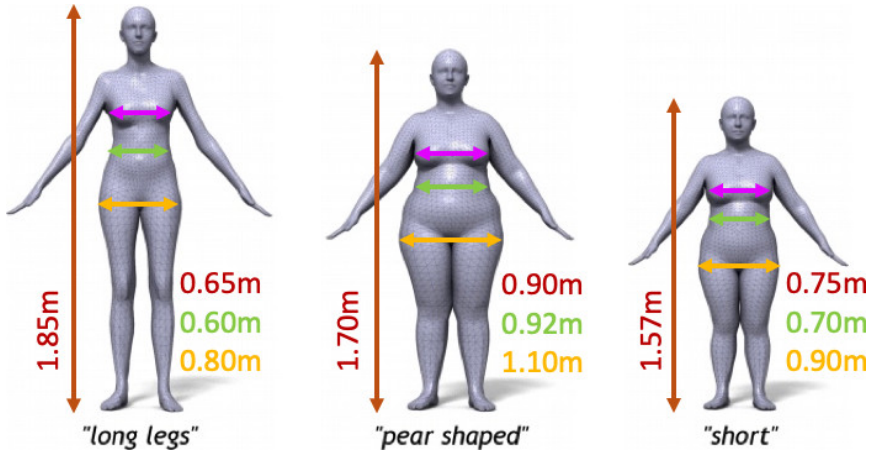


FIGURE 5.3: We crowd-source scores for linguistic body-shape attributes [329] and compute anthropometric measurements for CAESAR [290] body meshes. We also crowd-source linguistic shape attribute scores for model images, like those in Fig. 5.2

We study each term in detail to arrive at the final method. Evaluation is challenging because existing benchmarks with GT shape either contain too few subjects [235] or have limited clothing complexity and only pseudo-GT shape [310]. 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 shows that SHAPY estimates much more accurate 3D shape.

Models, data and code are available at <https://shapy.is.tue.mpg.de>.

5.2 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 [15], SMPL [222], SMPL-X, see Sec. 2.3.1, Adam [166], GHUM [387], and (2) **non-parametric methods** that predict a free-form representation of the human body [155, 300, 359, 385]. Parametric approaches lack details w.r.t. 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 in 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** [23, 34, 114, 270] search for model configurations that best explain image evidence, usually 2D landmarks [43], 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 [140, 195] to estimate shape. However, this needs special care to handle clothing [22]; without this, erroneous solutions emerge that “inflate” body shape to explain the “clothed” silhouette. **Regression-based methods** [56, 109, 156, 169, 182, 187, 211, 249, 401] 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 [150, 317], with model parameters estimated from MoCap markers [232], and in-the-wild image collections, such as COCO [215], that contain 2D keypoint annotations. Optimization and regression can be combined, for example via in-the-network model fitting [187, 249].

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 [34, 187, 249, 270]. Third, datasets with in-the-wild images have noisy 3D bodies, recovered by fitting a model to 2D keypoints [34, 270]. 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 [135, 310, 312, 360, 373], 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 [269, 299, 300], which are realistic but limited in variety.

Second, *2D shape cues* for in-the-wild images, (body-part segmentation masks [78, 262, 298], silhouettes [4, 140, 272]) are attractive, as these can be manually annotated or automatically detected [112, 128]. 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. [310, 311, 312] 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-GT 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. 5.1.

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 [10, 313, 314]. Tsoli et al. [354] register a body model to multiple high-resolution body scans to extract body measurements. The “Virtual Caliper” [281] allows users to build metrically accurate avatars of themselves using measurements or VR game controllers. ViBE [137] 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. [329] 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 [290] 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 [64] 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. [117] introduce the IMDB-23K dataset by gathering publicly available celebrity images and their height information. Zhu et al. [424] use this dataset to learn to predict

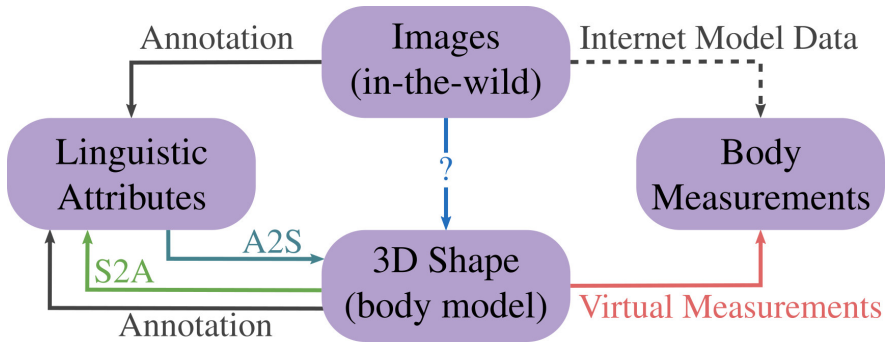


FIGURE 5.4: Shape representations and data collection. Our goal is 3D body shape estimation from in-the-wild images. Collecting data for direct supervision is difficult and does not scale. We explore two alternatives. **Linguistic Shape Attributes:** We annotate attributes (“A”) for CAESAR meshes, for 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 [281] 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 height of people in images. Dey et al. [74] 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. [31] use gravity as a prior to convert pixel measurements extracted from a video to metric height. These methods do not address body shape.

5.3 REPRESENTATIONS & 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. 5.4. 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.

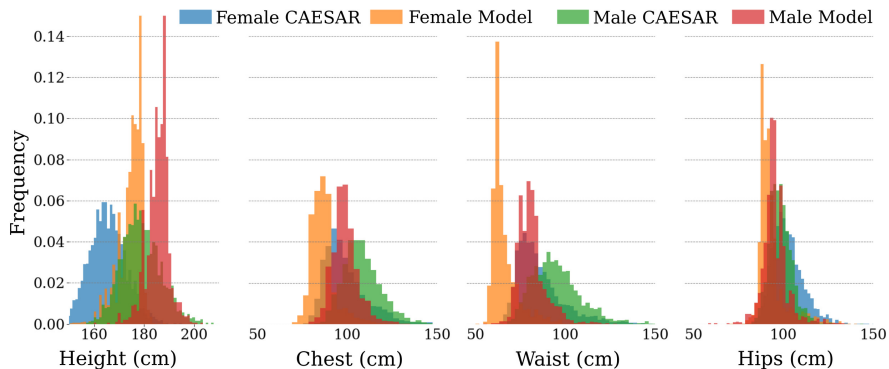


FIGURE 5.5: Histogram of height and chest/waist/hips circumference for data from model-agency websites (Sec. 5.3.2) and CAESAR. Model-agency data is diverse, yet not as much as CAESAR data.

5.3.1 SMPL-X Body Model

We use SMPL-X [270], described in Sec. 2.3.1, to represent the human body and adopt the notation of Sec. 3.3.1.

5.3.2 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](https://www.curve-models.com), [cocainemodels.com](https://www.cocainemodels.com), [nemesismodels.com](https://www.nemesismodels.com), [jayjay-models.de](https://www.jayjay-models.de), [kultmodels.com](https://www.kultmodels.com), [modelwerk.de](https://www.modelwerk.de), [models1.co.uk](https://www.models1.co.uk), [showcast.de](https://www.showcast.de), [the-models.de](https://www.the-models.de), and [ullamodels.com](https://www.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 Sec. D.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. 5.5, reveal a bias for “fashion model” body shapes, while other body types are under-represented in comparison

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

to CAESAR [290]. 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.

5.3.3 Linguistic Shape Attributes

Human body shape can be described by linguistic shape attributes [132]. We draw inspiration from Streuber et al. [329] 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 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 CAESAR [290] bodies 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. 5.3.2.

Annotation: To keep annotation tractable, we use $A = 15$ linguistic shape attributes per gender (subset of BodyTalk’s [329] attributes); see Tab. 5.1. 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”; for an example, see Sec. D.1.2. Annotations range on a discrete 5-level Likert scale from 1 (strongly disagree) to 5 (strongly agree). We get a rating matrix $A \in \{1, 2, 3, 4, 5\}^{N \times A \times K}$, where N is the number of subjects. In the following, a_{ijk} denotes an element of A .

5.4 MAPPING SHAPE REPRESENTATIONS

In Sec. 5.3 we introduce three body-shape representations: (1) SMPL-X’s PCA shape space (Sec. 5.3.1), (2) anthropometric measurements (Sec. 5.3.2), and (3) linguistic shape attribute scores (Sec. 5.3.3). Here we learn mappings between these, so that in Sec. 5.5 we can define new losses for training body shape regressors using multiple data sources.

5.4.1 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. [380] and the “Virtual Caliper” [281]. For details on how we compute these measurements, see Sec. D.2.1.

5.4.2 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^{A2S} , by averaging for each of the A attributes the corresponding K scores:

$$\mathbf{x}_i^{A2S} = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}], \quad \bar{a}_{i,j} = \frac{1}{K} \sum_{k=1}^K a_{ijk}, \quad (5.1)$$

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:

$$\mathbf{X}^{A2S} = [\phi(\mathbf{x}_1^{A2S}), \dots, \phi(\mathbf{x}_N^{A2S})]^\top, \quad (5.2)$$

where $\phi(\mathbf{x}_i^{A2S})$ 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:

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \epsilon. \quad (5.3)$$

Empirically, the polynomial model performs better than several models that we evaluated; for details, see Tab. D.1.

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 w.r.t. A2S:

$$\mathbf{x}_i^{\text{S2A}} = [\beta_{i,1}, \dots, \beta_{i,B}], \quad (5.4)$$

$$\mathbf{y}_i = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}]^\top. \quad (5.5)$$

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

$$\mathbf{x}_i^{\text{AHWC2S}} = [h_i, w_i, c_{c_i}, c_{w_i}, c_{h_i}], \quad (5.6)$$

where c_c, c_w, c_h is the chest, waist, and hip circumference, respectively, h and w are the height and weight, and **AHWC2S** 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:

$$\mathbf{x}_i^{\text{AHWC2S}} = [\bar{a}_{i,1}, \dots, \bar{a}_{i,A}, h_i, w_i, c_{c_i}, c_{w_i}, c_{h_i}]. \quad (5.7)$$

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 Sec. D.4.2.

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.

5.5 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. 5.3.1), linguistic attribute scores (Sec. 5.3.3) and anthropometric measurements (Sec. 5.4.1). 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. 5.4.1) to the mesh $M(\hat{\beta})$ in the canonical T-pose. We also convert shape $\hat{\beta}$ to linguistic attribute scores, with $\hat{A} = S2A(\hat{\beta})$.

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

$$L_{\text{attr}} = \|A - \hat{A}\|_2^2, \quad (5.8)$$

$$L_{\text{height}} = \|H - \hat{H}\|_2^2, \quad (5.9)$$

$$L_{\text{circ}} = \sum_{i \in \{c, w, h\}} \|C_i - \hat{C}_i\|_2^2 \quad (5.10)$$

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. 5.6. 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 [56] network weights and use curated fits [56], H3.6M [150], the SPIN [187] training data, and our model-agency dataset (Sec. 5.3.2) 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. (5.8) to (5.10) are applied only on the model-agency images. We use these on top of a standard base loss:

$$L_{\text{base}} = L_{\text{pose}} + L_{\text{shape}}, \quad (5.11)$$

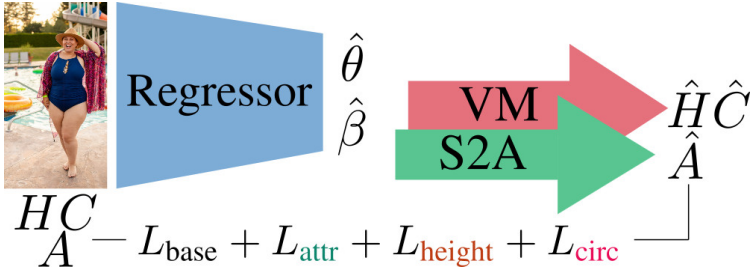


FIGURE 5.6: SHAPY first estimates shape, $\hat{\beta}$, and pose, $\hat{\theta}$. Shape is used by: (1) our virtual anthropometric measurement (VM) module to compute height, \hat{H} , and circumferences, \hat{C} , and (2) our S2A module to infer linguistic attribute scores, \hat{A} . There are several 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} .

where L_{joints}^{2D} and L_{joints}^{3D} are 2D and 3D joint losses, defined in Eqs. (3.8) and (3.9):

$$L_{pose} = L_{joints}^{2D} + L_{joints}^{3D} + L_{\theta}, \quad (5.12)$$

$$L_{shape} = L_{\beta} + L_{\beta}^{prior}, \quad (5.13)$$

L_{θ} and L_{β} are losses on pose and shape parameters, see Eq. (3.10), and L_{β}^{prior} is PIXIE’s [86] “gendered” shape prior, defined in Eq. (4.16). All losses are L2, unless otherwise explicitly specified. Losses on SMPL-X parameters are applied only on the pose data [56, 150, 187]. For more implementation details, see Sec. D.3.

5.6 EXPERIMENTS

5.6.1 Evaluation Datasets

3D Poses in the Wild (3DPW) [235]: 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 Sec. D.4.3.

Sports Shape and Pose 3D (SSP-3D) [310]: 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 [171], 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. 5.3.2

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. 5.4. 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. 5.3.3.

Human Bodies in the Wild (HBW): The field is missing a dataset with varied bodies, 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 [278]. (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 5.7 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.

5.6.2 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

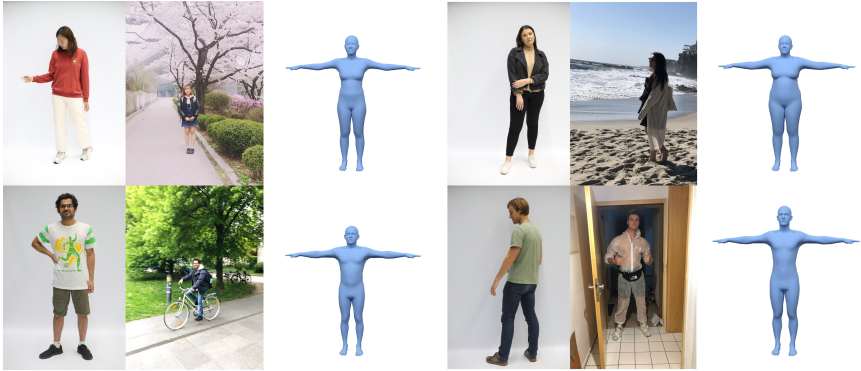


FIGURE 5.7: “Human Bodies in the Wild” (HBW) color images, taken in the lab and in the wild, and the SMPL-X ground-truth shape.

described in Sec. 5.4.1. When weight is available, we report the mean absolute error in kg.

MPJPE and V2V metrics: We report in Sec. D.4.3 the mean per-joint point error (MPJPE) and mean vertex-to-vertex error (V2V), when SMPL-X meshes are available. 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 evenly sample 20K points on SMPL-X’s surface, and report the mean point-to-point (P2P_{20K}) error. For details, see Sec. D.4.1.

5.6.3 Shape-Representation Mappings

We evaluate the models A2S and S2A, which map between the various body shape representations (Sec. 5.4).

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. 5.2, 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 per-

	Method	P2P _{20K}	Height	Weight	Chest	Waist	Hips
	-	(mm)	(mm)	(kg)	(mm)	(mm)	(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 5.2: Results of A2S variants on CMTS for male subjects, using the male SMPL-X model. For females, see Tab. D.2.

forming model, **AHWC**, uses everything, with P2P_{20K}-errors of 5.8 ± 2.0 mm (males) and 6.2 ± 2.4 mm (females).

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 Tab. D.5.

5.6.4 3D Shape from an Image

We evaluate all of our model’s variations (see Sec. 5.5) on the HBW validation set and find, perhaps surprisingly, that SHAPY-A outperforms other variants. We refer to this below (and Fig. 5.1) simply as “SHAPY” and report its performance in Tab. 5.3 for HBW, Tab. 5.4 for MMTS, and Tab. 5.5 for SSP-3D. For images with natural and varied clothing (HBW, MMTS), SHAPY significantly outperforms all other methods (Tabs. 5.3 and 5.4) using only weak 3D shape supervision (Attributes). On these images, Sengupta et al.’s method [311] struggles with the natural clothing.



FIGURE 5.8: Qualitative results from HBW. From left to right: RGB, ground-truth shape, SHAPY and Sengupta et al. [311]. 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 [231]	SMPL	182	267	309	305	69
STRAPS [310]	SMPL	135	167	145	102	47
SPIN [187]	SMPL	59	92	78	101	29
TUCH [249]	SMPL	58	89	75	57	26
Sengupta et al. [311]	SMPL	82	133	107	63	32
ExPose	SMPL-X	85	99	92	94	35
SHAPY (ours)	SMPL-X	51	65	69	57	21

TABLE 5.3: 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.

In contrast, their method is more accurate than SHAPY on SSP-3D (Tab. 5.5), which has tight “sports” clothing, in terms of V2V-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. 5.5. Note that, as HBW has true ground-truth 3D shape, it does not need SSP-3D’s scaling for evaluation.

Method	Model	Mean absolute error (mm) ↓			
		Height	Chest	Waist	Hips
Sengupta et al. [311]	SMPL	84	186	263	142
TUCH [249]	SMPL	82	92	129	91
SPIN [187]	SMPL	72	91	129	101
STRAPS [310]	SMPL	207	278	326	145
ExPose	SMPL-X	107	107	136	92
SHAPY (ours)	SMPL-X	71	64	98	74

TABLE 5.4: Evaluation on MMTS. We report the mean absolute error between ground-truth and estimated measurements.

Method	Model	V ₂ V-T-SC	mIOU
HMR [169]	SMPL	22.9	0.69
SPIN [187]	SMPL	22.2	0.70
STRAPS [310]	SMPL	15.9	0.80
Sengupta et al. [311]	SMPL	13.6	-
SHAPY (ours)	SMPL-X	19.2	-

TABLE 5.5: Evaluation on the SSP-3D test set [310]. We report the scaled mean vertex-to-vertex error in T-pose [310], and mIOU.

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 Sec. D.4.2. Figure 5.8 shows SHAPY’s qualitative results.

5.7 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. 5.3.2) 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.

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.

PART III

LEARNED OPTIMIZATION FOR 3D MORPHABLE MODEL FITTING

LEARNING TO FIT MORPHABLE MODELS

6.1 INTRODUCTION

Fitting parametric models [15, 79, 166, 270, 293, 387] to noisy input data is one of the most common tasks in computer vision. Notable examples include fitting the 3D body [34, 183, 187, 381], face [79], and hands [20, 36, 126, 316].

Direct regression using neural networks, such as the methods presented in Chapters 3 to 5, is the de facto tool to estimate model parameters from observations. While the obtained predictions are robust and accurate to a large extent, they often fail to tightly fit the observations [410] and require large quantities of annotated data. Classic optimization methods, e.g. the Levenberg-Marquardt algorithm [199, 236] or SMPLify-X from Chapter 2, can tightly fit the parametric model to the data by iteratively minimizing a hand-crafted energy function, but are getting dragged to local minima and require good starting points for fast convergence. Hence, practitioners combine these two approaches to benefit from their complementary strengths, initializing the model parameters from a regressor, followed by energy minimization using a classic optimizer.

If we look one level deeper, optimization-based model fitting methods have another disadvantage of often requiring hand-crafted energy functions that are difficult to define and non-trivial to tune. Besides the data terms, which have clear definitions, each fitting problem effectively requires the definition of their own prior terms and regularization terms. Besides the work required to formulate these terms and train the priors, domain experts needs to spend significant amounts of time to balance the effect of each term. Since these priors are often hand-defined or assumed to follow distributions that are tractable / easy to optimize, the resulting fitting energy usually contains biases that can limit the accuracy of the resulting fits.

To get the best of both regression using deep learning and classical numerical optimization, we turn to the field of machine learning based continuous optimization [14, 59, 306, 307, 324, 402]. Here, instead of updating the model parameters using a first or second order model fitter, a network learns to iteratively update the parameters that minimize the target loss, with the added benefit of optimized ML back-ends for fast inference. End-to-end

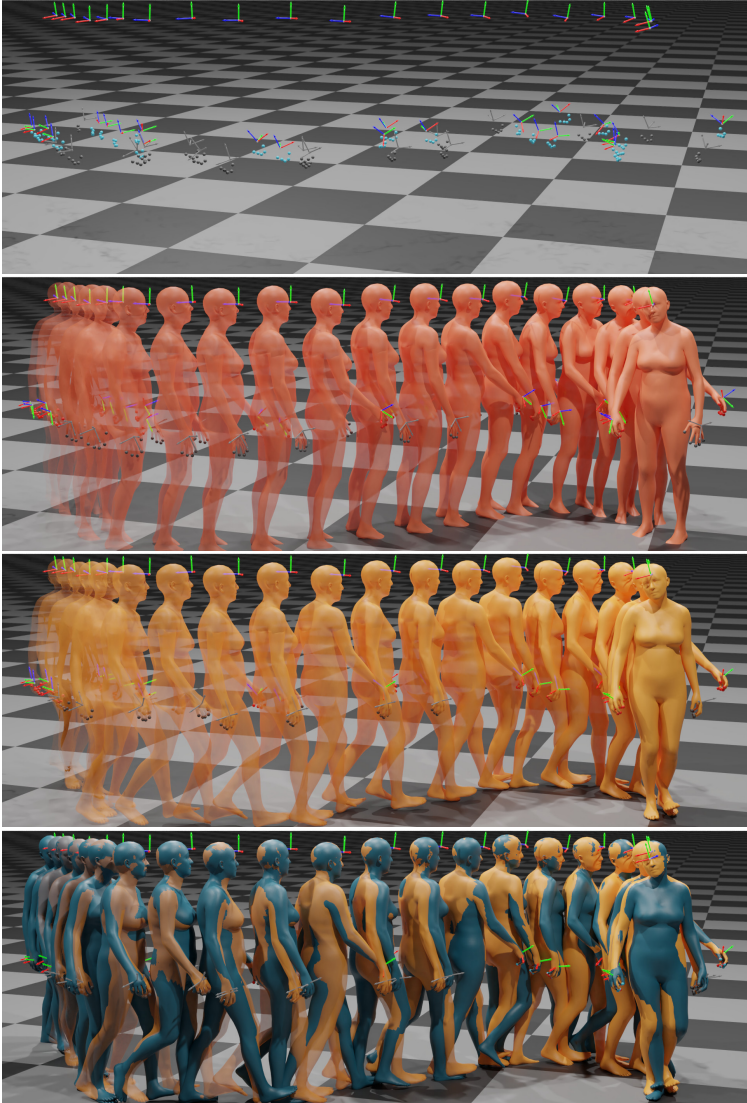


FIGURE 6.1: Top to bottom: (i) Head and hand tracking signals from AR/VR devices, (ii) the corresponding body model fit obtained from regression followed by iterative mathematical optimization, (iii) body model fit obtained from our learned optimizer (iv) and our estimate overlaid with the ground-truth. Learned optimizers are fast, able to tightly fit the input data and require significantly less manual labor to achieve this result. All results are estimated independently per-frame.

network training removes the need for hand-crafted priors, since the model learns them directly from data.

Inspired by the properties of the popular Levenberg-Marquardt and Adam [180] algorithms, our main contribution extends the system presented in [324] with an iterative machine learning solver which (i) keeps information from previous iterations, (ii) controls the learning rate of each variable independently and (iii) combines updates from gradient descent and from a network that is capable of swiftly reducing the fitting energy, for robustness and convergence speed. We evaluate our approach on different challenging scenarios: full-body tracking from head and hand inputs only, e.g. given by a device like the HoloLens 2, body estimation from 2D keypoints and face tracking from 2D landmarks, demonstrating both high quality results and versatility of the proposed framework.

6.2 RELATED WORK

Learning to optimize [14, 306, 307] is a field that, casts optimization as a learning problem. The goal is to create models that learn to exploit the problem structure, producing faster and more effective energy minimizers. In this way, we can remove the need for hand-designed parameter update rules and priors, since we can learn them directly from the data. This approach has been used for image denoising and depth-from-stereo estimation [363], rigid motion estimation [226], view synthesis [95], joint estimation of motion and scene geometry [59], non-linear tomographic inversion problem with simulated data [3], face alignment [384] and object reconstruction from a single image [186].

Parametric human model fitting: The seminal work of Blanz and Vetter [33] introduced a parametric model of human faces and a user-assisted method to fit the model to images. Since then, the field has evolved and produced better face models and faster, more accurate and more robust estimation methods [79]. With the introduction of SMPL [222], the field of 3D body pose and shape estimation has been rapidly progressing. The community has created large motion databases [232] from motion capture data, as well as datasets, both real and synthetic, with images and corresponding 3D body ground-truth [124, 235, 269]. Thanks to these, we can now train neural network regressors that can reliably predict SMPL parameters from images [163, 169, 187, 189, 203, 410] and videos [53, 182]. With the introduction of expressive models [166, 270, 387], the latest regression approaches [56, 86, 295] can now predict the 3D body, face and hands. However, one com-

mon issue, present in all regression scenarios, is the misalignment of the predictions and the input data [309, 410]. Thus, they often serve as the initial point for an optimization-based method [34, 270, 381], which refines the estimated parameters until some convergence criterion is met. This combination produces system that are effective, robust and able to work in real-time and under challenging conditions [248, 316, 338]. These hybrid regression-optimization systems are also effective pseudo annotators for in-the-wild images [187], where standard capture technologies are not applicable. However, formulating the correct energy terms and finding the right balance between them is a challenging and time-consuming task. Furthermore, adapting the optimizer to run in real-time is a non-trivial operation, even when using popular algorithms such as the Levenberg-Marquardt algorithm [146, 199, 236] which has a cubic complexity. Thus, explicitly computing the Jacobian [59, 226] is often prohibitive in practice, either in terms of memory or runtime. The most common and practical way to speed up the optimization is to utilize the sparsity of the problem or make certain assumptions to simplify it [82]. Learned optimizers promise to overcome these issues, by learning the parametric model priors directly from the data and taking more aggressive steps, thus converging in fewer iterations. The effectiveness of these approaches has been demonstrated in different scenarios, such as fitting a body model [222, 387] to images [324, 402] and videos [399], to sparse sensor data from electromagnetic sensors [174] and multi-body estimation from multi-view images [76].

We propose a new update rule, computed as a weighted combination of the gradient descent step and the network update [324], where their relative weights are a function of the residuals. Many popular optimizers have an internal memory, such as Adam’s [180] running averages, Clark et al.’s [59] and Neural Descent’s [402] RNN. We adopt this insight, using an RNN to predict the network update and the combination weights. In this way, the network can choose to follow either the gradient or the network direction more, using both the current and past residual values.

Estimating 3D human pose from a head-mounted device: is a difficult problem, due to self-occlusions caused by the position of the headset and the sparsity of the input signals [390]. Yuan and Kitani [397, 398] cast this as a control problem, where a model learns to produce target joint angles for a Proportional-Derivative (PD) controller. Other methods [350, 351] tackle this as a learning problem, where a neural network learns to predict the 3D pose from the cameras mounted on the HMD. Guzov et al. [119] use sensor data from IMUs placed on the subject’s body and combine them

with camera self-localization. They formulate an optimization problem with scene constraints, enabling the capture of long-term motions that respect scene constraints, such as foot contact with the ground. Finally, Dittadi et al. [75] propose a likelihood model that maps head and hand signals to full body poses. In our work, we focus on this scenario and empirically show that the proposed optimizer rule is competitive, both with a classic optimization baseline and a state-of-the-art likelihood model [75].

6.3 METHOD

6.3.1 *Neural Fitter*

Levenberg-Marquardt (LM) [146, 199, 236] and Powell’s dog leg method (PDL) [280] are examples of popular iterative optimization algorithms used in applications that fit either faces or full human body models to observations. These techniques employ the Gauss-Newton algorithm for both its convergence rate approaching the quadratic regime and its computational efficiency, enabling real-time model fitting applications, e.g. generative face [346, 429] and hand [316, 338] tracking. For robustness, LM and PDL both combine the Gauss-Newton algorithm and gradient descent, leading to implicit and explicit trust regions being used when calculating updates, respectively. In LM, the relative contribution of the approximate Hessian and the identity matrix is weighted by a single scalar that is changing over iterations with its value carried over from one iteration to the next. Given an optimization problem over a set of parameters Θ , LM computes the parameter update $\Delta\Theta$ as the solution of the system $(J^T J + \lambda \text{diag}(J^T J))\Delta\Theta = J^T R$, where J is the Jacobian and R are the current residual values. It is interesting to note that several popular optimizers, including ADAGRAD [77] and Adam [180], also carry over information about previous iteration(s), in this case to help control the learning rate for each parameter.

Inspired by the success of these algorithms, we aim at constructing a novel neural optimizer that (i) is easily applicable to different fitting problems, (ii) can run at interactive rates without requiring significant effort, (iii) does not require hand-crafted priors. (iv) carries over information about previous iterations of the solve, (v) controls the learning rate of each parameter independently, (vi) for robustness and convergence speed, combines updates from gradient descent and from a method capable of very quickly reducing the fitting energy. Note that the Learned Gradient

Algorithm 1 Neural fitting

Require: Input data D

$$\Theta_0 = \Phi(D)$$

$$h_0 = \Phi_h(D)$$

while not converged **do**

$$\Delta\Theta_n, h_n \leftarrow f([g_{n-1}, \Theta_{n-1}], D, h_{n-1})$$

$$\Theta_n \leftarrow \Theta_{n-1} + u(\Delta\Theta_n, g_{n-1}, \Theta_{n-1})$$

end while

Descent (LGD) proposed in [324] achieves (a), (b), and (c), but does not consider (d), (e), and (f). As demonstrated experimentally in Sec. 6.4, each of these additional properties leads to improved results compared to [324], and the best results are achieved when combined together.

