Democratizing 3D Human Digitization

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

vorgelegt von

Yuliang Xiu

aus Shandong, China

Tübingen
2025

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

Tag der mündlichen Prüfung:	17.03.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Michael J. Black
2. Berichterstatter:	Prof. Dr. Gerard Pons-Moll

To Mom, Dad, Cathy, and Midou

ABSTRACT

Reproducing real humans in virtual space, has been a long-standing topic since the early 20th century [i]. This field has advanced significantly over the last century, with realistic digital twins transforming the film industry, enabling immersive telepresence experiences, revolutionizing assistive techniques, and improving autonomous driving safety through simulations. Creating high-quality digital twins, whether scanned or artist-created, requires significant effort. Capturing 3D models requires expensive equipment (*e.g.*, 4D scanners, light stage), and involves extensive post-processing of raw scans. Additionally, creating 3D characters with photorealistic appearance, lifelike movements, and natural clothing dynamics can take several months even for skilled modelers and animators. Both of these creation processes present challenges that restrict accessibility to the wider public and hinder scalability to everyday data, underscoring the need for democratization.

As cameras have become standard on mobile devices, the cost of taking photos has decreased significantly. Consequently, reconstructing 3D humans from in-the-wild photos has emerged as one of the most scalable and affordable methods to digitize virtual 3D humans. This can be done with just a single full-body shot. Previous methods have made great progress, but limitations remain. They may struggle with unseen poses, loose clothing, and often produce overly smooth shapes with blurry textures for non-visible regions, like backsides. Additionally, full-body shots are still limited; most daily photos are casual and unconstrained, featuring varied poses, viewpoints, cropping, and occlusions. Few studies have attempted to reconstruct 3D clothed humans from such unstructured data, as commonly used dependencies, like body and camera estimation, are not reliable enough for these unconstrained photos. A paradigm shift is necessary to understand such unstructured data and produce structured output.

This Ph.D. thesis addresses the challenges mentioned above, and contributes to advancing image-based 3D clothed human reconstruction through the development of novel benchmarks (PuzzleIOI) and algorithms (ICON, ECON, TeCH, and PuzzleAvatar).

ICON represents a critical advancement in this field. Previous mainstream works (*e.g.*, PIFu, PaMIR, ARCH), regardless of using a parametric body template, always struggle with varied human poses and often produce broken limbs, missing details, or non-human shapes. This is mainly due to the spatial correlation of convolutional neural networks (CNN). To address these limitations, ICON replaces the CNN-based image

[i] Father of Boeing Man – William Fetter

global encoder with pixel-aligned hand-crafted local features. Using these locally queried features, a shallow multilayer perceptron (MLP) is trained to regress the occupancy field of the 3D human, achieving superior accuracy on unseen out-of-distribution poses, with even less than 10% of the training data used by other methods.

To enhance pose robustness during reconstruction, ICON utilizes geometric features derived locally from a 3D body prior. However, this body prior introduces bias as well, reducing geometric flexibility on free-form surfaces, especially evident with loose clothing like skirts and dresses that deviate from the body underneath. To tackle the "topological flexibility vs. pose robustness" dilemma, we introduce ECON. ECON revisits the traditional surface normal integration technique and uses a 3D body model as a "canvas" to stitch together 2.5D front and back clothed surfaces. These 2.5D surfaces are integrated from the dual normal maps estimated from the input image. ECON then inpaints the gaps between the dual surfaces to achieve a watertight 3D full-body human.

We now have ICON for reconstructing challenging poses and ECON for recovering loose clothing. However, since results are derived from a single input image, the back-side surfaces tend to be overly smooth with blurry textures. How can we fully capture the all-around visual attributes of an individual from just a single image? TeCH addresses this challenge by – reconstructing the visible and generating the invisible. Such imagination comes from both "descriptive" texts captioned from the input image, together with the "indescribable" appearance learned through DreamBooth, by personalizing Text-to-Image diffusion models via fine-tuning. Ultimately, employing Score Distillation Sampling (SDS), TeCH can produce high-fidelity 3D humans with consistent textures and detailed full-body geometry, even for the unseen backside.

All these efforts aim to reconstruct 3D humans from full-body shots. But is it feasible to use a collection of truly unconstrained photos — taken from any angle, in any pose, and with any cropping or occlusion — to accurately recreate digital twins? We call this new task "Album2Human" and introduce PuzzleAvatar to address it. PuzzleAvatar is the first work capable of reconstructing 3D textured clothed humans from unconstrained photo collections without needing body or camera pose estimation and re-projection terms. It can be seen as a more compositional and scalable version of TeCH — learning compositional asset IDs (*e.g.*, face, garments, accessories) by fine-tuning PuzzleBooth and sculpting the A-pose clothed human from an A-pose body via SDS, similar to TeCH. Additionally, we introduce a benchmark, PuzzleIOI, for quantitative comparison.

In conclusion, this Ph.D. thesis revisits the task of image-based 3D clothed human reconstruction and generalizes the reconstruction algorithms for challenging poses (ICON), loose clothing (ECON), non-visible backsides (TeCH), and unconstrained photo collections (PuzzleAvatar). Additionally, TeCH and PuzzleAvatar frame the reconstruction task as a conditional generation task, unifying the fields of 3D Reconstruction and 3D Generation into a single framework.

ZUSAMMENFASSUNG

Die Reproduktion realer Menschen im virtuellen Raum ist ein langjähriges Thema seit dem frühen 20. Jahrhundert [ii]. Dieses Feld hat sich im Laufe des letzten Jahrhunderts erheblich weiterentwickelt, wobei realistische digitale Zwillinge die Filmindustrie transformieren, immersive Telepräsenzerfahrungen ermöglichen, assistive Techniken revolutionieren und die Sicherheit beim autonomen Fahren durch Simulationen verbessern. Die Erstellung hochwertiger digitaler Zwillinge, ob gescannt oder von Künstlern erstellt, erfordert erheblichen Aufwand. Die Erfassung von 3D-Modellen erfordert teure Ausrüstung (*e.g.*, 4D-Scanner, Lichtbühne) und umfangreiche Nachbearbeitung der Rohscans. Darüber hinaus kann die Erstellung von 3D-Charakteren mit fotorealistischem Aussehen, lebensechten Bewegungen und natürlicher Kleidungsdynamik selbst für erfahrene Modellierer und Animator:innen mehrere Monate dauern. Beide Erstellungsprozesse stellen Herausforderungen dar, die den Zugang für die breite Öffentlichkeit einschränken und die Skalierbarkeit auf alltägliche Daten behindern, was die Notwendigkeit der Demokratisierung unterstreicht.

Da Kameras auf mobilen Geräten Standard geworden sind, sind die Kosten für das Fotografieren erheblich gesunken. Daraus ergibt sich, dass die Rekonstruktion von 3D-Menschen aus Fotos in freier Wildbahn als eine der skalierbarsten und kostengünstigsten Methoden zur Digitalisierung virtueller 3D-Menschen entstanden ist. Dies kann bereits mit nur einem Ganzkörperfoto erfolgen. Frühere Methoden haben große Fortschritte gemacht, aber es bestehen weiterhin Einschränkungen. Sie können Schwierigkeiten mit ungewöhnlichen Posen, lockerer Kleidung haben und produzieren oft übermäßig glatte Formen mit verschwommenen Texturen für nicht sichtbare Regionen, wie Rückseiten. Darüber hinaus sind Ganzkörperaufnahmen immer noch begrenzt; die meisten täglichen Fotos sind zwanglos und unbeschränkt, zeigen verschiedene Posen, Blickwinkel, Beschnitt und Verdeckungen. Nur wenige Studien haben versucht, 3D-bekleidete Menschen aus solchen unstrukturierten Daten zu rekonstruieren, da häufig verwendete Abhängigkeiten wie Körper- und Kamerabschätzung für diese unbeschränkten Fotos nicht zuverlässig genug sind. Ein Paradigmenwechsel ist erforderlich, um solche unstrukturierten Daten zu verstehen und strukturierte Ausgaben zu erzeugen.

Diese Doktorarbeit behandelt die oben genannten Herausforderungen und trägt zur Weiterentwicklung der bildbasierten 3D-Rekonstruktion bekleideter Menschen durch die

[ii] Vater des Boeing-Mannes – William Fetter

Entwicklung neuer Benchmarks (PuzzleIOI) und Algorithmen (ICON, ECON, TeCH und PuzzleAvatar) bei.

ICON stellt einen entscheidenden Fortschritt auf diesem Gebiet dar. Frühere Hauptwerke (*e.g.*, PIFu, PaMIR, ARCH), unabhängig davon, ob sie eine parametrische Körperschablone verwenden, haben immer Schwierigkeiten mit verschiedenen menschlichen Posen und produzieren oft gebrochene Gliedmaßen, fehlende Details oder nicht-menschliche Formen. Dies liegt hauptsächlich an der räumlichen Korrelation von Convolutional Neural Networks (CNN). Um diese Einschränkungen zu überwinden, ersetzt ICON den CNN-basierten globalen Bild-Encoder durch pixelausgerichtete handgefertigte lokale Merkmale. Unter Verwendung dieser lokal abgefragten Merkmale wird ein flaches Multilayer-Perzeptron (MLP) trainiert, um das Besetzungsfeld des 3D-Menschen zu regressieren, wodurch eine überlegene Genauigkeit bei nicht gesehenen Out-of-Distribution-Posen erreicht wird, selbst wenn weniger als 10% der Trainingsdaten im Vergleich zu anderen Methoden verwendet werden.

Um die Poserobustheit während der Rekonstruktion zu verbessern, verwendet ICON geometrische Merkmale, die lokal aus einem 3D-Körperprior abgeleitet sind. Dieser Körperprior führt jedoch auch zu einer Verzerrung, die die geometrische Flexibilität auf freiformigen Oberflächen verringert, insbesondere bei lockerer Kleidung wie Röcken und Kleidern, die sich vom darunter liegenden Körper abheben. Um das Dilemma "topologische Flexibilität vs. Poserobustheit" zu bewältigen, führen wir ECON ein. ECON überprüft die traditionelle Technik der Oberflächennormalenintegration erneut und verwendet ein 3D-Körpermodell als "Leinwand", um 2,5D-vordere und hintere bekleidete Oberflächen zusammenzufügen. Diese 2,5D-Oberflächen werden aus den aus dem Eingabebild geschätzten dualen Normalenkarten integriert. ECON füllt dann die Lücken zwischen den dualen Oberflächen aus, um einen dichten 3D-Ganzkörpermenschen zu erreichen.

Wir haben nun ICON zur Rekonstruktion herausfordernder Posen und ECON zur Wiederherstellung lockerer Kleidung. Da die Ergebnisse jedoch aus einem einzigen Eingabebild abgeleitet sind, neigen die Rückseitenoberflächen dazu, übermäßig glatt mit verschwommenen Texturen zu sein. Wie können wir alle visuellen Merkmale einer Person vollständig aus nur einem Bild erfassen? TeCH adressiert diese Herausforderung, indem es das Sichtbare rekonstruiert und das Unsichtbare generiert. Diese Vorstellung stammt sowohl aus "beschreibenden" Texten, die aus dem Eingabebild untertitelt sind, als auch aus dem "unerklärlichen" Aussehen, das durch DreamBooth gelernt wird, indem Text-zu-Bild-Diffusionsmodelle durch Feinabstimmung personalisiert werden. Letztendlich kann TeCH mithilfe von Score Distillation Sampling (SDS) hochwertige 3D-Menschen mit konsistenten Texturen und detaillierter Ganzkörpergeometrie produzieren, selbst für die unsichtbare Rückseite.

All diese Bemühungen zielen darauf ab, 3D-Menschen aus Ganzkörperaufnahmen zu rekonstruieren. Aber ist es machbar, eine Sammlung wirklich unbeschränkter Fotos —

aufgenommen aus beliebigen Blickwinkeln, in jeder Pose und mit beliebigem Beschnitt oder Verdeckung — zu verwenden, um digitale Zwillinge genau nachzubilden? Wir nennen diese neue Aufgabe "Album2Human" und stellen PuzzleAvatar vor, um sie zu bewältigen. PuzzleAvatar ist die erste Arbeit, die in der Lage ist, 3D-texturierte bekleidete Menschen aus unbeschränkten Fotosammlungen ohne die Notwendigkeit von Körper- oder Kameraposerfassung und Rückprojektionstermen zu rekonstruieren. Es kann als eine kompositionellere und skalierbarere Version von TeCH betrachtet werden — indem es kompositionelle Asset-IDs lernt (*e.g.*, Gesicht, Kleidungsstücke, Accessoires) durch Feinabstimmung von PuzzleBooth und das A-Pose bekleidete Individuum aus einem A-Pose-Körper formt, ähnlich wie TeCH. Darüber hinaus stellen wir einen Benchmark, PuzzleIOI, für quantitative Vergleiche vor.

Zusammenfassend überdenkt diese Doktorarbeit die Aufgabe der bildbasierten 3D-Rekonstruktion bekleideter Menschen und verallgemeinert die Rekonstruktionsalgorithmen für herausfordernde Posen (ICON), lockere Kleidung (ECON), nicht sichtbare Rückseiten (TeCH) und unbeschränkte Fotosammlungen (PuzzleAvatar). Darüber hinaus rahmen TeCH und PuzzleAvatar die Rekonstruktionsaufgabe als eine bedingte Generationsaufgabe, die die Bereiche 3D-Rekonstruktion und 3D-Generierung in ein einheitliches Rahmenwerk vereint.

ACKNOWLEDGEMENTS

"I once believed a PhD was essential for my dream life, but now I see it is the dream itself."

Looking back on my five-year Ph.D. journey from the US to Germany, those words truly resonate. It's time to awaken from this remarkable "dream" and fondly bid farewell to the incredible individuals and groups with whom I've shared unforgettable moments.

Advisors. I want to express my deepest gratitude to my Ph.D. supervisors, Prof. Michael J. Black, Prof. Dimitris Tzionas, and Prof. Hao Li, for giving me the incredible opportunity to pursue my Ph.D. with two of the world's leading avatar research teams (USC-VGL, and MPI-IS). **Michael**, as my research godfather, I've learned from you not only how to conduct professional research that withstands cycles, but also how to be a respected and trustworthy individual. When I start supervising others, I often prompt myself, "What would Michael do in this tricky situation?" It might not always yield the perfect answer, but it comes close. I believe we'll see each other often in the future, and I'll share your story with all my students – of a dedicated scholar and a sincere person who truly makes a significant impact on the world. **Dimitris**, I've become accustomed to your Greek accent! Your insightful critiques often prove to be the deciding factor in the project, consistently keeping our work on track and of high quality. I've learned the importance of details from you, and that "perfect" isn't in your vocabulary. Over the past four years, your invaluable emotional support and guidance in relationship management, communication, idea presentation, and conflict resolution have been essential. I admire your patience and see you as an excellent role model who enriches my journey by demonstrating how to be a thoughtful advisor without sacrificing productivity. **Hao**, I chose and left USC because of you. Your departure from USC felt like my world collapsed. I understand your decision, but starting over is incredibly tough as a first-year PhD student. Reflecting on it now, this decision provided me with a unique experience. Your passion and unconventional approach taught me the value of uniqueness and standing out. If I'm not living life to the fullest, then I'm not truly living. I aim to keep this mindset until I'm 50 – to remain vibrant, imaginative, move fast, and break things. From Michael, I've learned something quite different: certain things in this world are constant, and enduring success demands stability and depth. Only by concentrating and delving deeper, we can discover true worth. As I age, I aspire to be like him – down-to-earth and maintain a slim figure with healthy lifestyles.

Collaborators. I'm deeply grateful to Shunsuke Saito and Jinlong Yang for guiding me into the world of digital humans at the start of my PhD. They laid the foundation for my research and sparked a passion that has shaped my entire journey.

Weiyang, your in-depth insights have broadened my perspectives and sparked conversations about humanity's future. I wouldn't have pursued faculty positions without your encouragement. Each discussion was enlightening, and you always offered solutions to my doubts. It's been a privilege having you as a collaborator, role model, office mate, and dear friend during my PhD. To Yao, Yandong, and Zhen – your presence made this journey vibrant and joyful. Yao, your brilliance is unmatched; Yandong, you're less nerdy than you seem; and Zhen, despite your relentless awkward humor, working with you is always delightful. My PhD experience wouldn't have been so colorful without each of you. Thank you for your sincere advice and steadfast support during tough times. Additionally, Yandong and Anpei, see you in Westlake! What could be more exciting than collaborating with like-minded friends who share excellent academic taste?

I extend my heartfelt thanks to Yangyi Huang, Tingting Liao, Siyuan Bian, Xueting Yang, and Daiheng Gao. Working with such talented and self-motivated students has been an honor, especially as you welcomed me into my first advisory roles. Our delightful collaboration has significantly influenced my decision to remain in academia.

Special thanks to Xinxin Zhang, Chongyang Ma, Jun Xing, Zhengyi Luo, Ailing Zeng, Jingbo Wang, Tim Z. Xiao, Xu Chen, Songyou Peng, Zehao Yu, Gengshan Yang, Ruihan Gao, Xiaoyu Xiang, Soyong Shin, Yong-lu Li, Wenqiang Xu, and many other brilliant minds who have brainstormed with me and shared their unique thoughts without hesitation. We had many fruitful technical discussions, even though some of you didn't collaborate with me directly. I look forward to future opportunities together.

My internships were unforgettable thanks to the incredible support from collaborators like Shunsuke Saito, Timur Bagautdinov, Rawal Khirodkar, Yash Kant, Chen Guo, and Junxuan Li at Meta Reality Labs Research. The show must go on.

Lastly, immense gratitude to Xu Cao, Yufei Ye, Justus Thies, Hongwei Yi, Jiaxiang Tang, Ruilong Li, Kyle Olszewski, Zeng Huang, Xiaoguang Han, Lingteng Qiu, Chongjie Ye, Lixin Yang, Kailin Li, Hao-shu Fang, Jiefeng Li, Yihao Luo, Yuxuan Xue, Longhui Yu, Haven Feng, Zeju Qiu, Chenghao Xu, and many other brilliant collaborators. You are the reason my collaborative spirit thrives. One can go fast, but together we go far.

Colleagues, and Friends. Reflecting on my four-year PhD journey, I've realized that research is just part of the story. I couldn't have made it this far without the incredible support from colleagues and dear friends.

At MPI-IS, I'm deeply grateful to my peers: Peter Kulits, Artur Grigorev, Vanessa Sklyarova, Markos Diomataris, Sai Kumar Dwivedi, Shashank Tripathi, Nikos Athanasiou, Kiran Chhatre, Mirela Ostrek, Marilyn Keller, Muhammad Kocabas, Maria Paola Forte, Soubhik Sanyal, Ahmed Osman, Omid Taheri, Lea Müller, Nadine Rüegg, Mohamed

Hassan, Vassilis Choutas, Partha Ghosh, and Victoria Fernandez Abrevaya. I'll miss our coffee breaks and engaging conversations. Beyond academics at MPI-IS are incredible colleagues who supported us PhD students: Melanie Feldhofer, Nicole Overbaugh, Tsvetelina Alexiadis, Giorgio Becherini, Taylor Obersat, Priyanka Patel, Benjamin Pellkofer. Your presence made my time in Tuebingen vibrant and unforgettable.

Special thanks to the members of so-called "secret Chinese dinner party" from MPI, ETH Zürich, and Tuebingen: Qianli Ma, Yinghao Huang, Yufeng Zheng, Zicong Fan, Le Chen, Gege Gao, Yamei Chen, Siyuan Guo, Anpei Chen, Haolong Li, Tairan Yin, and Haoran Yun, Xianghui Xie, Xiaohan Zhang, Haofei Xu, Haven Huang, Shaofei Wang, Chuqiao Li, Yannan He, Huanbo Sun, Jiduan Wu, Guanhua Zhang, and Jiahua Xu. The countless brainstorming sessions, inspiring discussions, coffee breaks, late-night drinks, all-day brainstorms, and ski retreats have created memories I'll cherish forever. Tuebingen has become more than just a village; it's a place where I've formed lasting bonds and will repeatedly appear in my future sweet dreams.

I'm also grateful to my friends at USC: Tianye Li, Jiaman Li, Zhengfei Kuang, Xinglei Ren, Jing Yang, Yuming Gu, Mingming He, Hanyuan Xiao, Yajie Zhao, Yi Zhou, Yunhao Ge, Sidi Lv, and many cool Trojans. I truly enjoyed all the Avalon games, lunches, trips, and hikes we shared. Outside of Tuebingen, my life would have felt incomplete without incredible friends from Zurich, Saarbrücken, France, the US, and China. The list is too long to name everyone, but you are in my mind, and I cherish all the precious moments we shared. I sincerely wish you all the best.