Our proposed neural fitter estimates the values of the parameters Θ by iteratively updating an initial estimate Θ_0 , see Alg. 1. While the initial estimate Θ_0 obtained from a deep neural network Φ might be sufficiently accurate for some applications, we will show that a careful construction of the update rule ($u(\cdot)$ in Alg. 1) leads to significant improvements after only a few iterations. It is important to note that we do not focus on building the best possible initializer Φ for the fitting tasks at hand, which is the focus of e.g. VIBE [182] and SPIN [187]. That being said, note that these regressors could be leveraged to provide Θ_0 from Alg. 1. h_0 and h_n are the hidden states of the optimization process. At the n -th iteration in the loop of Alg. 1, we use a neural network f to predict $\Delta\Theta_n$, and then apply the following update rule:

$$u(\Delta\Theta_n, g_{n-1}, \Theta_{n-1}) = \lambda\Delta\Theta_n + (-\gamma g_{n-1}) \quad (6.1)$$

$$\lambda, \gamma = f_{\lambda, \gamma}(\mathbf{R}(\Theta_{n-1}), \mathbf{R}(\Theta_{n-1} + \Delta\Theta_n)), \lambda, \gamma \in \mathbb{R}^{|\Theta|} \quad (6.2)$$

Note that LGD [324] is a special case of Eq. (6.1), with $\lambda = 1, \gamma = 0$, and with no knowledge preserved across fitting iterations. g_n is the gradient of the target data term \mathcal{L}^D w.r.t. to the problem parameters: $g_n = \nabla \mathcal{L}^D$.

The proposed neural fitter satisfies the requirements (a), (b) and (c) in a similar fashion to LGD [324]. In the following, we describe how the properties (d), (e), and (f) outlined earlier in this section are satisfied.

(d): keeping track of past iterations.: The functions $f, f_{\lambda, \gamma}$ are implemented with a Gated Recurrent Unit (GRU) [52]. Previous methods only store past parameter values and the total loss [324]. Thanks to the GRU modules, our model can learn how to best incorporate past and current information.

(e): independent learning rate.: When fitting face or body models to data, the variables being optimized over are of different nature. For instance, rotations might be expressed in Euler angles while translation in meters. Since each of these parameter has a different scale and / or unit, it is useful to have per-parameter step size values. Here, we propose to predict vectors λ and γ independently to scale the relative contribution of $\Delta\Theta_n$ and g_n to the update applied to each entry of Θ_n . It is interesting to note that $f_{\lambda,\gamma}$, having knowledge about the current value of residuals at Θ_n and the residual at $\Theta_n + \Delta\Theta_n$, effectively makes use of an estimate of the step direction before setting a step size that is analogous to how line-search operates. Motivated by this observation we tried a few learned versions of line search which yielded similar or inferior results to what we propose here. The alternatives we tried are described in Sec. E.2.

(f): combining gradient descent and network updates.: LM interpolates between Gradient Descent (GD) and Gauss-Newton (GN) using an iteration-dependent scalar. LM combines the benefits of both approaches, namely fast convergence near the minimum like GN and large descent steps away from the minimum like GD. Here, we replace the GN direction, which is often prohibitive to compute, with a network-predicted update, described in Eq. (6.1). The neural optimizer should learn the optimal descent direction and the relative weights to minimize the data term in as few steps as possible. In Sec. E.2 we provide alternative combinations, e.g. via convex combination, which yielded inferior results in our experiments.

6.3.2 Human Body Model and Fitting Tasks

The 3D joints, $J(\beta)$, of a kinematic skeleton are computed from the shape parameters. The pose parameters $\theta \in \mathbb{R}^{J \times D + 3}$ contain the parent-relative rotations of each joint and the root translation, where D is the dimension of the rotation representation and J is the number of skeleton joints. We represent rotations using the 6D rotation parameterization of Zhou et al. [422], thus $\theta \in \mathbb{R}^{J \times 6 + 3}$. The world transformation $T_j(\theta) \in SE(3)$ of each joint j is computed by following the transformations of its parents in the kinematic tree: $T_j(\theta) = T_{p(j)}(\theta) * T(\theta_j, J_j(\beta))$, where $p(j)$ is the index of the parent of joint j and $T(\theta_j, J_j(\beta))$ is the rigid transformation of joint j relative to its parent. In the following sections, variables with a *hat* denote observed quantities.

We focus on two 3D human body estimation problems: 1) fitting a body model [222] to 2D keypoints and 2) inferring the body, including hand

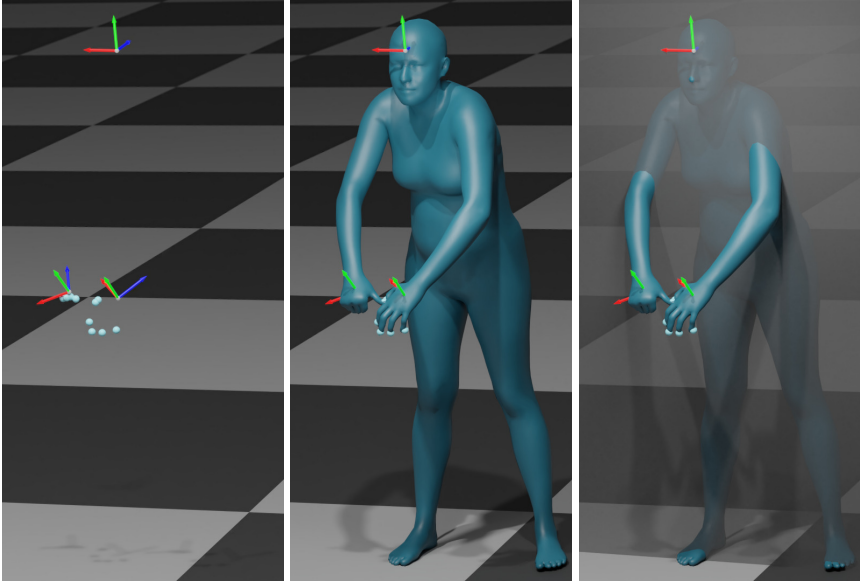


FIGURE 6.2: Left to right: 1) Input 6-DOF transformations T_H, T_L, T_R and fingertip positions $P_{i=1,\dots,5}^L, P_{i=1,\dots,5}^R$, given by the head-mounted device, 2) ground-truth mesh, 3) half-space visibility, everything behind the headset is not visible.

articulation [293], from head and hand signals returned by AR/VR devices, shown in Fig. 6.2. The first is by now a standard problem in the Computer Vision community. The second, which uses only head and hand signals in the AR/VR scenario, is a significantly harder task which requires strong priors, in particular to produce plausible results for the lower body and hands. The design of such priors is not trivial, requires expert knowledge and a significant investment of time.

2D keypoint fitting: We follow the setup of Song et al. [324], computing the projection of the 3D SMPL joints J with a weak-perspective camera Π_o with scale $s \in \mathbb{R}$, translation $t \in \mathbb{R}^2$: $j = \Pi_o(J(\theta, \beta), s, t)$. Our goal is to estimate SMPL and camera parameters $\Theta^B = \{\theta, \beta\}$, $K^B = \{s, t\}$, such that the projected joints j match the detected keypoints $D^B = \{\hat{j}\}$, e.g. from OpenPose [43].

Fitting SMPL+H to AR/VR device signals: We make the following assumptions: (i) the device head tracking system provides a 6-DOF transformation \hat{T}^H , that contains the position and orientation of the *headset* in the world coordinate frame. (ii) the device hand tracking system gives us the

orientation and position of the left and right wrist, $\hat{\mathbf{T}}^L, \hat{\mathbf{T}}^R \in SE(3)$, and the positions of the fingertips $\hat{P}_{1,\dots,5}^L, \hat{P}_{1,\dots,5}^R \in \mathbb{R}^3$ in the world coordinate frame, if and when they are in the field of view (FOV) of the HMD. In order to estimate the SMPL+H parameters that best fit the above observations, we compute the estimated headset position and orientation from the SMPL+H world transformations as $\mathbf{T}^H(\Theta) = \mathbf{T}^{\text{HMD}}\mathbf{T}_{j_H}(\Theta)$, where j_H is the index of the head joint of SMPL+H. \mathbf{T}^{HMD} is a fixed transform from the SMPL+H head joint to the headset, obtained from an offline calibration phase.

Visibility is represented by $v_L, v_R \in \{0, 1\}$ for the left and right hand respectively. We examine two scenarios: (i) full visibility, where the hands are always visible, (ii) half-space visibility, where only the area in front of the HMD is visible. Specifically, we transform the points into the coordinate frame of the headset, using \mathbf{T}^H . All points with $z \geq 0$ are behind the headset and thus invisible. Figure 6.2 right visualizes the plane that defines what is visible or not.

To sum up, the sensor data are: $D^{\text{HMD}} = \{\hat{\mathbf{T}}^H, \hat{\mathbf{T}}^L, \hat{\mathbf{T}}^R, \hat{P}_{i=1,\dots,5}^L, \hat{P}_{i=1,\dots,5}^R, v_L, v_R\}$. The goal is to estimate the parameters $\Theta^{\text{HMD}} = \{\theta\} \in \mathbb{R}^{315}$, which contain the $J = 52 \times 6$ joint rotation and 3 translation parameters, that best fit D^{HMD} . Note that we assume we are given body shape β from a separate enrollment step, only for the HMD fitting problem.

6.3.3 Human Face Model and Fitting Task

We represent the human face using the parametric face model proposed by Wood et al. [377]. It is a blendshape model [79], with $v = 7667$ vertices, 4 skeleton joints (head, neck and two eyes), with their rotations and translations denoted with θ , identity $\beta \in \mathbb{R}^{256}$ and expression $\psi \in \mathbb{R}^{233}$ blendshapes. The deformed face mesh is obtained with standard linear blend skinning.

For face fitting, we select a set of mesh vertices as the face landmarks $\mathcal{P}(\theta, \psi, \beta) \in \mathbb{R}^{P \times 3}$, $P = 669$ (see Fig. 6.3 right). The input data are the corresponding 2D face landmarks $\hat{p} \in \mathbb{R}^P \times 2$, detected using the landmark neural network proposed by Wood et al. [377].

For this task, our goal is to estimate translation, joint rotations, expression and identity coefficients $\Theta^F = \{\theta, \psi, \beta\} \in \mathbb{R}^{516}$ that best fit the 2D landmarks $D^F = \hat{p}$. We assume we are dealing with calibrated cameras and thus have access to the camera intrinsics K . $\Pi_p(\mathcal{P}; K)$ is the perspective camera projection function used to project the 3D landmarks \mathcal{P} onto the image plane.

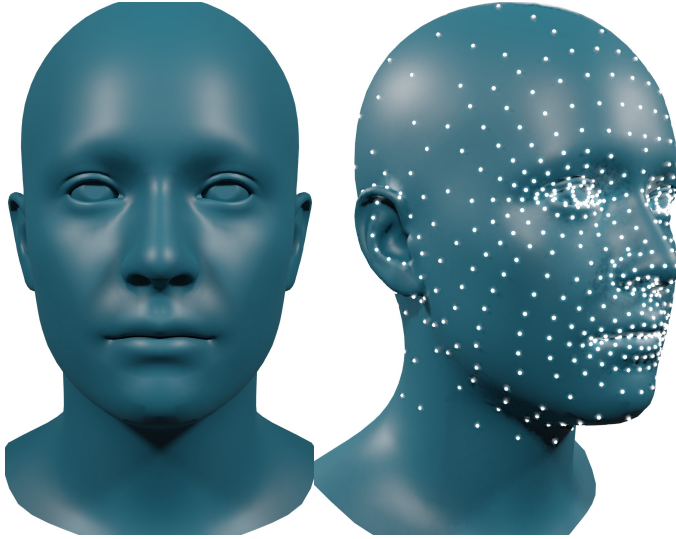


FIGURE 6.3: Blue: The face model template of Wood et al. [377]. White: 669 dense landmarks.

6.3.4 Data Terms

The data term is a function $\mathcal{L}^D(\Theta; D)$ that measures the discrepancy between the observed inputs D and the parametric model evaluated at the estimated parameters Θ .

At the n -th iteration of the fitting process, we compute both (i) the array $R(\Theta_n)$ that contains all the corresponding residuals of the data term \mathcal{L}^D for the current set of parameters Θ_n , and (ii) the gradient $g_n = \nabla \mathcal{L}^D(\Theta_n)$.

Let $\|\cdot\|$ be any metric appropriate for $SE(3)$ [75] and $\|\cdot\|_\rho$ a robust norm [26]. To compute residuals, we use the Frobenius norm for $\|\cdot\|$ and $\|\cdot\|_\rho$. Note that any other norm choice can be made compatible with LM [400].

Body fitting to 2D keypoints: We employ the re-projection error between the detected joints and those estimated from the model as the data term:

$$\mathcal{L}^D(\Theta^B; D^B) = \|\hat{\mathbf{j}} - \Pi_o(J(\theta, \beta), s, t)\|_\rho \quad (6.3)$$

Here $J(\Theta^B)$ denotes the “posed” joints.

Body fitting to HMD signals: We measure the discrepancy between the observed data D^{HMD} and the estimated model parameters Θ^{HMD} with the following data term:

$$\mathcal{L}^D(\Theta^{\text{HMD}}; D^{\text{HMD}}) = \llbracket \hat{\mathbf{T}}^{\text{H}}, \mathbf{T}^{\text{H}}(\Theta^{\text{HMD}}) \rrbracket + \sum_{w \in \mathbf{L}, \mathbf{R}} v_w \left(\llbracket \hat{\mathbf{T}}^w, \mathbf{T}^w(\Theta^{\text{HMD}}) \rrbracket + \sum_{i=1}^5 \|\hat{p}_i^w - p_i^w(\Theta^{\text{HMD}})\|_{\rho} \right) \quad (6.4)$$

Face fitting to 2D landmarks: we use the landmark re-projection error as our data term:

$$\mathcal{L}^D(\Theta^F; D^F) = \|\hat{p} - \Pi_p(\mathcal{P}(\Theta^F); \mathbf{K}^F)\|_{\rho} \quad (6.5)$$

6.3.5 Training Details

Training losses: We train our learned fitter using a combination of model parameter and mesh losses. Their precise formulation can be found in the Sec. E.5.2.

Model structure: Unless otherwise specified, $f, f_{\lambda, \gamma}$ (in Alg. 1, Eq. (6.2)) use a stack of two GRUs with 1024 units each. The initialization Φ, Φ_h in Alg. 1 are MLPs with two layers of 256 units, ReLU [254] and Batch Normalization [149].

Datasets: For the body fitting tasks, we use AMASS [232] to train and test our fitters. When fitting SMPL to 2D keypoints, we use 3DPW’s [235] test set to evaluate the learned fitter’s accuracy, using the detected OpenPose [43] keypoints as the target. The face fitter is trained and evaluated on synthetic data. Please see Sec. E.5.3 for more details on the datasets.

6.4 EXPERIMENTS

6.4.1 Metrics

Metrics with a *PA* prefix are computed after undoing rotation, scale and translation, i.e. Procrustes alignment. Variables with a *tilde* are ground-truth values.

Vertex-to-Vertex (V2V): As we know the correspondence between ground-truth \tilde{M} and estimated vertices M , we are able to compute the mean per-vertex error: $\text{V2V}(\tilde{M}, M) = \frac{1}{V} \sum_{i=1}^V \|\tilde{M}_i - M_i\|_2^2$. For SMPL+H, in addition to the full mesh error (FB), we report error values for the head (H) and

Method	Type	Image	2D keypoints	Part segmentation	PA-MPJPE
SMPLify [34]	O	✗	✓	✗	106.1
SCOPE [82]	O	✗	✓	✗	68.0
SPIN [187]	R	✓	✗	✗	59.6
VIBE [182]	R	✓	✗	✗	55.9
Neural Descent [402]	R+O	✓	✓	✓	57.5
LGD [324]	R+O	✗	✓	✗	55.9
Ours, LGD + Eq. (6.1)	R+O	✗	✓	✗	53.9
Ours (full)	R+O	✗	✓	✗	52.2

TABLE 6.1: Using 3DPW [235] to compare different approaches that estimate SMPL from images, 2D keypoints and part segmentation masks. Replacing LGD’s [324] update rule with ours leads to a 2 mm PA-MPJPE improvement. Our full system, that uses GRUs, leads to a further 1.6 mm improvement. “O/R” denotes Optimization/Regression.

hands (L, R). A visualization of the selected parts is included in Fig. E.5. The **3D per-joint error (MPJPE)** is equal to: $\text{MPJPE}(\tilde{J}, J) = \frac{1}{J} \sum_{i=1}^J \|\tilde{J}_i - J_i\|_2^2$.

Ground penetration (GrPe): We report the average distance to the ground plane for all vertices below ground [399]:

$$\begin{aligned} \text{GrPe.}(M) &= \frac{1}{|S|} \sum_{n \in S} |d_{\text{gnd}}(M_i)| \\ d_{\text{gnd}}(M_i) &= M_i \cdot n_{\text{gnd}} \\ S &= \{i \mid d_{\text{gnd}}(M_i) < 0\}. \end{aligned} \tag{6.6}$$

Face landmark error (LdmkErr): We report the mean distance between estimated and ground-truth 3D landmarks:

$$\text{LdmkErr}(\tilde{\mathcal{P}}, \mathcal{P}) = \frac{1}{P} \sum_{i=1}^P \|\tilde{\mathcal{P}}_i - \mathcal{P}_i\|_2^2. \tag{6.7}$$

6.4.2 Quantitative Evaluation

Fitting the body to 2D keypoints: We compare our proposed update rule with existing regressors, classic and learned optimization methods on 3DPW [235]. For a fairer comparison with Song et al. [324], we train two

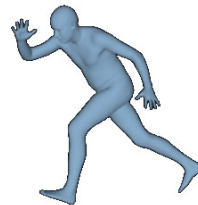
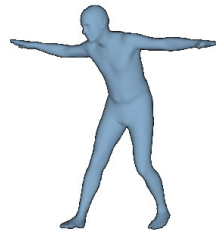


FIGURE 6.4: The input RGB image and the SMPL body predicted by our fitter.

Method	Vertex-to-Vertex (mm) ↓						MPJPE		GrPe.	
	Full body		Head		L / R hand		(mm) ↓		(mm) ↓	
	F	H	F	H	F	H	F	H	F	H
L-BFGS, GMM	73.1	116.2	2.9	3.4	3.2 / 3.0	5.6 / 5.3	49.7	137.26	70.8	74.0
L-BFGS, GMM, Tempo.	72.6	113.3	2.9	3.4	3.3 / 3.1	6.8 / 6.5	49.4	132.1	70.7	73.5
L-BFGS, VAE Enc.	76.1	119.3	3.9	4.1	5.3 / 4.7	8.7 / 7.6	52.6	140.5	63.6	66.7
Dittadi et al. [75]	N/A		N/A		N/A		43.3	N/A	N/A	
Ours Φ , ($N = 0$)	44.2	69.7	19.1	22.7	27.8 / 25.9	32.1 / 29.9	38.9	84.9	16.1	20.1
Ours ($N = 5$)	26.1	49.9	2.2	3.2	3.0 / 3.3	3.1 / 3.7	18.1	62.1	12.5	15.5

TABLE 6.2: Fitting SMPL+H to simulated sequences of HMD data. Our proposed fitter outperforms the classical optimization baselines (L-BFGS prefix) on the full body and ground penetration metrics, with similar or better performance on the part metrics, and the regressor baselines (the VAE predictor [75] and the regressor Φ), on all metrics. “F/H” denotes full / half-plane visibility.

versions of our proposed fitter, one where we change the update rule of LGD with Eq. (6.1), and our full system which also has network architecture changes. Table 6.1 shows that just by changing the update rule (Ours, LGD + Eq. (6.1)), we outperform all baselines. Figure 6.4 contains qualitative results of our method on images from the 3DPW test set.

Fitting the body to HMD data: In Tab. 6.2, we compare our proposed learned optimizer with a standard optimization pipeline, a variant of SMPLify [34, 270] adapted to the HMD fitting task (first 3 rows), and two neural network regressors (a VAE predictor [75] in the 4th row and our initializer Φ of Alg. 1 in the 5th row), on the task of fitting SMPL+H to sparse HMD signals, described in Sec. 6.3.2. The optimization baseline minimizes the energy with data term (\mathcal{L}^D in Equation (6.4)), gravity term \mathcal{L}^G , prior term $\mathcal{L}^{\theta}_{\text{prior}}$, without and with temporal term \mathcal{L}^T (first and second row of Tab. 6.2) to estimate the parameters $\Theta_{1,\dots,T}$ of a sequence of length T :

$$\begin{aligned}
\mathcal{L}^O(\Theta^{\text{HMD}}) &= \mathcal{L}^D(\Theta^{\text{HMD}}; D^{\text{HMD}}) + \mathcal{L}^G + \mathcal{L}^{\theta}_{\text{prior}} + \mathcal{L}^T \\
\mathcal{L}^G(\Theta^{\text{HMD}}) &= 1 - \frac{\mathbb{T}_{\text{pelvis}}(1, : 3) \cdot \mathbf{u}}{\left\| \mathbb{T}_{\text{pelvis}}(1, : 3) \right\|_2^2 \|\mathbf{u}\|_2^2}, \quad \mathbf{u} = (0, 1, 0) \\
\mathcal{L}^T(\Theta^{\text{HMD}}) &= \sum_{t=1}^{T-1} \left[\mathbb{T}_{t+1}(\Theta_{t+1}^{\text{HMD}}) - \mathbb{T}_t(\Theta_t^{\text{HMD}}) \right]
\end{aligned} \tag{6.8}$$

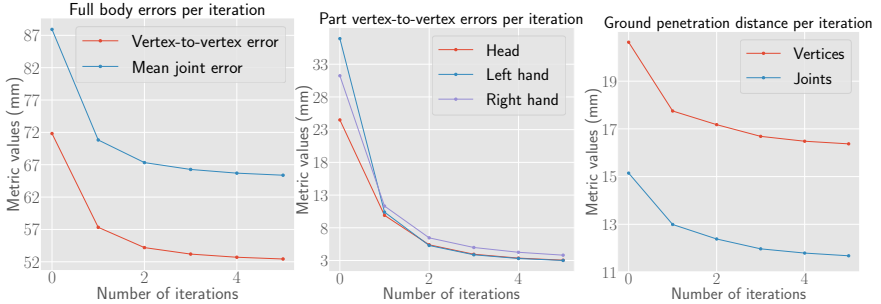


FIGURE 6.5: Errors per iteration when fitting SMPL+H to HMD data for the half-space visibility scenario, see Fig. E.2 for full visibility. Left to right: 1) full body vertex and joint errors, 2) head, left and right hand V2V errors and 3) vertex and joint ground distance, computed on the set of points below ground.

We use two different pose priors, a GMM [34] and a VAE encoder $E(\cdot)$ [270]:

$$\mathcal{L}_{\text{GMM}}^{\theta} = -\min_j \log(w_j \mathcal{N}(\theta; \mu_{\theta,j}, \Sigma_{\theta,j})) \quad (6.9)$$

$$\mathcal{L}_{\text{VAE}}^{\theta} = \text{Neg. Log-Likelihood}(\mathcal{N}(E(\theta), \mathcal{I})) \quad (6.10)$$

We minimize the loss above using L-BFGS [258] for 120 iterations on the test split of the MoCap data. We choose L-BFGS instead of Levenberg-Marquardt, since PyTorch currently lacks the feature to efficiently compute Jacobians, without having to resort to multiple backward passes for derivative computations. We report the results for both full and half-space visibility in Tab. 6.2 using the metrics of Section 6.4.1. Our method outperforms the baselines in terms of full-body and penetration metrics, and shows competitive performance w.r.t. to the part metrics. Regression-only methods [75] cannot tightly fit the data, due to the lack of a feedback mechanism.

Runtime: Our method (PyTorch) runs at 150 ms per frame on a P100 GPU, while the baseline L-BFGS method (PyTorch) above requires 520 ms , on the same hardware. We are aware that a highly optimized real-time version of the latter exists and runs at 0.8 ms per frame, performing at most 3 LM iterations, but it requires investing significant effort into a problem specific C++ codebase.

Fig. 6.5 contains the metrics per iteration of our method, averaged across the entire test dataset. It shows that our learned fitter is able to aggressively optimize the target data term and converge quickly.

Ablation study: We perform all our ablations on the problem of fitting SMPL+H to HMD signals, using the half-space visibility setting, see Fig. 6.2.

Weights	V2V (mm) ↓			MPJPE (mm) ↓	GrPe. (mm) ↓
	FB	H	L / R		
Shared	52.3	3.5	3.6 / 3.7	64.1	18.2
Per-step	49.9	3.2	3.1 / 3.7	62.1	15.5

TABLE 6.3: Using per-step network weights reduces head and ground penetration errors, albeit at an N-fold parameter increase.

Network Structure	V2V (mm) ↓			MPJPE (mm) ↓	GrPe. (mm) ↓
	FB	H	L / R		
ResNet50	65.3	6.8	7.3 / 7.6	73.1	16.2
GRU (1024)	53.6	3.7	3.4 / 4.0	66.1	15.1
GRU (1024, 1024)	49.9	3.2	3.1 / 3.7	62.1	15.5

TABLE 6.4: GRU vs a residual feed-forward network [130, 326]. GRU’s memory makes it more effective. Multiple layers bring further benefits, but increase runtime.

Update Rule	V2V (mm) ↓			MPJPE (mm) ↓	GrPe. (mm) ↓
	FB	H	L / R		
+ $\Delta\Theta_n$	53.8	14.7	7.8 / 7.9	66.3	15.8
+Eq. (6.1)	49.9	3.2	3.1 / 3.7	62.1	15.5

TABLE 6.5: Comparison of our update rule (Eq. (6.1)) with the pure network update $\Delta\Theta_n$. Our proposed combination improves the results for all metrics.

Learning rate γ	V2V (mm) ↓			MPJPE (mm) ↓	GrPe. (mm) ↓
	FB	H	L / R		
1e-4	51.9	3.5	3.8 / 4.6	64.2	15.5
Learned	49.9	3.2	3.1 / 3.7	62.1	15.5

TABLE 6.6: Learning to predict γ is better than a constant, with performance degrading gracefully, providing an option for a lower computational cost.

Unless otherwise stated, all numbers are reported after running the initial regressor and the learned fitter for 5 iterations.

We first compare two variants of the fitter, one with shared and the other with separate network weights per optimization step. Table 6.3 shows that the latter can help reduce the errors, at the cost of an N-fold increase in memory.

Secondly, we investigate the effect of the type and structure of the network, replacing the GRU with a feed-forward network with skip connections, i.e., ResNet [130, 326]. We also train a version of our fitter with a single GRU with 1024 units. Table 6.4 shows that the GRU is better suited to this type of problem, thanks to its internal memory.

Thirdly, we compare the update rule of Eq. (6.1) with a learned fitter that only uses the network update, i.e. $\gamma = 0, \lambda = 1$ in Eq. (6.1). This is an instantiation of LGD [324], albeit with a different network and task. Table 6.5 shows that the proposed weighted combination is better than the pure network update.

Fourthly, we investigate whether we need to learn the step size γ or if a constant value is enough. Table 6.6 shows that performance gracefully

Method	V2V (mm) ↓				LdmkErr	
	Face		Head		(mm) ↓	
	-	PA	-	PA	-	PA
LM	34.4	3.7	33.8	5.3	33.8	3.4
Ours	7.9	3.5	8.5	4.1	8.0	3.7

TABLE 6.7: Face fitting to 2D landmarks.

degrades when using a constant learning value. Therefore, it is an option for decreasing the computational cost, without a significant performance drop.

Finally, we present some qualitative results in Fig. 6.6. Notice how the learned fitter corrects the head pose and hand articulation of the initial predictions.

Face fitting to 2D landmarks: We compare our proposed learned optimizer with a C++ production grade solution that uses LM to solve the face fitting problem described in Sec. 6.3.3. Given the per-image 2D landmarks as input, the optimization baseline minimizes the energy with data term (\mathcal{L}^D in Eq. (6.5)) and a simple regularization term to estimate $\Theta^F = \{\theta, \psi, \beta\}$:

$$\mathcal{L}^O(\Theta^F) = \mathcal{L}^D(\Theta^F; D^F) + \mathbf{w} * \left\| \Theta^F \right\|_2^2 \quad (6.11)$$

\mathbf{w} contains the different regularization weights for θ, ψ, β , which are tuned manually for the best baseline result.

The quantitative comparison in Tab. 6.7 shows that our proposed fitter outperforms the LM baseline on almost all metrics. The large value in absolute errors (“-” columns) is due to the wrong estimation of the depth of the mesh. After alignment (PA columns), the gap is much smaller. See Fig. 6.7 for a qualitative comparison.

Runtime: For face fitting, the baseline optimization is in C++ and thus for a fair comparison, we only compare the time it takes to compute the parameter update given the residuals and Jacobians (per-iteration). Computing the values of the learned parameter update (ours, using PyTorch) takes *12 ms* on a P100 GPU, while computing the LM update (baseline, C++) requires *34.7 ms* (504 free variables). Note that the LM update only requires *0.8 ms* on a laptop CPU when optimizing over 100 free variables. The difference is due to the cubic complexity of LM w.r.t. the number of free variables of the problem.

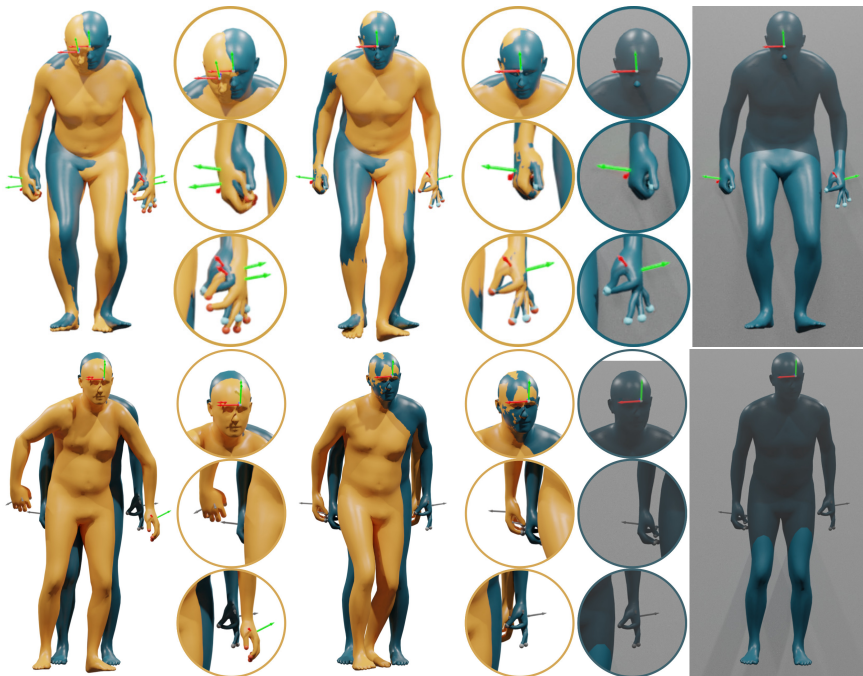


FIGURE 6.6: Estimates in yellow, ground-truth in blue, best viewed in color. From left to right: 1) Initial Φ output, 2) iteration $N = 5$ of our fitter, 3) ground-truth overlay. Our learned optimizer successfully fits the target data and produces plausible poses for the full 3D body. Points that are greyed out are outside of the field of view, e.g. the hands in the second row, and thus not perfectly fit.

6.4.3 Discussion

If we apply the proposed method to a sequence of data, we will get plausible per-frame results, but the overall motion will be implausible. Since the model is trained on a per-frame basis and lacks temporal context, it cannot learn the proper dynamics present in temporal data. Thus, limbs in successive frames will move unnaturally, with large jumps or jitter. Future extensions of this work should therefore explore how to best use past frames and inputs. This could be coupled with a physics based approach, either as part of a controller [399] or using explicit physical losses [285, 383, 413] in \mathcal{L}^D . Another interesting direction is the use of more effective parameterizations for the per-step weights [69, 136]. While all the problems we

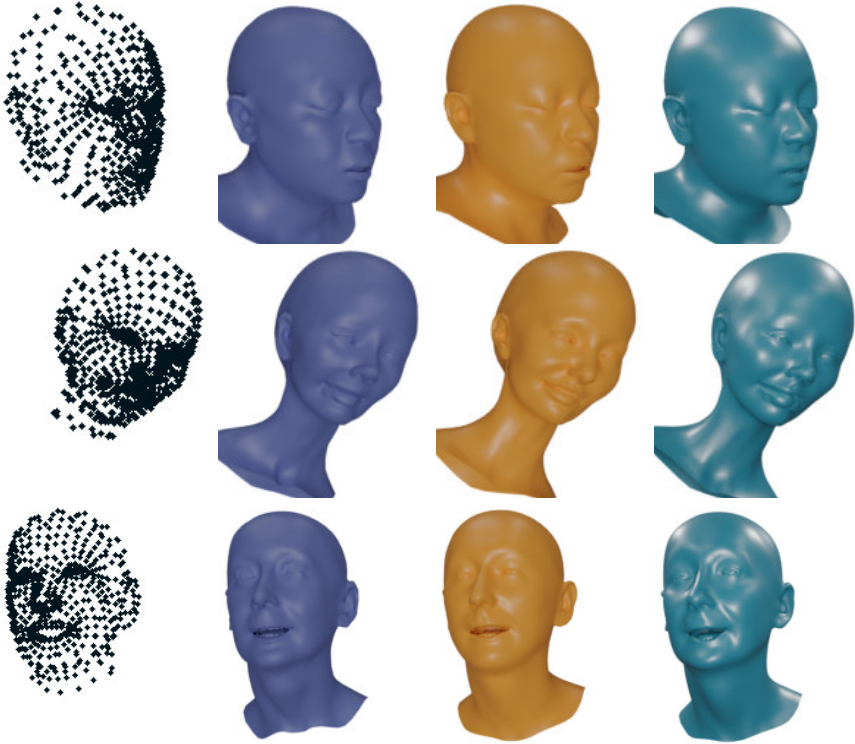


FIGURE 6.7: Face model [377] fitting to dense 2D landmarks 1) target 2D landmarks, 2) LM fitter, 3) ours, 4) ground-truth.

tackle here are under-constrained and could thus have multiple solutions, the current system returns only one. Therefore, combining the proposed system with multi-modal regressors [32, 189] is another possible extension.

Social impact: Accurate tracking is a necessary pre-requisite for the next generation of communication and entertainment through virtual and augmented reality. Learned optimizers represent a promising avenue to realize this potential. However, it can also be used for surveillance and tracking of private activities of an individual, if the corresponding sensor is compromised.

6.5 CONCLUSION

In this chapter, we propose a learned parameter update rule inspired from classic optimization algorithms that outperforms the pure network update and is competitive with standard optimization baselines. We demonstrate the utility of our algorithm on three different problem sets, estimating the 3D body from 2D joints, from sparse HMD signals and fitting the face to dense 2D landmarks. Learned optimizers combine the advantages of classic optimization and regression approaches. They greatly simplify the development process for new problems, since the parameter priors are directly learned from the data, without manual specification and tuning, and they run at interactive speeds, thanks to the development of specialized software for neural network inference. Thus, we believe that our proposed optimizer will be useful for any applications that involve generative model fitting.

SUMMARY

7.1 CONTRIBUTIONS

The contribution of this dissertation is a set of methods that increase the fidelity and accuracy of model-based 3D human reconstruction from images. Specifically, in Part I we describe optimization and regression approaches that jointly estimate 3D body pose and shape, hand articulation and facial expression from a single image, making the reconstructed model more expressive and realistic. Next, in Part II, we look at improving 3D shape reconstruction, which has received much less attention than pose, from easy-to-collect metric and semantic attributes. Last, Part III describes a learned optimization approach for 3D full body estimation from (i) 2D image keypoints, (ii) hand and head location and orientation, given by an AR/VR headset and (iii) face model fitting to dense 2D landmarks.

In Chapter 2 we introduce SMPL-X, a holistic 3D model of the body, hands, and face, and SMPLify-X, an approach that estimates SMPL-X parameters from a single image. Since this is an under-constrained problem, we employ a set of priors to regularize the fitting process. Rather than optimizing body joint rotations in axis-angle space, we optimize the latent code of a VAE trained on a large collection of body poses, obtained from MoCap. We encourage the L_2 norm of the latent code to be close to zero, to penalize unnatural poses. Furthermore, we detect self-collisions in the estimated body with the help of a Bounding Volume Hierarchy (BVH) for fast queries. To resolve the self-penetrations we formulate an energy term that assigns larger energy values to colliding triangles. We also collect a new dataset of images and ground-truth SMPL-X parameters, obtained by registering SMPL-X to 3D scans, called Expressive Hands and Faces (EHF). There we show that the joint estimation of body, hands, and face is more accurate than separate part estimation. Qualitative and quantitative results show that the use of an expressive model leads to more natural 3D reconstructions.

In Chapter 3 we introduce ExPose, a neural network regressor that predicts SMPL-X parameters from a single image. The use of the regressor is motivated by SMPLify-X's main weaknesses: (i) slow runtime and (ii) susceptibility to initialization. We can however use SMPLify-X to collect the necessary training data for the neural network, removing invalid estimates

with the help of human annotators. The hands and face occupy very few pixels compared to the full body, making their estimation very hard from the down-scaled images usually given to neural networks. We propose to use *body-driven attention* to overcome this issue. We start by predicting the 3D pose and shape of the body. The body estimate already localizes the hands and face well enough. We then use the detected hand and face location to extract high-resolution part crops and pass those to dedicated part networks to refine the initial part parameters. An added benefit of this approach is that we leverage part-only data to train and evaluate our part experts. ExPose estimates expressive 3D humans as accurately as SMPLify-X, at a fraction of the computation cost.

ExPose’s main weakness is its naive parameter integration mechanism that simply combines independent estimates from the body, hand, and face experts. PIXIE, introduced in Chapter 4, instead proposes to use moderators, neural networks that merge features from the part images weighted by a confidence value. To improve the realism of the reconstructed bodies, PIXIE also predicts lighting, facial albedo, and geometric details. Since body shape is highly correlated with gender, we label images as female, male and non-binary and train PIXIE to infer “gendered” 3D body shapes with an appropriate shape prior. Quantitative and qualitative results prove that PIXIE estimates more accurate and more realistic 3D bodies than prior methods.

In Chapter 5 we investigate the problem of estimating accurate 3D body shape from monocular RGB images. The main obstacle for training such a regressor is the lack of 3D shape data for in-the-wild images. To overcome this hurdle, we need to find alternate data that can be easily collected for in-the-wild images and constrain 3D body shape. Prior work has shown that linguistic shape attributes, e.g. “big”, “tall”, etc., and anthropometric measurements can be used to accurately predict 3D body shape. Motivated by these findings, we collect (i) images of fashion models with their anthropometric measurements and (ii) linguistic shape attributes for 3D body meshes and the model images. Using mapping functions from attributes and/or measurements to 3D body shape and vice-versa, we formulate shape-aware losses and use them to train SHAPY, a neural network that predicts 3D shape and pose parameters from an RGB image. We observe that existing benchmarks lack ground-truth shape annotations, subject, and clothing variation. Thus, we collect a new dataset, Human Bodies in the Wild (HBW), that contains images of people in natural clothing and settings, together with ground-truth 3D shape acquired from a body scanner.

SHAPY significantly outperforms existing work on this challenging new benchmark.

In Chapter 6 we revisit the problem of fitting parametric models of the human body using learned optimization. Similar to the classical Levenberg-Marquardt algorithm that computes the parameter update as a weighted combination of the gradient descent and Gauss-Newton directions, we propose an update rule that uses a weighted combination of gradient descent and a network-predicted update. We apply this neural optimizer on challenging 3D human model fitting problems: (i) 3D body estimation from 2D image landmarks, (ii) 3D body and hand estimation from a head-mounted device and (iii) 3D face fitting from dense 2D landmarks. The proposed method is versatile, being easy to apply to different problems, and offers a competitive alternative to well-tuned “traditional” model fitting pipelines, both in terms of accuracy and speed.

7.2 FUTURE WORK

We believe that this work opens up new and exciting avenues for future research. First, it motivates the rest of the community to switch from separate to joint 3D body, face and hand prediction [245, 246, 295, 381, 423]. Second, the work presented here unlocks other exciting applications, such as (i) sign language recognition [267] and (ii) reconstruction [96], (iii) forensic identification from body shape [345], (iv) generating perpetual motion [416], (v) action-/text-conditioned motion [17, 276] and (vi) full-body grasping motion [336, 378] to name only a few. Nevertheless, there remain important open questions and issues for the problem of 3D human body and motion modeling and reconstruction from different modalities, such as images, videos and 3D/4D scans.

Model fidelity: All existing holistic 3D body models, i.e. SMPL-X [270], Adam [166] and GHUM [387], represent only the surface of the body. Skin [62, 196, 197] and eye [29, 201] appearance models are key components for increased realism. Note that special care needs to be taken to avoid biases to a specific subset of the population [16, 85, 177]. Therefore, we need to move beyond indoor multi-view and light stage capture facilities [110, 168, 289], whose cost and difficulty of operation make scaling to a large number of subjects intractable. Moving to more lightweight capture methods is necessary to resolve this issue. For example, SunStage [368], captures the shape and appearance of a person’s head with an outdoor video in direct sunlight. Neural Radiance Fields (NeRF) [242], which represent scene

geometry and appearance with a single model, requiring only a few tens of calibrated multi-view images, are another promising alternative, especially with recent advances that train these models in minutes [251, 303]. Next, SMPL-X is “naked”, i.e. it does not model clothing and hair; adding clothing [63, 228, 301], and hair models [370] although not trivial, would significantly increase realism. Finally, applications in medicine, biomechanics and accurate physical simulation will require going beyond the surface [263, 270, 387] to capture [47] and model the skeleton, muscles and tissue of the human body [61, 145, 167, 175, 208, 209].

Temporal and physical plausibility: Despite the impressive results of 3D body estimation methods [183, 189, 203] on monocular images, reconstruction of motion sequences from RGB videos is still far from solved. Running these methods, including temporal extensions [53, 182, 224], on sequences produce results with high jitter, noticeable foot sliding and significant errors in the presence of occlusions. Adding physical constraints to monocular capture methods, either as explicit energy terms for an optimization problem [383], or by embedding our human models inside a physical simulation [225, 322, 356, 399] will help us overcome these problems, unlocking new applications such as physically stable grasp synthesis [57].

Multi-person estimation: All the work presented in this thesis deals with a single human subject at a time, assuming that a bounding box for each person is available. Although this is a valid assumption, given the accuracy of modern object detectors, single-person methods still suffer when multiple overlapping people are present in a scene [202]. More importantly, humans are, by nature, social beings, communicating and interacting with their fellows. Their pose, expression, and gestures depend on the actions of other persons in their surroundings. Therefore, in order to reason about their relations, their emotions, and their actions it is important to predict the 3D body, face, and hands of *all* persons in an image. Extending recent multi-person 3D pose and shape estimation methods [156, 253, 333, 334, 405] with the hand and face estimation using the techniques described in Chapters 2 to 4 is a promising first step. Of course, the more interesting question is how we can use models of interaction and conversation [164, 256] to improve our 3D body, hand, face estimates.

Human-scene estimation: Last, but not least, humans, with the notable exception of astronauts, do not float in space, but live, interact and move in their environment. The world constrains the body and vice-versa. Knowledge of the scene structure, e.g. a world model built by a SLAM system [425], can help us improve our estimates of the 3D body’s articulation, shape, and

location [124]. The converse is also true, i.e. body pose can help us improve the 3D reconstruction of a scene [375, 394].

Motion and interaction generation: Perpetual motion generation [416] is a useful tool, e.g. for animation purposes [101]. Humans however do not just stand or wander aimlessly in their surroundings, but interact with them to fulfill their goals. Consequently, to create virtual avatars that move and act like humans we need models that generate realistic (i) grasps and grasping motions [57, 336, 337], (ii) interactions with rigid [30, 123, 389, 411, 412] articulated [84, 157, 243, 382] and deformable objects [48], (iii) action-/text-conditioned motion [276, 277, 282], e.g. tasking an avatar to execute a cooking recipe.

7.3 CONCLUSION

The introduction of an expressive 3D model of the human body, described in Chapter 2, led to the creation of methods that jointly predict the 3D body, hands, and face from a single image, presented in Chapters 2 to 4. This has motivated the community to slowly shift from separate body [152, 169, 183, 187, 203], hand [120, 427] and face [79] prediction to holistic body reconstruction [295, 347, 409]. The increase in expressivity has also benefited applications, such as continuous sign language recognition [190, 267] and reconstruction [96], human-object interaction capture [143, 337] and synthesis [336, 378], human-scene interaction capture [124, 394] and synthesis [125, 414, 418], to name only a few. An important lesson here is that the different body parts complement and constrain each other, e.g. that the face and hands can be localized from the body, used by ExPose in Chapter 3, or that the body contains useful context information for the pose of the hand, as illustrated by PIXIE in Chapter 4.

Although the accuracy of methods that estimate 3D body poses from images has increased rapidly in the last few years, the same cannot be said for 3D shape estimation. SHAPY, presented in Chapter 5, makes a step towards reducing this performance gap by utilizing anthropometric measurements and easy-to-collect linguistic attributes for 3D shape supervision to predict more accurate body shapes than existing methods. Accurate estimates of the 3D body shape will be crucial for virtual try-on, augmented/virtual reality, and health applications. A key takeaway from Chapter 5 is that proxy data, which can be easily collected in large quantities, such as anthropometric measurements or linguistic attributes which only provide weak supervision 3D body shape estimation, can be an effective replacement for

full supervision, which might be significantly harder or even impossible to obtain.

Last, in Chapter 6 we described a learned optimization method that combines insights from classical gradient-based optimization and direct parameter regression. The proposed update rule follows the structure of the Levenberg-Marquardt algorithm, computing parameter updates as a combination of gradient descent and a higher order update predicted by a neural network, instead of the Gauss-Newton direction in LM, and controlling the learning rate of each variable independently. The integration of well-known and effective structures, in this case, a constraint on the form of the parameter update function, accelerates training and improves the final performance of the model. The proposed model is application-agnostic and easy to apply to different problem settings, as shown in Sec. 6.4. The effectiveness and versatility of our learned optimizer will benefit all applications that require generative model fitting.

Appendices

EXPRESSIVE BODY CAPTURE: 3D HANDS, FACE, AND BODY FROM A SINGLE IMAGE

A.1 QUALITATIVE RESULTS

Comparison of SMPL, SMPL+H & SMPL-X: In Tab. 2.1 from Sec. 2.4.2 we present a quantitative comparison between different models with different modeling capacities. In Fig. A.1 we present a similar comparison for SMPL (left), SMPL+H (middle) and SMPL-X (right) for an image of the EHF dataset. For fair comparison we fit all models with a variation of SMPLify-X to a single RGB image. The figure reflects the same findings as Tab. 2.1, but qualitatively; there is a clear increase in expressiveness from left to right, as model gets richer from body-only (SMPL) to include hands (SMPL+H) or hands and face (SMPL-X).

Holistic vs part models: In Sec. 2.4.2 and Fig. 2.4 we compare our holistic SMPL-X model to the hand-only approach of [266] on EHF. Figure A.2 shows a similar qualitative comparison, this time on the data of [266]. To further explore the benefit of holistic reasoning, we also focus on the head and we compare SMPL-X fitting to a head-only method by fitting FLAME [206] to 2D keypoints similar to our method. The context of the full body stabilizes head estimation for occlusions or non-frontal views, see Fig. A.3. This benefit is also quantitative, where the holistic SMPL-X improves over the head-only fitting by 17% in our EHF dataset in terms of vertex-to-vertex error.

Failure cases: Figure A.4 shows some representative failure cases; depth ambiguities can cause wrong estimation of torso pose or wrong ordinal depth estimation of body parts due to the simple 2D re-projection data term. Furthermore, occluded joints leave certain body parts unconstrained, which currently leads to failures and could be addressed by employing a visibility term in the objective.

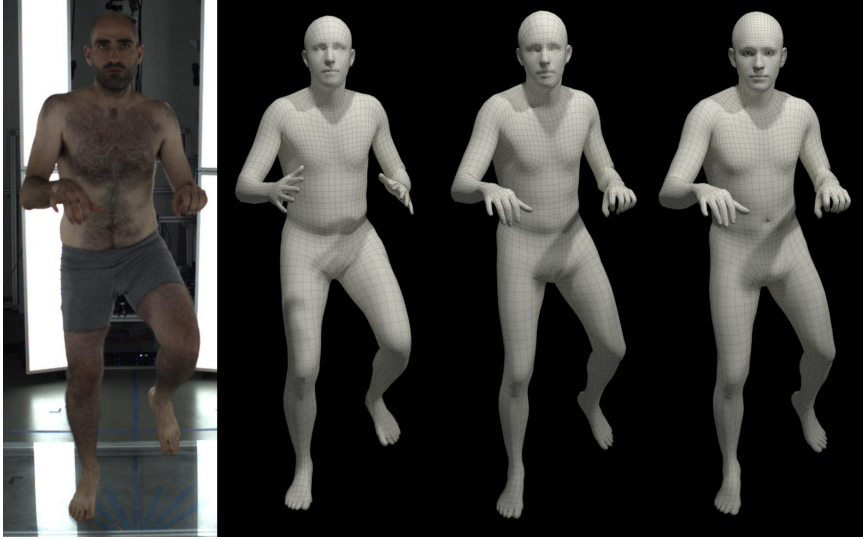


FIGURE A.1: Comparison of SMPL (left), SMPL+H (middle) and SMPL-X (right) on the EHB dataset, using the male models. For fair comparison we fit all models with a variation of SMPLify-X to a single RGB image. The results show a clear increase in *expressiveness* from left to right, as model gets richer from body-only (SMPL) to include hands (SMPL+H) or hands and face (SMPL-X).

A.2 COLLISION PENALIZER

Section 2.3.4 contains the description the collision penalizer. For technical details and visualizations the reader is redirected to [25, 356], but for the sake of completion we include the mathematical formulation also here.

We first detect a list of colliding triangles \mathcal{C} by employing Bounding Volume Hierarchies (BVH) [341] and compute local conic 3D distance fields $\Psi : \mathbb{R}^3 \rightarrow \mathbb{R}_+$ defined by the triangles \mathcal{C} and their normals $n \in \mathbb{R}^3$. Penetrations are then penalized by the depth of intrusion, efficiently computed by the position in the distance field. For two colliding triangles f_s and f_t intrusion is bi-directional; the vertices $v_t \in \mathbb{R}^3$ of f_t are the *intruders* in the distance field Ψ_{f_s} of the *receiver* triangle f_s and are penalized by $\Psi_{f_s}(v_t)$, and vice-versa. Thus, the collision term $E_{\mathcal{C}}$ is defined as

$$E_{\mathcal{C}}(\theta) = \sum_{(f_s(\theta), f_t(\theta)) \in \mathcal{C}} \left\{ \sum_{v_s \in f_s} \| -\Psi_{f_t}(v_s)n_s \|^2 + \sum_{v_t \in f_t} \| -\Psi_{f_s}(v_t)n_t \|^2 \right\}. \quad (\text{A.1})$$

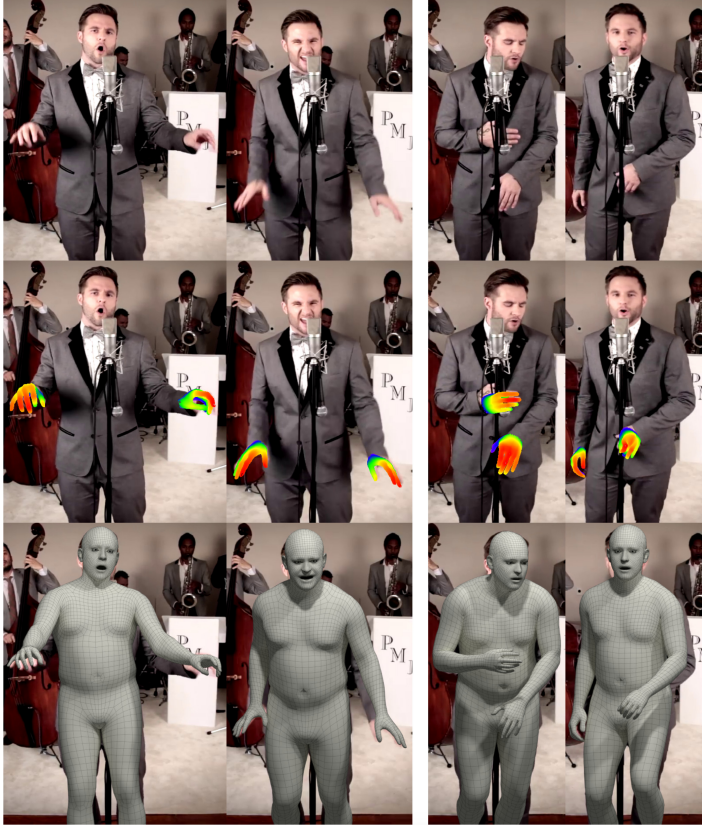


FIGURE A.2: Comparison of the hands-only approach of [266] (middle row) against SMPLify-X with the male SMPL-X (bottom row). Both approaches depend on OpenPose [43]. In case of good 2D detections both perform well (left group). In case of noisy detections (right group) fitting a holistic model is more robust.

For the case where f_t is the *intruder* and f_s is the *receiver* (similarly for the opposite case) the cone for the distance field Ψ_{f_s} is defined as

$$\Psi_{f_s}(v_t) = \begin{cases} |(1 - \Phi(v_t))Y(n_{f_s} \cdot (v_t - \mathbf{o}_{f_s}))|^2 & \Phi(v_t) < 1 \\ 0 & \Phi(v_t) \geq 1 \end{cases} \quad (\text{A.2})$$

The term

$$\Phi(v_t) = \frac{\|(v_t - \mathbf{o}_{f_s}) - (n_{f_s} \cdot (v_t - \mathbf{o}_{f_s}))n_{f_s}\|}{-\frac{r_{f_s}}{\sigma}(n_{f_s} \cdot (v_t - \mathbf{o}_{f_s})) + r_{f_s}} \quad (\text{A.3})$$

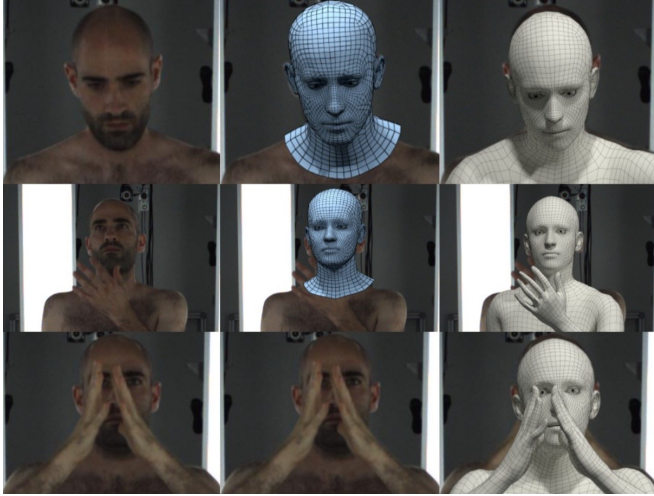


FIGURE A.3: Fitting SMPL-X (right) versus FLAME (middle). For minimal occlusions and frontal views (top) both methods perform well. For moderate (middle) or extreme (bottom) occlusions the body provides crucial context and improves fitting (bottom: missing FLAME model indicates a complete fitting failure).

projects the vertex v_t onto the axis of the cone defined by the triangle normal n_{f_s} and going through the circumcenter o_{f_s} . It then measures the distance to it, scaled by the radius of the cone at this point. If $\Phi(v) < 1$ the vertex is inside the cone and if $\Phi(v) = 0$ the vertex is on the axis. The term

$$Y(x) = \begin{cases} -x + 1 - \sigma & x \leq -\sigma \\ -\frac{1-2\sigma}{4\sigma^2}x^2 - \frac{1}{2\sigma}x + \frac{1}{4}(3-2\sigma) & x \in (-\sigma, +\sigma) \\ 0 & x \geq +\sigma \end{cases} \quad (\text{A.4})$$

measures how far the projected point is from the circumcenter to define the intensity of penalization. For $Y(x) < 0$ the projected point is behind the triangle. For $x \in (-\sigma, +\sigma)$ the penalizer is quadratic, while for $x > |\sigma|$ it becomes linear. The parameter σ also defines the field of view of the cone. In contrast to [25, 356] that use *mm* unit and $\sigma = 0.5$, we use *m* unit and $\sigma = 0.0001$. For the resolution of our meshes, we empirically find that this value allows for both penalizing penetrations, as well as for not over-penalizing in case of self-contact, e.g. arm resting on knee.

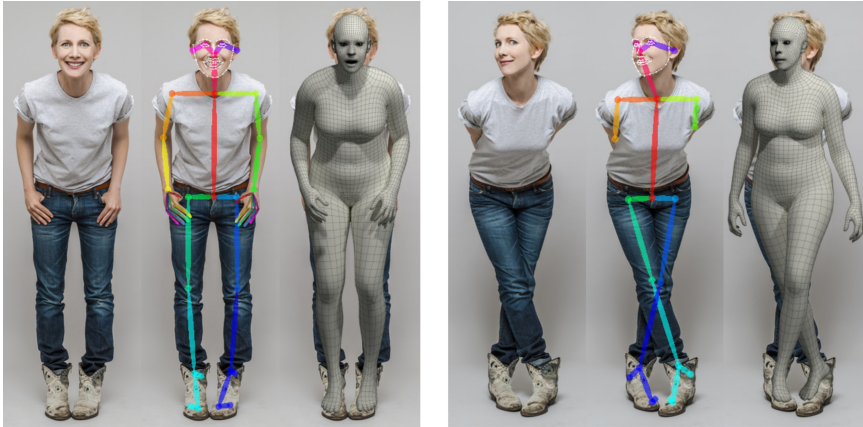


FIGURE A.4: Failure cases for *SMPLify-X* with the female *SMPL-X* for expressive RGB images similar to the ones in Figs. 2.1 and 2.2. In the left case, 2D keypoints are reasonable, but due to depth ambiguities the torso pose is wrong, while the head shape is under-estimated. In the right case, the arms and hands are occluded and due to lack of constraints the arm and hand pose is wrong. The ordinal depth for feet is estimated wrongly, while similarly to the left case the torso pose and head shape are not estimated correctly. *Left*: Input RGB image. *Middle*: Intermediate 2D keypoints from OpenPose. *Right*: *SMPL-X* fittings overlaid on the RGB image.