Families. Pursuing a Ph.D. felt like my battle, but it wasn't just mine. Over the past five years, so much has happened at home. You shielded me so I could study in peace, respecting my stubborn plans and always trusting my judgment. Your constant positive feedback meant a lot, and I know the emotional effort it took. Grandpa, grandma, and my maternal grandfather have passed away. Not seeing them one last time is a pain I'll carry forever. I often meet them in dreams, and I hope I've made them proud as they watch from above. To Kexin, my deepest love. We met during COVID-19 – *Love in the Time of Cholera*, and our wedding is this year. Amidst the world's noise, our time together brings true peace. Traveling with you has turned my life from black and white to colorful. Let's move forward to the next chapter.

The world has given me so much; now it's time to give in return.

Yuliang Xiu
Tübingen, Germany, March 18, 2025

List of Papers

I. *ICON: Implicit Clothed humans Obtained from Normals.* [229]
Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, Michael J. Black.
Conference on Computer Vision and Pattern Recognition (CVPR), **2022**.

II. *ECON: Explicit Clothed humans Optimized via Normal integration.* [228]
Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, Michael J. Black.
Conference on Computer Vision and Pattern Recognition (CVPR), **2023**.

III. *TeCH: Text-guided Reconstruction of Lifelike Clothed Humans.* [78]
Yangyi Huang*, **Yuliang Xiu***, Hongwei Yi*, Tingting Liao, Jiaxiang Tang, Deng Cai, Justus Thies.
International Conference on 3D Vision (3DV), **2024**.

IV. *PuzzleAvatar: Assembling 3D Avatars from Personal Albums.* [230]
Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, Michael J. Black.
ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia), **2024**.

Papers not included in the thesis:

V. *DART: Articulated Hand Model with Diverse Accessories and Rich Textures.* [53]
Daiheng Gao*, **Yuliang Xiu***, Kailin Li*, Lixin Yang*, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, Ping Tan.
Neural Information Processing Systems (NeurIPS, Datasets and Benchmarks Track), **2022**.

VI. *AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time.* [46]
Hao-shu Fang*, Jiefeng Li*, Hongyang Tang, Chao Xu, Haoyi Zhu, **Yuliang Xiu**, Yong-lu Li, Cewu Lu.
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), **2022**.

VII. *High-Fidelity Clothed Avatar Reconstruction from a Single Image.* [131]
Tingting Liao, Xiaomei Zhang, **Yuliang Xiu**, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, Zhen Lei
Conference on Computer Vision and Pattern Recognition (CVPR), **2023**.

VIII. *D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field.* [235]
Xueting Yang*, Yihao Luo*, **Yuliang Xiu**, Wei Wang, Hao Xu, Zhaoxin Fan
International Conference on Computer Vision(ICCV), **2023**.

IX. *TADA! Text to Animatable Digital Avatars.* [130]
Tingting Liao*, Hongwei Yi*, **Yuliang Xiu**, Jiaxiang Tang, Yangyi Huang, Justus Thies, Michael J. Black.
International Conference on 3D Vision (3DV), **2024**.

X. *Ghost on the Shell: An Expressive Representation of General 3D Shapes.* [138]
Zhen Liu, Yao Feng†, **Yuliang Xiu**†, Weiyang Liu, Liam Paull, Michael J. Black, Bernhard Schölkopf.
International Conference on Learning Representations (ICLR), **2024**.

XI. *Parameter-Efficient Orthogonal Finetuning via Butterfly Factorization.* [137]
Weiyang Liu*, Zeju Qiu*, Yao Feng†, **Yuliang Xiu**†, Yuxuan Xue†, Longhui Yu†, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, Bernhard Schölkopf.
International Conference on Learning Representations (ICLR), **2024**.

XII. *StableNormal: Reducing Diffusion Variance for Stable and Sharp Normal.* [237]
Chongjie Ye*, Lingteng Qiu*, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Leifeng Bo, **Yuliang Xiu**‡, Xiaoguang Han‡.
ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia), **2024**.

CONTRIBUTION REPORT

All the papers listed can be classified as three groups: 1) *Author projects*: I, II, III, IV, 2) *Supervised projects*: III, V, VII, VIII, IX, XII and 3) *Collaborated projects*: VI, X, XI. Notably, project III is primarily supervised by the author of this thesis, particularly contributing to core method design, experiment protocol, result analysis, story formulation, and writing. My efforts among the papers not included in the thesis, mainly involve core (V,XII) or non-core (VII,VIII,IX) technical novelty contributions, experiment design or evaluation protocol (V,VII,VIII,IX,XII), implementing baselines (X, XI), formulating the story (V,VIII,XII), results analysis (V,VII,VIII,IX,X,XI,XII), paper writing (V,VIII,IX,X,XI,XII), and rebuttal (ALL).

CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURE

The next list describes several symbols that will be later used within the body of the document.

Abbreviations

Here we list the frequently used abbreviations in the thesis.

BiNI	Bilateral Normal Integration
d-BiNI	Depth-aware Bilateral Normal Integration
$\mathcal{DR}$	Differetial Renderer
Poisson	Poisson surface reconstruction
AMT	Amazon Mechanical Turk
AR	Augmented Reality
CAPE-FP	CAPE Fashion Poses
CAPE-NFP	CAPE Non-Fashion Poses
CNN	Convolutional Neural Network
DoF	Degrees of Freedom
GHUM	Generative 3D Human Shape and Articulated Pose Models (moderate-resolution)
GHUML(ite)	Generative 3D Human Shape and Articulated Pose Models (low-resolution)
GPU	Graphics Processing Unit
GT	Ground-Truth
HMR	Human Mesh Recovery
LBS	Linear Blend Skinning
MANO	Hand Model with Articulated and Non-rigid Deformations
MC	Marching Cubes
MLP	Multilayer perceptron
MoCap	Motion Capture
MR	Mixed Reality
OOD	Out-of-Distribution
SCAPE	Shape Completion and Animation of People
SDF	Signed Distance Field
SDS	Score Distillation Sampling (SDS)
SMPL	Skinned Multi-Person Linear model
SMPL(-X)	SMPL or SMPL-X
SMPL+H	SMPL+Hands
SMPL-X	SMPL eXpressive
SOTA	State-of-the-art
STAR	Sparse Trained Articulated Human Body Regressor
VPoser	Variational Human Pose Prior

VR ..... Virtual Reality

Math Symbols

Here we list the mathematical symbols used in the thesis.

β ..... Body shape parameters
$\boldsymbol{E}$ ..... Energy function
ψ ..... Facial expression parameters
θ ..... Body Pose parameters
γ ..... Body translation parameters
$J(\beta)$ ..... 3D joints of the SMPLX kinematic skeleton

Chapter 3

$\mathcal{M}^{\text{b}}$ ..... Mesh of body
$\mathcal{F}_{\text{n}}^{\text{b}}$ ..... Extracted feature from 3D body normal field
$\mathcal{N}^{\text{b}}$ ..... Body normal maps (front+back)
$\mathcal{F}_{\text{s}}$ ..... Signed Distance Field of Body
$\mathcal{F}_{\text{v}}$ ..... Visibility Field of Body
$\mathcal{F}_{\text{n}}^{\text{c}}$ ..... Extracted feature from 2D clothed normal maps
$\mathcal{N}^{\text{c}}$ ..... Clothed normal map groundtruth (front+back)
$\mathcal{IF}$ ..... Implicit Function
$\mathcal{G}^{\text{N}}$ ..... Normal map estimator
$\mathcal{F}_{\text{P}}$ ..... Per-point feature of ICON
$\widehat{\mathcal{N}}^{\text{c}}$ ..... Clothed normal map predictions (front+back)
$\mathcal{R}$ ..... Clothed human reconstruction

Chapter 4

$\widehat{\mathcal{Z}}_{\{\text{F,B}\}}^{\text{c}}$ or $\widehat{\mathcal{Z}}^{\text{c}}$ .. d-BiNI clothed depth maps (front+back)
$\mathcal{M}_{\{\text{F,B}\}}^{\text{d-BiNI}}$ or $\mathcal{M}^{\text{d-BiNI}}$ d-BiNI mesh surfaces (front+back)
$\mathcal{Z}_{\{\text{F,B}\}}^{\text{b}}$ or $\mathcal{Z}^{\text{b}}$ .. Body depth maps (front+back)
$\mathcal{M}^{\text{cull}}$ ..... Remaining body mesh after visibility-aware culling
$\mathcal{R}_{\text{IF}}^{\text{cull}}$ ..... Remaining clothed mesh after visibility-based culling
$\mathcal{Z}_{\{\text{F,B}\}}^{\text{c}}$ or $\mathcal{Z}^{\text{c}}$. Ground-truth cloth depth maps (front+back)
$\mathcal{M}^{\text{f+h}}$ ..... Hand and face regions of body mesh
$\mathcal{R}_{\text{IF}}$ ..... Clothed mesh reconstructed via Implicit Function

Chapter 5

J_{est} ..... Estimated 3D joints
TeCH_{db} ..... TeCH w/ only DreamBooth prompte
TeCH_{vqa} ..... TeCH w/ only VQA prompt

Units

Here we list the units.

$^\circ$	Degrees
FPS	Frames per Second
m	Meter
mm	Millimeter

1

INTRODUCTION

Contents

1.1 Motivation

Human beings, as the bearers of intellect, shape the world to our will. If we all agree that, it is worth pursuing AGI (Artificial General Intelligence) — where agents learn to accomplish any intellectual task on par or even better than humans — then *reproducing a real human in the virtual space* could be the crucial first step to "approximate" it. This endeavor enables us to truly understand the nature of humans, and grasp the essence of intelligence, which will benefit various fields, and ultimately improve our lives.

As Feynman famously stated — *What I cannot create, I do not understand*. Vice versa, our understanding of human nature grows, as we enhance the realism of virtual humans we create. This carries **scientific significance**, as if offers a working model that mirrors our physiological structure, allowing for the analysis of medical indicators and exploration of ergonomics and biomechanics. Furthermore, integrating digital twins into a simulation environment can aid in investigating the optimal human-robot

interaction (*e.g.*, Caregiving robots (RCareWorld [*Cornell*], Assistive Gym [*Stanford*])), and inspiring the design of humanoid robots. On the other hand, simulating realistic-looking virtual beings has already facilitated numerous **practical applications**, such as the Metaverse Teleportation Platform (*e.g.*, Codec Avatar [*Meta*], Starline [*Google*]), AI-Generated Digital Content Creation, shorten as AIGC (*e.g.*, Omniverse [*NVIDIA*], Luma AI [*Berkeley*]), Robot-assisted Surgery (*e.g.*, STAR [*Johns Hopkins*]), and Human-centric Simulators for Autonomous Driving (*e.g.*, DRIVE Sim [*NVIDIA*]), covering a wide range of AI products.

1.2 Problem Statement

Challenges of Human Digitization. Compared to text or images, the level of digitization for the 3D world is relatively low due to the expensive and laborious process of capturing and synthesizing it. The capture of high-quality 3D human scans, in particular, requires costly 4D scanners or light stages. Similarly, the process of synthesizing realistic-looking avatars, such as MetaHumans [i], can be time-consuming. It takes months to rig the face and body or model intricate details like facial features and clothing wrinkles. These barriers considerably limit the potential size of the target user base. However, if there is a consensus that democratizing these techniques and making them accessible to everyone is worthwhile, then it becomes necessary to develop a scalable and affordable method for digitizing human beings.

Scalable Human Digitization from Pixels and Texts. Human Digitization refers to the process of converting aspects of a person's physical existence into a digital form. Specifically, the main focus of this thesis is on capturing and representing lifelike clothed humans from pixels (*e.g.*, image, video) and texts (*e.g.*, textural description). To be scalable, the reconstruction algorithms must generalize well in the wild, regardless of the variations in human pose, body shape, clothing, capturing view, occlusion, or cropping. Ideal reconstruction algorithms ought to be applicable to unconstrained photo collections without the need for additional off-the-shelf estimators. High-fidelity textured 3D humans are the ultimate goal, which should feature detailed geometry and high-quality textures for photorealistic rendering and natural clothing dynamics during animation.

[i] High-fidelity digital humans made easy, Unreal Engine

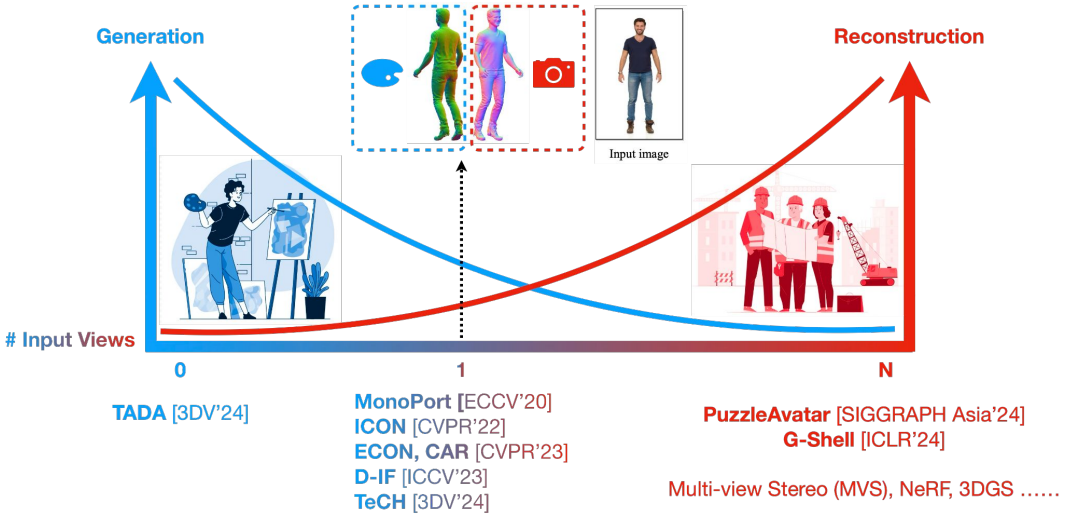


Figure 1.1: Overview of Thesis. X-axis refers to the number of input images.

1.3 Thesis Statement

To achieve this goal, the thesis first discusses **"Image-based pixel-aligned 3D clothed humans reconstruction"** in unconstrained poses (ICON in Chapter 3,), with loose clothing (ECON in Chapter 4), and at real-time speed (MonoPort [122]). However, due to the limited perspective provided by the monocular input, non-visible regions often have overly smooth surfaces with blurry textures. Given that **"Reconstruction can be seen as a form of conditional generation"** (see Fig. 1.1), could we use the imagination capacity of the diffusion-based generative model, *e.g.*, Stable Diffusion (SD) [181], to enhance the realism of unseen regions, under the guidance of visible pixels from the frontal view? TeCH [78] (Chapter 5) is our exploration in this direction. Focusing on the intersection of reconstruction and generation, TeCH significantly improves the fine details of non-visible regions, both in geometry and texture, guided by prompts derived from a single input frontal image.

TeCH is further extended into a more unconstrained setting — personal "OOTD" (Outfit Of The Day) photo collections. The challenge is that these casual photo collections contain diverse poses, challenging viewpoints, cropped views, and occlusion, making the pose estimation of cameras and body articulations difficult. We address this novel **"Album2Human"** task by developing PuzzleAvatar (Chapter 6), a novel paradigm that generates a 3D avatar (in a canonical pose), with faithful appearance and shape, from a personal OOTD album, bypassing the challenging estimation of body and camera pose.

1.3.1 ICON: Reconstructing Challenging Poses

Figure 1.2: Video to animatable clothed avatars with textures. ICON robustly reconstructs 3D clothed humans in unconstrained poses from individual video frames (Left). These are used to learn a fully textured and animatable clothed avatar with realistic clothing deformations (Right).

ICON's goal is to learn an avatar from only 2D images of people in *unconstrained* poses. Given a set of images, ICON estimates a detailed 3D surface from each image and then combines these into an animatable avatar. Implicit functions are well suited to the first task, as they can capture details like hair and clothes. Current methods, however, are not robust to varied human poses and often produce 3D surfaces with broken or disembodied limbs, missing details, or non-human shapes. The problem is that these methods use global feature encoders that are sensitive to global pose. ICON, instead, uses local features. ICON has two main modules, both of which exploit the SMPL-X body model. First, ICON infers detailed clothed-human normals (front/back) conditioned on the SMPL-X normals. Second, a visibility-aware implicit surface regressor produces an iso-surface of a human occupancy field. Importantly, at inference time, a feedback loop alternates between refining the SMPL-X mesh using the inferred clothed normals and then refining the normals. With multiple reconstructed frames in varied poses, an animatable avatar could be learned through SCANimate [187], see Fig. 1.2. Evaluation on the AGORA [164] and CAPE [144] datasets shows that ICON outperforms the state-of-the-art (SOTA) in reconstruction, even with heavily limited training data. Additionally, it is much more robust to out-of-distribution (OOD) samples, *e.g.*, in-the-wild poses/images and out-of-frame cropping. ICON takes a solid step towards robust 3D clothed human reconstruction from in-the-wild images. This enables avatar creation directly from video with personalized pose-dependent cloth deformation. See homepage at `icon.is.tue.mpg.de`.

1.3.2 ECON: Reconstructing Loose Clothing

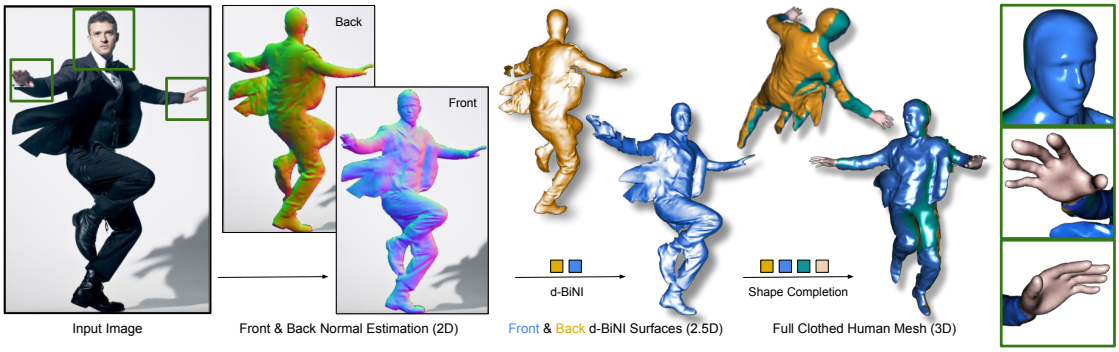


Figure 1.3: ECON reconstructs loose clothing. ECON combines the best aspects of free-form implicit representation, and explicit anthropomorphic regularization. It reconstructs 3D humans in three steps: (1) It infers detailed 2D normal maps for the front and back side (Sec. 4.2.1). (2) The normal maps are converted into detailed, yet incomplete, 2.5D **front** and **back** surfaces guided by a SMPL-X estimate (Sec. 4.2.2). (3) It then "inpaints" the **missing geometry** between two surfaces (Sec. 4.2.3). **Face or hands** can be optionally replaced with the parts from SMPL-X.

The use of an explicit parametric body model improves pose robustness but constrains topological flexibility, causing ICON to produce degenerate shapes for humans in loose clothing. Meanwhile, previous approaches based solely on implicit functions (IF), such as PIFuHD [185, 186], do not generalize well to challenging poses. What we want is a method that combines the best properties of implicit representation and explicit body regularization. To this end, we make two key observations: (1) current networks are better at inferring detailed 2D maps than full-3D surfaces, and (2) a parametric model can be seen as a "canvas" for stitching together detailed surface patches. Based on these, ECON, has three main steps, as Fig. 1.3 shows: (1) It infers detailed 2D normal maps for the front and back side of a clothed person. (2) From these, it recovers 2.5D front and back surfaces, called d-BiNI, that are equally detailed, yet incomplete, and registers these w.r.t. each other with the help of a SMPL-X body mesh recovered from the image. (3) It "inpaints" the missing geometry between d-BiNI surfaces. If the face and hands are noisy, they can optionally be replaced with the ones of SMPL-X. As a result, ECON infers high-fidelity 3D humans even in loose clothes and challenging poses. This goes beyond previous methods, according to the quantitative evaluation on the CAPE and RenderPeople [179] datasets. Perceptual studies also show that ECON's perceived realism significantly outperforms SOTA methods. See homepage at econ.is.tue.mpg.de

1.3.3 TeCH: Reconstructing Invisible from the Visible

Figure 1.4: TeCH generates avatars with all-around details. A 3D avatar with "all-around details" should have 1) a detailed full-body geometry, including facial features and clothing wrinkles, in both frontal and unseen regions, and 2) a high-quality texture with consistent color and detailed texture patterns. The key insight is to guide the reconstruction using a personalized Text-to-Image (T2I), and descriptive prompts derived via visual questioning answering (VQA).

ECON generalizes to both unseen poses and loose clothing. However, accurately restoring the "unseen regions" with high-level details remains an unsolved challenge. Given a frontal image, existing methods often generate overly smooth back-side surfaces with a blurry texture. But how can all visual attributes of an individual be effectively captured from a single image? Motivated by the power of foundation models, TeCH reconstructs the 3D human by leveraging 1) descriptive prompts (*e.g.* garments, colors, hairstyles) which are automatically generated via a garment parsing model and Visual Question Answering (VQA), 2) a personalized fine-tuned Text-to-Image diffusion model (T2I) which learns the "indescribable" appearance. To represent high-resolution 3D clothed humans at an affordable cost, TeCH presents a hybrid 3D representation based on DMTet [55, 190], which consists of an explicit body shape grid and an implicit distance field. Guided by the descriptive prompts + personalized T2I model, the geometry and texture of the 3D humans are optimized through multi-view Score Distillation Sampling (SDS) [169] and reconstruction losses based on the original observation. TeCH produces 3D clothed humans with consistent and detailed texture, and high-quality clothed human geometry, see Fig. 1.4. Quantitative and qualitative experiments demonstrate that TeCH outperforms the SOTA in both shape and color quality. See homepage at huangyangyi.github.io/TeCH

1.3.4 PuzzleAvatar: Humans from Casual Photos

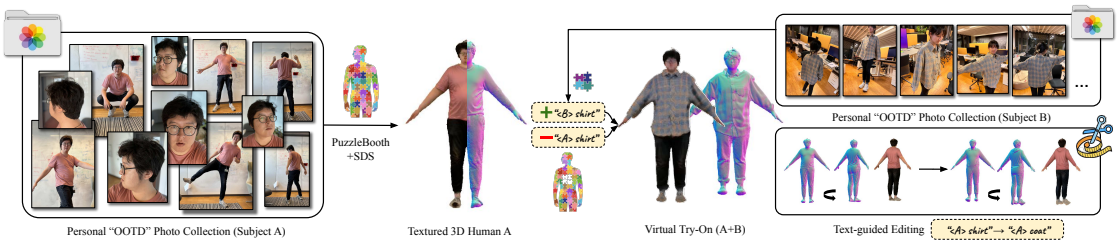


Figure 1.5: PuzzleAvatar reconstructs avatars from peronal album. It takes as input a set of "OOTD" (Outfit Of The Day) personal photos with unconstrained body poses, camera poses, framing, lighting and backgrounds, albeit with a consistent outfit and hairstyle. All these consistent factors are learned as separate unique tokens `<asset X>` in a compositional manner, like pieces of a puzzle. PuzzleAvatar allows one to easily inter-change tokens for downstream tasks, such as for customizing avatars and performing virtual try-on while preserving identity.

Generating **personalized** 3D avatars is crucial for AR/VR. However, recent text-to-3D methods that generate avatars for celebrities or fictional characters, struggle with everyday people. Methods for faithful reconstruction (*e.g.*, ICON, ECON, TeCH, *etc.*) typically require full-body images in controlled settings. What if users could just upload their personal "OOTD" (Outfit Of The Day) photo collection and get a faithful avatar in return? We call this novel task "Album2Human". The challenge is that such casual photo collections contain diverse poses, challenging viewpoints, cropped views, and occlusion (albeit with a consistent outfit, accessories and hairstyle). We address this task by developing **PuzzleAvatar**, a novel model that generates a faithful 3D avatar (in a canonical pose) from a personal OOTD album, bypassing the challenging estimation of body and camera pose. To this end, we fine-tune a foundational vision-language model (VLM) on such photos, encoding the appearance, identity, garments, hairstyles, and accessories of a person into separate learned tokens, instilling these cues into the VLM. In effect, we exploit the learned tokens as "puzzle pieces" from which we assemble a faithful, personalized 3D avatar. Importantly, we can customize avatars by simply inter-changing tokens, see Fig. 1.5. As a benchmark for this new task, we create a new dataset, called **PuzzleIOI**, with 41 subjects in a total of nearly 1k OOTD configurations, in challenging partial photos with paired ground-truth 3D bodies. Evaluation shows that PuzzleAvatar not only has high reconstruction accuracy, outperforming TeCH and MV-DreamBooth [192], but also a unique scalability to album photos, and has demonstrated strong robustness. See homepage at puzzleavatar.is.tue.mpg.de

1.4 Thesis Organization

This thesis is organized into several chapters, each focusing on a distinct aspect of reconstructing 3D clothed humans from images. The structure of the thesis is as follows.

- **Chapter 3: ICON** robustly reconstructs 3D clothed humans in challenging poses from a single image. It relies on two key strategies: (1) Regularizing the solution with a 3D body model and iteratively optimizing it. (2) Using local features to avoid spurious correlations with global pose. This chapter details its architecture, training process, and results, demonstrating generalization to unseen poses even with limited training data. Ablation studies confirm these design choices. The results are sufficient to create 3D avatars from monocular images or videos with challenging poses like parkour, martial arts, dance, and sports, featuring personalized and natural cloth dynamics.

- **Chapter 4: ECON** merges the strengths of implicit and explicit representations to handle unseen poses and loose garments. Its key component, d-BiNI optimization, simplifies the reconstruction process into a "sandwich-like" paradigm: integrating dual normals for front and back surfaces, then inpainting the missing parts as a shape completion process. This allows ECON to recover loose clothing without being constrained by body topology (unlike ICON) and to perform well on multi-person images with occlusions. Qualitative, quantitative, and perceptual studies demonstrate its effectiveness across various benchmarks.

- **Chapter 5: TeCH** is the first attempt at treating image-based reconstruction as a conditional generation task, using both the input image and its derived descriptions to generate the non-visible backside while maintaining overall consistency. The method section details how the descriptive prompt is built with descriptions (VQA) and indescribable tokens (DreamBooth), following a multi-view SDS process to create the final textured 3D model and ensure front-back consistency. Experiments validate TeCH's advantages over existing baselines, in producing detailed geometry together with consistent textures.

- **Chapter 6: PuzzleAvatar** introduces the problem of reconstructing articulated humans from personal, natural, and unconstrained photo collections – introducing the new "Album2Human" task. Meanwhile, PuzzleIOI offers a new benchmark that facilitates objective evaluation of puzzle-like 3D human reconstruction, and the analysis of design choices (*e.g.*, model customization, distilling sampling). Detailed evaluations and ablation studies show PuzzleAvatar's effectiveness and scalability. The application section also demonstrates modular tokens and text guidance through character editing and virtual try-on examples.

- **Chapter 7: Conclusion**: This final chapter summarizes the key findings and contributions of the thesis, highlighting novel approaches to reconstruct 3D clothed humans from pixels. We reflect on the implications of the work, discuss potential future directions, and conclude with a review of the thesis's broader impact on democratizing and scaling up 3D human digitization.

- **Appendices**: The appendices provide supplementary material that supports the main thesis content. This includes mathematical derivations, additional experimental results, technical specifications, and other relevant information that offers further insights into the methods and concepts discussed in the main chapters.

2

BACKGROUND

Contents

In this chapter, we provide a background overview of the research fields related to this thesis. We first discuss the development of 3D clothed human reconstruction methods in Sec. 2.1. The overview is categorized by 3D representations (*i.e.*, explicit, implicit, hybrid), along with a discussion on pose-free in-the-wild reconstructions. Then, in Sec. 2.2, we explore the development of generative modeling of 3D clothed humans, focusing on their training data. These works can be roughly grouped into generative models trained on 3D human data, 2D images, and image-text pairs. Advances in Vision-Language Foundation Models, like diffusion models, offer a new paradigm shift in avatar generation, making creation more controllable through text descriptions.

2.1 3D Clothed Human Reconstruction

2.1.1 Explicit-shape-based Reconstruction Methods

Explicit-shape-based approaches use either a mesh-based parametric body model [90, 140, 165, 182, 231], non-parametric depth maps [52, 194], normal maps [69, 102, 117], or point cloud [243], to reconstruct 3D humans. Among these approaches, the use of mesh-based parametric body models, such as SMPL(-X) [140, 165] and GHUM [231], to regress the underlying minimally clothed body from images or videos is a well-explored problem often referred to as "Human Mesh Recovery (HMR)" [35, 47, 92, 103–105, 107, 112, 113, 126, 194, 200, 206, 239, 247, 248]. Other work estimates clothed humans, instead, by modeling clothing geometry as 3D offsets on top of body geometry [3–6, 111, 144, 168, 223, 261]. This approach is favored not only because such models capture the statistics across a human population, but also because meshes are compatible with standard graphics pipelines. The resulting clothed 3D humans can be easily animated, as they naturally inherit the skeleton and surface skinning weights from the underlying statistical body model. Another advantage of mesh-based statistical models is that texture information can be easily accumulated through multi-view images or image sequences [5, 15], due to their consistent mesh topology.

An important limitation, though, is that the "body+offset" approach lacks the topological flexibility needed to model loose clothing that significantly deviates from body topology, such as dresses and skirts. To address this, some methods [15, 87] first identify clothing types from the input image, and then perform clothing-aware 3D reconstruction. However, scaling up these "cloth-aware" approaches to diverse clothing styles is challenging, which limits their generalization to variations in real-world outfits. Consequently, the estimated meshes often misalign with the pixels in the input images. To summarize, the 3D template prior regularizes the overall human shape but also introduces topological constraints.

Another line of template-free work involves using a "sandwich-like" monocular structure for 3D reconstruction, represented by Moduling Humans [52], FACSIMILE [194] and Any-Shot GIN [224]. Moduling Humans has two networks: a *generator* that

estimates the visible (front) and invisible (back) depth maps from RGB images, and a *discriminator* that helps regularize the estimation via an adversarial loss. FACSIMILE further improves the geometric details by leveraging a normal loss, which is directly computed from depth estimates via differentiable layers. Recently, Any-Shot GIN generalizes the sandwich-like scheme to novel classes of objects. Given RGB images, it predicts front and back depth maps as well, and then exploits IF-Nets [33] for shape completion. In Chapter 4 of this thesis, we introduce ECON, which follows a similar "sandwich-like" path and extends it with body template, to successfully reconstruct clothed human shapes with robust pose generalization and better geometric details from front+back normal predictions.

2.1.2 Implicit-function-based Reconstruction Methods

Unlike meshes, deep implicit functions [31, 148, 162] can represent detailed 3D shapes with arbitrary topology, such as open jackets and loose skirts, and have no resolution limitations. PIFu [185] introduces deep implicit functions for clothed 3D human reconstruction from RGB images, and later PIFuHD [186] significantly improves 3D geometric details with a multi-level architecture and normal maps predicted from the RGB image. The estimated shapes align well to image pixels. MonoPort [121, 122] speeds up inference through an efficient volumetric sampling scheme. PHORHUM [7] additionally decomposes the albedo and global illumination.

A limitation of the above methods is that the estimated 3D humans cannot be reposed because implicit shapes (unlike statistical models) lack a consistent mesh topology, a skeleton, and skinning weights. To address this, NeuralGraph [19] infers an embedded deformation graph to manipulate implicit functions, while S3 [236] also infers a skeleton and skinning fields. However, their shape reconstruction still lacks shape regularization and does not fully utilize knowledge of the human body structure. Consequently, these methods tend to overfit to the body poses in the training data, such as fashion poses, and fail to generalize to novel poses, resulting in issues like broken or disembodied limbs, missing details, or geometric noise. To address these issues, several methods introduce different geometric priors to regularize the deep implicit representation, which we refer to as "hybrid-based reconstruction methods".

2.1.3 Hybrid-based Reconstruction Methods

This direction often combines both implicit representations with parametric body models, such as SMPL [140], SMPL-X [165], and GHUM [231], which represent human body shape well, model the kinematic structure of the body, and can be reliably estimated from RGB images of clothed people. Such a hybrid representation can be viewed as a base explicit shape upon which to model expressive clothing implicitly, to get the best of both worlds. PaMIR [256], DIF [235], and DeepMultiCap [255] condition the pixel-aligned features on a posed and voxelized SMPL mesh. While S3F [39] conditions on GHUM, and additionally decomposes the albedo and global illumination. JIFF introduces a 3DMM face prior to improve the realism of the facial region. However, these methods are sensitive to global pose, due to the spatial correlation in their 2D or 3D convolutional encoders. Thus, when training data has limited pose variation, they struggle with out-of-distribution poses on in-the-wild images. Instead, ARCH [80], ARCH++ [68] and CAR [131] use SMPL to unpose the pixel-aligned query points from a posed space to a canonical space. However, training these models requires unposing scans into a canonical pose using an accurately fitted body model; inaccuracies in this process can cause artifacts. Moreover, unposing clothed scans with the skinning weights of an "undressed" model can alter shape details. Non-parametric 3D priors are also introduced to regularize the clothed human shape. GeoPIFu [67] introduces a coarse shape of volumetric humans, Self-Portraits [124], PINA [45], and S3 [236] use depth or LIDAR information to regularize shape and improve robustness to pose variation. Their primary limitation lies in the lack of cross-view information. Additionally, some works attempt to leverage the garment generative models as the garment prior. SMPLicit [38], ClothWild [151], DIG [119] and GarmentRecovery *et al*. [118] learn a generative clothing model with neural distance fields [34, 148, 162] from 3D clothing datasets. Given an image, the clothed human is reconstructed by estimating the body and then optimizing the latent space of the clothing model. However, the results usually do not align well with the image and lack geometric detail, due to the limited flexibility, diversity, and capacity of the generative garment models. There is a separate line of research that focuses on optimizing neural radiance fields (NeRF) [150] from a single image. SHERF [75] and ELICIT [77] optimize a generalized human NeRF, incorporating

SMPL-X. While SHERF complements missing information from partial 2D observations, ELICIT leverages appearance prior from CLIP [174].

Sharing the same spirit, our ICON (Chapter 3) combines the statistical body model SMPL with an implicit function, to reconstruct clothed 3D human shape from a single RGB image. What sets ICON apart is that, SMPL not only guides ICON's estimation, but is also optimized "in the loop" during inference to enhance its pose accuracy. Instead of relying on the global body features, ICON exploits local body features that are agnostic to global pose variations. As a result, even when trained on heavily limited data, ICON is still robust to out-of-distribution poses.

2.1.4 Pose-Free in-the-wild Reconstruction

The term "pose" refers not only to camera pose but also to body articulation. Camera pose plays a crucial role in 3D reconstruction, as it "anchors" 3D geometry onto 2D images; however, estimating it for in-the-wild images is highly challenging. Thus, to account for camera estimation errors, some work leverages joint optimization between the object and camera [133, 217, 221], off-the-shelf geometric cue estimates [16, 51, 149], or learning-based camera estimation [211, 213, 250]. Body pose is also difficult to accurately estimate from in-the-wild images and is much higher dimensional than camera pose. While some work can reconstruct static scenes from in-the-wild images with challenging illumination conditions and backgrounds [147, 197], they cannot be applied to articulated objects like humans. To address this, we introduce PuzzleAvatar in Chapter 6 of this thesis, to tackle all the above challenges for "pose-free" 3D human reconstruction. PuzzleAvatar requires neither camera nor body poses; thus it is uniquely capable of operating on unconstrained in-the-wild photos with unknown camera poses, unknown body poses, possibly truncated images (*e.g.*, headshots, random cropping), and diverse backgrounds and illumination conditions, which are highly challenging for existing methods. Total-Selfie [27] takes daily selfies to generate a full-body 2D selfie, which somewhat aligns with PuzzleAvatar's goal but not in the 3D domain.

2.2 3D Clothed Human Generation

2.2.1 3D Human Generators Trained on 3D Data

Statistical body models [90, 140, 165, 231] can be considered as 3D generative models of the human body. These models are trained on numerous 3D scans of minimally clothed bodies, and can generate posed bodies with varying shapes, but without clothing. To account for the outfits, CAPE [144] learns a clothing offset layer based on the SMPL-D model, from registered human scans, Chupa [99] "carves" the SMPL mesh by dual normal maps generated by a pose-conditioned diffusion model. Alternatively, gDNA [30], NPMs [160], NSF [233], and SPAMs [161], learn the implicit clothed avatars from normalized raw captures (*i.e.*, scans, depth maps). Unfortunately, all the aforementioned methods of learning generative 3D humans with diverse shapes and appearance require 3D data, which is both limited and expensive to acquire. Rodin [214] has recently employed large-scale 3D synthetic head avatars in combination with a diffusion model to develop a high-fidelity head avatar generator. However, the scarcity of datasets containing real 3D clothed humans [21, 32, 82, 227, 240, 255, 257] limits the model's generalization ability and may lead to overfitting on constrained datasets.

2.2.2 3D Human Generator from 2D Image Collections

In contrast to 3D data, large-scale datasets of 2D human images are widely available from DeepFashion [58, 139], SHHQ [50] and LAION-5B [188]. Related human generators represent 3D humans using meshes [60, 72, 88], DMTet [56], Tri-planes [13, 44, 156, 201, 252], implicit functions [226], or neural fields [23, 71, 106, 130, 244]. Some methods adapt GANs [93] by integrating diff-renderer [13, 44, 60, 156, 201, 202, 226, 252], an increasing number of recent works leverage diffusion models [23, 72, 79, 106, 246, 249], which are trained on a tremendous amount of 2D data, allowing their strong generalizability to be exploited for downstream tasks. Despite the demonstrated quality of these methods in generating textured avatars, a gap still exists in achieving "lifelike" avatars with detailed geometry and texture, consistent with the input.

2.2.3 3D Human Creation from Text Descriptions

Recent advancements in large vision-language foundation models [174, 181] have spurred numerous efforts to create human avatars based on textual descriptions. Early initiatives used the CLIP semantic consistency loss [72] to roughly shape the human body. More recent efforts [23, 79, 106, 130, 210] leverage pretrained 2D diffusion models for 3D content generation, capturing finer geometry and texture for clothed individuals or multiple humans. These developments utilize techniques such as Score Jacobian Chaining (SJC) [209] and Score Distillation Sampling (SDS) [169, 209]. Additionally, there has been a focus on model customization through fine-tuning pre-trained networks to introduce new concepts [9, 85, 109, 137, 183]. Besides the aforementioned "generate via multi-view optimization" scheme, which typically takes a few thousand iterations, recent attention has been drawn to the "generate via direct view-conditional generation" [24, 135, 136, 170, 218, 260] for quicker and more efficient reconstruction.

In addition to using text-only descriptions, when subject images are available, they could be used to fine-tune pretrained 2D diffusion models and to enhance fidelity via re-projection losses [57, 77, 234]. Our TeCH in Chapter 5 is the first to merge visual captions from input images with personalized tokens learned via DreamBooth [183], aiming to well preserve human identity. However, all image-conditioned methods assume reliable human pose estimation [165] as a proxy representation to draw correspondences between the input image and the reconstructed 3D avatar. Hence, they require images with clean backgrounds, common body poses, and full-body views without crops. Furthermore, the use of external controllers (*e.g.*, ControlNet [254], Zero123 [136]), along with additional geometric regularizers (*e.g.*, Laplacian and Eikonal [29]) is crucial to achieving high-quality results. These limitations are inherited by TeCH (Chapter 5) but overcome by PuzzleAvatar (Chapter 6).

3

ICON: IMPLICIT CLOTHED HUMANS OBTAINED FROM NORMALS

Contents

3.1 Introduction

Realistic virtual humans will play a central role in mixed and augmented reality, forming a key foundation for the "metaverse" and supporting remote presence, collaboration, education, and entertainment. To enable this, new tools are needed to easily create 3D virtual humans that can be readily animated. Traditionally, this requires significant artist effort and expensive scanning equipment. Therefore, such approaches do not scale easily. A more practical approach would enable individuals to create an avatar from one or more images. There are now several methods that take a single image and regress a

minimally clothed 3D human model [5, 6, 35, 47, 107, 165]. Existing parametric body models, however, lack important details like clothing and hair [90, 140, 165, 182, 231]. In contrast, we present a method that robustly extracts 3D scan-like data from images of people in arbitrary poses and uses this to construct an animatable avatar.