As seen in Fig. A.5, for certain parts of the body, like the eyes, toes, armpits and crotch, as well as neighboring parts in the kinematic chain, there is either always or frequently self-contact. For simplicity, since the model does not model deformations due to contact, we simply ignore collisions for neighboring parts in these areas. Our empirical observations suggest that collision detection for the other parts resolves most penetrations and helps prevent physically implausible poses. Figure A.6 shows the effect of the collision penalizer, by including or excluding it from optimization, and depicts representative success and failure cases.

For computational efficiency, we developed a custom PyTorch wrapper operator for our CUDA kernel based on the highly parallelized implementation of BVH [172].

A.3 OPTIMIZATION

In Sec. 2.3.6 we present the main information about optimizing our objective function, while in the following we present omitted details.

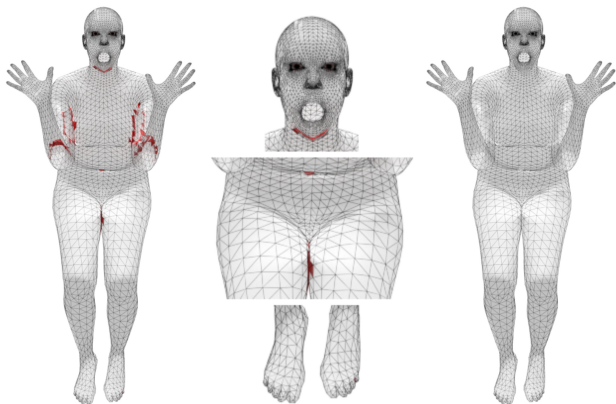


FIGURE A.5: For certain parts of the body, like the eyes, toes, armpits and crotch, as well as neighboring parts in the kinematic chain, there is either always or frequently self-contact. The triangles for which collisions are detected are highlighted with red (left, middle). Since the model does not model deformations due to contact, for simplicity we just ignore collisions for these areas (right).

To keep optimization tractable, we use a PyTorch implementation and the Limited-memory BFGS optimizer (L-BFGS) [258] with strong Wolfe line search. We use a learning rate of 1.0 and 30 maximum iterations. For the annealing scheme presented in Sec. 2.3.6 we take the following three steps. We start with high regularization to mainly refine the global body pose, ($\gamma_b = 1, \gamma_h = 0, \gamma_f = 0$) and gradually increase the influence of hand keypoints to refine the pose of the arms ($\gamma_b = 1, \gamma_h = 0.1, \gamma_f = 0$). After converging to a better pose estimate, we increase the influence of both hands and facial keypoints to capture expressivity ($\gamma_b = 1, \gamma_h = 2, \gamma_f = 2$). Throughout the above steps the weights $\lambda_\alpha, \lambda_\beta, \lambda_\psi$ in the objective function E start with high regularization that progressively lowers to allow for better fitting. The only exception is λ_C that progressively increases while the influence of hands and facial keypoints gets stronger in E_J , thus bigger pose changes and more collisions are expected.

Regarding the weights of the optimization, they are set empirically and the exact parameters for each stage of the optimization will be released with our code. For more intuition we performed sensitivity analysis by perturbing each weight λ separately by up to $\pm 25\%$. This resulted to relative changes smaller than 6% in the vertex-to-vertex error metric, meaning that

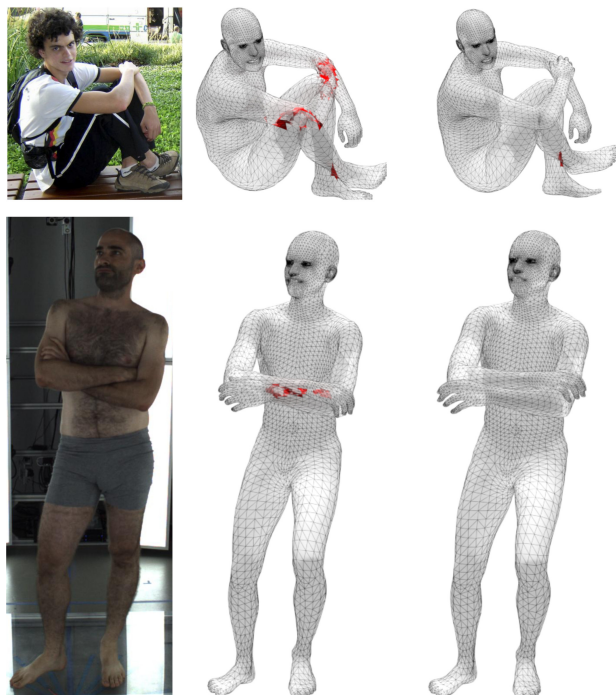


FIGURE A.6: Effect of the collision penalizer. The colliding triangles are highlighted to show penetrations at the end of optimization with SMPLify-X without (middle) and with (right) the collision term in the objective function. The top row shows a successful case, where optimization resolves most collisions and converges in a physically plausible pose that reflects the input image. The bottom row shows a failure case, for which arm crossing causes a lot of collisions due to self-touch. The final pose (right) is still physically plausible, but optimization gets trapped in a local minima and the pose does not reflect the input image.

our approach is robust for significant weight ranges and not sensitive to fine-tuning. The detailed results are presented in Fig. A.7.

A.4 QUANTITATIVE EVALUATION ON “TOTAL CAPTURE”

In Sec. 2.4.1 we present a curated dataset called *Expressive hands and faces dataset (EHF)* with ground-truth shape for bodies, hands and faces together.

Since the most relevant model is Frank [166], we also use the “Total Capture” dataset [66] of the authors, focusing on the “PtCloudDB” part that

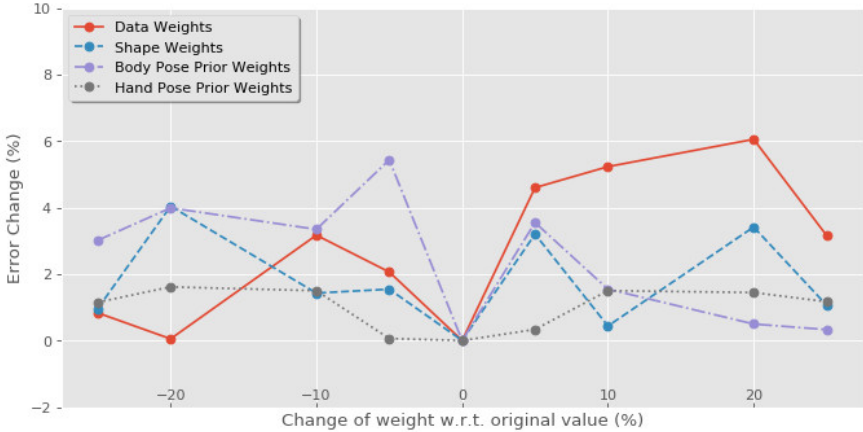


FIGURE A.7: Sensitivity of the weights for the different terms of the optimization. Each weight λ is perturbed separately up to $\pm 25\%$. The relative changes in the vertex-to-vertex error are smaller than 6%, indicating that our approach is robust for significant weight ranges and not sensitive to fine-tuning.

includes pseudo ground-truth for all body, face and hands. This pseudo ground-truth is created with triangulated 3D joint detection from multi-view with OpenPose [43]. We curate and pick 200 images, according to the degree of visibility of the body in the image, interesting hand poses and facial expressions. In the following, we refer to this data as “total hands and faces” (THF) dataset. Figure A.8 shows qualitative results on part of THF. For each group of images the top row shows a reference RGB image, the middle row shows SMPLify-X results using pseudo ground-truth OpenPose keypoints (projected on 2D for use by our method), while the bottom row shows SMPLify-X results using 2D OpenPose keypoints estimated with [43]. Quantitative results for this dataset are reported in Tab. A.1.

A.5 QUANTITATIVE EVALUATION ON HUMAN3.6M

Table 2.1 demonstrates that evaluating the reconstruction accuracy using 3D body joints is not representative of the accuracy and the detail of a method’s reconstruction. However, many approaches do evaluate quantitatively based on 3D body joints metrics, so here we compare our results with SMPLify [34] to demonstrate that our approach is not only more natural, expressive and detailed, but the results are also more accurate in the common metrics.



FIGURE A.8: Qualitative results on some of the data of the “total capture” dataset [66], focusing on the “PtCloudDB” part that includes pseudo ground-truth for all body, face and hands. We curate and pick 200 images, according to degree of body coverage in the image and interesting hand poses and facial expressions. We refer to this data as “total hands and faces” dataset (THF). *Top row*: Reference RGB image. *Middle row*: SMPLify-X results using pseudo ground-truth OpenPose keypoints (3D keypoints of [66] estimated from multi-view and projected on 2D). *Bottom row*: SMPLify-X results using 2D OpenPose keypoints estimated with [43]. Gray color depicts the gender-specific model for confident gender detections. Blue is the gender-neutral model that is used when the gender classifier is uncertain.

		SMPLify-X using	
Error Joints	Alignment Joints	GT 2D	pred 2D
Body	Body	92.6	117.5
Body+H+F	Body	101.2	136.2
Body+H+F	Body+H+F	71.2	93.4

TABLE A.1: Quantitative results on the selected frames from CMU Panoptic Studio, using SMPLify-X on the 2D re-projection of the ground-truth 3D joints, and the 2D joints detected by OpenPose respectively. The numbers are mean 3D joint errors after Procrustes alignment. First, we evaluate the error on the body-only keypoints after Procrustes alignment with the ground-truth body-only keypoints (row 1). Then, we consider the same alignment using body-only keypoints, but we evaluate the joint error across all the body+hands+face keypoints (row 2). Finally, we align the prediction using all body+hands+face keypoints and we report the mean error across all of them (row 3).

Method	Mean (mm)	Median (mm)
SMPLify [34]	82.3	69.3
SMPLify-X	75.9	60.8

TABLE A.2: Quantitative results on the Human3.6M dataset [150]. The numbers are mean 3D joint errors after Procrustes alignment. We use the evaluation protocol of SMPLify [34].

In Tab. A.2 we present our results using the Human3.6M [150] dataset. We follow the same protocol as SMPLify [34] and we report results after Procrustes alignment with the ground-truth 3D pose. Even though there are several factors that improve our approach over SMPLify and this experiment does not say which is more important (we direct the reader to the ablative study in Tab. 2.2 for this), we still outperform the original SMPLify using this crude metric based on 3D joints.

A.6 QUALITATIVE EVALUATION ON MPII

In Fig. A.13 we present qualitative results on the MPII dataset [13]. For this dataset we also include some cases with low resolution, heavily occluded or cropped people.

A.7 MODEL

Section 2.3.1 contains the description of the SMPL-X model. The model shape space is trained on the CAESAR database [290]. In Fig. A.9a we present the percentage of explained variance as a function of the number of PCA components used. All models explain more than 95% of the variance with 10 principle components.

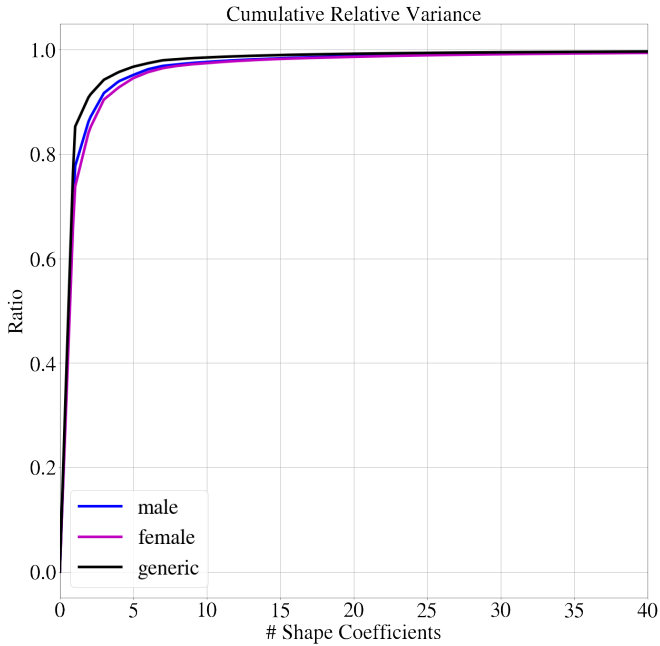
We further evaluate the model on a held out set of 180 alignments of male and female subjects in different poses. The male model is evaluated on the male alignments, the female model is evaluated on the female alignments, while the gender neutral is evaluated on both male and female alignments. We report the model alignment vertex-to-vertex (v2v) mean absolute error as a function of the number of principle components used, shown in Fig. A.9b.

A.8 VPOSER

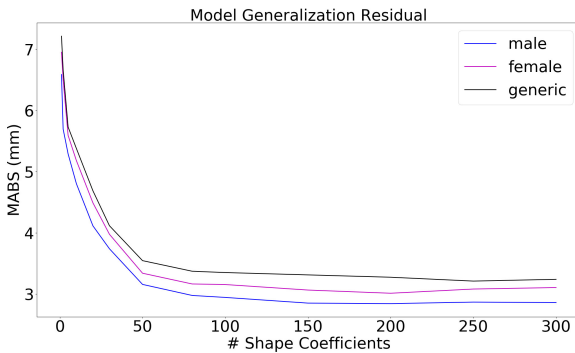
In Sec. 2.3.3 we introduce a new parametrization of the human pose and a prior on this parameterization, also referred to as VPoser. In this section, we present further details on the data preparation and implementation.

A.8.1 *Data preparation*

We use SMPL body pose parameters extracted with [221, 232] from human motion sequences of CMU [60], Human3.6M [150], and PosePrior [5] as our dataset. Subsequently, we hold out parameters for Subjects 9 and 11 of Human3.6M as our test set. We randomly select 5% of the training set as our validation set and use that to make snapshots of the model with minimum validation loss. We choose matrix rotations for our pose parameterization.



(a) Cumulative relative variance of the CAESAR dataset explained as a function of the number of shape coefficients for three SMPL-X models: male, female, gender neutral model.



(b) Evaluating SMPL-X generalization on a held out test set of male and female 3D alignments.

FIGURE A.9: SMPL-X evaluation on held-out test set.

A.8.2 Implementation details

For implementation we use TensorFlow [2] and later port the trained model and weights to PyTorch [268]. Figure A.10 shows the network architecture during training and test time. We use only fully-connected layers, with LReLU [230] non-linearity and keep the encoder and decoder symmetric. The encoder has two dense layers with 512 units each, and then one dense layer for mean and another for variance of the VAE’s posterior Normal distribution. The decoder weights have the same shape as the encoder, only in reverse order. We use the ADAM solver [180], and update the weights of the network to minimize the loss defined in Eq. (2.6). We empirically choose the values for loss weights as: $c_1 = 0.005$, $c_2 = 1.0 - c_2$, $c_3 = 1.0$, $c_4 = 1.0$, $c_5 = 0.0005$. We train for 60 epochs for each of the following learning rates: $[5e-4, 1e-4, 5e-5]$.

After training, the latent space describes a manifold of physically plausible human body poses, that can be used for efficient 2D-to-3D lifting. Figure A.12 shows a number of random samples drawn from the latent space of the model.

A.9 GENDER CLASSIFIER

Figure A.11 shows some qualitative results of the gender classifier on the test set.

A.9.1 Training data

For training data we employ the LSP [160], LSP-extended [161], MPII [13], COCO [215], LIP [212] datasets, respecting their original train and test splits. To curate our data for gender annotations we collect tight crops around persons and keep only the ones for which there is at least one visible joint with high confidence for the head, torso and for each limb. We further reject crops with size smaller than 200×200 pixels. The gathered samples are annotated with gender labels using Amazon Mechanical Turk. Each image is annotated by two Turkers and we keep only the ones with consistent labels.

A.9.2 *Implementation details*

For implementation we use Keras [54] with TensorFlow [2] backend. We use a pretrained ResNet18 [130] for feature extraction and append fully-connected layers for our classifier. We employ a cross entropy loss, augmented with an L2 norm on the weights. Each data sample is resized to 224×224 pixels to be compatible with the ResNet18 [130] architecture. We start by training the final fully-connected layers for two epochs with each of the following learning rate values [1e-3, 1e-4, 1e-5, 1e-6]. Afterwards, the entire network is finetuned end-to-end for two epochs using these learning rates [5e-5, 1e-5, 1e-6, 1e-7]. Optimization is performed using ADAM [180].

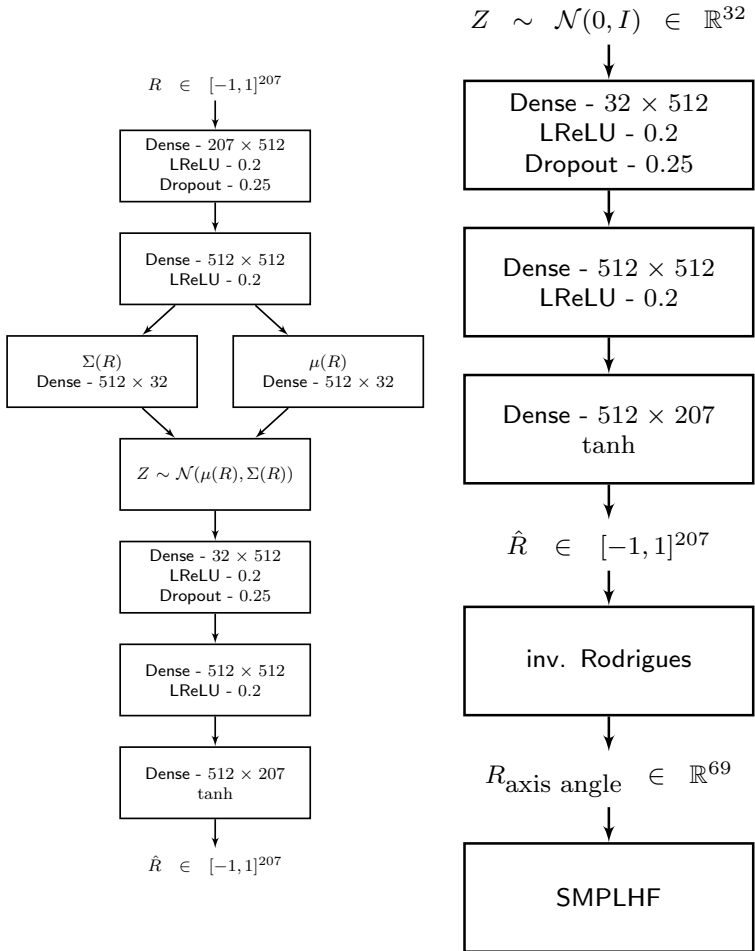


FIGURE A.10: VPoser model in different modes. For training the network consists of an encoder and a decoder. For testing we use the latent code instead of the body pose parameters, i.e. θ_b , of SMPL-X. By “inverse Rodrigues” we note the conversion from a rotation matrix to an axis-angle representation for posing SMPL-X.



FIGURE A.11: Gender classifier results on the test set. From left to right column: Successful predictions, predictions discarded due to low confidence (< 0.9), failure cases.

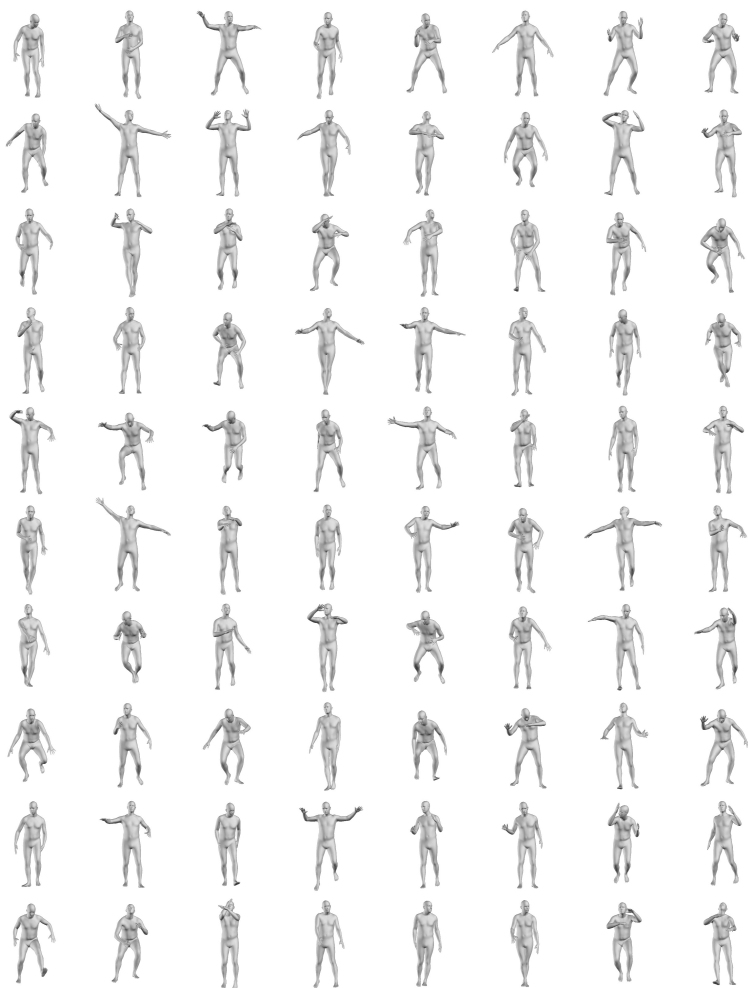


FIGURE A.12: Random pose samples from the latent space of VPoser. We sample from a 32 dimensional normal distribution and feed the value to the decoder of VPoser; shown in Sec. A.7. SMPL is then posed with the decoder output, after conversion to an axis-angle representation.

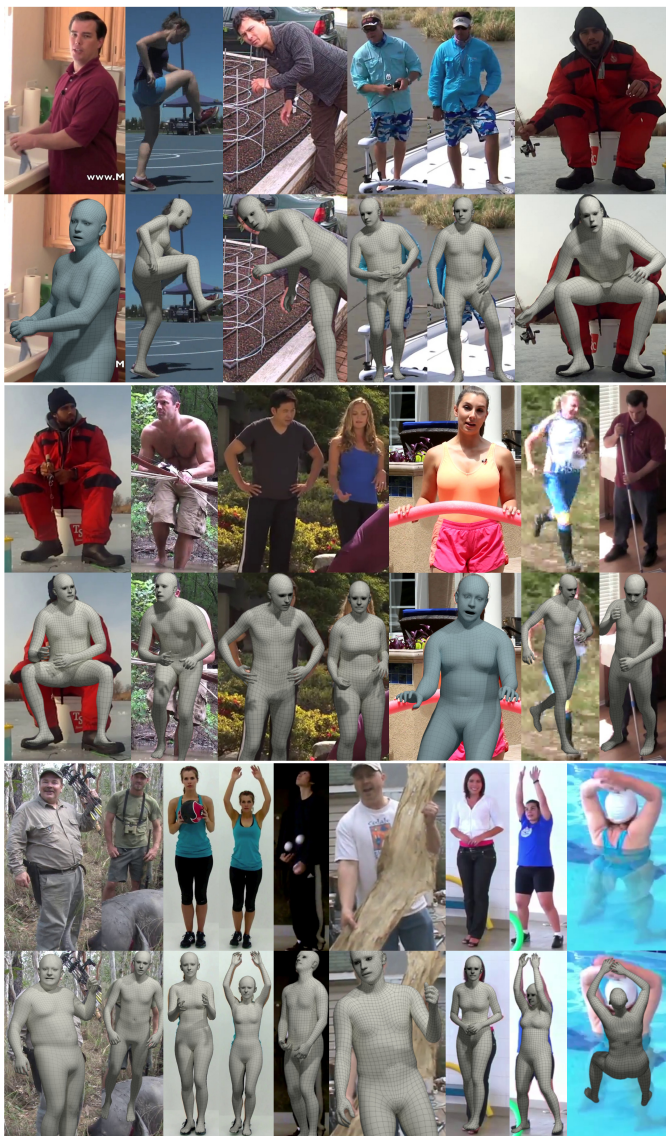


FIGURE A.13: Qualitative results of SMPLify-X with SMPL-X on the MPII dataset [13]. In this figure we also include images with some heavily occluded or cropped bodies. *Gray* color depicts the gender-specific model for confident gender detections. *Blue* is the gender-neutral model that is used when the gender classifier is uncertain or when cropping does not agree with the filtering criterion described in Sec. A.9.1.



FIGURE A.14: Results of SMPLify-X fitting for the LSP dataset. For each group of images we compare two body priors; the top row shows a reference RGB image, the bottom row shows results of SMPLify with VPoser, while the middle row shows results for which VPoser is replaced with the GMM body pose prior of SMPLify [34]. To eliminate factors of variation, for this comparison we use the gender neutral SMPL-X model.

MONOCULAR EXPRESSIVE BODY REGRESSION THROUGH BODY-DRIVEN ATTENTION

B.1 TRAINING DETAILS

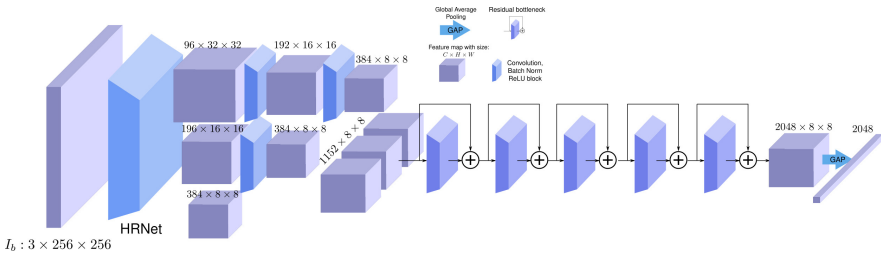


FIGURE B.1: Structure of the feature extractor used by the body prediction network. The image I_b is fed to HRNet [330] to extract multi-scale feature maps. These are then processed by extra convolutional blocks and downsampled to the same spatial resolution. All feature maps are subsequently concatenated and fed to 5 residual blocks [130], followed by a global average pooling operation that produces the final feature vector F_b .

Architecture: The features F_b are extracted from the body image I_b using the architecture of Fig. B.1. The parameters $\Theta = \{\beta, \theta, \psi, s, t\}$ are predicted by feeding the features F_b and the mean parameters $\bar{\Theta}$ to an iterative regression network, whose structure follows HMR [169]. The composition of the feature extraction network of Fig. B.1 and the iterative regressor forms the body network g .

Training: We pre-train the body network until validation performance on 3DPW [235] saturates, using ADAM [180], with batch size 48. The hand and head sub-networks are pre-trained as well on the FreiHAND [427] and FFHQ [173] data, again with ADAM [180] and a batch size of 64. Once validation performance saturates, we freeze the body network and fine-tune the hand and head sub-networks with all available training data to produce ExPose. The exact hyper-parameters will be included in the released code. The entire pipeline is implemented in PyTorch.



FIGURE B.2: **Illustrative examples.** The default global rotation of the hand is replaced by a random rotation with angle $r_{\text{global}} \sim \mathcal{U}(r_{\text{min}}, r_{\text{max}})$ around the ground truth axis of rotation given by the training data. We selected a range $(r_{\text{min}}, r_{\text{max}})_{\text{hand}} = (-90, 90)$ degrees. Blue is the ground-truth mesh used as a target for training, while gray is the starting point of the iterative hand regressor with a perturbed global rotation.

B.2 DATA AUGMENTATION

For hand and face-only data, shape and pose regression is done following the iterative scheme of [169], which computes offsets from a set of mean parameters. When we have access to full body information, we wish to condition the part specific sub-networks on the output of the body network. However, naively adding this conditioning is not enough, as this creates a domain gap between hands and face-only images and those coming from the body attention mechanism. To bridge this, we augment the training data by modifying the initial mean point to some random point. In this way, the part sub-network will be forced to learn to predict the correct offsets, no matter the initial point, that lead to the pose and shape that matches the image. As described in Sec. 3.3.3, we randomly perturb the global rotation of the hand and face data around the ground-truth axis of rotation, as illustrated in Figs. B.2 and B.3 respectively. We also modify the shape of the hand and the face by randomly sampling from normal distributions over the hand and face shape parameters, as illustrated in Figures B.5 and B.6 respectively. For the face-only data, we also augment the rotation of the jaw, by replacing the default value with a random rotation around the x-axis, seen in Fig. B.4. Finally, we replace the default mean expression with a sample drawn from a standard normal distribution, as seen in Fig. B.7.

B.3 CONVERTING SMPL TO SMPL-X

There exist a wide variety of SMPL annotations for training 3D body pose and shape estimation methods. It is therefore important to create an automated method to convert them to the corresponding SMPL-X parameters,

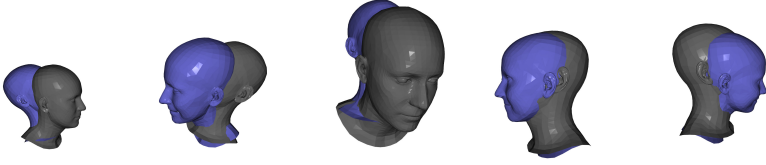


FIGURE B.3: The default global rotation of the head is replaced by a random rotation with angle $r_{\text{global}} \sim \mathcal{U}(r_{\text{min}}, r_{\text{max}})$ around the ground truth axis of rotation given by the training data. We selected a range $(r_{\text{min}}, r_{\text{max}})_{\text{head}} = (-45, 45)$ degrees. Blue is the ground-truth mesh used as a target for training, while gray is the starting point of the iterative face regressor with a perturbed global rotation.

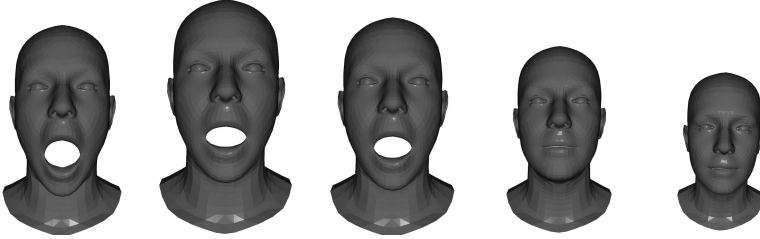


FIGURE B.4: The default rotation of the jaw, which corresponds to a closed mouth, is replaced by a random rotation around the x-axis. The angle of rotation is sampled randomly from the uniform distribution $r_{\text{jaw}} \sim \mathcal{U}(0, 45)$.

to use them as training data. To achieve this, we leverage the relation between SMPL and SMPL-X to build a correspondence map between the two models. SMPL and SMPL-X are articulated models of the human body that produce 3D triangle meshes:

$$\text{SMPL: } (M_{\text{SMPL}}, \mathcal{T}_{\text{SMPL}}) \quad (\text{B.1})$$

$$\text{SMPL-X: } (M_{\text{SMPL-X}}, \mathcal{T}_{\text{SMPL-X}}) \quad (\text{B.2})$$

$$M_{\text{SMPL}} \in \mathbb{R}^{6890 \times 3}, \mathcal{T}_{\text{SMPL}} \in \mathbb{N}^{13776 \times 3} \quad (\text{B.3})$$

$$M_{\text{SMPL-X}} \in \mathbb{R}^{10475 \times 3}, \mathcal{T}_{\text{SMPL-X}} \in \mathbb{N}^{20908 \times 3} \quad (\text{B.4})$$

We start by registering the SMPL template mesh to the SMPL-X template. Given the registered meshes, we compute for each SMPL-X vertex v_i its nearest point p_i on the SMPL mesh and store the index of the nearest SMPL triangle t_i , its vertex indices $t_i = [t_0^i, t_1^i, t_2^i]$ and the barycentric coordinates $[\alpha_i, \beta_i, \gamma_i]$ of point_i with respect to triangle t_i . We also store a binary mask

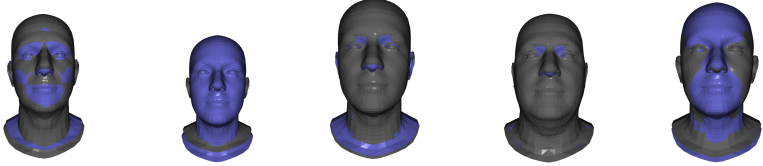


FIGURE B.5: The default mean shape of the head is replaced with a random vector $\beta \sim \mathcal{N}(\vec{0}, I), I \in \mathbb{R}^{10 \times 10}$. The blue mesh represents the mean shape, while the gray mesh has a random shape drawn from the above distribution.



FIGURE B.6: The default mean shape of the hand is replaced with a random vector $\beta \sim \mathcal{N}(\vec{0}, I), I \in \mathbb{R}^{10 \times 10}$. The blue mesh represents the mean shape, while the gray mesh has a random shape drawn from the above distribution.

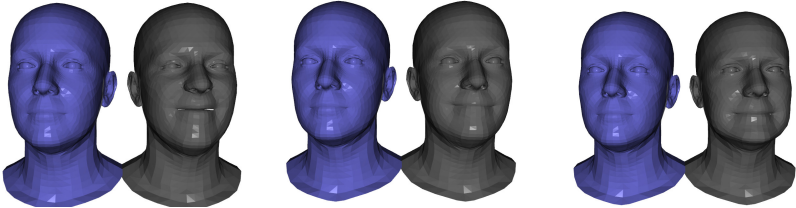


FIGURE B.7: The default neutral expression of the head is replaced with a random vector $\psi \sim \mathcal{N}(\vec{0}, I), I \in \mathbb{R}^{10 \times 10}$. The blue mesh represents the neutral expression, while the gray mesh has a random expression drawn from the above distribution.

$m_i \in \{0, 1\}$ for each vertex that is used to mask invalid correspondences between the two models, such as the eyes, inner lip region, etc..

Given a posed SMPL mesh (M', \mathcal{T}') , e.g. one sample from the fit data of SPIN [187], we build a mesh \hat{M} in SMPL-X topology. Vertex \hat{v}_i of the mesh \hat{M} is computed as:

$$\hat{v}_i = \alpha_i v'_{t'_i} + \beta_i v'_{t'_1} + \gamma_i v'_{t'_2} \quad (\text{B.5})$$

where $v'_{t'_i}$ is the SMPL vertex with index t'_i . We now have a mesh in SMPL-X topology, which we will use to find the corresponding pose θ , shape β ,

expression ψ and translation t parameters. Let v_i be the i -th vertex returned by posing SMPL-X using the current values of the parameters (θ, β, ψ, t) . We start by optimizing only over the pose θ using the following loss:

$$E_1(\theta) = \sum_{(i,j) \in E} m_i m_j \| (v_i - v_j) - (\hat{v}_i - \hat{v}_j) \|_2^2 \quad (\text{B.6})$$

where E is the set of 3D edges of the SMPL-X mesh. We use the binary masks m_i, m_j to compute the loss only on valid vertices. For the second stage, we optimize the translation vector t using a vertex-to-vertex loss:

$$E_2(t) = \sum_i m_i \| v_i - \hat{v}_i \|_2^2 \quad (\text{B.7})$$

By this point, we have rigidly aligned the two meshes and matched the articulation of the original SMPL mesh. All that remains is to also match the shape, to get the best possible fit. The final step is to optimize over all parameters (θ, β, ψ, t) using again a vertex-to-vertex loss:

$$E_3(\theta, \beta, \psi, t) = \sum_i m_i \| v_i - \hat{v}_i \|_2^2 \quad (\text{B.8})$$

We use a Trust Region Newton Conjugate Gradient optimizer [258] to search for minimize the objectives. The implementation for the transfer process can be found on our website: <https://expose.is.tue.mpg.de>.

B.4 SMPLIFY-X QUALITATIVE COMPARISON

As shown in Tab. 3.2, ExPose is almost $200\times$ times faster compared to SMPLify-X [270], and provides qualitatively similar results to the latter, as seen in Fig. B.8. Although the accuracy of ExPose is slightly lower than SMPLify-X, it can provide a better initialization to the latter, helping it overcome failures of its initialization heuristic and of the keypoint detector. Potentially, this could be done in a loop, similar to SPIN [187] to continuously improve the performance of ExPose using more in-the-wild data.

B.5 IN-THE-WILD QUALITATIVE RESULTS

A qualitative comparison of our method with the state-of-the-art SMPL regression methods shows the increase in expressivity offered by ExPose;



FIGURE B.8: 1. The input image, 2. OpenPose detections, 3. SMPLify-X fitting, with the neutral model and default focal length, 4. ExPose. When 2D keypoint detections are missing or wrong, optimization based methods, such as SMPLify-X are unable to avoid implausible poses. Furthermore, they heavily depend on their initialization and can produce unnatural poses and shapes, when their initialization heuristic fails. Regression methods, such as ExPose, avoid these problems and can provide better initialization points, closer to the actual solution, and accelerate convergence.

see Figs. B.9 and B.10. Figure B.11 compares the output of the naive regression approach with the body-driven attention mechanism of ExPose. Finally, Figs. B.12 to B.15 contain visualizations of ExPose predictions from multiple views.



FIGURE B.9: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [187] (in yellow), 3. ExPose (in blue). Our proposed method produces 3D body pose and shape results on par with SPIN [187] and captures more details for the hands and face. Best viewed in color.



FIGURE B.10: Comparison of ExPose with the state-of-the-art body regression method: 1. RGB image, 2. SPIN [187] (in yellow), 3. ExPose (in blue). Our proposed method produces 3D body pose and shape results on par with SPIN [187] and captures more details for the hands and face. Best viewed in color.



FIGURE B.11: *Left*: The input image. *Middle*: Naive regression from a body crop. *Right*: ExPose. The attention mechanism helps capture detailed hand articulation and facial expression.



FIGURE B.12: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis



FIGURE B.13: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis



FIGURE B.14: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis



FIGURE B.15: ExPose results visualized from multiple views. 1. RGB image, 2. overlay, 3. , 4. rotations around the vertical axis

COLLABORATIVE REGRESSION OF EXPRESSIVE BODIES USING MODERATION

C.1 IMPLEMENTATION DETAILS

Data augmentation: For training data, we use image crops around the body, face and hands. We augment our training image crops, following mainly [56], as described below. First, we use standard techniques, namely random horizontal flipping, random image rotations, color noise addition and random translation of the crop’s center. However this is not enough, as there is a significant domain gap between face-only and hand-only datasets, and the respective image crops extracted from full-body images; the former have significantly higher resolution. To account for this, we also randomly down-sample and up-sample the head and hand image crops, to simulate various lower resolutions. Finally, inspired by [295], we add synthetic motion blur to face and hand crops, to simulate the motion blur that is common in full-body images. Exact augmentation parameters can be found in our code website.

Training details: We use PyTorch [268] to implement our pipeline. We follow a three-step training procedure: (1) We pre-train the model with body-only, face-only and hand-only datasets; for each dataset we train only the respective parameters. Since these datasets are captured independently, there is no body image that corresponds to a face-only or hand-only image. Consequently, for this step we cannot apply feature fusion, and body-part features go directly to the respective regressor(s) (bypassing the moderators), to estimate the respective body-part parameters. Similar to existing work, we train only a right hand regressor; for images of a left hand, we flip the image horizontally to use the right hand regressor, and mirror the predictions to get a left hand. (2) Then, using the same data, we freeze the feature encoders and proceed with training the regressors and extractors, see Fig. 4.3 for a description of each module. This step encourages features F_b^h and F_b^f from body images to be in the same space as features F_h and F_f from part-only images, so that regressors $\mathcal{R}_f^{\text{fused}}$ and $\mathcal{R}_h^{\text{fused}}$ work for both feature types. (3) Finally, we train the full network, including the moderators \mathcal{M}_h and \mathcal{M}_f , but this time using training images with full

SMPL-X ground truth, to extract part crops from full-body images as well. However, there are two problems. First, for these images there is no skin mask available, consequently we remove the loss for body shape β and do not apply a photometric and identity loss on head crops. Second, localizing the hands with body-driven attention is much harder compared to the head, due to the longer kinematic chain, consequently we freeze the hand regressor $\mathcal{R}_h^{\text{fused}}$ to avoid fine-tuning it with invalid inputs.

All parameters are optimized using Adam [180] with a learning rate of 0.0001. For training the body, hand and face sub-networks, we use a batch size of 16, 16, and 8, respectively. The moderator is a fully connected network with the following structure: FC (2048, 1024), ReLU, FC (1024, 1). All input images are resized to 224×224 pixels before feeding them to our network. During inference, we extract the hand/face crops using the hand and face locations from \mathcal{R}_b 's output. Hand and face cameras are ignored when estimating full body pose.

Global to relative pose: The regressors $\mathcal{R}_f^{\text{fused}}$ and $\mathcal{R}_h^{\text{fused}}$ estimate the absolute head and wrist orientation θ_g , i.e. irrespective of the (parent) main body's pose. However, to "apply" these θ_g estimates on a SMPL-X body that is already posed by \mathcal{R}_b with θ_b (up to the wrist and neck, excluding them), we need to express them relative to their parent in the kinematic skeleton:

$$\theta_{\text{relative}} = \mathbf{\Gamma}(\theta_g, \theta_b), \quad (\text{C.1})$$

where $\mathbf{\Gamma}$ is the chain transformation function according to SMPL-X's kinematic skeleton hierarchy.

C.2 EVALUATION

C.2.1 Body-face correlations discussion

PIXIE gives more realistic body shapes, not only due to its gendered shape loss, but also thanks to the shared body, hand and face shape space of SMPL-X. This allows PIXIE's face expert to – uniquely – contribute to whole-body shape. To verify this, we apply our face expert on face-only images and get the whole-body shapes of Fig. C.1. These are not only correctly "gendered", but also have a plausible BMI. For the sumo wrestler in Fig. C.1, ExPose predicts a body with higher BMI (26.9) than the mean shape (26.1). PIXIE is the only 3D whole-body estimation method that explores such face-body

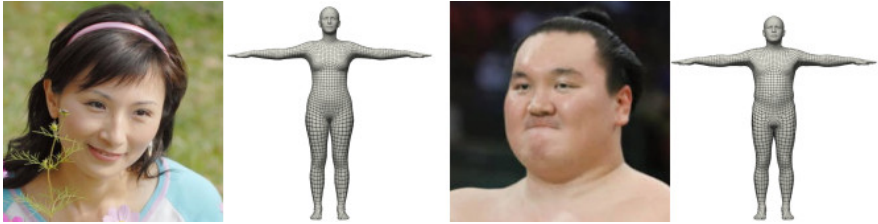


FIGURE C.1: Whole-body shape estimation from *only* our face expert, using SMPL-X’s joint shape space for all body parts.

shape correlations explicitly. We believe that this is a useful insight and points the community towards a new direction.

c.2.2 Qualitative Evaluation

Comparison with MTC: In Fig. C.2 we compare ExPose with MTC [381]. ExPose is two orders of magnitude faster and predicts more accurate 3D body shapes. However, when 2D joint estimations are accurate, optimization-based methods, such as MTC [381] and SMPLify-X, described in Sec. 2.3.2, tend to estimate bodies that are better aligned with the image.

Expressive body reconstruction: We compare our method, ExPose, with other state-of-the-art expressive body reconstruction methods in Fig. C.3. PIXIE is more robust to challenging ambiguities (blur, occlusion) than existing whole-body regressors [56, 295], since its moderators fuse “global” body and “local” part information.

Qualitative results: Finally, in Figs. C.4 to C.6 we provide more standalone ExPose results. Overall, ExPose produces visually plausible body shapes with detailed facial expressions.

Failure cases: Although the gender prior loss and the shared whole-body shape space result in better 3D shape predictions, they are not sufficient for perfectly estimating full-body 3D shape. Furthermore, the employed photometric term often causes the model to prefer to explain image evidence using lighting, rather than albedo, which leads to incorrect skin tone predictions. These points highlight important directions for improving PIXIE. Representative failure cases can be seen in Fig. C.7.

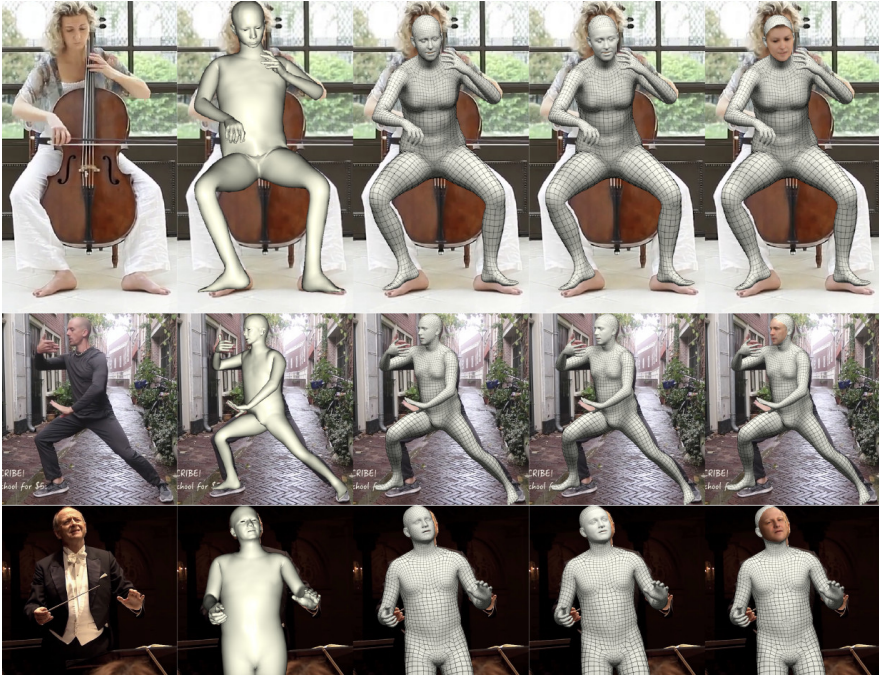


FIGURE C.2: Qualitative ExPose results and comparison to MTC [381]. From left to right: (1) RGB image, (2) MTC [381], (3) ExPose, (4) ExPose with facial geometric details, (5) ExPose with estimated face albedo and lighting. Overall, ExPose produces more visually plausible body shapes and more detailed facial expressions.



FIGURE C.3: Qualitative ExPose results and comparison to ExPose [56] and FrankMocap [295]. From left to right: (1) RGB images from video, (2) FrankMocap [295], (3) ExPose, (4) ExPose 3D body predictions with color-coded part-expert confidence. The moderator predicts the confidence of body/face/hand experts, redder means higher confidence in the body expert rather than the results from face/hand experts. Thanks to the moderators, PIXIE is more robust to low-quality part images. For example, when the hand is blurry, ExPose still predicts a plausible wrist pose, instead of an unnatural twist.

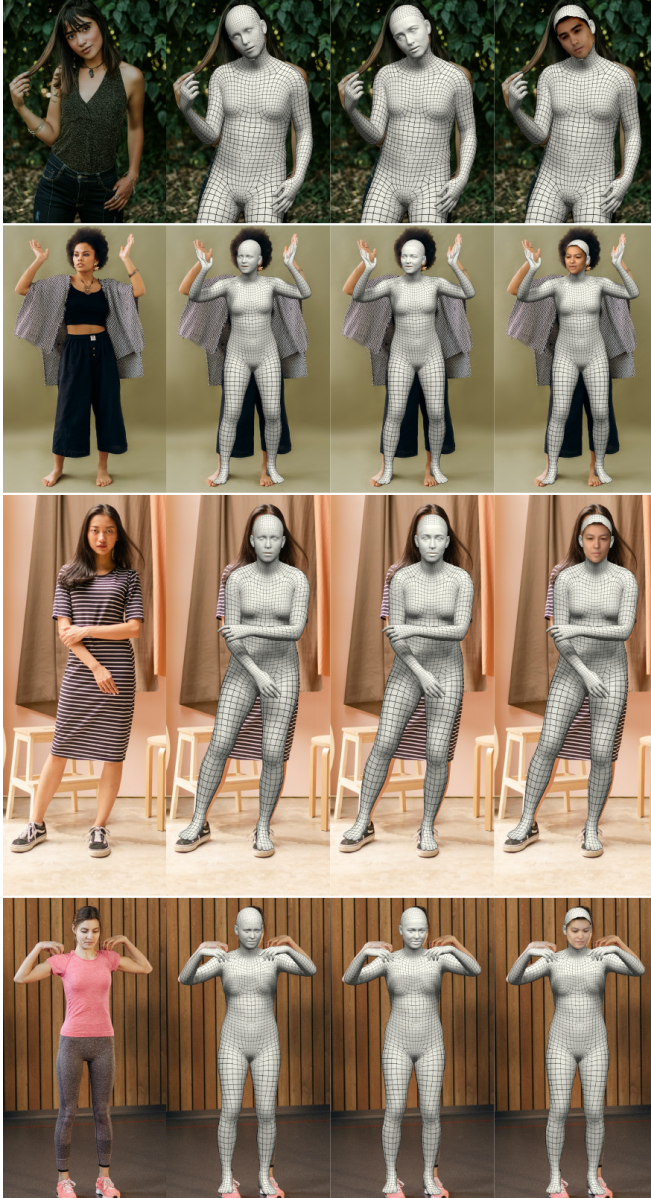


FIGURE C.4: Qualitative ExPose results. From left to right: (1) RGB image, (2) ExPose, (3) ExPose with facial geometric details, (4) ExPose with estimated face albedo and lighting.

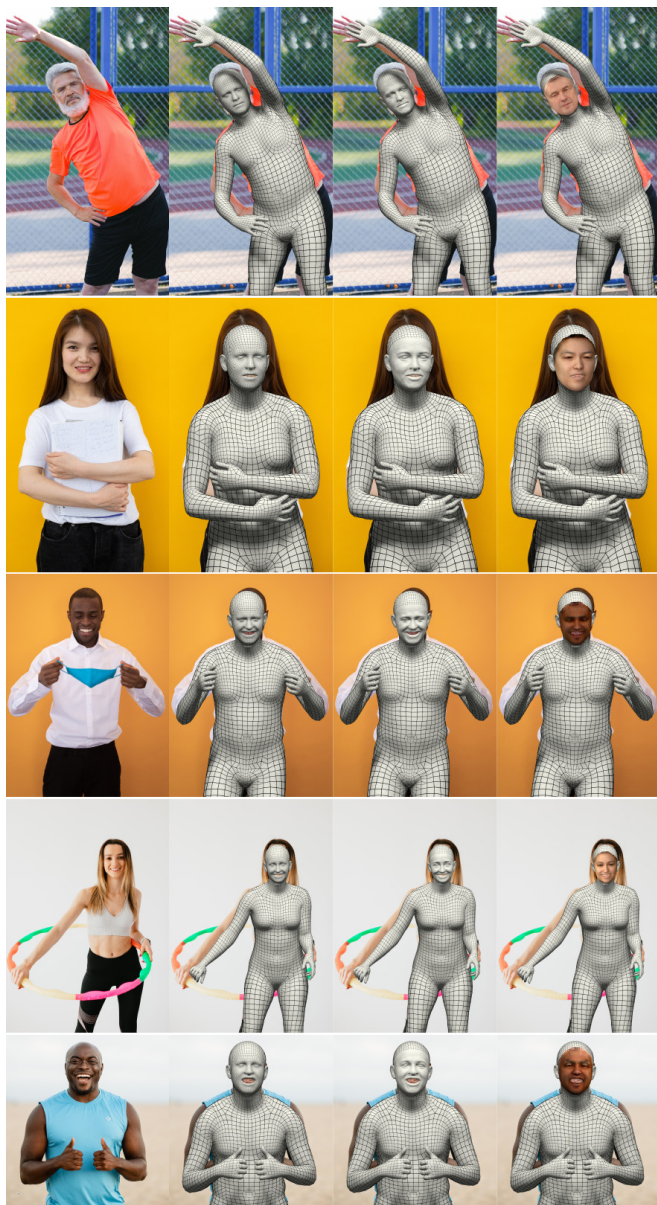


FIGURE C.5: Qualitative ExPose results. From left to right: (1) RGB image, (2) ExPose, (3) ExPose with facial geometric details, (4) ExPose with estimated face albedo and lighting.

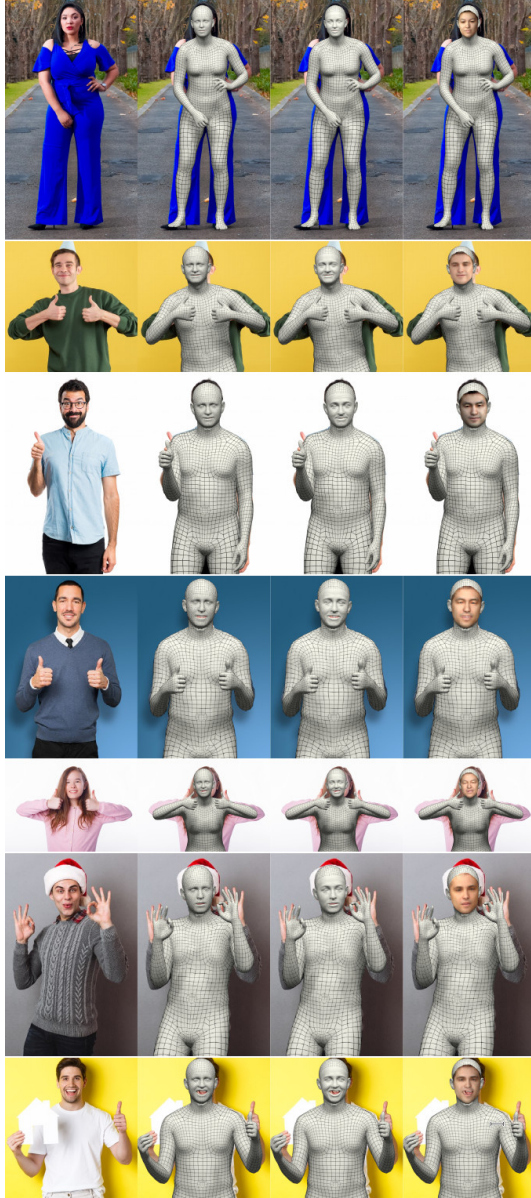


FIGURE C.6: Qualitative ExPose results. From left to right: (1) RGB image, (2) ExPose, (3) ExPose with facial geometric details, (4) ExPose with estimated face albedo and lighting.



FIGURE C.7: Failure cases for ExPose. In these examples, the implicit reasoning about gender and the face shape information are not enough to correctly infer the body shape. Furthermore, due to the formulation of the photometric term the model prefers to explain image evidence using lighting, rather than albedo, which leads to wrong skin tone predictions. Finally, replacing the weak-perspective camera with a perspective model would make the model more robust to extreme viewing angles and perspective distortion effects. Future work should look into denser forms of supervision, formulating a better photometric term and integrating a perspective camera to resolve these issues.

ACCURATE 3D BODY SHAPE REGRESSION USING METRIC AND SEMANTIC ATTRIBUTES

D.1 DATA COLLECTION

D.1.1 *Model-Agency Identity Filtering*

We collect internet data consisting of 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 [71] for face detection and RetinaNet [72] 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:

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

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

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.

D.1.2 *Crowd-sourced Linguistic Shape Attributes*

To collect human ratings of how much a word describes a body shape, we conduct a human intelligence task (hit) on amazon mechanical turk (AMT). In this task, we show an image of a person along with 15 different gender-specific attributes. We then ask participants to indicate how strongly they agree or disagree that the provided words describe the shape of this

person’s body. We arrange the rating buttons from strong disagreement to strong agreement with equal distances to create a 5-point 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).

We ask multiple persons to rate each body and image, to “average out” the subjectivity of individual ratings [329]. 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”.

The layout of our CAESAR annotation task is visualized in Fig. D.1. 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 across raters.

D.2 MAPPING SHAPE REPRESENTATIONS

D.2.1 Shape to Anatomical Measurements (S2M)

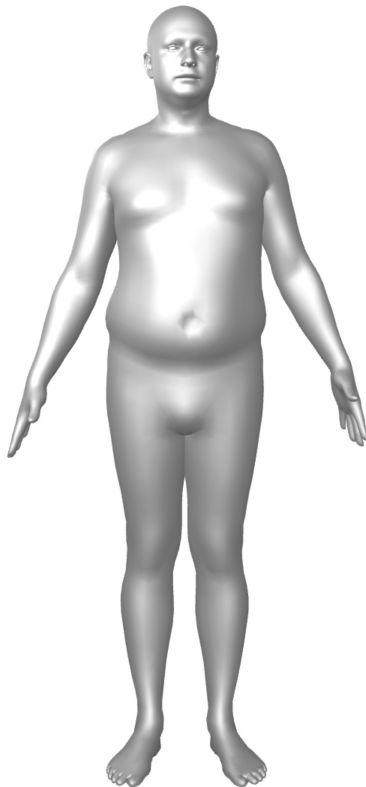
An important part of our project is the computation of body measurements. Following “Virtual Caliper” [281], we present a method to compute anatomical measurements from a 3D mesh in the canonical T-pose, i.e. after “undoing” the effect of pose. Specifically, we measure the height, $H(\beta)$, weight, $W(\beta)$, 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. [380].

Here, $\mathcal{T} \in \mathbb{R}^{\text{T} \times 3 \times 3}$, where $\text{T} = 20,908$ 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, 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 $\mathcal{I} = \{p_i\}_{i=1}^N$ be the set of points of P that intersect the body mesh (red points in Fig. D.4). We store their barycentric coordinates (u_i, v_i, w_i) and the

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 D.1: Layout of the AMT task for a male subject. **Left:** the 3D body mesh in A-pose. **Right:** the attributes and ratings buttons.

corresponding body-triangle index t_i . Let \mathcal{CH} be the convex hull of \mathcal{I} (black lines in Fig. D.4), and \mathcal{HE} the set of edge indices of \mathcal{CH} . $C_c(\beta)$ is equal to the length of the convex hull:

$$C_c(\beta) = \sum_{(i,j) \in \mathcal{HE}} \left\| \begin{pmatrix} \mathbf{u}_i \\ \mathbf{v}_i \\ \mathbf{w}_i \end{pmatrix}^\top \mathcal{T}_{t_i} - \begin{pmatrix} \mathbf{u}_j \\ \mathbf{v}_j \\ \mathbf{w}_j \end{pmatrix}^\top \mathcal{T}_{t_j} \right\|_2, \quad (\text{D.3})$$

where i, j are point indices for line segments of \mathcal{HE} and \mathcal{T}_{t_j} contains the coordinates of the triangle with index t_j . 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.