We base ICON, short for "Implicit Clothed humans Obtained from Normals", on implicit functions (IFs), which go beyond parametric body models to represent fine shape details and varied topology. IFs allow recent methods to infer detailed shape from an image [67, 80, 185, 186, 236, 256]. Despite promising results, state-of-the-art (SOTA) methods struggle with in-the-wild data and often produce humans with broken or disembodied limbs, missing details, high-frequency noise, or non-human shape; see Fig. 3.1 for examples.

The issues with previous methods are twofold: (1) Such methods are typically trained on small, hand-curated, 3D human datasets (*e.g.* RenderPeople [179]) with very limited pose, shape and clothing variation. (2) They typically feed their implicit-function module with features of a global 2D image or 3D voxel encoder, but these are sensitive to global pose. While more, and more varied, 3D training data would help, such data remains limited. Hence, we take a different approach and improve the model.

Specifically, the goal of ICON is to reconstruct a detailed clothed 3D human from a single RGB image with a method that is training-data efficient and robust to in-the-wild images and out-of-distribution poses. Our method, called *ICON*, stands for *Implicit Clothed humans Obtained from Normals*. ICON replaces the global encoder of existing methods with a more data-efficient local scheme; Fig. 3.2 shows a model overview. ICON takes as input an RGB image of a segmented clothed human and a SMPL body estimated from the image [104]. The SMPL body is used to guide two of ICON's modules: one infers detailed clothed-human surface normals (front and back views), and the other infers a visibility-aware implicit surface (iso-surface of an occupancy field). Errors in the initial SMPL estimate, however, might misguide inference. Thus, at inference time, an iterative feedback loop refines SMPL (*i.e.*, its 3D shape, pose, and translation) using the inferred detailed normals, and vice versa, leading to a refined implicit shape with better 3D details.

We evaluate ICON quantitatively and qualitatively on challenging datasets, namely AGORA [164] and CAPE [144], as well as on in-the-wild images. Results show that ICON has two advantages w.r.t. the state of the art: **(1) Generalization:**. ICON's locality helps it generalize to in-the-wild images and out-of-distribution poses and clothes better

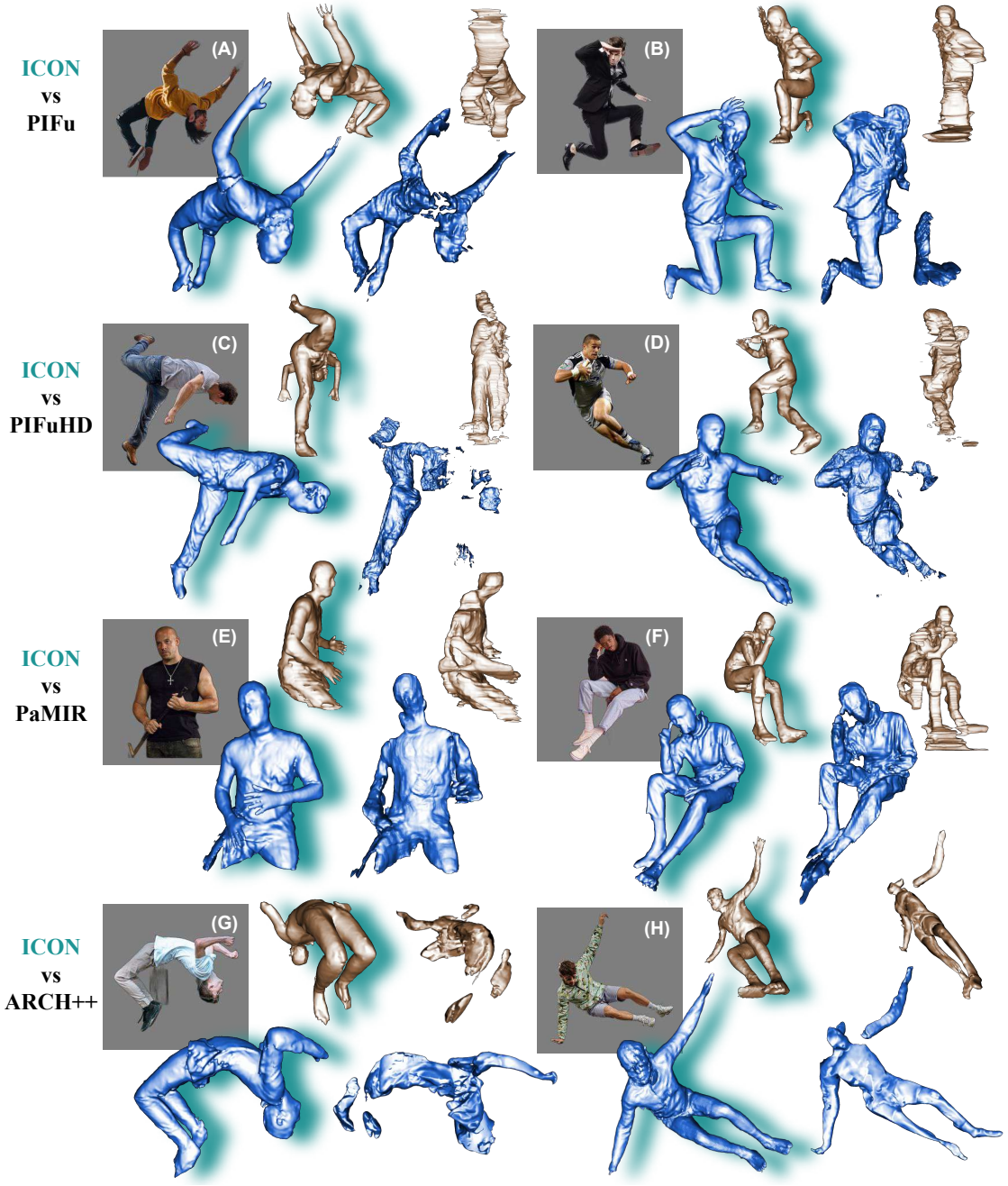


Figure 3.1: SOTA methods for inferring 3D humans from in-the-wild images, *e.g.*, PIFu [185], PIFuHD [186], PaMIR [256], and ARCH++ [68], struggle with challenging poses and out-of-frame cropping (E), resulting in various artifacts including non-human shapes (A,G), disembodied parts (B,H), missing body parts (C,D), missing details (E), and high-frequency noise (F). ICON deals with these challenges and produces high-quality results, highlighted with a green shadow Front view (blue) and rotated view (bronze).

than previous methods. Representative cases are shown in Fig. 3.1; notice that, although ICON is trained on full-body images only, it can handle images with out-of-frame cropping, with no fine tuning or post processing. **(2) Data efficacy:**. ICON's locality helps it avoid spurious correlations between pose and surface shape. Thus, it needs less data for training. ICON significantly outperforms baselines in low-data regimes, as it reaches SOTA performance when trained with as little as 12% of the data.

We provide an example application of ICON for creating an animatable avatar; see Fig. 1.2 for an overview. We first apply ICON on the individual frames of a video sequence, to obtain 3D meshes of a clothed person in various poses. We then use these to train a poseable avatar using a modified version of SCANimate [187]. Unlike 3D scans, which SCANimate takes as input, ICON's estimated shapes are not equally detailed and reliable from all views. Consequently, we modify SCANimate to exploit visibility information in learning the avatar. The output is a 3D clothed avatar that moves and deforms naturally; see Fig. 1.2-right and Fig. 3.6b.

ICON takes a step towards robust reconstruction of 3D clothed humans from in-the-wild photos. Based on this, fully textured and animatable avatars with personalized pose-aware clothing deformation can be created directly from video frames.

3.2 Method

ICON is a deep-learning model that infers a 3D clothed human from a color image. Specifically, ICON takes as input an RGB image with a segmented clothed human (following the suggestion of PIFuHD's repository [167]), along with an estimated human body shape "under clothing" (SMPL), and outputs a pixel-aligned 3D shape reconstruction of the clothed human. ICON has two main modules (see Fig. 3.2) for: (1) SMPL-guided clothed-body normal prediction and (2) local-feature based implicit surface reconstruction.

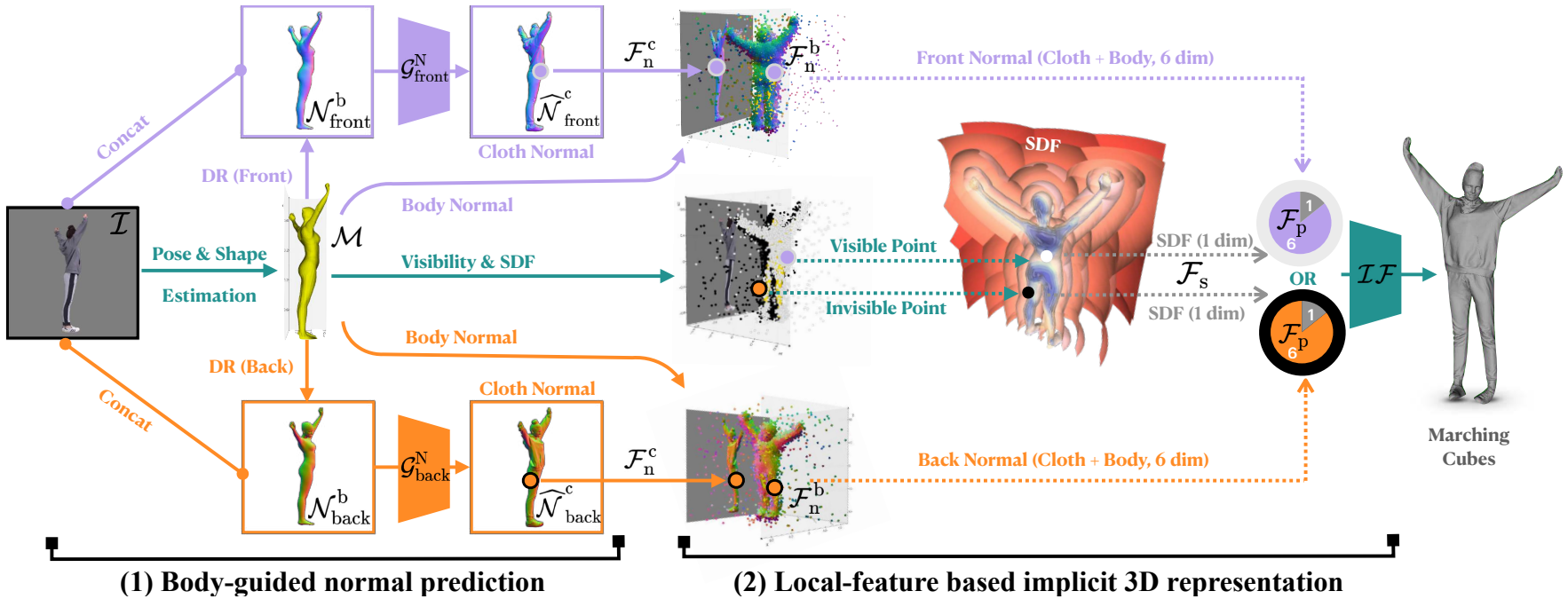


Figure 3.2: ICON's architecture. It contains two main modules for: (1) body-guided normal prediction, and (2) local-feature based implicit 3D reconstruction. The dotted line with an arrow is a 2D or 3D query function. The two $\mathcal{G}^{\mathrm{N}}$ networks (purple/orange) have different parameters. More explanations of each module could be seen at Sec. 3.2.

3.2.1 Body-guided normal prediction

Inferring full-360° 3D normals from a single RGB image of a clothed person is challenging; normals for the occluded parts need to be hallucinated based on the observed parts. This is an ill-posed task and is challenging for deep networks. Unlike model-free methods [84, 186, 204], ICON takes into account a SMPL [140] "body-under-clothing" mesh to reduce ambiguities and guide front and (especially) back clothed-body normal prediction. To estimate the SMPL mesh $\mathcal{M}^{\text{b}}(\beta, \theta) \in \mathbb{R}^{N\times 3}$ from image $\mathcal{I}$, we use PyMAF [248] due to its better mesh-to-image alignment compared to other methods. SMPL is parameterized by shape, $\beta \in \mathbb{R}^{10}$, and pose, $\theta \in \mathbb{R}^{3\times K}$, where $N = 6,890$ vertices and $K = 24$ joints. ICON is also compatible with SMPL-X [165].

Under a weak-perspective camera model, with scale $s \in \mathbb{R}$ and translation $t \in \mathbb{R}^3$, we use the PyTorch3D [176] differentiable renderer, denoted as $\mathcal{DR}$, to render $\mathcal{M}^{\text{b}}$ from two opposite views, obtaining "front" (*i.e.*, observable side) and "back" (*i.e.*, occluded side) SMPL-body normal maps $\mathcal{N}^{\text{b}} = \{\mathcal{N}^{\text{b}}_{\text{front}}, \mathcal{N}^{\text{b}}_{\text{back}}\}$. Given $\mathcal{N}^{\text{b}}$ and the original color image $\mathcal{I}$, ICON's normal networks $\mathcal{G}^{\text{N}} = \{\mathcal{G}^{\text{N}}_{\text{front}}, \mathcal{G}^{\text{N}}_{\text{back}}\}$ predict clothed-body normal maps, denoted as $\widehat{\mathcal{N}}^{\text{c}} = \{\widehat{\mathcal{N}}^{\text{c}}_{\text{front}}, \widehat{\mathcal{N}}^{\text{c}}_{\text{back}}\}$:

$$\mathcal{DR}(\mathcal{M}^{\text{b}}) \rightarrow \mathcal{N}^{\text{b}}, \tag{3.1}$$

$$\mathcal{G}^{\text{N}}(\mathcal{N}^{\text{b}}, \mathcal{I}) \rightarrow \widehat{\mathcal{N}}^{\text{c}}. \tag{3.2}$$

We train the normal networks $\mathcal{G}^{\text{N}}$, with the following loss:

$$\mathcal{L}_{\text{N}} = \mathcal{L}_{\text{pixel}} + \lambda_{\text{VGG}}\mathcal{L}_{\text{VGG}}, \tag{3.3}$$

where $\mathcal{L}_{\text{pixel}} = |\mathcal{N}^{\text{c}}_{\text{v}} - \widehat{\mathcal{N}}^{\text{c}}_{\text{v}}|$, $\text{v} = \{\text{front}, \text{back}\}$, is a loss (L1) between ground-truth and predicted normals (the two $\mathcal{G}^{\text{N}} = \{\mathcal{G}^{\text{N}}_{\text{front}}, \mathcal{G}^{\text{N}}_{\text{back}}\}$ in Fig. 3.2 have different parameters), and $\mathcal{L}_{\text{VGG}}$ is a perceptual loss [89] weighted by λ_{VGG}. With only $\mathcal{L}_{\text{pixel}}$, the inferred normals are blurry, but adding $\mathcal{L}_{\text{VGG}}$ helps recover details.

Refining SMPL: Intuitively, a more accurate SMPL body fit provides a better prior that helps infer better clothed-body normals. However, in practice, human pose and shape (HPS) regressors do not give pixel-aligned SMPL fits. To account for this, during inference, the SMPL fits are optimized based on the difference between the rendered

SMPL-body normal maps, $\mathcal{N}^{\text{b}}$, and the predicted clothed-body normal maps, $\widehat{\mathcal{N}}^{\text{c}}$, as shown in Sec. 3.2.1. Specifically we optimize over SMPL's shape, β, pose, θ, and translation, t, parameters to minimize:

$$\mathcal{L}_{\text{SMPL}} = \min_{\theta,\beta,t}(\lambda_{\text{N\_diff}}\mathcal{L}_{\text{N\_diff}} + \mathcal{L}_{\text{S\_diff}}), \tag{3.4}$$

$$\mathcal{L}_{\text{N\_diff}} = |\mathcal{N}^{\text{b}} - \widehat{\mathcal{N}}^{\text{c}}|, \quad \mathcal{L}_{\text{S\_diff}} = |\mathcal{S}^{\text{b}} - \widehat{\mathcal{S}}^{\text{c}}|, \tag{3.5}$$

where $\mathcal{L}_{\text{N\_diff}}$ is a normal-map loss (L1), weighted by $\lambda_{\text{N\_diff}}$; $\mathcal{L}_{\text{S\_diff}}$ is a loss (L1) between the silhouettes of the SMPL body normal-map $\mathcal{S}^{\text{b}}$ and the human mask $\widehat{\mathcal{S}}^{\text{c}}$ segmented [177] from $\mathcal{I}$. We ablate $\mathcal{L}_{\text{N\_diff}}$, $\mathcal{L}_{\text{S\_diff}}$ in Appendix A.1.2.

Refining normals.: The normal maps rendered from the refined SMPL mesh, $\mathcal{N}^{\text{b}}$, are fed to the $\mathcal{G}^{\text{N}}$ networks. The improved SMPL-mesh-to-image alignment guides $\mathcal{G}^{\text{N}}$ to infer more reliable and detailed normals $\widehat{\mathcal{N}}^{\text{c}}$.

Refinement loop.: During inference, ICON alternates between: (1) refining the SMPL mesh using the inferred $\widehat{\mathcal{N}}^{\text{c}}$ normals and (2) re-inferring $\widehat{\mathcal{N}}^{\text{c}}$ using the refined SMPL. Experiments show that this feedback loop leads to more reliable clothed-body normal maps for both (front/back) sides.

3.2.2 Local-feature based implicit 3D reconstruction

Given the predicted clothed-body normal maps, $\widehat{\mathcal{N}}^{\text{c}}$, and the SMPL-body mesh, $\mathcal{M}^{\text{b}}$, we regress the implicit 3D surface of a clothed human based on local features $\mathcal{F}_{\text{P}}$:

$$\mathcal{F}_{\text{P}} = [\mathcal{F}_{\text{s}}(\text{P}), \mathcal{F}^{\text{b}}_{\text{n}}(\text{P}), \mathcal{F}^{\text{c}}_{\text{n}}(\text{P})], \tag{3.6}$$

where $\mathcal{F}_{\text{s}}$ is the signed distance from a query point P to the closest body point $\text{P}^{\text{b}} \in \mathcal{M}^{\text{b}}$, and $\mathcal{F}^{\text{b}}_{\text{n}}$ is the barycentric surface normal of P^{b}; both provide strong regularization against self occlusions. Finally, $\mathcal{F}^{\text{c}}_{\text{n}}$ is a normal vector extracted from $\widehat{\mathcal{N}}^{\text{c}}_{\text{front}}$ or $\widehat{\mathcal{N}}^{\text{c}}_{\text{back}}$ depending on the visibility of P^{b}:

$$\mathcal{F}^{\text{c}}_{\text{n}}(\text{P}) = \begin{cases} \widehat{\mathcal{N}}^{\text{c}}_{\text{front}}(\pi(\text{P})) & \text{if } \text{P}^{\text{b}} \text{ is visible} \\ \widehat{\mathcal{N}}^{\text{c}}_{\text{back}}(\pi(\text{P})) & \text{else,} \end{cases} \tag{3.7}$$

where $\pi(\text{P})$ denotes the 2D projection of the 3D point P.