D.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. D.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). We. With these additions, the 2nd degree polynomial achieves the best performance.

D.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 measure I2A’s accuracy on inferring the correct Likert score. I2A achieves 60.7 / 69.3% (fe-/male) of correctly predicted attributes, while our S2A

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 D.1: Comparison of models for A2S and AHW2S regression.

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 GT shape [203]. “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.

D.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 [56] 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 [268] and we use the ADAM [180] 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.

D.4 EXPERIMENTS

D.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 $N_p = 20k$ mesh surface points; see Fig. D.5 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 $\mathbf{H}_{\text{SMPL-X}} \in \mathbb{R}^{N_p \times V}$ that regresses the mesh surface points from SMPL-X vertices V , as $\mathbf{P} = \mathbf{H}_{\text{SMPL-X}}M$.

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, $\mathbf{H}_{\text{SMPL}} \in \mathbb{R}^{N_p \times 6,890}$. 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 with triangulation \mathcal{T}_1 and \mathcal{T}_2 we obtain the mesh surface points $\mathbf{P}_1 = \mathbf{H}_{\mathcal{T}_1}M_{S,1}$ and $\mathbf{P}_2 = \mathbf{H}_{\mathcal{T}_2}M_{S,2}$, where $M_{S,1}$ and $M_{S,2}$ denote the vertices of each mesh, at rest pose (t-pose), with only the shape blend shapes applied. To compute the P2P_{20K} error we correct for translation $t = \bar{\mathbf{P}}_2 - \bar{\mathbf{P}}_1$, where $\bar{\mathbf{P}}$ is the center point of \mathbf{P} , and define

$$\text{P2P}_{20\text{K}}(M_{S,1}, M_{S,2}) = \|\mathbf{H}_{\mathcal{T}_1}M_{S,1} + t - \mathbf{H}_{\mathcal{T}_2}M_{S,2}\|_2^2. \quad (\text{D.4})$$

D.4.2 Shape Estimation

A2S and its variations: For completeness, Table D.2 shows the results of the female A2S models in addition to the male ones. The male results can also be found in Tab. 5.2. Note that attributes improve shape reconstruction across the board. For example, in terms of P2P_{20K}, AH2S is better than just H2S, AHW2S is better than just HW2S. It should be emphasized that even

	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	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 D.2: Results of A2S and its variations on CMTS test set, in mm or kg. Trained with gender-specific SMPL-X model.

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**.

Method	Mean absolute error (mm) ↓								
	HBW					MMTS			
	Height	Chest	Waist	Hips	P2P _{20K}	Height	Chest	Waist	Hips
SHAPY- H	54	90	77	54	22	52	113	172	108
SHAPY- HA	49	62	71	58	20	60	64	96	77
SHAPY- C	72	65	77	60	26	119	66	70	70
SHAPY- CA	54	69	78	58	22	74	60	82	69
SHAPY- HC	53	61	77	55	23	54	62	72	69
SHAPY- HCA	47	66	75	52	20	57	61	85	73

TABLE D.3: Leave-one-out evaluation on HBW and MMTS.

Attribute/Masurement 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 Sec. D.4.2 (MMTS) and Sec. D.4.2 (HBW). The tables show that attributes are an adequate replacement for measurements. For example, in Sec. D.4.2, 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 Sec. D.4.2. Surprisingly, seeing attributes improves the height error in all three variations. This suggests that training on model images introduces a bias that A2S antagonizes.

D.4.3 Pose evaluation

3D Poses in the Wild (3DPW) [235]: 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. D.4). SHAPY’s pose accuracy is slightly behind ExPose which also uses SMPL-X. SHAPY’s performance is better than HMR [169] and STRAPS [310]. However, SHAPY does not outperform recent pose estimation methods, e.g. HybrIK [203]. 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. [311] who sample poses from 3DPW training set) and (2) creating pseudo ground-truth fits for the model data.

	Model	MPJPE	PA-MPJPE
HMR [169]	SMPL	130	81.3
SPIN [187]	SMPL	96.9	59.2
TUCH [249]	SMPL	84.9	55.5
EFT [163]	SMPL	-	54.2
HybrIK [203]	SMPL	80.0	48.8
STRAPS [310]*	SMPL	-	66.8
Sengupta et al. [312]*	SMPL	-	61.0
Sengupta et al. [311]*	SMPL	84.9	53.6
ExPose	SMPL-X	93.4	60.7
SHAPY (ours)	SMPL-X	95.2	62.6

TABLE D.4: Evaluation on 3DPW [235]. * uses body poses sampled from the 3DPW training set for training.

S2A: Table D.5 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.

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 D.5: S2A evaluation. We report mean, standard deviation and percentage of correctly predicted classes per attribute on CMTS test set.

D.4.4 Qualitative Results

We show additional qualitative results in Figs. D.2 and D.3. Failure cases are shown in Fig. D.6. 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.



FIGURE D.2: Qualitative results of SHAPY predictions for female bodies



FIGURE D.3: Qualitative results of SHAPY predictions for male bodies.

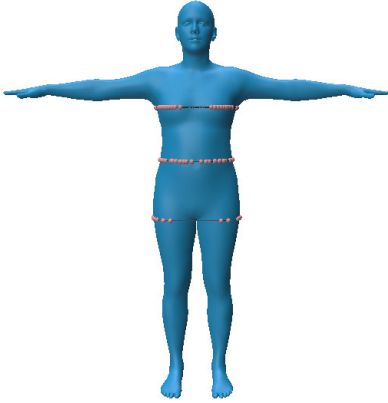


FIGURE D.4: Automatic anthropometric measurements on a 3D mesh. The red points lie on the intersection of planes at chest/waist/hip height with the mesh, while their convex hull is shown with black lines.

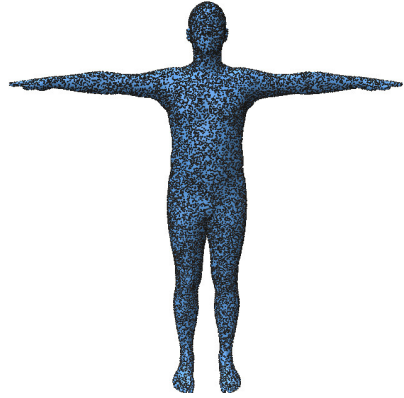


FIGURE D.5: The 20K body mesh surface points (in black) used to evaluate body shape estimation accuracy.



FIGURE D.6: 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).

LEARNING TO FIT MORPHABLE MODELS

E.1 ERRORS PER ITERATION

Figure E.2 shows the metric values per iteration, averaged across the test set, for our fitter on the task of fitting SMPL+H to HMD head and hand signals. Different to Fig. 6.5, this figure corresponds to the full visibility scenario, i.e. the hands are always visible. The learned fitter aggressively optimizes the target data term and quickly converges to the minimum.

E.2 UPDATE RULE

In addition to the update rule described in Eq. (6.1), we investigated two other alternatives, based on the convex combination of the network update and gradient descent. The first is a simple re-formulation of Eq. (6.1), with $\lambda \in [0, 1]$, selecting either the network update or the gradient descent direction. In the second, we first compute a convex combination between the normalized network update and gradient descent, i.e. selecting a direction, and then scale the computed direction according to γ .

$$\begin{aligned}
 u(\Delta\Theta_n, g_n, \Theta_n) &= \lambda\Delta\Theta_n + (1 - \lambda)(-\gamma g_n) \\
 u(\Delta\Theta_n, g_n, \Theta_n) &= \gamma \left[\lambda \left(\frac{\Delta\Theta_n}{\|\Delta\Theta_n\|} \right) + (1 - \lambda) \left(\frac{-g_n}{\|g_n\|} \right) \right] \\
 \lambda &= \sigma \left(f_\lambda(\mathbf{R}(\Theta_n), \mathbf{R}(\Theta_n + \Delta\Theta_n)), \lambda \in \mathbb{R}^{|\Theta|} \right)
 \end{aligned} \tag{E.1}$$

Here, $\sigma(\cdot)$ is the sigmoid function: $\sigma(x) = \frac{1}{1 + \exp(-x)}$. The learning rate of the gradient descent term is the same as Eq. (6.2):

$$\gamma = f_\gamma(\mathbf{R}(\Theta_n), \mathbf{R}(\Theta_n + \Delta\Theta_n)), \gamma \in \mathbb{R}^{|\Theta|} \tag{E.2}$$

We empirically found that the performance of these two variants is inferior to the proposed update rule, but we nevertheless list them for completeness.

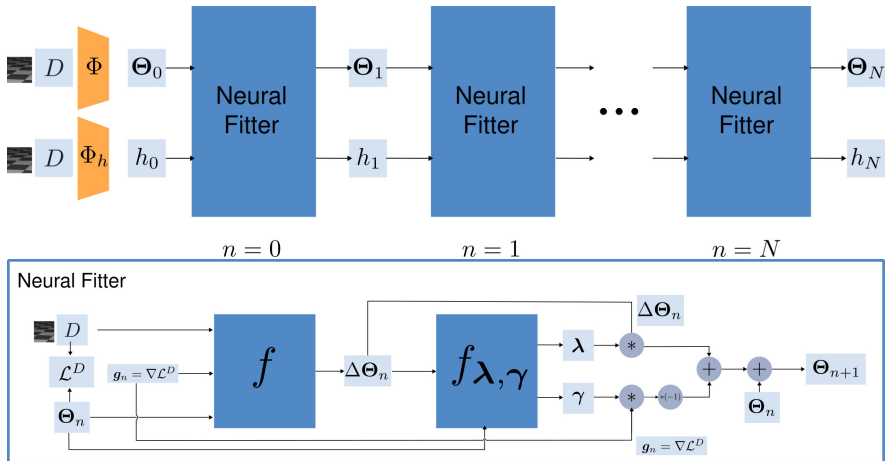


FIGURE E.1: Top: the general fitting process described in Alg. 1. Bottom: A schematic representation of our update rule, described Eqs. (6.1) and (6.2).

E.3 ADDITIONAL ABLATION

Table E.1 contains an additional ablation experiment, where we compare different options for the type of variable for λ , γ , namely whether to use a scalar or a vector variable, and whether to use a common network predictor for λ , γ . We use the problem of fitting SMPL to 2D keypoint predictions, evaluating our results using the 3DPW test set.

E.4 QUALITATIVE COMPARISONS

We present a qualitative comparison of the proposed learned optimizer with a classic optimization-based method in Fig. E.3. Without explicit hand-crafted constraints, the classic approach cannot resolve problems such as ground-floor penetration. Formulating a term to represent this constraint is not a trivial process. Furthermore, tuning the relative weight of this term to avoid under-fitting the data term is not a trivial process. Our proposed method on the other hand can learn to handle these constraints directly from data, without any heuristics.

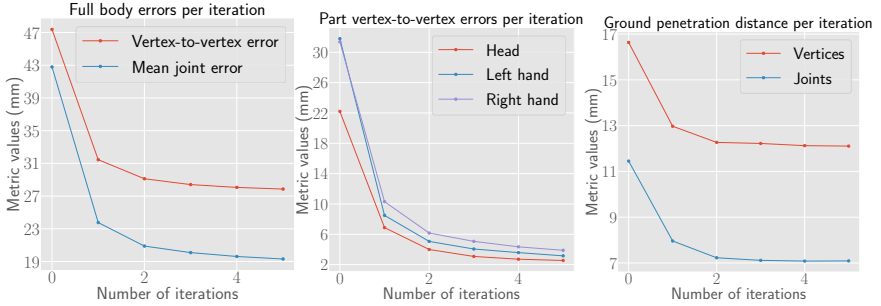


FIGURE E.2: Errors per iteration when fitting SMPL+H to HMD data, assuming that the hands are always visible. From left to right: 1) Full body vertex and joint errors, 2) head, left and right hand V2V errors and 3) vertex and joint ground distance, computed on the set of points below ground.

E.5 TRAINING DETAILS

E.5.1 GRU formulation

All our recurrent networks are implemented with Gated Recurrent Units (GRU) [52], with layer normalization [19]:

$$\begin{aligned}
 z_n &= \sigma_g (LN(W_z x) + LN(U_z h_{n-1})) \\
 r_n &= \sigma_g (LN(W_r x) + LN(U_r h_{n-1})) \\
 \hat{h}_n &= \phi_h (LN(W_h x) + LN(U_h (r_n \odot h_{n-1}))) \\
 h_n &= (1 - z_n) \odot h_{n-1} + z_n \odot \hat{h}_n, \quad h_0 = \Phi_h(D)
 \end{aligned} \tag{E.3}$$

We also tried replacing the GRUs with LSTMs [134], but did not observe significant performance benefit. Hence we chose the computationally lighter GRUs.

E.5.2 Training losses

We apply a loss on the output of every step of our network:

$$\mathcal{L}(\{\Theta_n\}_{n=0}^N, \{\hat{\Theta}_n\}_{n=0}^N; D) = \sum_{i=0}^N \mathcal{L}_i(\Theta_i, \hat{\Theta}_i; D) \tag{E.4}$$

Vector λ	Vector γ	Shared network for λ, γ	PA-MPJPE (mm)
✓	✗	✗	52.8
✗	✓	✗	52.7
✓	✓	✗	52.3
✓	✓	✓	52.2

TABLE E.1: Predicting vector values for λ, γ is always better than scalars. This is expected, since each variable to be optimized has different scale and the learned fitter must adapt its predicted updates accordingly. Having a shared network for λ, γ improves performance and lowers the number of parameters of the learned fitter.

The loss \mathcal{L}_i contains the following terms:

$$\mathcal{L}_i = \lambda_M \mathcal{L}_i^M + \lambda_E \mathcal{L}_i^E + \lambda_T \mathcal{L}_i^T + \lambda_\theta \mathcal{L}_i^\theta \quad (\text{E.5})$$

$$\mathcal{L}_i^M = \|\hat{M} - M\|_1 \quad (\text{E.6})$$

$$\mathcal{L}_i^E = \sum_{(i,j) \in E} \|(\hat{M}_i - \hat{M}_j) - (M_i - M_j)\|_1 \quad (\text{E.7})$$

$$\mathcal{L}_i^T = \sum_{j=1}^J \|\hat{\mathbb{T}}_j - \mathbb{T}_j\|_1 \quad (\text{E.8})$$

$$\mathcal{L}_i^\theta = \|\hat{R}_\theta - R_\theta\|_1 + \|\hat{t} - t\|_1 \quad (\text{E.9})$$

M represents the mesh vertices deformed by parameters Θ . E is the set of vertex indices of the mesh edges. T denotes the transformations in world coordinate while R_θ denotes the rotation matrices (in the parent-relative coordinate frame) computed from the pose values θ . t is the root translation vector. We use the following values for the weights of the training losses: $\lambda_M = 1000$, $\lambda_E = 1000$, $\lambda_T = 100$, $\lambda_\theta = 1$, $\lambda_t = 100$.

E.5.3 Datasets

For body fitting from HMD signals, we use a subset of AMASS [232] to train and test our method. Specifically, we use CMU [60], KIT [233] and MPI_HDM05 [250], adopting the same pre-processing and training, test splits as [75]. An important difference is that we fit the neutral SMPL+H to

the gendered SMPL+H data found in AMASS, to preserve correct contact with the ground and avoid the use of heuristics [285]. We attach random hand poses from the MANO [293] training set to simulate hand articulation. In all our experiments that involve SMPL+H, we use the ground-truth shape parameters β . Future work could include estimating a subset of the shape parameters corresponding to height from the position of the headset. For the learned fitter that estimates body parameters from 2D joints, we use the data, augmentation and evaluation protocol of Song et al. [324]. To be more precise, we use AMASS [232] to train the fitter and evaluate the resulting model on 3DPW [235], which contains sequences of subjects in complex poses in outdoor scenes, along with SMPL parameters captured using RGB cameras and IMUs.

For face fitting from 2D landmarks, we use the face model proposed in [377] to generate a synthetic face dataset by sampling 50000 sets of parameters from the model space. For each sample, we vary pose, identity and expression. We use a perspective camera with focal length (512, 512) and principal point (256, 256) (in pixels) to project the 3D landmarks onto the image for 2D landmarks. Afterwards, we randomly split this by 80/20 into training and testing sets.

E.5.4 *Training schedule*

We implement our model in [268] and train it with a batch size of 512 on 4 GPUs using Adam [180]. We anneal the learning rate by a factor of 0.1 after 400 epochs. We apply dropout with a probability of $p = 0.5$ on the hidden states of the GRUs. We initialize the weights of the output linear layer of Eq. (E.3) with a gain equal to 0.01 [111].

E.5.5 *Edge loss*

We empirically observed that the loss between the 3D edges of the predicted and ground-truth meshes helps training converge faster.

E.5.6 *Runtimes*

We measure time on the 2D keypoint fitting problem on a Quadro P5000 GPU and with a batch size of 512 data points. Our extra networks and update rule add 6 (ms) per iteration to LGD’s [324] runtime. Using a common network for γ and λ reduces this to 4 (ms).

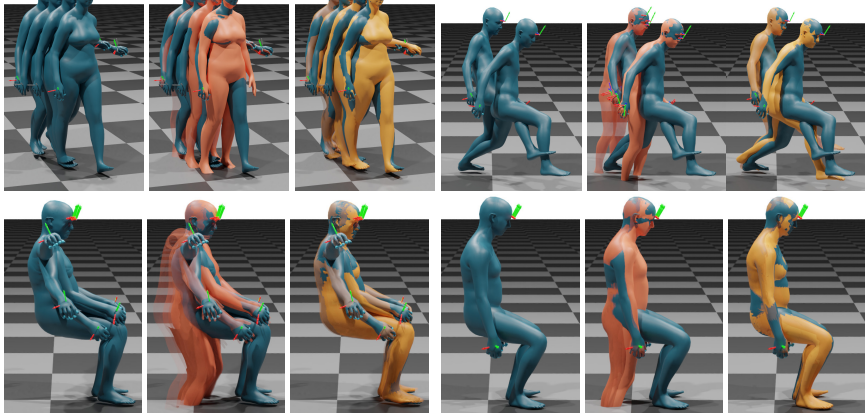


FIGURE E.3: Comparison of our learned fitter with a Levenberg-Marquardt based optimization method. 1) Input HMD data and Ground-Truth mesh (blue), 2) LM solution (orange) overlaid on the GT, 3) our solution (yellow) overlaid on the GT. While the classic LM optimization successfully fits the input data, it still needs hand-crafted priors to prevent ground floor penetration. In contrast, our proposed fitter learns from the data to avoid such penetrations. Best viewed in color.

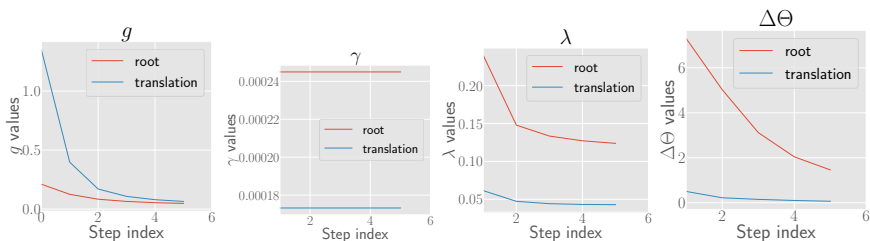


FIGURE E.4: Average norm for (left to right) 1) $\|g_n\|_2^2$, 2) $\|\gamma\|_2^2$, 3) $\|\lambda\|_2^2$ and 4) $\|\Delta\Theta_n\|_2^2$, computed across the test set, for the root rotation and translation. The learned optimizer slows down as it approaches a minimum of the target data term.

E.5.7 Number of iterations

Similar to LGD [324], we observe limited gains beyond 5 iterations. Training with more iterations, e.g. 10 or 20, leads to similar performance, at the cost of increased training time. Picking a random number of iterations during training, e.g. 5 to 20, does not affect the final result.



FIGURE E.5: A visualization of the different body parts used to compute metrics.

BIBLIOGRAPHY

- [1] *3DPeople*. 3dpeople.com (See p. 56).
- [2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. “TensorFlow: A System for Large-Scale Machine Learning.” In: *OSDI*. Vol. 16. 2016, pp. 265–283 (See pp. 130, 131).
- [3] Jonas Adler and Ozan Öktem. “Solving ill-posed inverse problems using iterative deep neural networks”. In: *Inverse Problems* 33.12 (2017), p. 124007 (See p. 91).
- [4] Ankur Agarwal and Bill Triggs. “Recovering 3D human pose from monocular images”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 28.1 (2006), pp. 44–58 (See pp. 30, 49, 73).
- [5] Ijaz Akhter and Michael J. Black. “Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1446–1455 (See pp. 31, 49, 128).
- [6] Oswald Aldrian and William AP Smith. “Inverse Rendering of Faces with a 3D Morphable Model”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.5 (2013), pp. 1080–1093 (See pp. 2, 49).
- [7] Tiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. “Learning to Reconstruct People in Clothing From a Single RGB Camera”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1175–1186 (See p. 2).
- [8] Tiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. “imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 5461–5470 (See p. 2).
- [9] Tiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. “Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 1506–1515 (See p. 2).

- [10] Brett Allen, Brian Curless, and Zoran Popović. “The space of human body shapes: Reconstruction and parameterization from range scans”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)* 22.3 (2003), pp. 587–594 (See pp. 10, 12, 73).
- [11] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. “Learning a Correlated Model of Identity and Pose-dependent Body Shape Variation for Real-time Synthesis”. In: *ACM SIGGRAPH/Eurographics Symposium on Computer Animation. SCA '06*. Eurographics Association, 2006, pp. 147–156 (See pp. 10, 12).
- [12] Brian Amberg, Reinhard Knothe, and Thomas Vetter. “Expression Invariant 3D Face Recognition with a Morphable Model”. In: *International Conference on Automatic Face & Gesture Recognition (FG)*. 2008 (See p. 10).
- [13] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. “2D human pose estimation: New benchmark and state of the art analysis”. In: *CVPR*. 2014, pp. 3686–3693 (See pp. 18, 24, 29, 31, 36, 40, 128, 130, 135).
- [14] Marcin Andrychowicz, Misha Denil, Sergio Gómez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. “Learning to learn by gradient descent by gradient descent”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. 2016, pp. 3988–3996 (See pp. 89, 91).
- [15] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. “SCAPE: Shape Completion and Animation of PEople”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)* 24.3 (2005), pp. 408–416 (See pp. 2, 10, 12, 30, 71, 89).
- [16] *Anti-Racist Graphics Research*. https://s2021.siggraph.org/presentation/?id=div_127&sess=sess271. 2021 (See p. 112).
- [17] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. “TEACH: Temporal Action Composition for 3D Humans”. In: *International Conference on 3D Vision (3DV)*. 2022 (See p. 112).
- [18] *AXYZ*. secure.axyz-design.com (See p. 56).
- [19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016) (See p. 175).

- [20] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. “Pushing the Envelope for RGB-Based Dense 3D Hand Pose Estimation via Neural Rendering”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1067–1076 (See pp. 2, 31, 49, 89).
- [21] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. “Driving-Signal Aware Full-Body Avatars”. In: *Transactions on Graphics (TOG)* 40.4 (2021) (See p. 1).
- [22] Alexandru Balan and Michael J. Black. “The naked truth: Estimating body shape under clothing”. In: *European Conference on Computer Vision (ECCV)*. 2008, pp. 15–29 (See p. 72).
- [23] 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)*. 2007, pp. 1–8 (See pp. 2, 72).
- [24] Luca Ballan and Guido Maria Cortelazzo. “Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes”. In: *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*. 2008 (See p. 13).
- [25] Luca Ballan, Aparna Taneja, Juergen Gall, Luc Van Gool, and Marc Pollefeys. “Motion Capture of Hands in Action using Discriminative Salient Points”. In: *European Conference on Computer Vision (ECCV)*. 2012, pp. 640–653 (See pp. 12, 18, 119, 121).
- [26] Jonathan T. Barron. “A General and Adaptive Robust Loss Function”. In: *Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4326–4334 (See p. 98).
- [27] Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. “Fitting a 3D Morphable Model to Edges: A Comparison Between Hard and Soft Correspondences”. In: *Asian Conference on Computer Vision Workshops (ACCVw)*. 2017, pp. 377–391 (See p. 49).
- [28] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. “High-Quality Passive Facial Performance Capture Using Anchor Frames”. In: *Transactions on Graphics (TOG)* 30.4 (2011) (See p. 1).
- [29] Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus Gross. “High-Quality Capture of Eyes”. In: *Transactions on Graphics (TOG)* 33.6 (2014) (See p. 112).

- [30] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. "BEHAVE: Dataset and Method for Tracking Human Object Interactions". In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15935–15946 (See p. 114).
- [31] 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)*. 2019, pp. 8568–8576 (See p. 74).
- [32] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. "3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data". In: *Conference on Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 20496–20507 (See p. 107).
- [33] Volker Blanz and Thomas Vetter. "A morphable model for the synthesis of 3D faces". In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)*. 1999, pp. 187–194 (See pp. 2, 10, 11, 30, 49, 91).
- [34] 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)*. 2016, pp. 561–578 (See pp. 2, 9, 13, 15, 16, 19, 21, 30, 31, 46, 48, 49, 58, 68, 72, 89, 92, 100, 102, 103, 125, 127, 136).
- [35] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. "Large scale 3D morphable models". In: *International Journal of Computer Vision (IJCV)* 126.2-4 (2018), pp. 233–254 (See p. 10).
- [36] Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. "3D Hand Shape and Pose From Images in the Wild". In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10843–10852 (See pp. 2, 31, 46, 49, 89).
- [37] Christoph Bregler, Jitendra Malik, and Katherine Pullen. "Twist Based Acquisition and Tracking of Animal and Human Kinematics". In: *International Journal of Computer Vision (IJCV)* 56.3 (2004), pp. 179–194 (See p. 14).

- [38] Alan Brunton, Augusto Salazar, Timo Bolkart, and Stefanie Wuhrer. “Review of statistical shape spaces for 3D data with comparative analysis for human faces”. In: *Computer Vision and Image Understanding (CVIU)* 128.0 (2014), pp. 1–17 (See pp. 10, 11).
- [39] Adrian Bulat and Georgios Tzimiropoulos. “How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 1021–1030 (See pp. 2, 30, 36, 48, 54, 56).
- [40] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. “HuMMan: Multi-Modal 4D Human Dataset for Versatile Sensing and Modeling”. In: (2022), pp. 557–577 (See p. 13).
- [41] Chen Cao, Yanlin Weng, Shun Zhou, Yiyong Tong, and Kun Zhou. “FaceWarehouse: a 3D Facial Expression Database for Visual Computing”. In: *Transactions on Visualization and Computer Graphics (TVCG)* 20.3 (2014), pp. 413–425 (See pp. 10, 12).
- [42] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. “VGGFace2: A dataset for recognising faces across pose and age”. In: *International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 67–74 (See pp. 49, 55, 56).
- [43] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43.1 (2019), pp. 172–186 (See pp. 9, 28, 30, 32, 40, 48, 50, 72, 96, 99, 120, 125, 126).
- [44] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. “Realtime multi-person 2D pose estimation using part affinity fields”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1302–1310 (See pp. 9, 10, 16).
- [45] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. “End-to-end object detection with transformers”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 213–229 (See p. 2).

- [46] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. “Attention-Driven Cropping for Very High Resolution Facial Landmark Detection”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5861–5870 (See p. 32).
- [47] He Chen, Hyojoon Park, Kutay Macit, and Ladislav Kavan. “Capturing Detailed Deformations of Moving Human Bodies”. In: *Transactions on Graphics (TOG)* 40.4 (2021) (See p. 113).
- [48] Hsiao-yu Chen, Edith Tretschk, Tuur Stuyck, Petr Kadlecek, Ladislav Kavan, Etienne Vouga, and Christoph Lassner. “Virtual Elastic Objects”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15827–15837 (See p. 114).
- [49] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. “gDNA: Towards Generative Detailed Neural Avatars”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20427–20437 (See p. 2).
- [50] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. “SNARF: Differentiable Forward Skinning for Animating Non-Rigid Neural Implicit Shapes”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11574–11584 (See p. 2).
- [51] Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. “Tensor-based human body modeling”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 105–112 (See p. 12).
- [52] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1724–1734 (See pp. 94, 175).
- [53] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. “Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1964–1973 (See pp. 91, 113).
- [54] François Chollet et al. *Keras*. <https://keras.io>. 2015 (See p. 131).
- [55] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. “Accurate 3D Body Shape Regression using Metric and Semantic Attributes”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 2718–2728 (See p. 231).