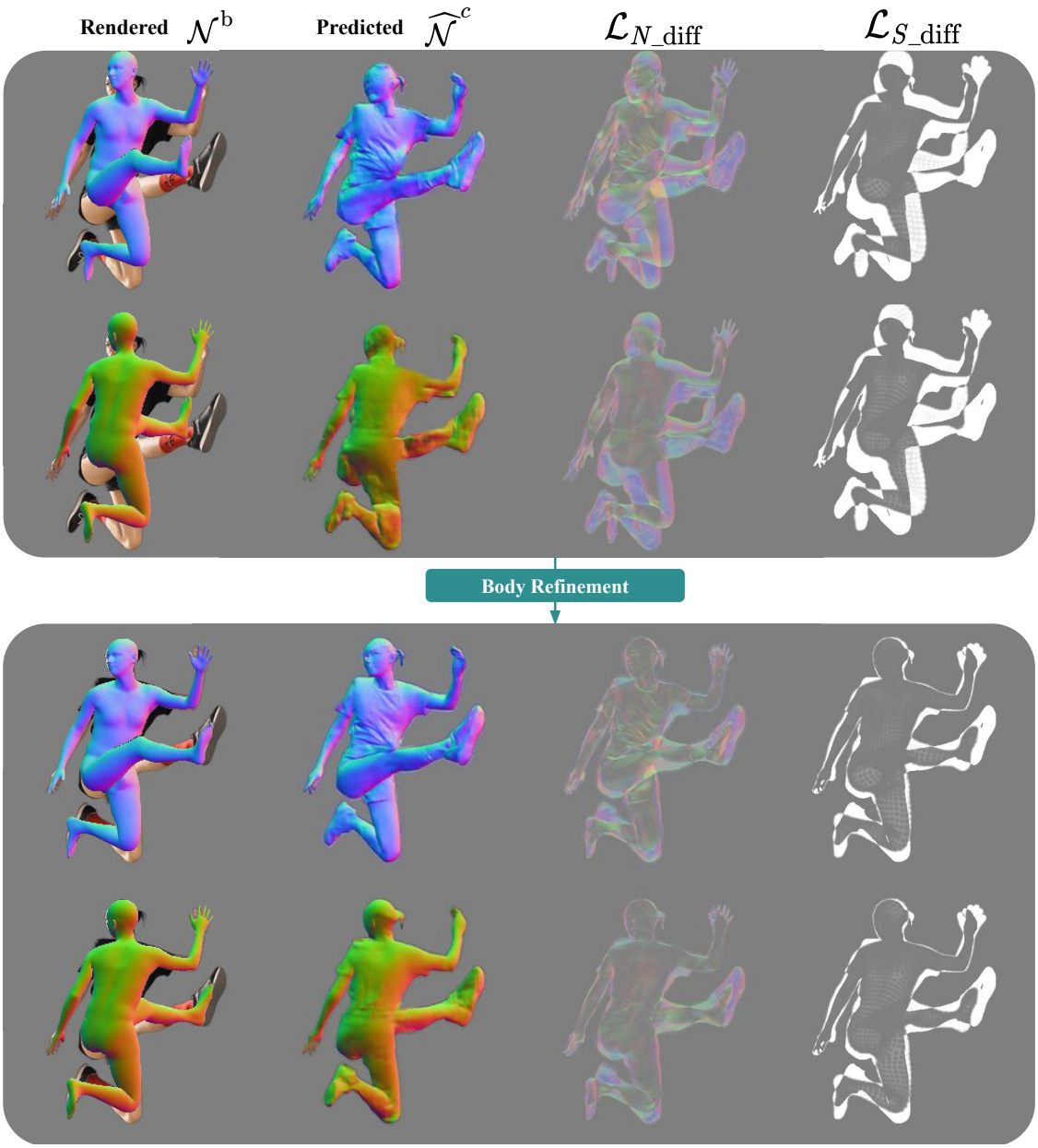


Figure 3.3: SMPL refinement using a feedback loop. When HPS regressors do not give pixel-aligned SMPL fits, during inference, ICON alternates between: (1) refining the SMPL mesh using the inferred $\widehat{\mathcal{N}}^{\mathrm{c}}$ normals and (2) re-inferring $\widehat{\mathcal{N}}^{\mathrm{c}}$ using the refined SMPL.

Please note that, in comparison with the features extracted by a CNN-based global image encoder, the locally queried $\mathcal{F}_{\mathrm{P}}$ has less spatial correlation. Experiments indicate this is crucial for robustness to out-of-distribution poses and efficacy w.r.t. training data.

We feed $\mathcal{F}_{\mathrm{P}}$ into an implicit function, $\mathcal{IF}$, parameterized by a Multi-Layer Perceptron (MLP) to estimate the occupancy at point P, denoted as $\hat{o}(\mathrm{P})$. A mean squared error loss is used to train $\mathcal{IF}$ with ground-truth occupancy, $o(\mathrm{P})$. Then the fast surface localization algorithm [122, 141] is used to extract meshes from the 3D occupancy inferred by $\mathcal{IF}$.

3.3 Experiments

3.3.1 Baseline models

We compare ICON primarily with PIFu [185] and PaMIR [256]. These methods differ from ICON and from each other w.r.t. the training data, the loss functions, the network structure, the use of the SMPL body prior, *etc*. To isolate and evaluate each factor, we re-implement PIFu and PaMIR by "simulating" them based on ICON's architecture. This provides a unified benchmarking framework, and enables us to easily train each baseline with the exact same data and training hyper-parameters for a fair comparison. Since there might be small differences w.r.t. the original models, we denote the "simulated" models with a "star" as:

- PIFu* : $\{f_{2\mathrm{D}}(\mathcal{I},\mathcal{N})\} \rightarrow \mathcal{O}$,
- PaMIR*: $\{f_{2\mathrm{D}}(\mathcal{I},\mathcal{N}), f_{3\mathrm{D}}(\mathcal{V})\} \mapsto \mathcal{O}$,
- ICON : $\{\mathcal{N}, \gamma(\mathcal{M}^{\mathrm{b}})\} \rightarrow \mathcal{O}$,

Where $f_{2\mathrm{D}}$ denotes the 2D image encoder, $f_{3\mathrm{D}}$ denotes the 3D voxel encoder, $\mathcal{V}$ denotes the voxelized SMPL, $\mathcal{O}$ denotes the entire predicted occupancy field, and γ is the mesh-based local feature extractor described in Sec. 3.2.2. The results are summarized in Tab. 3.2-A, and discussed in Sec. 3.3.3-A. For reference, we also report the performance of the original PIFu [185], PIFuHD [186], and PaMIR [256]; ICON's "simulated" models perform well, and even outperform the original ones.

3.3.2 Datasets

Several public or commercial 3D clothed-human datasets are used in the literature, but each method uses different subsets and combinations of these, as shown in Tab. 3.1.

Training data. To compare models fairly, we factor out differences in training data as explained in Sec. 3.3.1. Following previous work [185, 186], we retrain all baselines on the same 450 RenderPeople scans (subset of AGORA [164]). Methods that require the 3D body prior (*i.e.*, PaMIR, ICON) use the SMPL-X meshes provided by AGORA. ICON's $\mathcal{G}^{\text{N}}$ and $\mathcal{IF}$ modules are trained on the same data.

Testing data. We evaluate primarily on CAPE [144], which no method uses for training, to test their generelization ability. Specifically, we divide the CAPE dataset into the "CAPE-FP" and "CAPE-NFP" sets that have "fashion" and "non-fashion" poses, respectively, to better analyze the generalization to complex body poses; for details on data splitting, please see Appendix A.1.1. To evaluate performance without a domain gap between train/test data, we also test all models on "AGORA-50" [185, 186], which contains 50 samples from AGORA that are different from the 450 used for training.

Generating synthetic data. We use the OpenGL scripts of MonoPort [122] to render photo-realistic images with dynamic lighting. We render each clothed-human 3D scan ($\mathcal{I}$ and $\mathcal{N}^{\text{c}}$) and their SMPL-X fits ($\mathcal{N}^{\text{b}}$) from multiple views by using a weak perspective camera and rotating the scan in front of it. In this way we generate 138,924 samples, each containing a 3D clothed-human scan, its SMPL-X fit, an RGB image, camera parameters, 2D normal maps for the scan and the SMPL-X mesh (from two opposite views) and SMPL-X triangle visibility information w.r.t. the camera.

	Train & Validation Sets				Test Set	
	Renderp. [179]	Twindom [207]	AGORA [164]	THuman [257]	BUFF [245]	CAPE [144, 168]
Free & public	✗	✗	✗	✓	✓	✓
Diverse poses	✗	✗	✗	✓	✗	✓
Diverse identities	✓	✓	✓	✗	✗	✗
SMPL(-X) poses	✗	✗	✓	✓	✓	✓
High-res texture	✓	✓	✓	✗	✓	✓
Number of scans	450 [185, 186] 375 [80]	1000 [256]	450 [IC] 3109 [IC†]	600 [IC†] 600 [256]	5 [185, 186] 26 [80] 300 [122, 256]	150 [IC]

Table 3.1: Datasets for 3D clothed humans. Gray color indicates datasets used by ICON. The bottom "number of scans" row indicates the number of scans each method uses. The cell format is `number_of_scans [method]`. ICON is denoted as `[IC]`. The symbol † corresponds to the "8x" setting in Fig. 3.5. SMPL(-X) means the 3D human scans are registered as either SMPL or SMPL-X.

	Methods	SMPL-X condition.	AGORA-50 Chamfer ↓	AGORA-50 P2S ↓	AGORA-50 Normals ↓	CAPE-FP Chamfer ↓	CAPE-FP P2S ↓	CAPE-FP Normals ↓	CAPE-NFP Chamfer ↓	CAPE-NFP P2S ↓	CAPE-NFP Normals ↓	CAPE Chamfer ↓	CAPE P2S ↓	CAPE Normals ↓
Ours	ICON	✓	1.204	1.584	0.060	1.233	**1.170**	0.072	1.096	**1.013**	0.063	**1.142**	**1.065**	0.066
A	PIFu [185]	✗	3.453	3.660	0.094	2.823	2.796	0.100	4.029	4.195	0.124	3.627	3.729	0.116
	PIFuHD[186]	✗	3.119	3.333	0.085	2.302	2.335	0.090	3.704	3.517	0.123	3.237	3.123	0.112
	PaMIR [256]	✓	2.035	1.873	0.079	1.936	1.263	0.078	2.216	1.611	0.093	2.122	1.495	0.088
	SMPL-X GT	N/A	1.518	1.985	0.072	1.335	1.259	0.085	**1.070**	1.058	0.068	1.158	1.125	0.074
	PIFu*	✗	2.688	2.573	0.097	2.100	2.093	0.091	2.973	2.940	0.111	2.682	2.658	0.104
	PaMIR*	✓	1.401	1.500	0.063	1.225	1.206	**0.055**	1.413	1.321	0.063	1.350	1.283	**0.060**
B	$\text{ICON}_{\text{N}^\dagger}$	✓	**1.153**	1.545	0.057	1.240	1.226	0.069	1.114	1.097	0.062	1.156	1.140	0.064
	ICON w/o $\mathcal{F}_{\text{n}}^{\text{b}}$	✓	1.259	1.667	0.062	1.344	1.336	0.072	1.180	1.172	0.064	1.235	1.227	0.067
C	$\text{ICON}_{\text{enc}(\mathcal{I},\widehat{\mathcal{N}}^{\text{c}})}$	✓	1.172	**1.350**	**0.053**	1.243	1.243	0.062	1.254	1.122	0.060	1.250	1.229	0.061
	$\text{ICON}_{\text{enc}(\widehat{\mathcal{N}}^{\text{c}})}$	✓	1.180	1.450	0.055	**1.202**	1.196	0.061	1.180	1.067	**0.059**	1.187	1.110	**0.060**
D	ICON	✓	1.583	1.987	0.079	1.364	1.403	0.080	1.444	1.453	0.083	1.417	1.436	0.082
	ICON + BR	✓	1.554	1.961	0.074	1.314	1.356	0.070	1.351	1.390	0.073	1.339	1.378	0.072
	PaMIR*	✓	1.674	1.802	0.075	1.608	1.625	0.072	1.803	1.764	0.079	1.738	1.718	0.077
	SMPL-X perturbed	N/A	1.984	2.471	0.098	1.488	1.531	0.095	1.493	1.534	0.098	1.491	1.533	0.097

Table 3.2: Quantitative evaluation for ICON and SOTA. **(A)** performance w.r.t. SOTA; **(B)** body-guided normal prediction; **(C)** local-feature based implicit reconstruction; and **(D)** robustness to SMPL-X noise. Inference conditioned on: (✓) SMPL-X ground truth (GT); (✓) perturbed SMPL-X GT; (✗) no SMPL-X condition. SMPL-X ground truth is provided by each dataset. CAPE is not used for training, and tests generalization.

3.3.3 Evaluation

We use 3 evaluation metrics, described in the following:

"Chamfer" distance: We report the Chamfer distance between ground-truth scans and reconstructed meshes. For this, we sample points uniformly on scans/meshes, to factor out resolution differences, and compute average bi-directional point-to-surface distances. This metric captures large geometric differences, but misses smaller geometric details.

"P2S" distance: CAPE has raw scans as ground truth, which can contain large holes. To factor holes out, we additionally report the average point-to-surface (P2S) distance from scan points to the closest reconstructed surface points. This metric can be viewed as a 1-directional version of the above metric.

"Normals" difference: We render normal images for reconstructed and ground-truth surfaces from fixed viewpoints (Sec. 3.3.2, "generating synthetic data"), and calculate the L2 error between them. This captures errors for high-frequency geometric details, when Chamfer and P2S errors are small.

A. ICON -vs- SOTA. ICON outperforms all original SOTA methods, and is competitive to our "simulated" versions of them, as shown in Tab. 3.2-A. We use AGORA's SMPL-X [164] ground truth (GT) as a reference. We notice that our re-implemented PaMIR* outperform the SMPL-X GT for images with in-distribution body poses ("AGORA-50" and "CAPE-FP"), However, this is not the case for images with out-of-distribution poses ("CAPE-NFP"). This shows that, although conditioned on GT SMPL-X fits, PaMIR* is still sensitive to global body pose due to its global feature encoder, and fails to generalize to out-of-distribution poses. On the contrary, ICON generalizes well to out-of-distribution poses due to its lower spatial correlation (see Sec. 3.2.2).

B. Body-guided normal prediction. We evaluate the conditioning on SMPL-X-body normal maps, $\mathcal{N}^{\text{b}}$, for guiding inference of clothed-body normal maps, $\widehat{\mathcal{N}^{\text{c}}}$ (Sec. 3.2.1). Table 3.2-B shows performance with ("ICON") and without ("ICON$_{\text{N}^{\dagger}}$") conditioning. With no conditioning, errors on "CAPE" increase slightly. Qualitatively, guidance by body normals heavily improves the inferred normals, especially for occluded body regions; see Fig. 3.4. We also ablate the effect of the body-normal feature (Sec. 3.2.2), $\mathcal{F}_{\text{n}}^{\text{b}}$, by removing it; this worsens results, see "ICON w/o $\mathcal{F}_{\text{n}}^{\text{b}}$" in Tab. 3.2-B.

C. Local-feature based implicit reconstruction. To evaluate the importance of our "local" features (Sec. 3.2.2), $\mathcal{F}_{\text{P}}$, we replace them with "global" features produced by 2D

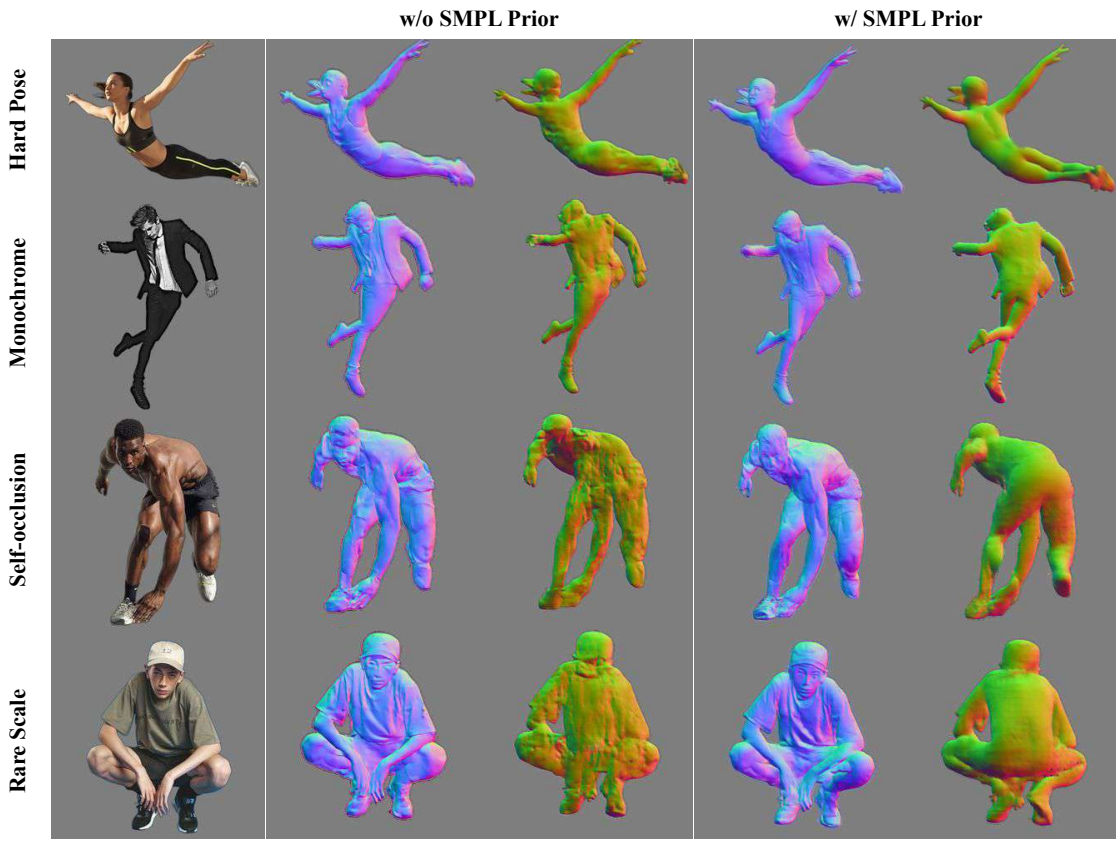


Figure 3.4: Normal prediction ($\widehat{\mathcal{N}}^{\mathbf{c}}$) w/ and w/o SMPL prior ($\mathcal{N}^{\mathbf{b}}$).

convolutional filters. These are applied on the image and the clothed-body normal maps ("ICON$_{\text{enc}(\mathcal{I},\widehat{\mathcal{N}}^{\text{c}})}$" in Tab. 3.2-C), or only on the normal maps ("ICON$_{\text{enc}(\widehat{\mathcal{N}}^{\text{c}})}$" in Tab. 3.2-C). We use a 2-stack hourglass model [83], whose receptive field expands to 46% of the image size. This takes a large image area into account and produces features sensitive to global body pose. This worsens reconstruction performance for out-of-distribution poses, such as in "CAPE-NFP". For an evaluation of PaMIR's receptive field size, see Appendix A.1.4

We compare ICON with SOTA models for a varying amount of training data in Fig. 3.5. The "Dataset scale" axis reports the data size as the ratio of the 450 scans of the original PIFu methods [185, 186]; the left-most side corresponds to 56 scans and the right-most side corresponds to $3,709$ scans, *i.e.*, all the scans of AGORA [164] and THuman [257]. ICON consistently outperforms all methods. Importantly, ICON achieves SOTA performance even when trained on just a *fraction* of the data. We attribute this to 1) the local nature of ICON's queried pointwise features; this helps ICON generalize well in the pose space and be data efficient; and 2) the normal-based body refinement in a feedback loop; this is critical when HPS regressors do not give pixel-aligned body fits.

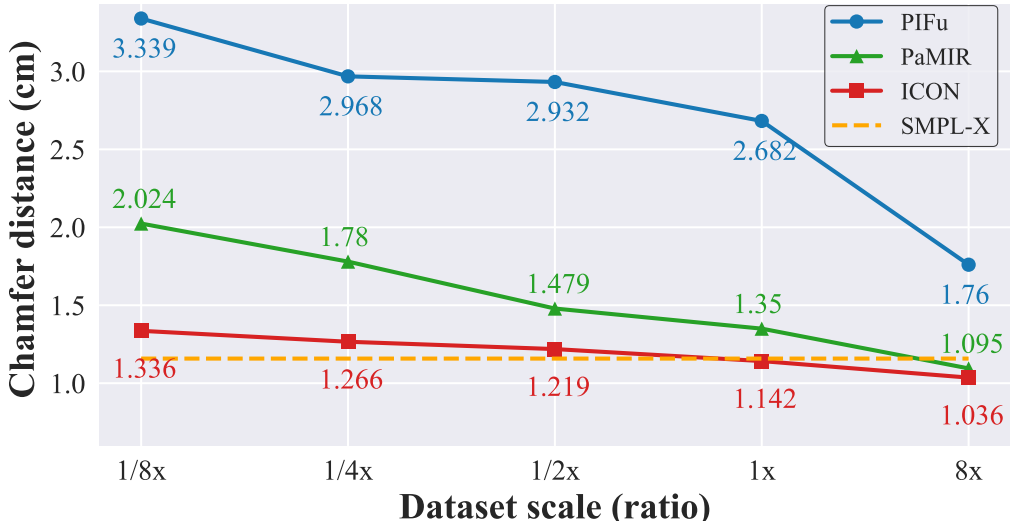


Figure 3.5: Reconstruction error w.r.t. training-data size. "Dataset size" is defined as the ratio of the 450 scans used in [185, 186]. The "8x" setting is all 3,709 scans of AGORA [164] and THuman [257]. Please note that the "SMPL-X" refers to the ground-truth SMPL-X fits, but in practice, achieving such accuracy is challenging.

D. Robustness to SMPL-X noise. SMPL-X estimated from an image might not be perfectly aligned with body pixels in the image. However, PaMIR and ICON are conditioned on this estimation. Thus, they need to be robust against various noise levels in SMPL-X shape and pose. To evaluate this, we feed PaMIR* and ICON with ground-truth and perturbed SMPL-X, denoted with (✓) and (✓) in Tab. 3.2-A,D. ICON conditioned on perturbed (✓) SMPL-X produces larger errors compared with conditioning on ground truth (✓). However, adding the body refinement module ("ICON +BR") of Sec. 3.2.1, refines SMPL-X and improves performance. As a result, "ICON +BR" conditioned on noisy SMPL-X (✓) performs comparably to PaMIR* conditioned on ground-truth SMPL-X (✓); it is slightly worse/better for in-/out-of-distribution poses.

3.4 Applications

3.4.1 Reconstruction from in-the-wild images

We collect 200 in-the-wild images from Pinterest that show people performing parkour, sports, street dance, and kung fu. These images are unseen during training. We show

	PIFu*	PIFuHD [186]	PaMIR*
Preference	30.9%	22.3%	26.6%
P-value	1.35e-33	1.08e-48	3.60e-54

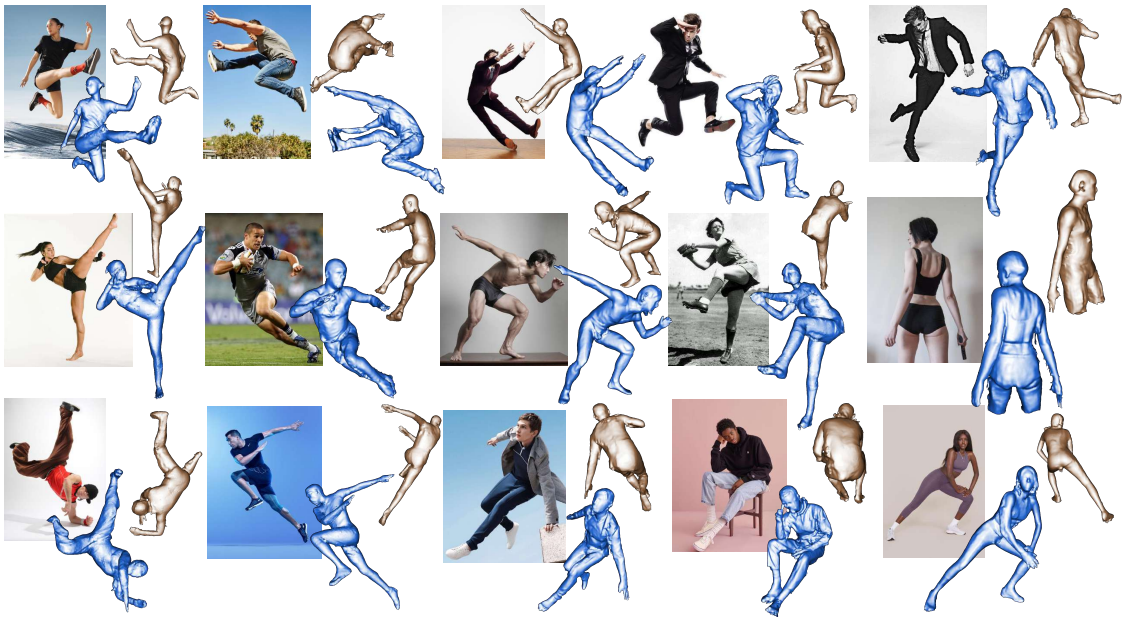
Table 3.3: Perceptual study. Numbers denote the chance that participants prefer the result of a competitor over ICON for internet images. ICON is judged significantly more realistic.

qualitative results for ICON in Fig. 3.6a and comparisons to SOTA in Fig. 3.1; for more results see Appendix A.3.

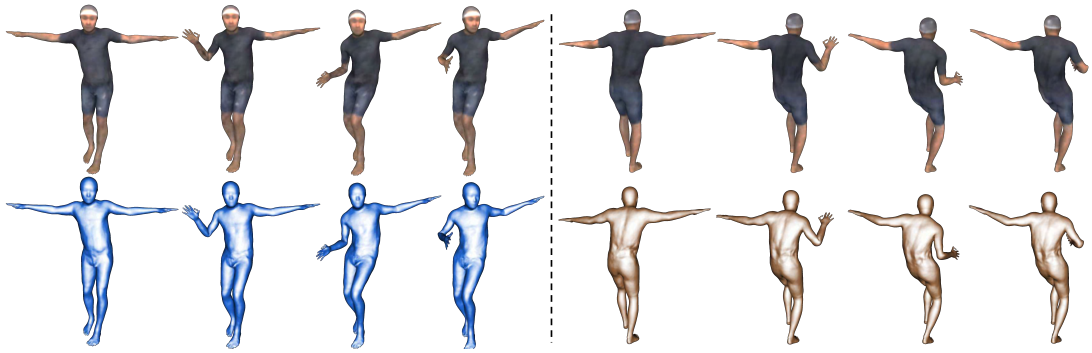
To evaluate the perceived realism of ICON's results, we compare ICON to PIFu*, PaMIR*, and the original PIFuHD [186] in a perceptual study. ICON, PIFu* and PaMIR* are trained on all 3,709 scans of AGORA [164] and THuman [257] ("8x" setting in Fig. 3.5). For PIFuHD we use its pre-trained model. In the study, participants were shown an image and either a rendered result of ICON or of another method. Participants were asked to choose the result that best represents the shape of the human in the image. We report the percentage of trails in which participants preferred the baseline methods over ICON in Tab. 3.3; p-values correspond to the null-hypothesis that two methods perform equally well. For details on the study, example stimuli, catch trials, *etc*., see Appendix A.1.3

3.4.2 Animatable avatar creation from video

Given a sequence of images with the same subject in various poses, we create an animatable avatar with the help of SCANimate [187]. First, we use ICON to reconstruct a 3D clothed-human mesh per frame. Then, we feed these meshes to SCANimate. ICON's robustness to diverse poses enables us to learn a clothed avatar with pose-dependent clothing deformation. Unlike raw 3D scans, which are taken with multi-view systems, ICON operates on a single image and its reconstructions are more reliable for observed body regions than for occluded ones. Thus, we reformulate the loss of SCANimate to downweight occluded regions depending on camera viewpoint. Results are shown in Fig. 1.2 and Fig. 3.6b;



(a) ICON reconstructions for in-the-wild images with extreme poses (Sec. 3.4.1).



(b) Avatar creation from images with SCANimate (Sec. 3.4.2). The input per-frame meshes are reconstructed with ICON.

Figure 3.6: ICON results for two applications (Sec. 3.4). We show two views for each mesh, *i.e.*, a front (blue) and a rotated (bronze) view.

Figure 3.7: Failure cases of ICON for extreme clothing, pose, or camera view. We show the front (blue) and rotated (bronze) views.

3.5 Discussion

Limitations and future work: Due to the strong body prior exploited by ICON, loose clothing that is far from the body may be difficult to reconstruct; see Fig. 3.7. Although ICON is robust to small errors of body fits, significant failure of body fits leads to reconstruction failure. Because it is trained on orthographic views, ICON has trouble with strong perspective effects, producing asymmetric limbs or anatomically improbable shapes. A key future application is to use images alone to create a dataset of clothed avatars. Such a dataset could advance research in human shape generation [30], be valuable to fashion industry, and facilitate graphics applications.

Possible negative impact: While the quality of virtual humans created from images is not at the level of facial "deep fakes", as this technology matures, it will open up the possibility for full-body deep fakes, with all the attendant risks. These risks must also be balanced by the positive use cases in entertainment, tele-presence, and future metaverse applications. Clearly regulation will be needed to establish legal boundaries for its use. In lieu of societal guidelines today, we have made code available with an appropriate license.

4

ECON: Explicit Clothed humans Optimized via Normal integration

Contents

4.1 Introduction

In Chapter 3, we introduce ICON, which demonstrates promising pose robustness in the task of image-based 3D clothed human reconstruction. However, the complexity of 3D clothed humans lies not only in diverse poses but also in various clothing with flexible topologies. People wear all kinds of different clothing and accessories, and they pose their bodies in many, often imaginative, ways. A good reconstruction method must accurately capture these, while also being robust to novel clothing and poses.

Initial, promising, results have been made possible by using artist-curated scans as training data, and implicit functions (IF) [148, 162] as the 3D representation. Seminal work on PIFu(HD) [185, 186] uses "pixel-aligned" IF and reconstructs clothed 3D humans with unconstrained topology. However, these methods tend to overfit to the poses seen in the training data, and have no explicit knowledge about the human body's structure. Consequently, they produce disembodied limbs or degenerate shapes for images with novel poses; see the 2nd row of Fig. 4.1. Follow-up works [67, 256], including ICON in Chapter 3, accounts for such artifacts by regularizing the IF using a shape prior provided by an explicit body model [140, 165], but regularization introduces a topological constraint, restricting generalization to novel clothing while attenuating shape details; see the 3rd and 4th rows of Fig. 4.1. In a nutshell, there are trade-offs between robustness, generalization and detail.

What we actually want is the *best of both worlds*; that is, the robustness of explicit anthropomorphic body models, same as ICON (Chapter 3), and the flexibility of IF to capture arbitrary clothing topology. To that end, we make two key observations: (1) While inferring detailed 2D normal maps from color images is relatively easy [84, 186, 229], inferring 3D geometry with equally fine details is still challenging [30]. Thus, we exploit networks to infer detailed "geometry-aware" 2D maps that we then lift to 3D. (2) A body model can be seen as a low-frequency "canvas" that "guides" the stitching of detailed surface parts. What sets ECON and ICON apart is that, ECON's the geometry modeling process is purely optimization-based, rather than using MLP to implicitly regress the occupancy field.

With these in mind, we develop ECON, which stands for "Explicit Clothed humans Optimized via Normal integration". It takes, as input, an RGB image and a SMPL-X body inferred from the image. Then, it outputs a 3D human in free-form clothing with a level of detail and robustness that goes beyond the state of the art (SOTA); see the bottom of Fig. 4.1. Specifically, ECON has *three steps*.

Step 1: Front & back normal reconstruction: We predict front- and back-side clothed-human normal maps from the input RGB image, conditioned on the body estimate, with a standard image-to-image translation network.

Step 2: Front & back surface reconstruction: We take the previously predicted normal maps, and the corresponding depth maps that are rendered from the SMPL-X mesh, to produce detailed and coherent front-/back-side 3D surfaces, $\{\mathcal{M}_{\mathrm{F}}, \mathcal{M}_{\mathrm{B}}\}$. To this

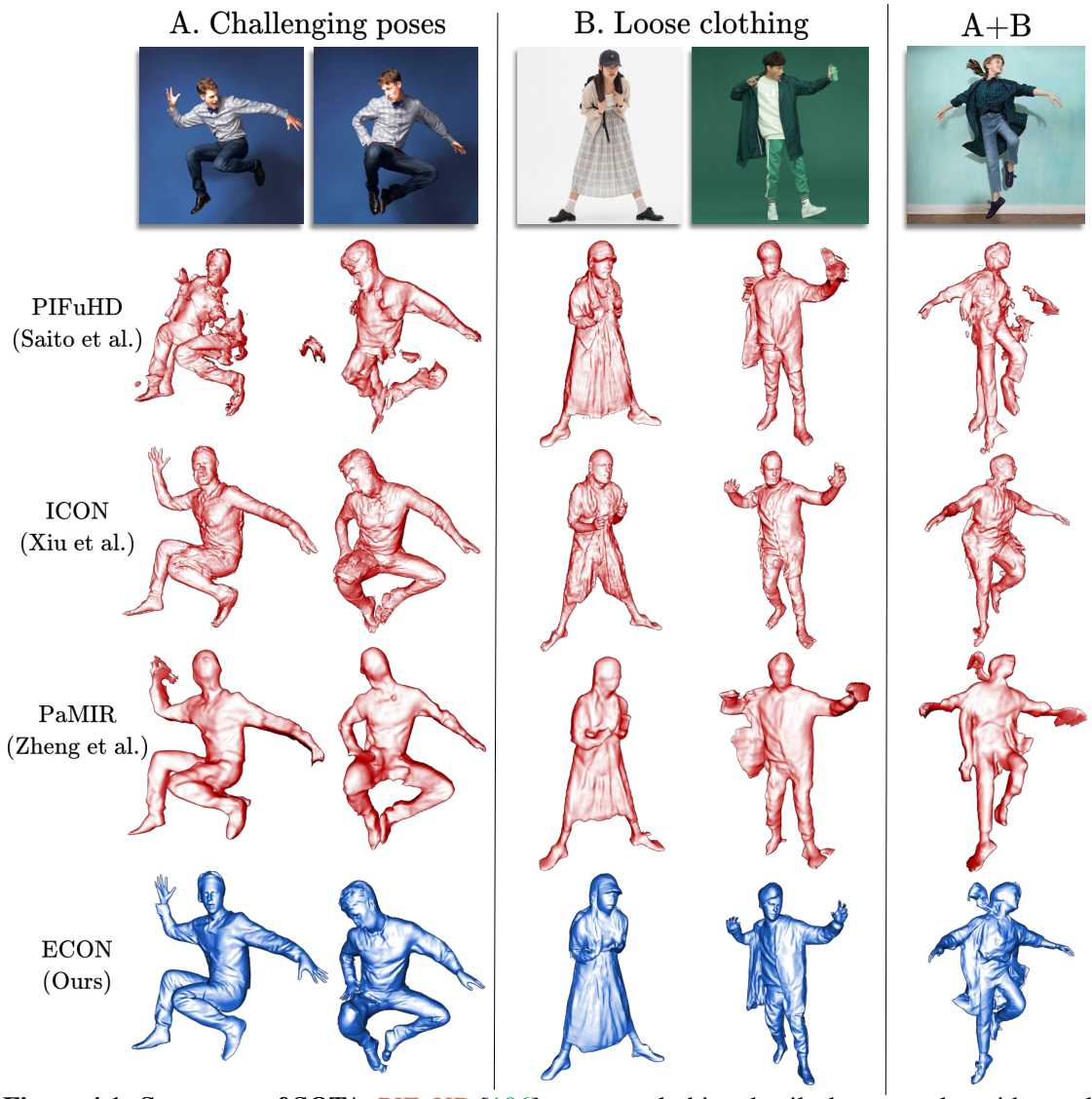


Figure 4.1: Summary of SOTA. PIFuHD [186] recovers clothing details, but struggles with novel poses. ICON (Chapter 3) and PaMIR [256] regularize shape to a body shape, but over-constrain the skirts, or over-smooth the wrinkles. ECON combines their best aspects.

end, we extend the recent BiNI method [22], and develop a novel optimization scheme that is aimed at satisfying three goals for the resulting surfaces: (1) their high-frequency components agree with clothed-human normals, (2) their low-frequency components and the discontinuities agree with the SMPL-X ones, and (3) the depth values on their silhouettes are coherent with each other and consistent with the SMPL-X-based depth maps. The two output surfaces, $\{\mathcal{M}_{\mathrm{F}}, \mathcal{M}_{\mathrm{B}}\}$, are detailed yet incomplete, *i.e.*, there is missing geometry in occluded and "profile" regions.

Step 3: Full 3D shape completion: This module takes two inputs: (1) the SMPL-X mesh, and (2) the two d-BiNI surfaces, $\{\mathcal{M}_{\mathrm{F}}, \mathcal{M}_{\mathrm{B}}\}$, where "d" refers to "depth-aware".

The goal is to "inpaint" the missing geometry. Existing methods struggle with this problem. On one hand, Poisson reconstruction [96] produces "blobby" shapes and naively "infills" holes without exploiting a shape distribution prior. On the other hand, data-driven approaches, such as IF-Nets [33], struggle with missing parts caused by (self-)occlusions, and fail to keep the fine details present on two d-BiNI surfaces, producing degenerate geometries.

We address above the limitations in two steps: (1) We extend and re-train IF-Nets to be conditioned on the SMPL-X body, so that SMPL-X regularizes shape "infilling". We discard the triangles that lie close to $\{\mathcal{M}_{\mathrm{F}}, \mathcal{M}_{\mathrm{B}}\}$, and keep the remaining ones as "infilling patches". (2) We stitch together the front- and back-side surfaces and infilling patches via Poisson reconstruction; note that holes between these are small enough for a general purpose method. The result is a full 3D shape of a clothed human; see Fig. 4.1, bottom.

We evaluate ECON both on established benchmarks (CAPE [144] and RenderPeople [179]) and in-the-wild images. Quantitative analysis reveals ECON's superiority. A perceptual study echos this, showing that ECON is significantly preferred over competitors on challenging poses and loose clothing, and competitive with PIFuHD on fashion images. Qualitative results show that ECON generalizes better than the SOTA to a wide variety of poses and clothing, even with extreme looseness or complex topology; see Fig. 4.12.

With both pose-robustness and topological flexibility, ECON recovers 3D clothed humans with a good level of detail and realistic pose. Code and models are available for research purposes at econ.is.tue.mpg.de

4.2 Method

Given an RGB image, ECON first estimates front and back normal maps (Sec. 4.2.1), then converts them into front and back partial surfaces (Sec. 4.2.2), and finally "inpaints" the missing geometry with the help of IF-Nets+ (Sec. 4.2.3). See ECON's overview in Fig. 4.2.

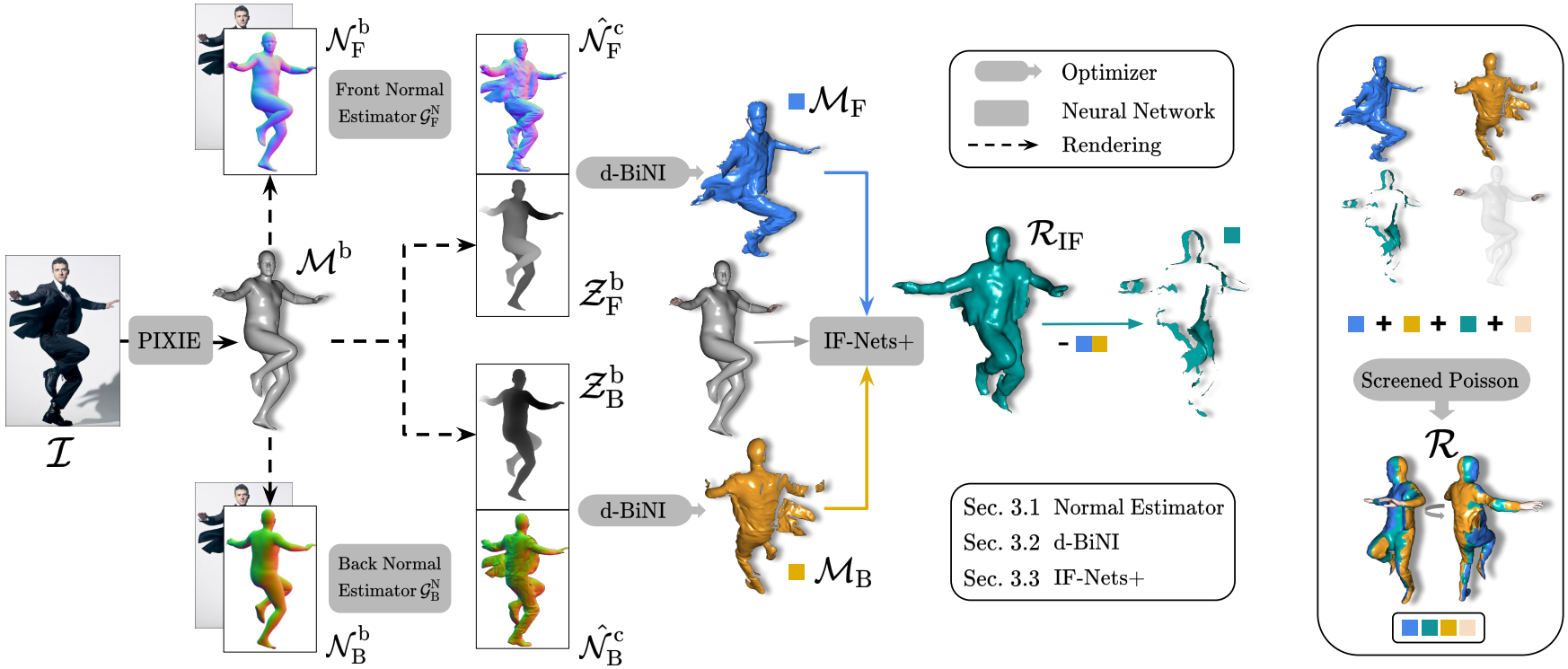


Figure 4.2: Architecture of ECON. ECON takes as input an RGB image, $\mathcal{I}$, and a SMPL-X body, $\mathcal{M}^{\text{b}}$. Conditioned on the rendered front and back body normal images, $\mathcal{N}^{\text{b}}$, ECON first predicts front and back clothing normal maps, $\widehat{\mathcal{N}}^{\text{c}}$. These two maps, along with body depth maps, $\mathcal{Z}^{\text{b}}$, are fed into a d-BiNI optimizer to produce front and back surfaces, $\{\mathcal{M}_{\text{F}}, \mathcal{M}_{\text{B}}\}$. Based on such partial surfaces, and body estimate $\mathcal{M}^{\text{b}}$, IF-Nets+ implicitly completes $\mathcal{R}_{\text{IF}}$. With optional Face or hands from $\mathcal{M}^{\text{b}}$, screened Poisson reconstruction combines everything as final watertight $\mathcal{R}$.