- [56] 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)*. Vol. LNCS 12355. 2020, pp. 20–40 (See pp. 47, 49, 50, 53, 56, 59, 61, 72, 79, 80, 91, 150, 152, 154, 163).
- [57] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. “D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20577–20586 (See pp. 113, 114).
- [58] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. “Multi-context attention for human pose estimation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5669–5678 (See p. 31).
- [59] Ronald Clark, Michael Bloesch, Jan Czarnowski, Stefan Leutenegger, and Andrew J. Davison. “Learning to Solve Nonlinear Least Squares for Monocular Stereo”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 291–306 (See pp. 89, 91, 92).
- [60] CMU. *CMU MoCap dataset* (See pp. 128, 176).
- [61] Matthew Cong, Michael Bao, Jane L. E, Kiran S. Bhat, and Ronald Fedkiw. “Fully Automatic Generation of Anatomical Face Simulation Models”. In: *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. SCA '15. Los Angeles, California, 2015, pp. 175–183 (See p. 113).
- [62] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. “LISA: Learning Implicit Shape and Appearance of Hands”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20533–20543 (See p. 112).
- [63] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. “SMPLicit: Topology-aware Generative Model for Clothed People”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11870–11880 (See pp. 63, 113).
- [64] Antonio Criminisi, Ian Reid, and Andrew Zisserman. “Single view metrology”. In: *International Journal of Computer Vision (IJCV)* 40.2 (2000), pp. 123–148 (See p. 73).

- [65] Radek Daněček, Michael J. Black, and Timo Bolkart. “EMOCA: Emotion Driven Monocular Face Capture and Animation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20311–20322 (See p. 2).
- [66] Total Capture Dataset. <http://domedb.perception.cs.cmu.edu> (See pp. 21, 124, 126).
- [67] Martin De La Gorce, David J. Fleet, and Nikos Paragios. “Model-Based 3D Hand Pose Estimation from Monocular Video”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33.9 (2011), pp. 1793–1805 (See p. 12).
- [68] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. “Acquiring the Reflectance Field of a Human Face”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)*. SIGGRAPH ’00. 2000, pp. 145–156 (See p. 1).
- [69] Javier Dehesa, Andrew Vidler, Julian Padget, and Christof Lutteroth. “Grid-Functioned Neural Networks”. In: *International Conference on Machine Learning (ICML)*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2559–2567 (See p. 106).
- [70] Quentin Delamarre and Olivier D. Faugeras. “3D Articulated Models and Multiview Tracking with Physical Forces”. In: *Computer Vision and Image Understanding (CVIU)* 81.3 (2001), pp. 328–357 (See p. 13).
- [71] 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)*. 2019, pp. 4690–4699 (See p. 159).
- [72] 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)*. 2020, pp. 5202–5211 (See p. 159).
- [73] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. “Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set”. In: *Computer Vision and Pattern Recognition Workshops (CVPRw)*. 2019, pp. 285–295 (See pp. 49, 54, 55, 62).

- [74] 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)*. 2014, pp. 227–238 (See p. 74).
- [75] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lundell, Thomas J. Cashman, and Jamie Shotton. “Full-Body Motion From a Single Head-Mounted Device: Generating SMPL Poses From Partial Observations”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11687–11697 (See pp. 5, 93, 98, 102, 103, 176).
- [76] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. “Shape-aware Multi-Person Pose Estimation from Multi-View Images”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11158–11168 (See pp. 1, 92).
- [77] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization.” In: *Journal of machine learning research* 12.7 (2011) (See p. 93).
- [78] 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)*. 2021, pp. 11250–11259 (See p. 73).
- [79] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. “3D Morphable Face Models—Past, Present and Future”. In: *Transactions on Graphics (TOG)* 39.5 (2020), pp. 1–38 (See pp. 2, 30, 46, 49, 89, 91, 97, 114).
- [80] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978 (See p. 11).
- [81] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. “Vision-based hand pose estimation: A review”. In: *Computer Vision and Image Understanding (CVIU)* 108.1-2 (2007), pp. 52–73 (See p. 28).
- [82] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. “Revitalizing Optimization for 3D Human Pose and Shape Estimation: A Sparse Constrained Formulation”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11457–11466 (See pp. 92, 100).

- [83] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael Black, and Otmar Hilliges. “Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-pixel Part Segmentation”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 1–10 (See p. 2).
- [84] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. “Articulated Objects in Free-form Hand Interaction”. In: *arXiv preprint arXiv: Arxiv-2204.13662* (2022) (See p. 114).
- [85] Haiwen Feng, Timo Bolkart, Joachim Tesch, Michael J. Black, and Victoria Abrevaya. “Towards Racially Unbiased Skin Tone Estimation via Scene Disambiguation”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 72–90 (See p. 112).
- [86] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. “Collaborative Regression of Expressive Bodies using Moderation”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 792–804 (See pp. 80, 91, 231).
- [87] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. “Learning an Animatable Detailed 3D Face Model from In-the-Wild Images”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)* 40.4 (2021), 88:1–88:13 (See pp. 2, 47–49, 51–54, 61, 62).
- [88] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. “Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 557–574 (See p. 2).
- [89] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. “Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 557–574 (See pp. 49, 62).
- [90] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Räscht. “Evaluation of dense 3D reconstruction from 2D face images in the wild”. In: *International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 780–786 (See pp. 38, 40–42).
- [91] 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)*. 2020, pp. 7212–7221 (See pp. 42, 49).

- [92] 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 (See p. 63).
- [93] Martin A. Fischler and Robert C. Bolles. “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”. In: *Communications of the ACM (CACM)* 24.6 (1981), pp. 381–395 (See p. 50).
- [94] models FLAME website: dataset and code. <http://flame.is.tue.mpg.de> (See p. 12).
- [95] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. “Deepview: View synthesis with learned gradient descent”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 2367–2376 (See p. 91).
- [96] Maria-Paola Forte, Chun-Hao P. Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. “SGNify: Full-body 3D Reconstruction of Sign Language Signs from RGB Videos”. In: (2022) (See pp. 112, 114).
- [97] Oren Freifeld and Michael J. Black. “Lie Bodies: A Manifold Representation of 3D Human Shape”. In: *European Conference on Computer Vision (ECCV)*. 2012, pp. 1–14 (See p. 12).
- [98] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. “Moulding Humans: Non-Parametric 3D Human Shape Estimation From Single Images”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2232–2241 (See pp. 31, 49).
- [99] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. “Motion capture using joint skeleton tracking and surface estimation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2009, pp. 1746–1753 (See p. 13).
- [100] Alessio Gallucci, Dmitry Znamenskiy, and Milan Petkovic. “Prediction of 3D Body Parts from Face Shape and Anthropometric Measurements”. In: *Journal of Image and Graphics* 8.3 (2020) (See p. 47).

- [101] *GAMMA in the exhibition Motion, Autos, Art, Architecture at the Guggenheim Museum Bilbao curated by Norman Foster Foundation*. <https://vlg.inf.ethz.ch/news/Our-model-GAMMA-gained-life-in-the.html> (See p. 114).
- [102] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. “Reconstructing Detailed Dynamic Face Geometry from Monocular Video”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH Asia)*. Vol. 32. 6. 2013, 158:1–158:10 (See p. 2).
- [103] Darius M. Gavrilă. “The Visual Analysis of Human Movement: A Survey”. In: *Computer Vision and Image Understanding (CVIU)* 73.1 (1999), pp. 82–98 (See p. 28).
- [104] Lihao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. “3D Hand Shape and Pose Estimation From a Single RGB Image”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10825–10834 (See p. 31).
- [105] Lihao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. “3D Hand Shape and Pose Estimation From a Single RGB Image”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10833–10842 (See pp. 46, 49).
- [106] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. “GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1155–1164 (See p. 55).
- [107] Stuart Geman and Donald E. McClure. “Statistical methods for tomographic image reconstruction”. In: *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*. Vol. 52. 1987 (See p. 16).
- [108] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. “Unsupervised Training for 3D Morphable Model Regression”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8377–8386 (See p. 49).
- [109] 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)*. 2020, pp. 768–784 (See pp. 2, 72).

- [110] Yuliya Gitlina, Daljit Singh Dhillon, Giuseppe Claudio Guarnera, and Abhijeet Ghosh. “Practical Measurement and Modeling of Spectral Skin Reflectance”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)*. Los Angeles, California: Association for Computing Machinery, 2019 (See p. 112).
- [111] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings*. 2010, pp. 249–256 (See p. 177).
- [112] 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)*. 2019, pp. 7450–7459 (See p. 73).
- [113] Kristen Grauman, Gregory Shakhnarovich, and Trevor Darrell. “Inferring 3D structure with a statistical image-based shape model”. In: *International Conference on Computer Vision (ICCV)*. 2003, pp. 641–647 (See pp. 30, 49).
- [114] 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)*. 2009, pp. 1381–1388 (See pp. 30, 49, 72).
- [115] Riza Alp Guler and Iasonas Kokkinos. “HoloPose: Holistic 3D Human Reconstruction In-The-Wild”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10876–10886 (See pp. 2, 32, 63).
- [116] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. “DensePose: Dense Human Pose Estimation in the Wild”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7297–7306 (See p. 30).
- [117] 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)*. 2019, pp. 1819–1827 (See pp. 47, 73).
- [118] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. “Towards Fast, Accurate and Stable 3D Dense Face Alignment”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 152–168 (See p. 62).

- [119] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. “Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-Localization in Large Scenes From Body-Mounted Sensors”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 4318–4329 (See p. 92).
- [120] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. “HONotate: A Method for 3D Annotation of Hand and Object Poses”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 3196–3206 (See pp. 31, 49, 114).
- [121] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. “A statistical model of human pose and body shape”. In: *Computer Graphics Forum* 28.2 (2009), pp. 337–346 (See pp. 10, 12).
- [122] Nils Hasler, Thorsten Thormählen, Bodo Rosenhahn, and Hans-Peter Seidel. “Learning Skeletons for Shape and Pose”. In: *Symposium on Interactive 3D Graphics (SI3D)*. I3D ’10. New York, NY, USA: ACM, 2010, pp. 23–30 (See p. 12).
- [123] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. “Stochastic Scene-Aware Motion Prediction”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11374–11384 (See pp. 1, 114).
- [124] 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)*. 2019, pp. 2282–2292 (See pp. 42, 91, 114).
- [125] 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)*. 2021, pp. 14708–14718 (See p. 114).
- [126] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. “Learning Joint Reconstruction of Hands and Manipulated Objects”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 11807–11816 (See pp. 2, 31, 46, 49, 89).
- [127] *HDRI Haven*. hdrihaven.com (See p. 56).

- [128] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. “Mask R-CNN”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 42.2 (2020), pp. 386–397 (See pp. 2, 41, 73).
- [129] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 1026–1034 (See p. 2).
- [130] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778 (See pp. 18, 36, 51, 104, 131, 137).
- [131] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. “Single-Network Whole-Body Pose Estimation”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 6981–6990 (See p. 28).
- [132] Matthew Hill, Stephan Streuber, Carina Hahn, Michael Black, and Alice O’Toole. “Exploring the relationship between body shapes and descriptions by linking similarity spaces”. In: *Journal of Vision (JOV)* 15.12 (2015), pp. 931–931 (See p. 76).
- [133] David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. “Coregistration: Simultaneous alignment and modeling of articulated 3D shape”. In: *European Conference on Computer Vision (ECCV)*. 2012, pp. 242–255 (See p. 12).
- [134] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780 (See p. 175).
- [135] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. “Learning to Train with Synthetic Humans”. In: *German Conference on Pattern Recognition (GCPR)*. 2019, pp. 609–623 (See p. 72).
- [136] Daniel Holden, Taku Komura, and Jun Saito. “Phase-Functioned Neural Networks for Character Control”. In: *Transactions on Graphics (TOG)* 36.4 (2017) (See p. 106).
- [137] Wei-Lin Hsiao and Kristen Grauman. “ViBE: Dressing for diverse body shapes”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11056–11066 (See p. 73).
- [138] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. “Unconstrained realtime facial performance capture”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1675–1683 (See p. 2).

- [139] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. “Capturing and Inferring Dense Full-Body Human-Scene Contact”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 13274–13285 (See pp. 1, 13).
- [140] 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)*. 2017, pp. 421–430 (See pp. 30, 31, 49, 72, 73).
- [141] 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)*. 2017, pp. 421–430 (See p. 13).
- [142] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. “Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH Asia)* 37 (2018), 185:1–185:15 (See p. 2).
- [143] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. “InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction”. In: *German Conference on Pattern Recognition (GCPR)*. Springer. 2022 (See p. 114).
- [144] *HumanAlloy*. humanalloy.com (See p. 56).
- [145] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. “Phace: Physics-Based Face Modeling and Animation”. In: *Transactions on Graphics (TOG)* 36.4 (2017) (See p. 113).
- [146] Christian Igel, Marc Toussaint, and Wan Weishui. “Rprop Using the Natural Gradient”. In: *Trends and Applications in Constructive Approximation*. Ed. by Detlef H. Mache, József Szabados, and Marcel G. de Bruin. Basel: Birkhäuser Basel, 2005, pp. 259–272 (See pp. 92, 93).
- [147] *INRIA Kinovis platform*. <https://kinovis.inria.fr/> (See p. 1).

- [148] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. “DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 34–50 (See p. 9).
- [149] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International Conference on Learning Representations (ICLR)*. PMLR. 2015, pp. 448–456 (See p. 99).
- [150] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. “Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36.7 (2014), pp. 1325–1339 (See pp. 1, 31, 36, 72, 79, 80, 127, 128).
- [151] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. “Hand Pose Estimation via Latent 2.5D Heatmap Regression”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 125–143 (See pp. 2, 31, 49).
- [152] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. “KAMA: 3D Keypoint Aware Body Mesh Articulation”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 689–699 (See pp. 2, 114).
- [153] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. “Large pose 3D face reconstruction from a single image via direct volumetric CNN regression”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 1031–1039 (See p. 49).
- [154] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. “Spatial Transformer Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2015, pp. 2017–2025 (See p. 33).
- [155] 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)*. 2021, pp. 12753–12762 (See p. 71).
- [156] 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)*. 2020, pp. 5579–5588 (See pp. 42, 72, 113).

- [157] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. “Ditto: Building Digital Twins of Articulated Objects from Interaction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5616–5626 (See p. 114).
- [158] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. “Whole-Body Human Pose Estimation in the Wild”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 196–214 (See pp. 50, 56).
- [159] Gunnar Johansson. “Visual perception of biological motion and a model for its analysis”. In: *Perception & psychophysics* 14.2 (1973), pp. 201–211 (See p. 2).
- [160] Sam Johnson and Mark Everingham. “Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation”. In: *British Machine Vision Conference (BMVC)*. 2010, pp. 12.1–12.11 (See pp. 10, 18, 24, 29, 31, 36, 40, 58, 130).
- [161] Sam Johnson and Mark Everingham. “Learning effective human pose estimation from inaccurate annotation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2011, pp. 1465–1472 (See pp. 18, 24, 29, 31, 36, 40, 130).
- [162] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. “Panoptic Studio: A Massively Multiview System for Social Motion Capture”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 3334–3342 (See p. 13).
- [163] 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)*. 2021, pp. 42–52 (See pp. 2, 36, 48, 49, 59, 91, 167).
- [164] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. “Towards Social Artificial Intelligence: Nonverbal Social Signal Prediction in a Triadic Interaction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10865–10875 (See p. 113).
- [165] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. “Panoptic Studio: A Massively Multiview System for Social Interaction Capture”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 41.1 (2019), pp. 190–204 (See p. 1).

- [166] 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)*. 2018, pp. 8320–8329 (See pp. [2](#), [10](#), [12](#), [13](#), [20–23](#), [28](#), [30](#), [32](#), [46](#), [47](#), [50](#), [68](#), [71](#), [89](#), [91](#), [112](#), [124](#)).
- [167] Petr Kadleček and Ladislav Kavan. “Building Accurate Physics-Based Face Models from Data”. In: *Proceedings of the ACM on Computer Graphics and Interactive Techniques 2.2* (2019) (See p. [113](#)).
- [168] Christos Kampouris and Abhijeet Ghosh. “ICL Multispectral Light Stage: Building a Versatile LED Sphere with Off-the-shelf Components”. In: *Workshop on Material Appearance Modeling*. Ed. by Reinhard Klein and Holly Rushmeier. The Eurographics Association, 2018 (See pp. [1](#), [112](#)).
- [169] 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)*. 2018, pp. 7122–7131 (See pp. [2](#), [9](#), [10](#), [13](#), [28–31](#), [33](#), [36](#), [38–41](#), [46](#), [48](#), [49](#), [59](#), [61](#), [72](#), [86](#), [91](#), [114](#), [137](#), [138](#), [167](#)).
- [170] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. “Learning 3D Human Dynamics From Video”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5607–5616 (See pp. [31](#), [42](#), [49](#), [68](#)).
- [171] 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)*. 2014, pp. 1725–1732 (See p. [80](#)).
- [172] Tero Karras. “Maximizing Parallelism in the Construction of BVHs, Octrees, and K-d Trees”. In: *Proceedings of the Fourth ACM SIGGRAPH / Eurographics Conference on High-Performance Graphics*. 2012, pp. 33–37 (See pp. [18](#), [122](#)).
- [173] Tero Karras, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4396–4405 (See pp. [29](#), [36](#), [137](#)).

- [174] Manuel Kaufmann, Yi Zhao, Chengcheng Tang, Lingling Tao, Christopher Twigg, Jie Song, Robert Wang, and Otmar Hilliges. “EM-POSE: 3D Human Pose Estimation From Sparse Electromagnetic Trackers”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11510–11520 (See p. 92).
- [175] Marilyn Keller, Silvia Zuffi, Michael J. Black, and Sergi Pujades. “OSSO: Obtaining Skeletal Shape from Outside”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20492–20501 (See p. 113).
- [176] Sameh Khamis, Jonathan Taylor, Jamie Shotton, Cem Keskin, Shahram Izadi, and Andrew Fitzgibbon. “Learning an Efficient Model of Hand Shape Variation From Depth Images”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 2540–2548 (See pp. 10, 12, 30).
- [177] Theodore Kim, Holly Rushmeier, Julie Dorsey, Derek Nowrouzezahrai, Raqi Syed, Wojciech Jarosz, and AM Darke. “Countering Racial Bias in Computer Graphics Research”. In: *arXiv preprint arXiv:2103.15163* (2021) (See p. 112).
- [178] Davis E. King. “Dlib-ml: A Machine Learning Toolkit”. In: *Journal of Machine Learning Research (JMLR)* 10 (2009), pp. 1755–1758 (See p. 65).
- [179] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *ICLR*. 2014 (See pp. 12, 17).
- [180] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *International Conference on Learning Representations (ICLR)*. 2015 (See pp. 37, 91–93, 130, 131, 137, 151, 163, 177).
- [181] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. “Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction”. In: *Transactions on Graphics (TOG)* 36.4 (2017), pp. 1–13 (See p. 39).
- [182] 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)*. 2020, pp. 5252–5262 (See pp. 42, 68, 72, 91, 94, 100, 113).

- [183] 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)*. 2021, pp. 11127–11137 (See pp. 2, 48, 49, 59, 60, 63, 68, 89, 113, 114).
- [184] 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)*. 2021, pp. 11035–11045 (See p. 68).
- [185] Enes Kocabey, Mustafa Camurcu, Ferda Ofli, Yusuf Aytar, Javier Marín, Antonio Torralba, and Ingmar Weber. "Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media". In: *International Conference on Web and Social Media (ICWSM)*. 2017, pp. 572–575 (See p. 47).
- [186] Filippos Kokkinos and Iasonas Kokkinos. "To The Point: Correspondence-driven monocular 3D category reconstruction". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021, pp. 7760–7772 (See p. 91).
- [187] 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)*. 2019, pp. 2252–2261 (See pp. 28, 29, 31, 36, 38–42, 46, 49, 57, 59, 61, 68, 72, 79, 80, 85, 86, 89, 91, 92, 94, 100, 114, 140, 141, 143, 144, 167).
- [188] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. "Convolutional Mesh Regression for Single-Image Human Shape Reconstruction". In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4496–4505 (See pp. 30, 31, 48, 49).
- [189] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. "Probabilistic Modeling for Human Mesh Recovery". In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11605–11614 (See pp. 91, 107, 113).
- [190] Agelos Kratimenos, Georgios Pavlakos, and Petros Maragos. "Independent Sign Language Recognition with 3d Body, Hands, and Face Reconstruction". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 4270–4274 (See p. 114).

- [191] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2012, pp. 1097–1105 (See p. 29).
- [192] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Vol. 25. 2012 (See p. 2).
- [193] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. “Weakly-Supervised Mesh-Convolutional Hand Reconstruction in the Wild”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 4990–5000 (See pp. 2, 31, 49).
- [194] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. “Single Image 3D Hand Reconstruction with Mesh Convolutions”. In: *British Machine Vision Conference (BMVC)*. 2019 (See pp. 31, 49).
- [195] 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)*. 2017, pp. 4704–4713 (See pp. 13, 58, 72).
- [196] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. “AvatarMe: Realistically Renderable 3D Facial Reconstruction “In-the-Wild””. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 760–769 (See p. 112).
- [197] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. “AvatarMe++: Facial Shape and BRDF Inference with Photorealistic Rendering-Aware GANs”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2021), pp. 1–1 (See p. 112).
- [198] Hsi-Jian Lee and Zen Chen. “Determination of 3D human body postures from a single view”. In: *Computer Vision, Graphics, and Image Processing (CGIP)* 30.2 (1985), pp. 148–168 (See pp. 31, 49).
- [199] Kenneth Levenberg. “A method for the solution of certain non-linear problems in least squares”. In: *Quarterly of applied mathematics* 2.2 (1944), pp. 164–168 (See pp. 5, 89, 92, 93).

- [200] John P. Lewis, Matt Cordner, and Nickson Fong. “Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-driven Deformation”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)*. 2000, pp. 165–172 (See pp. 14, 15).
- [201] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C. Buehler, Otmar Hilliges, and Thabo Beeler. “EyeNeRF: A Hybrid Representation for Photorealistic Synthesis, Animation and Relighting of Human Eyes”. In: *Transactions on Graphics (TOG)* 41.4 (2022) (See p. 112).
- [202] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. “CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10855–10864 (See p. 113).
- [203] 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)*. 2021, pp. 3383–3393 (See pp. 91, 113, 114, 163, 167).
- [204] Kun Li, Yali Mao, Yunke Liu, Ruizhi Shao, and Yebin Liu. “Full-body motion capture for multiple closely interacting persons”. In: *Graphical Models* 110 (2020), p. 101072 (See p. 42).
- [205] Sijin Li, Weichen Zhang, and Antoni B Chan. “Maximum-margin structured learning with deep networks for 3D human pose estimation”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 2848–2856 (See pp. 31, 49).
- [206] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. “Learning a model of facial shape and expression from 4D scans”. In: *Transactions on Graphics (TOG)* 36.6 (2017), p. 194 (See pp. 10, 11, 15, 118).
- [207] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. “Learning a model of facial shape and expression from 4D scans”. In: *Transactions on Graphics (TOG)* 36.6 (2017), 194:1–194:17 (See pp. 36, 53).
- [208] Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. “PIANO: A Parametric Hand Bone Model from Magnetic Resonance Imaging”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. 2021, pp. 816–822 (See p. 113).

- [209] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. “NIMBLE: A Non-rigid Hand Model with Bones and Muscles”. In: *Transactions on Graphics (TOG)* 41.4 (2022) (See p. 113).
- [210] 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)*. 2019, pp. 8632–8641 (See p. 42).
- [211] Junbang Liang and Ming C. Lin. “Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 4351–4361 (See pp. 69, 72).
- [212] Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Si Liu, Jinhui Tang, Liang Lin, and Shuicheng Yan. “Human parsing with contextualized convolutional neural network”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 1386–1394 (See pp. 18, 130).
- [213] Kevin Lin, Lijuan Wang, and Zicheng Liu. “End-to-End Human Pose and Mesh Reconstruction with Transformers”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1954–1963 (See pp. 48, 49, 59).
- [214] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. “Feature Pyramid Networks for Object Detection”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 936–944 (See p. 41).
- [215] 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)*. 2014, pp. 740–755 (See pp. 18, 29, 31, 72, 130).
- [216] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. “Markerless motion capture of multiple characters using multiview image segmentation”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.11 (2013), pp. 2720–2735 (See p. 13).
- [217] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. “A ConvNet for the 2020s”. In: 2022, pp. 11976–11986 (See p. 2).

- [218] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. “Deep Learning Face Attributes in the Wild”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 3730–3738 (See p. 30).
- [219] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. “Neural Volumes: Learning Dynamic Renderable Volumes from Images”. In: *Transactions on Graphics (TOG)* 38.4 (2019) (See p. 2).
- [220] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. “Mixture of Volumetric Primitives for Efficient Neural Rendering”. In: *Transactions on Graphics (TOG)* 40.4 (2021) (See p. 2).
- [221] Matthew Loper, Naureen Mahmood, and Michael J Black. “MoSh: Motion and shape capture from sparse markers”. In: *Transactions on Graphics (TOG)* 33.6 (2014), p. 220 (See pp. 10, 14, 128).
- [222] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. “SMPL: A Skinned Multi-Person Linear Model”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH Asia)* 34.6 (2015), 248:1–248:16 (See pp. 1, 2, 10, 12, 20, 30, 31, 49, 69, 71, 91, 92, 95).
- [223] Matthew M Loper and Michael J Black. “OpenDR: An approximate differentiable renderer”. In: *European Conference on Computer Vision (ECCV)*. 2014, pp. 154–169 (See pp. 19, 49).
- [224] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. “3D Human Motion Estimation via Motion Compression and Refinement”. In: *Asian Conference on Computer Vision (ACCV)*. 2020 (See p. 113).
- [225] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. “Dynamics-Regulated Kinematic Policy for Egocentric Pose Estimation”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021 (See p. 113).
- [226] Zhaoyang Lv, Frank Dellaert, James M Rehg, and Andreas Geiger. “Taking a deeper look at the inverse compositional algorithm”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4581–4590 (See pp. 91, 92).
- [227] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. “SCALE: Modeling Clothed Humans with a Surface Codec of Articulated Local Elements”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 16077–16088 (See p. 2).

- [228] 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: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6468–6477 (See pp. 63, 113).
- [229] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. “The Power of Points for Modeling Humans in Clothing”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 10954–10964 (See p. 2).
- [230] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. “Rectifier nonlinearities improve neural network acoustic models”. In: *International Conference on Machine Learning Workshops (ICMLw)*. 2013 (See p. 130).
- [231] Meysam Madadi, Hugo Bertiche, and Sergio Escalera. “SMPLR: Deep learning based SMPL reverse for 3D human pose and shape recovery”. In: *Pattern Recognition (PR)* 106 (2020), p. 107472 (See p. 85).
- [232] 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)*. 2019, pp. 5441–5450 (See pp. 1, 10, 17, 72, 91, 99, 128, 176, 177).
- [233] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour. “The KIT whole-body human motion database”. In: *2015 International Conference on Advanced Robotics (ICAR)*. 2015, pp. 329–336 (See p. 176).
- [234] MANO, models SMPL+H website: dataset, and code. <http://mano.is.tue.mpg.de> (See pp. 12, 20).
- [235] Timo von Marcard, Roberto Henschel, Michael J. 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)*. 2018, pp. 614–631 (See pp. 2, 37, 39, 56, 59, 61, 71, 80, 91, 99, 100, 137, 166, 167, 177).
- [236] Donald W Marquardt. “An algorithm for least-squares estimation of nonlinear parameters”. In: *Journal of the society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441 (See pp. 5, 89, 92, 93).

- [237] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. “A simple yet effective baseline for 3D human pose estimation”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 2659–2668 (See pp. 30, 31, 49).
- [238] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiří Matas, and Viktoriia Sharmanska. “DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20942–20952 (See p. 2).
- [239] Stan Melax, Leonid Keselman, and Sterling Orsten. “Dynamics Based 3D Skeletal Hand Tracking”. In: *Graphics Interface*. 2013, pp. 63–70 (See pp. 10, 12).
- [240] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. “COAP: Compositional Articulated Occupancy of People”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 13201–13210 (See p. 2).
- [241] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. “LEAP: Learning Articulated Occupancy of People”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 10456–10466 (See p. 2).
- [242] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 405–421 (See p. 112).
- [243] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. “PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 909–918 (See p. 114).
- [244] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. “A survey of advances in vision-based human motion capture and analysis”. In: *Computer Vision and Image Understanding (CVIU)* 104.2 (2006), pp. 90–126 (See pp. 13, 28).
- [245] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. “Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation”. In: *Computer Vision and Pattern Recognition Workshops (CVPRw)*. 2022, pp. 2308–2317 (See p. 112).

- [246] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. “NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets”. In: *Computer Vision and Pattern Recognition Workshops (CVPRw)*. 2022, pp. 2299–2307 (See p. 112).
- [247] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. “GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 49–59 (See pp. 2, 31, 49).
- [248] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. “Real-Time Pose and Shape Reconstruction of Two Interacting Hands with a Single Depth Camera”. In: *Transactions on Graphics (TOG)* 38.4 (2019) (See p. 92).
- [249] 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)*. 2021, pp. 9990–9999 (See pp. 63, 72, 85, 167).
- [250] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. *Documentation Mocap Database HDM05*. Tech. rep. CG-2007-2. Universität Bonn, 2007 (See p. 176).
- [251] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding”. In: *Transactions on Graphics (TOG)* 41.4 (2022), 102:1–102:15 (See p. 113).
- [252] Richard M. Murray, Li Zexiang, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC press, 1994 (See p. 14).
- [253] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. “Multi-Person Implicit Reconstruction From a Single Image”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14474–14483 (See p. 113).
- [254] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *International Conference on Machine Learning (ICML)*. 2010 (See p. 99).