4.2.1 Detailed normal map prediction

Trained on abundant pairs of RGB images and normal images, a “front" normal map, $\widehat{\mathcal{N}}_{\mathrm{F}}^{\mathrm{c}}$, can be accurately estimated from an RGB image using image-to-image translation networks, as demonstrated in PIFuHD [186] or ICON (Chapter 3). Both methods also infer a “back” normal map, $\widehat{\mathcal{N}}_{\mathrm{B}}^{\mathrm{c}}$, from the image. But, the absence of image cues leads to over-smooth $\widehat{\mathcal{N}}_{\mathrm{B}}^{\mathrm{c}}$. To address this, we fine-tune ICON’s backside normal predictor, $\mathcal{G}_{\mathrm{B}}^{\mathrm{N}}$, with an additional MRF loss [216] to enhance the local details by minimizing the difference between the predicted $\widehat{\mathcal{N}}^{\mathrm{c}}$ and ground truth (GT) $\mathcal{N}^{\mathrm{c}}$ in feature space.

To guide the normal map prediction and make it robust to various body poses, ICON conditions the normal map prediction module on the body normal maps, $\mathcal{N}^{\mathrm{b}}$, rendered from the estimated body $\mathcal{M}^{\mathrm{b}}$. Thus, it is important to accurately align the estimated body and clothing silhouette. Apart from the $\mathcal{L}_{\mathrm{N\_diff}}$ and $\mathcal{L}_{\mathrm{S\_diff}}$ used in ICON (Chapter 3), we also apply 2D body landmarks in an additional loss term, $\mathcal{L}_{\mathrm{J\_diff}}$, to further optimize the SMPL-X body, $\mathcal{M}^{\mathrm{b}}$, inferred from PIXIE [47] or PyMAF-X [247]. Specifically, we optimize SMPL-X’s shape, β, pose, θ, and translation, t, to minimize:

$$\begin{aligned} \mathcal{L}_{\text{SMPL-X}} &= \mathcal{L}_{\mathrm{N\_diff}} + \mathcal{L}_{\mathrm{S\_diff}} + \mathcal{L}_{\mathrm{J\_diff}}, \\ \mathcal{L}_{\mathrm{J\_diff}} &= \lambda_{\mathrm{J\_diff}} |\mathcal{J}^{\mathrm{b}} - \widehat{\mathcal{J}}^{\mathrm{c}}|, \end{aligned} \tag{4.1}$$

where $\mathcal{L}_{\mathrm{N\_diff}}$ and $\mathcal{L}_{\mathrm{S\_diff}}$ are the normal-map loss and silhouette loss introduced in ICON (Chapter 3), and $\mathcal{L}_{\mathrm{J\_diff}}$ is the joint loss (L2) between 2D landmarks $\widehat{\mathcal{J}}^{\mathrm{c}}$, which are estimated by a 2D keypoint estimator from the RGB image $\mathcal{I}$, and the corresponding re-projected 2D joints $\mathcal{J}^{\mathrm{b}}$ from $\mathcal{M}^{\mathrm{b}}$. For more implementation details, see Appendix B.1.1.

4.2.2 Front and back surface reconstruction

We now lift the clothed normal maps to 2.5D surfaces. We expect these 2.5D surfaces to satisfy three conditions: (1) high-frequency surface details agree with predicted clothed normal maps, (2) low-frequency surface variations, including discontinuities, agree with SMPL-X’s ones, and (3) the depth of the front and back silhouettes are close to each other.

Unlike PIFuHD [186] or ICON (Chapter 3), which train a neural network to regress the implicit surface from normal maps, we explicitly model the depth-normal relationship

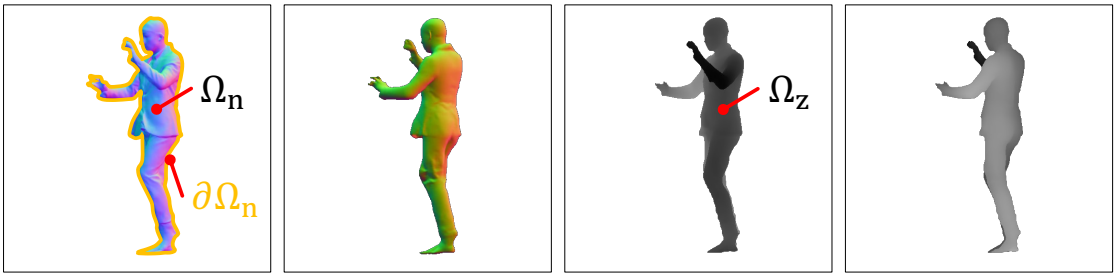


Figure 4.3: Four inputs to d-BiNI. Ω_n and Ω_z are the domains of clothed and body regions, respectively. $\partial\Omega_n$ is the silhouette of Ω_n.

using variational normal integration methods [22, 173]. Specifically, we tailor the recent bilateral normal integration (BiNI) method [22] to full-body mesh reconstruction by harnessing the coarse prior, depth maps, and silhouette consistency.

To satisfy the three conditions, we propose a depth-aware silhouette-consistent bilateral normal integration (d-BiNI) method to jointly optimize for the front and back clothed depth maps, $\widehat{\mathcal{Z}}^c_F$ and $\widehat{\mathcal{Z}}^c_B$:

$$\text{d-BiNI}(\widehat{\mathcal{N}}^c_F, \widehat{\mathcal{N}}^c_B, \mathcal{Z}^b_F, \mathcal{Z}^b_B) \rightarrow \widehat{\mathcal{Z}}^c_F, \widehat{\mathcal{Z}}^c_B. \tag{4.2}$$

Here, $\widehat{\mathcal{N}}^c_*$ is the front or back clothed normal map predicted by $\mathcal{G}^N_{F,B}$ from $\{\mathcal{I}, \mathcal{N}^b\}$, and $\mathcal{Z}^b_*$ is the front or back coarse body depth image rendered from the SMPL-X mesh, $\mathcal{M}^b$.

Specifically, the objective function consists of five terms:

$$\begin{aligned}\min_{\widehat{\mathcal{Z}}^c_F, \widehat{\mathcal{Z}}^c_B} & \mathcal{L}_n(\widehat{\mathcal{Z}}^c_F; \widehat{\mathcal{N}}^c_F) + \mathcal{L}_n(\widehat{\mathcal{Z}}^c_B; \widehat{\mathcal{N}}^c_B) + \\ & \lambda_d \mathcal{L}_d(\widehat{\mathcal{Z}}^c_F; \mathcal{Z}^b_F) + \lambda_d \mathcal{L}_d(\widehat{\mathcal{Z}}^c_B; \mathcal{Z}^b_B) + \\ & \lambda_s \mathcal{L}_s(\widehat{\mathcal{Z}}^c_F, \widehat{\mathcal{Z}}^c_B),\end{aligned} \tag{4.3}$$

where $\mathcal{L}_n$ is the BiNI loss term introduced by BiNI [22], $\mathcal{L}_d$ is a depth prior applied to the front and back depth surfaces, and $\mathcal{L}_s$ is a front-back silhouette consistency term. For a more detailed discussion on these terms, see Appendix B.1.2.

With Eq. (4.3), we make two technical contributions beyond BiNI [22]. First, we use the coarse depth prior rendered from the SMPL-X body mesh, $\mathcal{Z}^b_i$, to regularize BiNI:

$$\mathcal{L}_d(\widehat{\mathcal{Z}}^c_i; \mathcal{Z}^b_i) = |\widehat{\mathcal{Z}}^c_i - \mathcal{Z}^b_i|_{\Omega_n \bigcap \Omega_z} \quad i \in \{F, B\}. \tag{4.4}$$

This addresses the key problem of putting the front and back surfaces together in a coherent way to form a full body. Optimizing BiNI terms $\mathcal{L}_n$ leaves an arbitrary

global offset between the front and back surfaces. The depth prior terms $\mathcal{L}_{\mathrm{d}}$ encourage the surfaces with undecided offsets to be consistent with the SMPL-X body, and is computed in the domains $\Omega_{\mathrm{n}} \bigcap \Omega_{\mathrm{z}}$ (Fig. 4.3). For further intuitions on $\mathcal{L}_{\mathrm{n}}$ and $\mathcal{L}_{\mathrm{d}}$, see Fig. 4.4 and Fig. 4.5.

Secondly, we use a silhouette consistency term to ensure that the front and back depth values match at the silhouette boundary, computed within the domain $\partial\Omega_{\mathrm{n}}$ (Fig. 4.3):

$$\mathcal{L}_{\mathrm{s}}(\widehat{\mathcal{Z}}_{\mathrm{F}}^{\mathrm{c}}, \widehat{\mathcal{Z}}_{\mathrm{B}}^{\mathrm{c}}) = |\widehat{\mathcal{Z}}_{\mathrm{F}}^{\mathrm{c}} - \widehat{\mathcal{Z}}_{\mathrm{B}}^{\mathrm{c}}|_{\partial\Omega_{\mathrm{n}}}. \tag{4.5}$$

The silhouette term improves the physical consistency of the reconstructed front and back clothed depth maps. Without this term, d-BiNI produces intersections of the front and back surfaces around the silhouette, causing "blobby" artifacts and hurting reconstruction quality; see Fig. 4.6.

Hyper-parameters: d-BiNI has three hyper-parameters: λ_{d}, λ_{s}, and k. λ_{d} and λ_{s} are used in the objective function Eq. (4.3), which control the influence of coarse depth prior term Eq. (4.4) and silhouette consistency term Eq. (4.5) separately. k is used in the original BiNI [22] to control the surface stiffness (See Sup.Mat-A in BiNI [22] for more explanation of k). Empirically, we set $\lambda_{\mathrm{d}} = 1e^{-4}$, $\lambda_{\mathrm{s}} = 1e^{-6}$, and $k = 2$.

Discussion of hyper-paramters: Figure 4.4 shows the d-BiNI integration results under different values of k. It can be seen that a small k leads to tougher d-BiNI surfaces where discontinuities are not accurately recovered, while a large k softens the surface, and redundant discontinuities and noisy artifacts are introduced. Figure 4.5 shows the effects of λ_{d}, which controls how much d-BiNI surfaces agree on the SMPL-X mesh. Small λ_{d} causes misalignment between the d-BiNI surface and the SMPL-X mesh, which will produce stitching artifacts. While an excessively large λ_{d} enforces d-BiNI to rely overly on SMPL-X, thus smoothing out the high-frequency details obtained from normals. Figure 4.6 justifies the necessity of the silhouette consistency term. Without this term, the front and back d-BiNI surfaces intersect each other around the silhouettes, which will cause "blobby" artifacts after screened Poisson reconstruction [96].

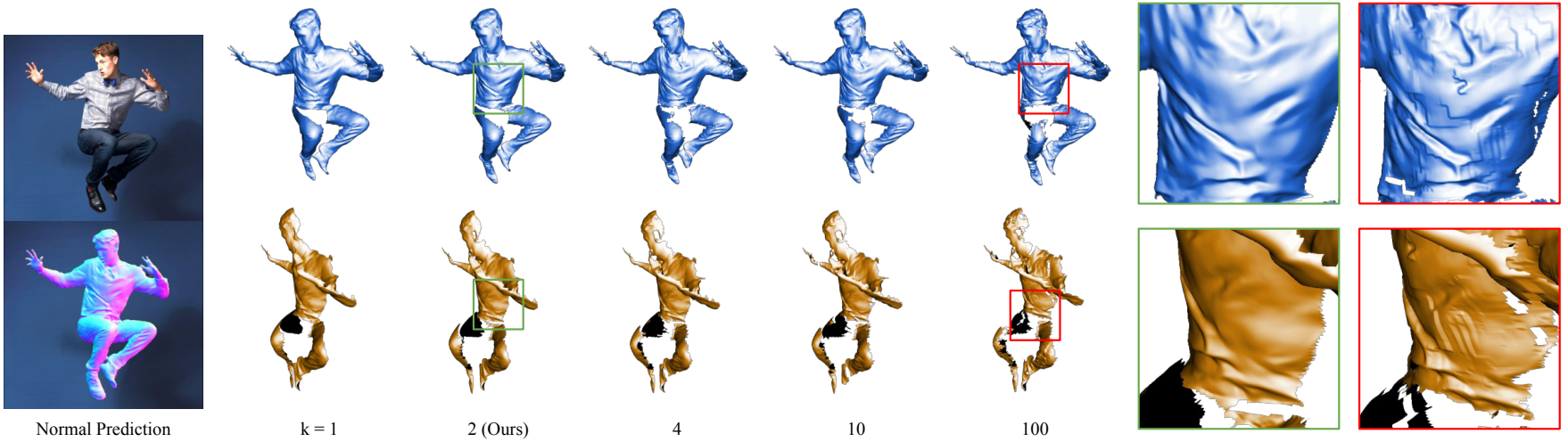


Figure 4.4: The effects of the hyper-parameter k on d-BiNI results. k controls the stiffness of the target surface [22]. A smaller k leads to smooth d-BiNI surfaces, while a large k introduces unnecessary discontinuities and noise artifacts.

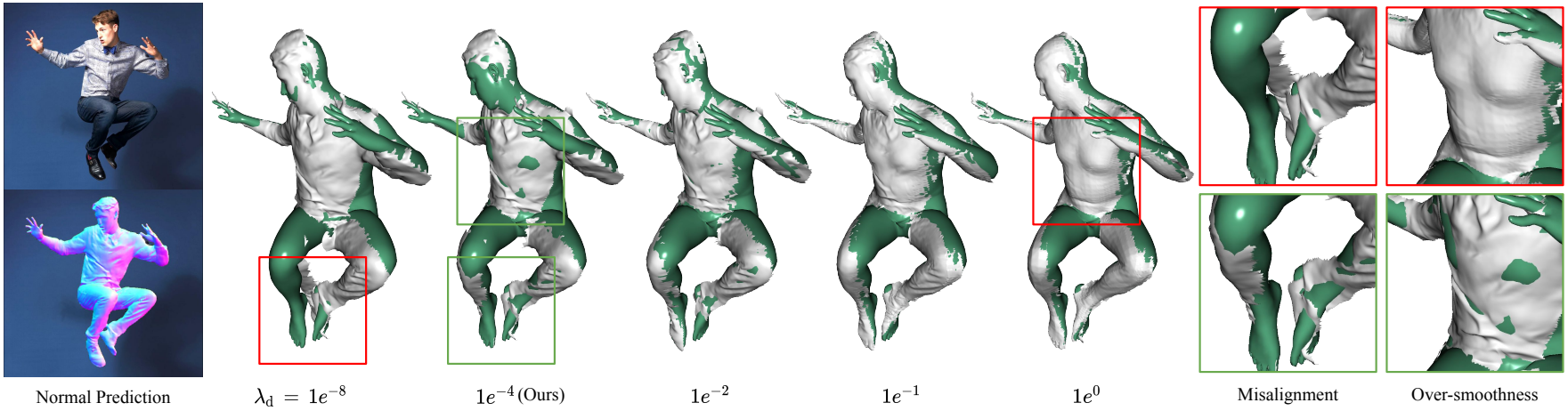


Figure 4.5: The effects of the hyperparameter λ_d on d-BiNI results. λ_d controls how much d-BiNI surfaces agree with the SMPL-X mesh. A small λ_d causes a misalignment between the d-BiNI surface and the SMPL-X mesh, thus it produces stitching artifacts. An excessively large λ_d enforces d-BiNI to rely too heavily on SMPL-X, thus it smooths out the high-frequency details obtained from normals.

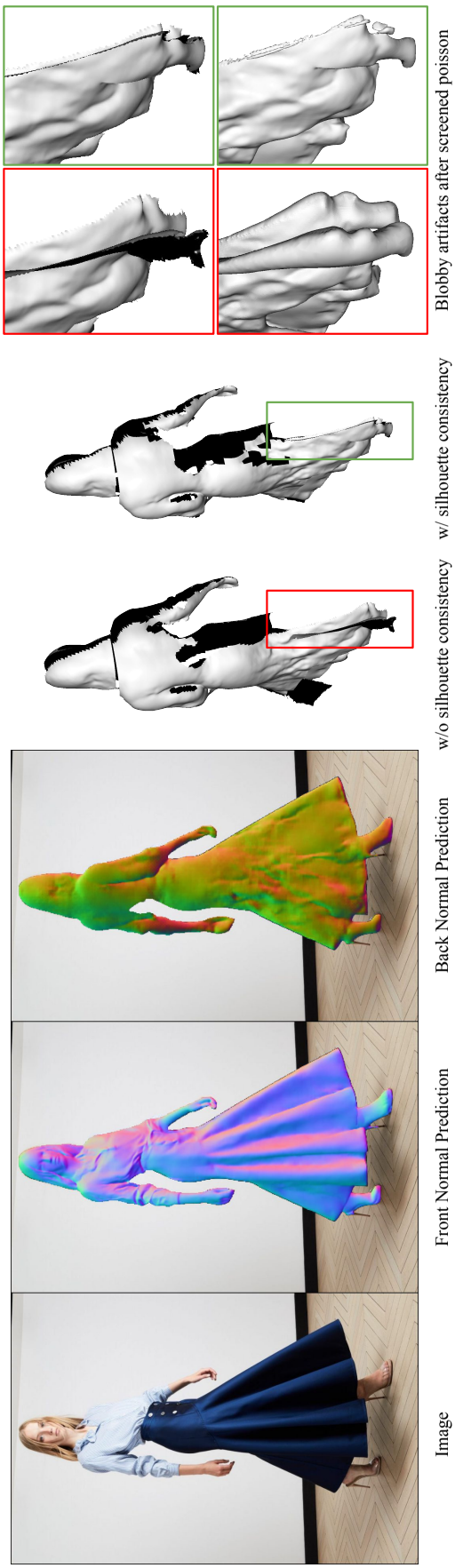


Figure 4.6: Necessity of silhouette consistency. This term can be regarded as the mediator between front and back d-BiNI surfaces, preventing these surfaces from intersecting. Such intersection causes blobby artifacts after screened Poisson reconstruction [96].

4.2.3 Human shape completion

For simple body poses without self-occlusions, merging front and back d-BiNI surfaces in a straightforward way, as done in FACSIMILE [194] and Moduling Humans [52], can result in a complete 3D clothed scan. However, often poses result in self-occlusions, which cause large portions of the surfaces to be missing. In such cases, Poisson Surface Reconstruction (PSR) [95] leads to blobby artifacts.

PSR completion with SMPL-X (ECON$_{EX}$). A naive way to "infill" the missing surface is to make use of the estimated SMPL-X body. We remove the triangles from $\mathcal{M}^{b}$ that are visible to front or back cameras. The remaining triangle "soup" $\mathcal{M}^{cull}$ contains both side-view boundaries and occluded regions. We apply PSR [95] to the union of $\mathcal{M}^{cull}$ and d-BiNI surfaces $\{\mathcal{M}_{F}, \mathcal{M}_{B}\}$ to obtain a watertight reconstruction $\mathcal{R}$. This approach is denoted as ECON$_{EX}$. Although ECON$_{EX}$ avoids missing limbs or sides, it does not produce a coherent surface for the originally missing clothing and hair surfaces because of the discrepancy between SMPL-X and actual clothing or hair; see ECON$_{EX}$ in Fig. 4.7.