- [255] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked hourglass networks for human pose estimation”. In: *European Conference on Computer Vision (ECCV)*. 2016, pp. 483–499 (See p. 31).
- [256] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. “You2Me: Inferring Body Pose in Egocentric Video via First and Second Person Interactions”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 9887–9897 (See p. 113).
- [257] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. “On face segmentation, face swapping, and face perception”. In: *International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 98–105 (See p. 54).
- [258] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. second. New York, NY, USA: Springer, 2006 (See pp. 19, 103, 123, 141).
- [259] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. “Neural Articulated Radiance Field”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 5762–5772 (See p. 2).
- [260] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. “Training a Feedback Loop for Hand Pose Estimation”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 3316–3324 (See p. 10).
- [261] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. “Efficient model-based 3D tracking of hand articulations using Kinect”. In: *British Machine Vision Conference (BMVC)*. 2011 (See pp. 10, 12).
- [262] 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)*. 2018, pp. 484–494 (See pp. 2, 9, 10, 13, 30, 31, 48, 49, 73).
- [263] 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, pp. 598–613 (See p. 113).
- [264] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. “NPMs: Neural Parametric Models for 3D Deformable Shapes”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 12675–12685 (See p. 2).
- [265] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. “SPAMs: Structured Implicit Parametric Models”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 12851–12860 (See p. 2).

- [266] Paschalis Panteleris, Jason Oikonomidis, and Antonis Argyros. “Using a single RGB frame for real time 3D hand pose estimation in the wild”. In: *Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 436–445 (See pp. [2](#), [23](#), [24](#), [118](#), [120](#)).
- [267] Maria Parelli, Katerina Papadimitriou, Gerasimos Potamianos, Georgios Pavlakos, and Petros Maragos. “Spatio-Temporal Graph Convolutional Networks for Continuous Sign Language Recognition”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2022, pp. 8457–8461 (See pp. [112](#), [114](#)).
- [268] 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)*. 2019, pp. 8024–8035 (See pp. [36](#), [130](#), [150](#), [163](#), [177](#)).
- [269] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. “AGORA: Avatars in Geography Optimized for Regression Analysis”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 13468–13478 (See pp. [56](#), [59](#), [60](#), [72](#), [91](#)).
- [270] 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)*. 2019, pp. 10975–10985 (See pp. [28](#), [32](#), [36](#), [42](#), [46](#), [47](#), [49–51](#), [56](#), [58](#), [59](#), [64](#), [68](#), [72](#), [75](#), [89](#), [91](#), [92](#), [102](#), [103](#), [112](#), [113](#), [141](#), [231](#)).
- [271] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. “Coarse-to-fine volumetric prediction for single-image 3D human pose”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1263–1272 (See pp. [31](#), [49](#)).
- [272] 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)*. 2018, pp. 459–468 (See pp. [2](#), [9](#), [10](#), [13](#), [20](#), [30](#), [31](#), [48](#), [49](#), [73](#)).

- [273] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. “A 3D face model for pose and illumination invariant face recognition”. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance*. 2009, pp. 296–301 (See p. 10).
- [274] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. “Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 14314–14323 (See p. 2).
- [275] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. “Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 9054–9063 (See p. 2).
- [276] Mathis Petrovich, Michael J. Black, and Gül Varol. “Action-Conditioned 3D Human Motion Synthesis With Transformer VAE”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 10985–10995 (See pp. 112, 114).
- [277] Mathis Petrovich, Michael J. Black, and Gül Varol. “TEMOS: Generating diverse human motions from textual descriptions”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 480–497 (See p. 114).
- [278] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. “Dyna: A Model of Dynamic Human Shape in Motion”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)* 34.4 (2015), 120:1–120:14 (See pp. 12, 81).
- [279] Gerard Pons-Moll and Bodo Rosenhahn. “Model-Based Pose Estimation”. In: *Visual Analysis of Humans: Looking at People*. Springer, 2011. Chap. 9, pp. 139–170 (See p. 14).
- [280] M. J. D. Powell. “A Hybrid Method for Nonlinear Equations”. In: *Numerical Methods for Nonlinear Algebraic Equations*. Ed. by P. Rabinowitz. Gordon and Breach, 1970 (See p. 93).
- [281] 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”. In: *Transactions on Visualization and Computer Graphics (TVCG)* 25.5 (2019), pp. 1887–1897 (See pp. 4, 73, 74, 77, 160).

- [282] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. “BABEL: Bodies, Action and Behavior with English Labels”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 722–731 (See p. 114).
- [283] Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. “Pixel-Aligned Volumetric Avatars”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 11733–11742 (See p. 2).
- [284] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. “Accelerating 3D Deep Learning with PyTorch3D”. In: *arXiv:2007.08501* (2020) (See pp. 49, 54).
- [285] 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, pp. 11488–11499 (See pp. 106, 177).
- [286] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)* 28 (2015) (See p. 2).
- [287] *RenderPeople*. renderpeople.com (See p. 56).
- [288] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. “Learning Monocular 3D Human Pose Estimation from Multi-view Images”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8437–8446 (See p. 13).
- [289] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. “Single-Shot High-Quality Facial Geometry and Skin Appearance Capture”. In: *Transactions on Graphics (TOG)* 39.4 (2020) (See pp. 1, 112).
- [290] 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*. Tech. rep. AFRL-HE-WP-TR-2002-0169. US Air Force Research Laboratory, 2002 (See pp. 12, 15, 30, 55, 71, 73, 76, 128).

- [291] Chris Rockwell and David F. Fouhey. “Full-Body Awareness from Partial Observations”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 522–539 (See p. 2).
- [292] Grégory Rogez and Cordelia Schmid. “MoCap-guided data augmentation for 3D pose estimation in the wild”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2016, pp. 3108–3116 (See p. 31).
- [293] Javier Romero, Dimitrios Tzionas, and Michael J. Black. “Embodied Hands: Modeling and Capturing Hands and Bodies Together”. In: *Transactions on Graphics (TOG)* 36.6 (2017), 245:1–245:17 (See pp. 10, 12, 15, 16, 20, 30–32, 49, 50, 89, 96, 177).
- [294] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. “Delving Deep Into Hybrid Annotations for 3D Human Recovery in the Wild”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 5339–5347 (See pp. 30, 49).
- [295] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. “FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration”. In: *International Conference on Computer Vision Workshops (ICCVw)*. 2021, pp. 1749–1759 (See pp. 47, 50, 57, 59, 61, 64, 91, 112, 114, 150, 152, 154).
- [296] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. “Monocular 3D Reconstruction of Interacting Hands via Collision-Aware Factorized Refinements”. In: *International Conference on 3D Vision (3DV)*. 2021 (See p. 2).
- [297] Rasmus Rothe, Radu Timofte, and Luc Van Gool. “DEX: Deep Expectation of apparent age from a single image”. In: *International Conference on Computer Vision Workshops (ICCVw)*. 2015, pp. 252–257 (See p. 56).
- [298] 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)*. 2020, pp. 5561–5569 (See pp. 30, 49, 73).
- [299] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. “PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2304–2314 (See pp. 2, 30, 31, 48, 49, 63, 72).

- [300] 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)*. 2020, pp. 81–90 (See pp. 2, 31, 48, 49, 71, 72).
- [301] Igor Santesteban, Miguel A Otaduy, and Dan Casas. “SNUG: Self-Supervised Neural Dynamic Garments”. In: *Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 8140–8150 (See p. 113).
- [302] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. “Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 7763–7772 (See pp. 2, 36, 42, 49, 56, 61, 62).
- [303] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. “Plenoxels: Radiance Fields without Neural Networks”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5501–5510 (See p. 113).
- [304] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. “3D Human pose estimation: A review of the literature and analysis of covariates”. In: *Computer Vision and Image Understanding (CVIU)* 152 (2016), pp. 1–20 (See p. 28).
- [305] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. “PiGraphs: Learning Interaction Snapshots from Observations”. In: *Transactions on Graphics (TOG)* 35.4 (2016), pp. 1–12 (See p. 42).
- [306] Jürgen Schmidhuber. “Learning to control fast-weight memories: An alternative to dynamic recurrent networks”. In: *Neural Computation* 4.1 (1992), pp. 131–139 (See pp. 89, 91).
- [307] Jürgen Schmidhuber. “A neural network that embeds its own meta-levels”. In: *IEEE International Conference on Neural Networks*. IEEE. 1993, pp. 407–412 (See pp. 89, 91).
- [308] Tanner Schmidt, Richard Newcombe, and Dieter Fox. “DART: Dense Articulated Real-Time Tracking”. In: *RSS*. 2014 (See pp. 10, 12).
- [309] Michael Seeber, Roi Poranne, Marc Pollefeys, and Martin Oswald. “RealisticHands: A Hybrid Model for 3D Hand Reconstruction”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 22–31 (See p. 92).

- [310] 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 (See pp. 69, 71–73, 80, 85, 86, 167).
- [311] 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)*. 2021, pp. 11219–11229 (See pp. 68, 69, 73, 83–86, 167).
- [312] 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)*. 2021, pp. 16094–16104 (See pp. 72, 73, 167).
- [313] Hyewon Seo, Frederic Cordier, and Nadia Magnenat-Thalmann. “Synthesizing animatable body models with parameterized shape modifications”. In: *Symposium on Computer Animation (SCA)*. 2003, pp. 120–125 (See p. 73).
- [314] Hyewon Seo and Nadia Magnenat-Thalmann. “An automatic modeling of human bodies from sizing parameters”. In: *Symposium on Interactive 3D Graphics (SI3D)*. 2003, pp. 19–26 (See p. 73).
- [315] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. “DoubleField: Bridging the Neural Surface and Radiance Fields for High-fidelity Human Reconstruction and Rendering”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15872–15882 (See p. 2).
- [316] Jingjing Shen, Thomas J. Cashman, Qi Ye, Tim Hutton, Toby Sharp, Federica Bogo, Andrew William Fitzgibbon, and Jamie Shotton. “The Phong Surface: Efficient 3D Model Fitting using Lifted Optimization”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 687–703 (See pp. 89, 92, 93).
- [317] Leonid Sigal, Alexandru Balan, and Michael J Black. “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion”. In: *International Journal of Computer Vision (IJCV)* 87.1 (2010), pp. 4–27 (See pp. 1, 31, 72).
- [318] Leonid Sigal and Michael J Black. “Predicting 3D people from 2D pictures”. In: *International Conference on Articulated Motion and Deformable Objects*. 2006, pp. 185–195 (See pp. 30, 49).

- [319] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. “Hand Keypoint Detection in Single Images using Multiview Bootstrapping”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4645–4653 (See pp. 9, 10, 16, 30, 48).
- [320] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations (ICLR)*. 2015 (See p. 2).
- [321] Cristian Sminchisescu and Bill Triggs. “Covariance scaled sampling for monocular 3D body tracking”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2001, pp. I–I (See p. 2).
- [322] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K. Hodgins, and Takaaki Shiratori. “Constraining Dense Hand Surface Tracking with Elasticity”. In: *Transactions on Graphics (TOG)* 39.6 (2020) (See pp. 2, 113).
- [323] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. “FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 5329–5338 (See pp. 31, 49).
- [324] Jie Song, Xu Chen, and Otmar Hilliges. “Human Body Model Fitting by Learned Gradient Descent”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 744–760 (See pp. 2, 89, 91, 92, 94, 96, 100, 104, 177, 178).
- [325] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. “Interactive Markerless Articulated Hand Motion Tracking using RGB and Depth Data”. In: *International Conference on Computer Vision (ICCV)*. 2013, pp. 2456–2463 (See pp. 10, 12).
- [326] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. “Training Very Deep Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Vol. 28. 2015 (See p. 104).
- [327] Jonathan Starck and Adrian Hilton. “Surface capture for performance-based animation”. In: *IEEE Computer Graphics and Applications* 27.3 (2007) (See p. 13).
- [328] Neal Stephenson. *Snow Crash*. Bantam Spectra, 1992 (See p. 1).

- [329] 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”. In: *ACM Transactions on Graphics (ToG), (Proc. SIGGRAPH)* 35.4 (2016), 54:1–54:14 (See pp. 3, 70, 71, 73, 76, 160).
- [330] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5686–5696 (See pp. 36, 51, 137).
- [331] Xiao Sun, Jiayang Shang, Shuang Liang, and Yichen Wei. “Compositional human pose regression”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 2621–2630 (See pp. 31, 49).
- [332] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. “Integral Human Pose Regression”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 536–553 (See pp. 31, 49).
- [333] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. “Monocular, One-stage, Regression of Multiple 3D People”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11179–11188 (See pp. 48, 49, 59, 113).
- [334] 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)*. 2022, pp. 13243–13252 (See pp. 2, 113).
- [335] James S. Supančič III, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. “Depth-Based Hand Pose Estimation: Data, Methods, and Challenges”. In: *International Conference on Computer Vision (ICCV)*. 2015, pp. 1868–1876 (See p. 28).
- [336] 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)*. 2022, pp. 13263–13273 (See pp. 112, 114).
- [337] 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)*. 2020, pp. 581–600 (See pp. 1, 42, 114).

- [338] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. “Efficient and Precise Interactive Hand Tracking through Joint, Continuous Optimization of Pose and Correspondences”. In: *Transactions on Graphics (TOG)* (2016) (See pp. 92, 93).
- [339] Bugra Tekin, Federica Bogo, and Marc Pollefeys. “H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4506–4515 (See pp. 31, 49).
- [340] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. “Structured Prediction of 3D Human Pose with Deep Neural Networks”. In: *British Machine Vision Conference (BMVC)*. 2016, pp. 130.1–130.11 (See pp. 31, 49).
- [341] Matthias Teschner, Stefan Kimmerle, Bruno Heidelberger, Gabriel Zachmann, Laks Raghupathi, Arnulph Fuhrmann, Marie-Paule Cani, François Faure, Nadia Magnenat-Thalmann, Wolfgang Strasser, and Pascal Volino. “Collision Detection for Deformable Objects”. In: *Eurographics*. 2004 (See pp. 18, 119).
- [342] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. “FML: Face Model Learning From Videos”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10812–10822 (See p. 2).
- [343] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeonwoo Kim, Patrick Pérez, and Christian Theobalt. “Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 2549–2559 (See p. 49).
- [344] Ayush Tewari, Michael Zollhöfer, Hyeonwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. “MoFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 3735–3744 (See pp. 49, 50).

- [345] Neerja Thakkar, Georgios Pavlakos, and Hany Farid. “The Reliability of Forensic Body-Shape Identification”. In: *Computer Vision and Pattern Recognition Workshops (CVPRw)*. 2022 (See p. 112).
- [346] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. “Face2Face: Real-time face capture and reenactment of RGB videos”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2387–2395 (See pp. 2, 49, 93).
- [347] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. “Recovering 3D Human Mesh from Monocular Images: A Survey”. In: *arXiv preprint arXiv:2203.01923* (2022) (See p. 114).
- [348] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. “Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11688–11698 (See p. 2).
- [349] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. “Spheremeshes for real-time hand modeling and tracking”. In: *Transactions on Graphics (TOG)* 35.6 (2016) (See pp. 10, 12).
- [350] Denis Tomè, Thiemo Alldieck, Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernán Badino, and Fernando De la Torre. “Self-Pose: 3D Egocentric Pose Estimation from a Headset Mounted Camera”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2020), pp. 1–1 (See p. 92).
- [351] Denis Tomè, Patrick Peluse, Lourdes Agapito, and Hernan Badino. “xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 7728–7738 (See p. 92).
- [352] Denis Tomè, Chris Russell, and Lourdes Agapito. “Lifting from the deep: Convolutional 3D pose estimation from a single image”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5689–5698 (See pp. 30, 31, 49).
- [353] Luan Tran, Feng Liu, and Xiaoming Liu. “Towards High-Fidelity Nonlinear 3D Face Morphable Model”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1126–1135 (See p. 49).
- [354] 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)*. 2014, pp. 83–90 (See p. 73).

- [355] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. “Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1493–1502 (See p. 62).
- [356] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. “Capturing Hands in Action using Discriminative Salient Points and Physics Simulation”. In: *International Journal of Computer Vision (IJCV)* 118.2 (2016), pp. 172–193 (See pp. 10, 12, 18, 113, 119, 121).
- [357] *Unreal Engine*. unrealengine.com (See p. 56).
- [358] *Unreal Engine Marketplace*. unrealengine.com/marketplace/en-US/store (See p. 56).
- [359] Gül 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)*. 2018, pp. 20–38 (See pp. 2, 30, 31, 48, 49, 71).
- [360] 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)*. 2017, pp. 4627–4635 (See p. 72).
- [361] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017 (See p. 2).
- [362] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. “Face transfer with multilinear models”. In: *Transactions on Graphics (TOG)* 24.3 (2005), pp. 426–433 (See p. 10).
- [363] Christoph Vogel and Thomas Pock. “A Primal Dual Network for Low-Level Vision Problems”. In: *Pattern Recognition (PR)*. Ed. by Volker Roth and Thomas Vetter. Springer International Publishing, 2017, pp. 189–202 (See p. 91).
- [364] Haoyang Wang, Riza Alp Guler, Iasonas Kokkinos, George Papan-dreou, and Stefanos Zafeiriou. “BLSM: A Bone-Level Skinned Model of the Human Mesh”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 1–17 (See p. 2).

- [365] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. “RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video”. In: *Transactions on Graphics (TOG)* 39.6 (2020) (See p. 2).
- [366] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. “MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021 (See p. 2).
- [367] Yangang Wang, Cong Peng, and Yebin Liu. “Mask-pose Cascaded CNN for 2D Hand Pose Estimation from Single Color Images”. In: *Transactions on Circuits and Systems for Video Technology (TCSVT)* 29.11 (2019), pp. 3258–3268 (See p. 48).
- [368] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Cecilia Zhang. “SunStage: Portrait Reconstruction and Relighting using the Sun as a Light Stage”. In: *arXiv preprint arXiv:2204.03648* (2022) (See p. 112).
- [369] Ziyang Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. “Learning Compositional Radiance Fields of Dynamic Human Heads”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 5704–5713 (See p. 2).
- [370] Ziyang Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhofer, Jessica Hodgins, and Christoph Lassner. “HVV: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 6143–6154 (See p. 113).
- [371] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. “Convolutional Pose Machines”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4724–4732 (See pp. 10, 16, 31).
- [372] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. “DOPE: Distillation Of Part Experts for whole-body 3D pose estimation in the wild”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 380–397 (See p. 50).
- [373] Andrew Weitz, Lina Colucci, Sidney Primas, and Brinnae Bent. “InfiniteForm: A synthetic, minimal bias dataset for fitness applications”. In: *arXiv:2110.01330* (2021) (See p. 72).

- [374] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. “HumanNeRF: Free-viewpoint Rendering of Moving People from Monocular Video”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16210–16220 (See p. 2).
- [375] Zhenzhen Weng and Serena Yeung. “Holistic 3D Human and Scene Mesh Estimation From Single View Images”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 334–343 (See p. 114).
- [376] Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien Valentin. “3D face reconstruction with dense landmarks”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 160–177 (See p. 5).
- [377] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Matthew Johnson, Virginia Estellers, Thomas J. Cashman, and Jamie Shotton. “Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 3681–3691 (See pp. 5, 97, 98, 107, 177).
- [378] 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)*. 2022, pp. 257–274 (See pp. 112, 114).
- [379] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. *Detectron2*. <https://github.com/facebookresearch/detectron2>. 2019 (See p. 41).
- [380] Stefanie Wuhrer and Chang Shu. “Estimating 3D human shapes from measurements”. In: *Machine Vision and Applications (MVA)* 24.6 (2013), pp. 1133–1147 (See pp. 77, 160).
- [381] 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)*. 2019, pp. 10957–10966 (See pp. 28, 32, 38, 40, 41, 50, 56–58, 89, 92, 112, 152, 153).
- [382] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. “SAPIEN: A SimulATED Part-based Interactive ENvironment”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 11094–11104 (See p. 114).

- [383] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. “Physics-Based Human Motion Estimation and Synthesis From Videos”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11532–11541 (See pp. 106, 113).
- [384] Xuehan Xiong and Fernando De la Torre. “Supervised Descent Method and Its Applications to Face Alignment”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2013, pp. 532–539 (See p. 91).
- [385] 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, pp. 13296–13306 (See pp. 2, 71).
- [386] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. “H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021, pp. 14955–14966 (See p. 2).
- [387] 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)*. 2020, pp. 7214–7223 (See pp. 2, 12, 28, 30, 32, 46, 47, 50, 68, 71, 89, 91, 92, 112, 113).
- [388] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. “Monoperfcap: Human performance capture from monocular video”. In: *Transactions on Graphics (TOG)* 37.2 (2018), p. 27 (See p. 13).
- [389] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. “D3D-HOI: Dynamic 3D Human-Object Interactions from Videos”. In: *arXiv preprint arXiv:2108.08420* (2021) (See p. 114).
- [390] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. “LoBSTr: Real-time Lower-body Pose Prediction from Sparse Upper-body Tracking Signals”. In: *Computer Graphics Forum* (2021) (See p. 92).
- [391] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. “Expression flow for 3D-aware face component transfer”. In: *Transactions on Graphics (TOG)* 30.4 (2011), p. 60 (See p. 10).

- [392] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Vedaldi Andrea, and Joo Hanbyul. “BANMo: Building Animatable 3D Neural Models from Many Casual Videos”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 2863–2873 (See p. 2).
- [393] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. “FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 598–607 (See p. 47).
- [394] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. “Human-Aware Object Placement for Visual Environment Reconstruction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 3959–3970 (See p. 114).
- [395] Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Park. “HUMBI: A Large Multiview Dataset of Human Body Expressions and Benchmark Challenge”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2021), pp. 1–1 (See pp. 1, 13).
- [396] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuha0 Ge, et al. “Depth-based 3D hand pose estimation: From current achievements to future goals”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 2636–2645 (See pp. 13, 28, 31).
- [397] Ye Yuan and Kris Kitani. “Ego-Pose Estimation and Forecasting as Real-Time PD Control”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 10082–10092 (See p. 92).
- [398] Ye Yuan and Kris M. Kitani. “3D Ego-Pose Estimation via Imitation Learning”. In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 763–778 (See p. 92).
- [399] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. “SimPoE: Simulated Character Control for 3D Human Pose Estimation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 7159–7169 (See pp. 92, 100, 106, 113).
- [400] Christopher Zach. “Robust Bundle Adjustment Revisited”. In: *European Conference on Computer Vision (ECCV)*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. 2014, pp. 772–787 (See p. 98).

- [401] 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)*. 2020, pp. 465–481 (See p. 72).
- [402] 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)*. 2021, pp. 14484–14493 (See pp. 89, 92, 100).
- [403] 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)*. 2018, pp. 2148–2157 (See pp. 2, 31, 42, 49).
- [404] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. "Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2018, pp. 8410–8419 (See p. 42).
- [405] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. "THUNDR: Transformer-Based 3D Human Reconstruction With Markers". In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 12971–12980 (See p. 113).
- [406] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. "3D Human Mesh Regression with Dense Correspondence". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 7052–7061 (See p. 49).
- [407] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. "Detailed, accurate, human shape estimation from clothed 3D scan sequences". In: *Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5484–5493 (See p. 56).
- [408] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. "Learning 3D Human Shape and Pose From Dense Body Parts". In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44.5 (2022), pp. 2610–2627 (See p. 63).

- [409] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. “PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images”. In: *arXiv preprint arXiv:2207.06400* (2022) (See p. 114).
- [410] Hongwen Zhang, Yating Tian, Xinchi 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)*. 2021, pp. 11446–11456 (See pp. 4, 42, 68, 89, 91, 92).
- [411] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. “Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild”. In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 34–51 (See p. 114).
- [412] Siwei Zhang, Ma Qianli, Yan Zhang, Qian Zhiyin, Pollefeys Marc, Federica Bogo, and Siyu Tang. “EgoBody: Human Body Shape, Motion and Social Interactions from Head-Mounted Devices”. In: *European Conference on Computer Vision (ECCV)*. 2022, pp. 180–200 (See p. 114).
- [413] Siwei Zhang, Yan Zhang, Federica Bogo, Pollefeys Marc, and Siyu Tang. “Learning Motion Priors for 4D Human Body Capture in 3D Scenes”. In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11343–11353 (See p. 106).
- [414] 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)*. Vol. 1. 2020, pp. 642–651 (See p. 114).
- [415] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. “End-to-End Hand Mesh Recovery From a Monocular RGB Image”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2354–2364 (See pp. 31, 46, 49).
- [416] Yan Zhang and Siyu Tang. “The Wanderings of Odysseus in 3D Scenes”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 20481–20491 (See pp. 112, 114).
- [417] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. “4D Association Graph for Realtime Multi-Person Motion Capture Using Multiple Video Cameras”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 1324–1333 (See p. 1).

- [418] 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 (ECCV)*. 2022, pp. 311–327 (See p. 114).
- [419] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. “Semantic Graph Convolutional Networks for 3D Human Pose Regression”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 3420–3430 (See pp. 30, 49).
- [420] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. “Structured Local Radiance Fields for Human Avatar Modeling”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15893–15903 (See p. 2).
- [421] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. “DeepHuman: 3D Human Reconstruction From a Single Image”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 7738–7748 (See pp. 2, 31, 49).
- [422] 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)*. 2019, pp. 5738–5746 (See pp. 32, 95).
- [423] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. “Monocular Real-time Full Body Capture with Inter-part Correlations”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 4811–4822 (See pp. 50, 61, 62, 65, 112).
- [424] 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)*. 2020, pp. 316–333 (See p. 73).
- [425] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. “NICE-SLAM: Neural Implicit Scalable Encoding for SLAM”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 12786–12796 (See p. 113).
- [426] Christian Zimmermann and Thomas Brox. “Learning to Estimate 3D Hand Pose from Single RGB Images”. In: *International Conference on Computer Vision (ICCV)*. 2017, pp. 4913–4921 (See pp. 31, 49).

- [427] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. “FreiHAND: A Dataset for Markerless Capture of Hand Pose and Shape From Single RGB Images”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 813–822 (See pp. [2](#), [29](#), [30](#), [36](#), [38–42](#), [56](#), [61](#), [62](#), [114](#), [137](#)).
- [428] Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, and Tony Tung. “Data-Driven 3D Reconstruction of Dressed Humans From Sparse Views”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 494–504 (See p. [2](#)).
- [429] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. “State of the art on monocular 3D face reconstruction, tracking, and applications”. In: *Computer Graphics Forum* 37.2 (2018), pp. 523–550 (See pp. [11](#), [13](#), [28](#), [31](#), [93](#)).

CURRICULUM VITAE

PERSONAL DATA

Name	Vasileios Choutas
Date of Birth	September 30, 1993
Place of Birth	Thessaloniki, Greece
Citizen of	Greece

EDUCATION

2018 – 2022	Perceiving Systems and Computer Vision Lab MPI for Intelligent Systems and ETH Zürich Doctoral Studies
2011 – 2017	Aristotle University of Thessaloniki Thessaloniki, Greece Diploma (M.Eng) in Electrical and Computer Engineering

PROFESSIONAL EXPERIENCE

July – December 2022	Research Intern Meta Reality Labs, Zürich, Switzerland
June – November 2021	Research Intern Microsoft, Zürich, Switzerland
December 2021 – June 2022	Research Assistant ETH Zürich, Switzerland
May 2018 – April 2021	Research Assistant Max Planck Institute for Intelligent Systems, Tübingen, Germany
February – April 2018	Research Intern TUM, München, Germany
April – November 2017	Research Intern INRIA and Naver Labs Europe, Grenoble, France

PUBLICATIONS

The following publications are included as whole or in parts in this dissertation:

- [1] 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)*. 2019, pp. 10975–10985.
- [2] 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)*. Vol. LNCS 12355. 2020, pp. 20–40.
- [3] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. “Collaborative Regression of Expressive Bodies using Moderation”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 792–804.
- [4] Vasileios Choutas, Federica Bogo, Jingjing Shen, and Julien Valentin. “Learning to Fit Morphable Models”. In: *European Conference on Computer Vision (ECCV)*. 2022.
- [5] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. “Accurate 3D Body Shape Regression using Metric and Semantic Attributes”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 2718–2728.

The first two authors in [55, 86, 270] contributed equally.

The following conference publications are not included in this dissertation:

- [1] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. “PoTion: Pose MoTion Representation for Action Recognition”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7024–7033.
- [2] 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)*. 2019, pp. 2282–2292.

- [3] 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)*. 2022, pp. 13263–13273.
- [4] Gurkirt Singh, Vasileios Choutas, Suman Saha, Fisher Yu, and Luc Van Gool. "Spatio-Temporal Action Detection Under Large Motion". In: *Winter Conference on Applications of Computer Vision (WACV)*. 2023.

Articles under submission:

- [1] Maria-Paola Forte, Chun-Hao P. Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. "SGNify: Full-body 3D Reconstruction of Sign Language Signs from RGB Videos". In: (2022).