Inpainting with IF-Nets+ ($\mathcal{R}_{IF}$). To improve reconstruction coherence, we use a learned implicit-function (IF) model to "inpaint" the missing geometry given front and back d-BiNI surfaces. Specifically, we tailor a general-purpose shape completion method, IF-Nets [33], to a SMPL-X-guided one, denoted as IF-Nets+. IF-Nets [33] completes the 3D shape from a deficient 3D input, such as an incomplete 3D human shape or a low-resolution voxel grid. Inspired by Li *et al*. [115], we adapt IF-Nets by conditioning it on a voxelized SMPL-X body to deal with pose variation; for details see Appendix B.1.3. IF-Nets+ is trained on voxelized front and back ground-truth clothed depth maps, $\{\mathcal{Z}^{c}_{F}, \mathcal{Z}^{c}_{B}\}$, and a voxelized (estimated) body mesh, $\mathcal{M}^{b}$, as input, and is supervised with ground-truth 3D shapes. During training, we randomly mask $\{\mathcal{Z}^{c}_{F}, \mathcal{Z}^{c}_{B}\}$ for robustness to occlusions. During inference, we feed the estimated $\widehat{\mathcal{Z}}^{c}_{F}$, $\widehat{\mathcal{Z}}^{c}_{B}$ and $\mathcal{M}^{b}$ into IF-Nets+ to obtain an occupancy field, from which we extract the inpainted mesh, $\mathcal{R}_{IF}$, with Marching cubes [141].

PSR completion with $\mathcal{R}_{IF}$ (ECON$_{IF}$). To obtain the final mesh, $\mathcal{R}$, we apply PSR to stitch (1) d-BiNI surfaces, (2) sided and occluded triangle soup $\mathcal{M}^{cull}$ from $\mathcal{R}_{IF}$, and optionally, (3) face or hands cropped from the estimated SMPL-X body $\mathcal{M}^{b}$. The

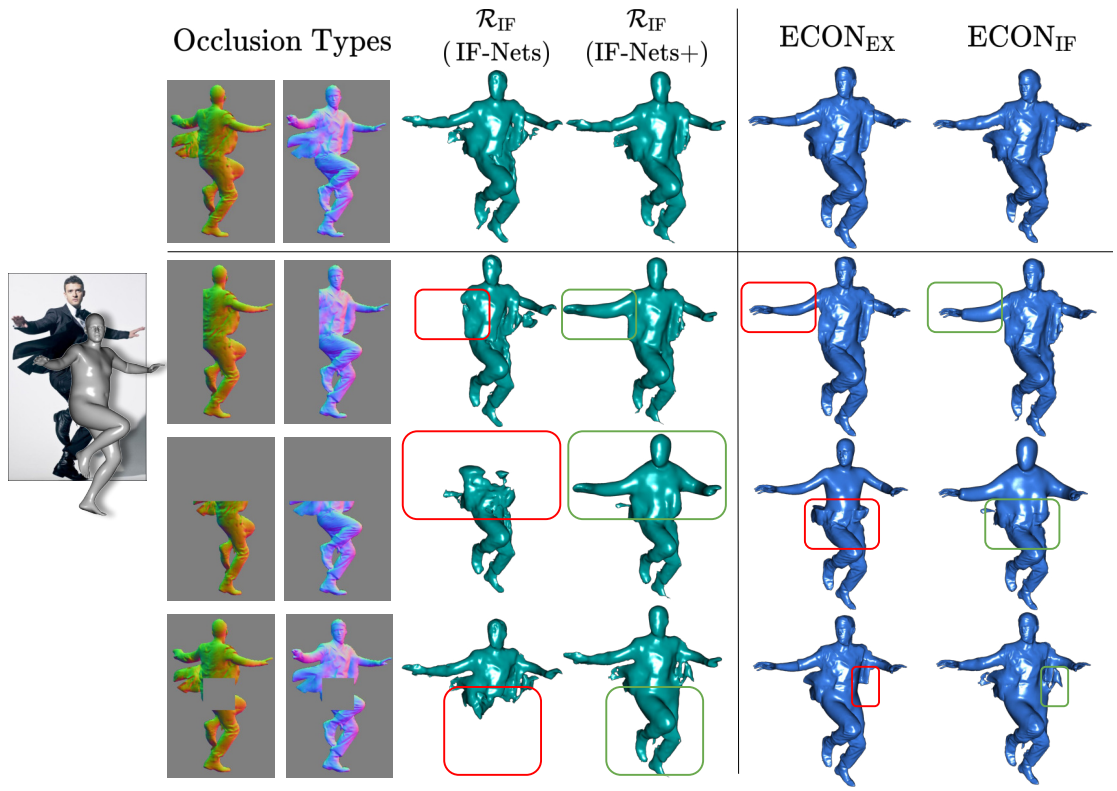


Figure 4.7: "Inpainting" the missing geometry. We simulate different occlusions by masking the normal images and present the intermediate and final 3D reconstruction of different design choices. While IF-Nets misses certain body parts, IF-Nets+ produces a plausible overall shape. ECON$_{\text{IF}}$ produces more consistent clothing surfaces than ECON$_{\text{EX}}$ due to a learned shape distribution.

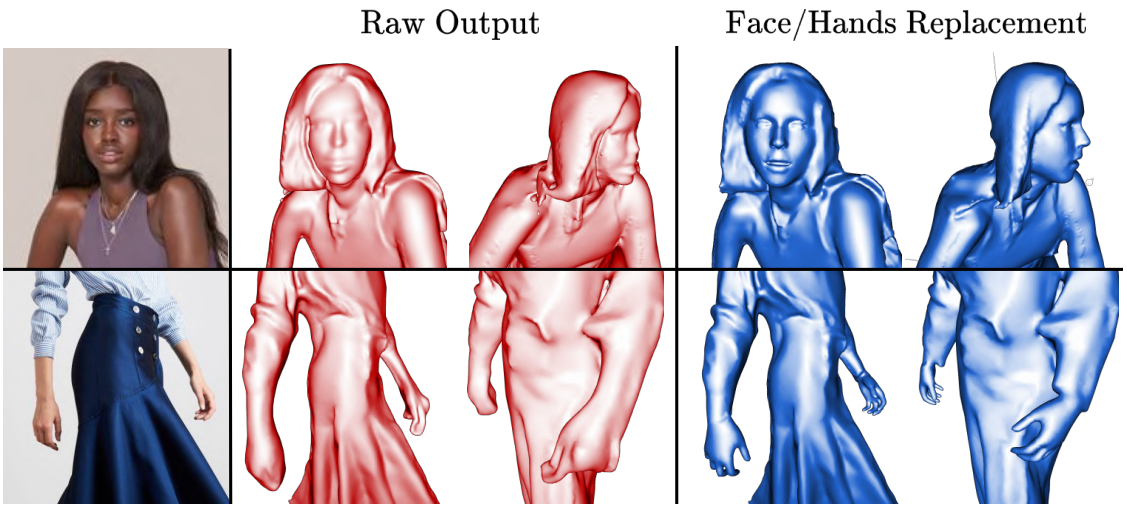


Figure 4.8: Face and hand details. The face and hands of the raw reconstruction (left), can be replaced with the ones of the SMPL-X body (right).

Figure 4.9: Datasets for numerical evaluation. We evaluate ECON on images with unseen poses (left) and unseen outfits (right) on the CAPE [144] and RenderPeople [179] datasets, respectively.

necessity of (3) arises from the poorly reconstructed hands/face in $\mathcal{R}_{\text{IF}}$; see difference in Fig. 4.8. The approach is denoted as ECON_{IF}.

Notably, although $\mathcal{R}_{\text{IF}}$ is already a complete human mesh, due to the lossy voxelization of inputs and limited resolution of the Marching cubes algorithm, it somehow smooths out the details of $\widehat{\mathcal{Z}}^{\text{c}}_{\{\text{F,B}\}}/\mathcal{M}^{\text{d-BiNI}}$, which are optimized via d-BiNI (see $\mathcal{R}_{\text{IF}}$ vs $\text{ECON}_{\{\text{IF,EX}\}}$ in Fig. 4.7). While $\text{ECON}_{\{\text{IF,EX}\}}$ preserves d-BiNI details better, only the side-views and occluded parts of $\mathcal{R}_{\text{IF}}$ are fused in the Poisson step. In Tabs. 4.1 and 4.4, we use $\text{ECON}_{\{\text{IF,EX}\}}$ instead of $\mathcal{R}_{\text{IF}}$ for evaluation.

4.3 Experiments

4.3.1 Datasets

Training on THuman: We use the same THuman [240] dataset as ICON (Chapter 3) to train ECON_{IF} (IF-Nets+), IF-Nets, PIFu and PaMIR.

Quantitative evaluation on CAPE & RenderPeople: Same as ICON Chapter 3, we primarily evaluate on CAPE [144] and RenderPeople [179]. Specifically, we use the "CAPE-NFP" set (100 scans), which is used by ICON (Chapter 3) to analyze robustness to complex human poses. Moreover, we select another 100 scans from RenderPeople, containing loose clothing, such as dresses, skirts, robes, down jackets, costumes, *etc*. With such clothing variance, RenderPeople helps numerically evaluate the flexibility of reconstruction methods w.r.t. shape topology. Samples of the two datasets are shown in Fig. 4.9.

Methods	Data-driven	OOD poses (CAPE)			OOD outfits (RenderPeople)		
		Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓
w/o SMPL-X body prior							
PIFu *	✓	1.722	1.548	0.0674	1.706	1.642	0.0709
PIFuHD†	✓	3.767	3.591	0.0994	1.946	1.983	0.0658
w/ GT SMPL-X body prior							
PaMIR *	✓	0.989	0.992	0.0422	1.296	1.430	0.0518
ICON	✓	0.971	0.909	0.0409	1.373	1.522	0.0566
ECON_{IF}	✓	0.996	0.967	0.0413	1.401	1.422	0.0516
ECON_{EX}	✗	0.926	0.917	0.0367	1.342	1.458	0.0478

Table 4.1: Evaluation against the state of the art. All models use a resolution of 256 for marching cubes. *Methods are re-implemented in [229] for a fair comparison in terms of network settings and training data. †Official model is trained on the RenderPeople dataset. ECON_{EX} is optimization-based, thus requires no training (✗). "OOD" is short for "out-of-distribution".

4.3.2 Metrics

Chamfer and P2S distance (cm): To capture large geometric errors, *e.g.* occluded parts or wrongly positioned limbs, we report the commonly used Chamfer (bi-directional point-to-surface) and P2S distance (1-directional point-to-surface) between ground-truth and reconstructed meshes.

Normal difference (L2): To measure the fineness of reconstructed local details, as well as projection consistency from the input image, we also report the L2 error between normal images rendered from reconstructed and ground-truth surfaces, by rotating a virtual camera around these by $\{0°, 90°, 180°, 270°\}$ w.r.t. to a frontal view.

4.3.3 Evaluation

Quantitative evaluation: We compare ECON with body-agnostic methods, *i.e.*, PIFu [185] and PIFuHD [186], and body-aware methods, *i.e.*, PaMIR [256] and ICON (Chapter 3); see in Tab. 4.1. For fair comparison, we use re-implementations of PIFu and PaMIR as in Chapter 3, because they have the same network settings and input data. ECON_{EX} performs on par with ICON, and outperforms other methods on images containing out-of-distribution (OOD) poses (CAPE), with a distance error below 1cm. In terms of out-of-distribution outfits (RenderPeople), $\text{ECON}_{\textbf{EX/IF}}$ performs on par with PaMIR,

	ICON (Chapter 3)	PIFuHD [186]	PaMIR [256]
Challenging poses	0.283	0.108	0.132
Loose clothing	0.147	0.362	0.232
Fashion images	0.199	0.551	0.290

Table 4.2: Perceptual study. Numbers denote the chance that participants prefer the reconstruction of a competing method over ECON for in-the-wild images. A value of 0.5 indicates equal preference. A value of < 0.5 favors ECON, while of > 0.5 favors competitors.

and much better than PIFuHD. When it comes to high-frequency details measured by normals, ECON_{EX} achieves SOTA performance on both datasets.

Perceptual study: Due to the lack of ground-truth geometry (clothed scan + underneath SMPL-X), we further conduct a perceptual study to evaluate ECON on in-the-wild images. Test images are divided into three categories: "challenging poses", "loose clothing", and "fashion images". Examples of challenging poses and loose clothing can be seen in Fig. 4.12, and fashion poses are in Fig. 4.10.

Figure 4.10: SHHQ 3D reconstruction. For each image we show a **front** and **side** view of ECON's reconstruction and a **SMPL-X** fit.

Participants are asked to choose the reconstruction they perceive as more realistic, between a baseline method and ECON. We compute the chances that each baseline is preferred over ECON in Tab. 4.2. The results of the perceptual study confirm the quantitative evaluation in Tab. 4.1. For "challenging poses" images, ECON is significantly preferred over PIFuHD and outperforms ICON. On images of people wearing loose clothing, ECON is preferred over ICON by a large margin and outperforms PIFuHD. The

Methods	OOD poses (CAPE [144])		OOD outfits (RenderPeople) [179]		Speed
	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	FPS ↑
BiNI [22]	27.64	21.11	20.61	16.07	0.52
d-BiNI	**13.43**	**10.29**	**14.43**	**11.26**	**0.69**

Table 4.3: BiNI vs d-BiNI. Comparison between BiNI and d-BiNI surfaces w.r.t. reconstruction accuracy and optimization speed.

Methods	OOD poses (CAPE)			OOD outfits (RenderPeople)		
	Chamfer ↓	P2S ↓	Normals ↓	Chamfer ↓	P2S ↓	Normals ↓
IF-Nets [33]	2.116	**1.233**	0.075	1.883	1.622	0.070
IF-Nets+	**1.401**	1.353	**0.056**	**1.477**	**1.564**	**0.055**
ECON_{IF}	0.996	0.967	0.0413	1.401	1.422	0.0516

Table 4.4: Evaluation for shape completion. Same metrics as Tab. 4.1, and ECON_{IF} is added as a reference.

reasons for a slight preference of PIFuHD over ECON on fashion images are discussed in Sec. 4.4. Figure 4.1 visualizes some comparisons. More examples are provided in Figs. B.3 to B.5.

4.3.4 Ablation study

d-BiNI vs BiNI: We compare d-BiNI with BiNI using 600 samples (200 scans x 3 views) from CAPE and RenderPeople where ground-truth normal maps and meshes are available. Table 4.3 reports the "root mean squared error" (RMSE) and "mean absolute error" (MAE) between the estimated and rendered depth maps. d-BiNI significantly improves the reconstruction accuracy by about 50% compared to BiNI. This demonstrates the efficacy of using the coarse body mesh as regularization and taking the consistency of both the front and back surface into consideration. Additionally, d-BiNI is 33% faster than BiNI.

IF-Nets+ vs IF-Nets: Following the metrics of Sec. 4.3.2, we compare IF-Nets [33] with the updated IF-Nets+ used by ECON on $\mathcal{R}_{\text{IF}}$. We show the quantitative comparison in Tab. 4.4. The improvement for out-of-distribution ("OOD") poses shows that IF-Nets+ is more robust to pose variations than IF-Nets, as it is conditioned on the SMPL-X body. Figure 4.7 compares the geometry "inpainting" of both methods in the case of occlusions.

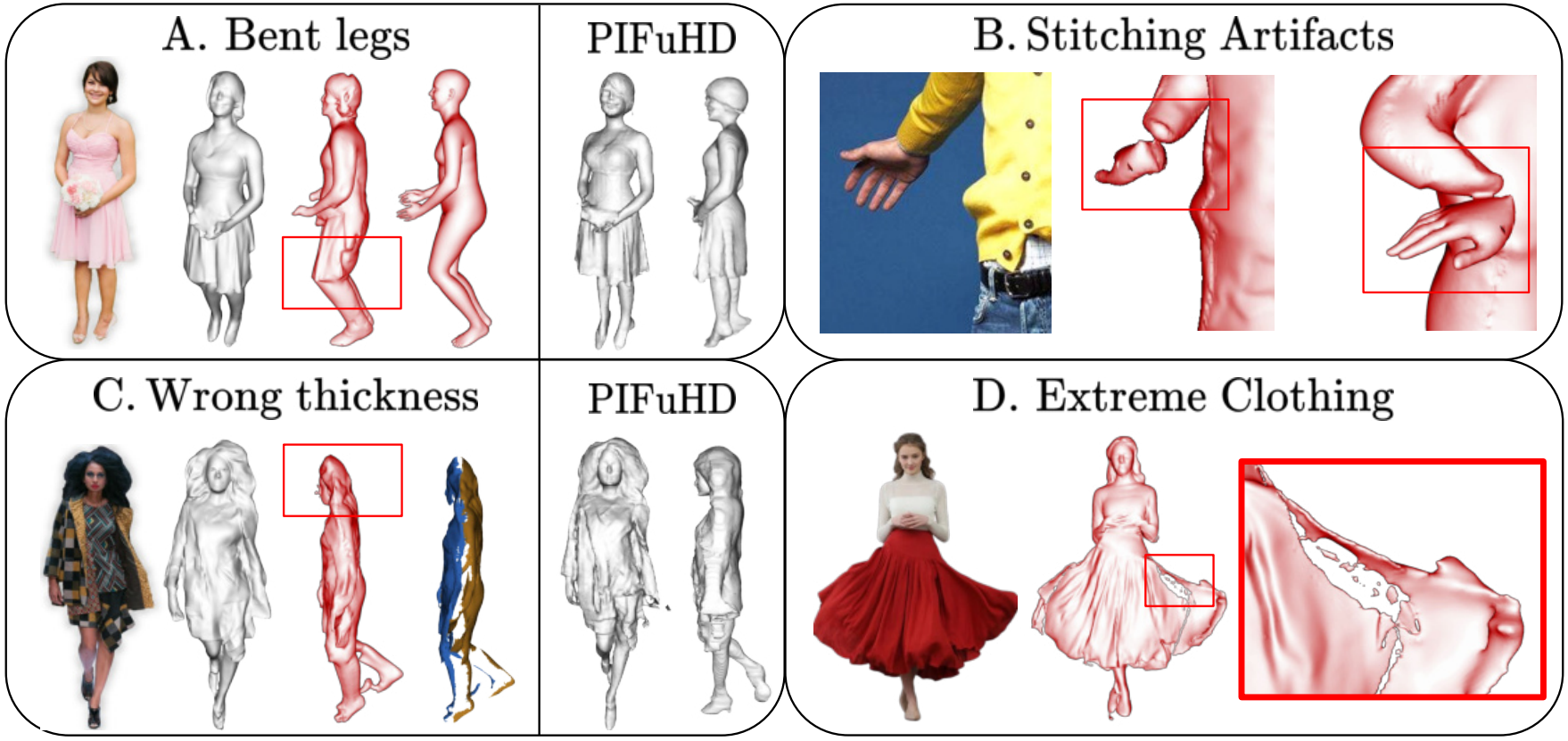
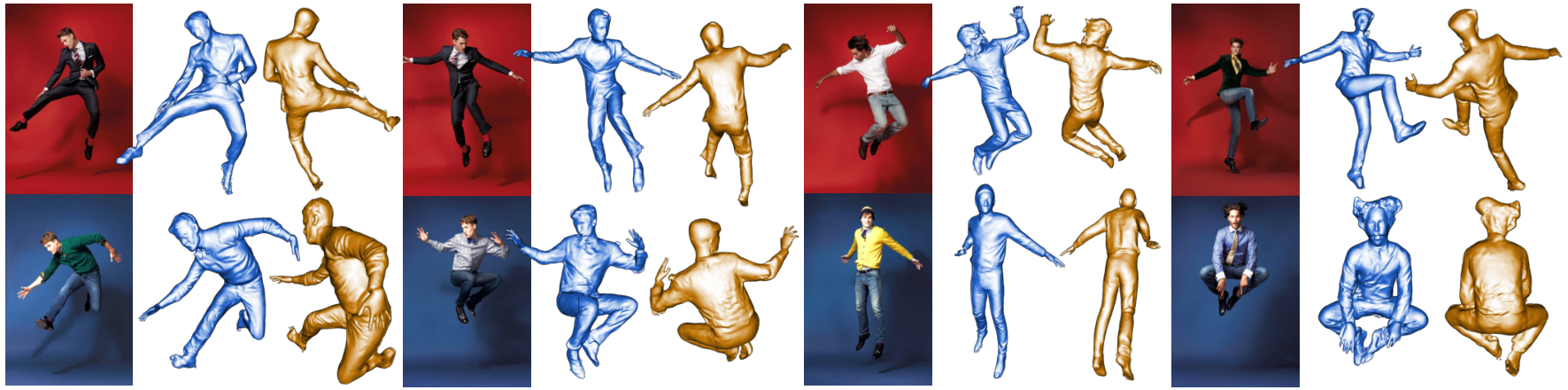


Figure 4.11: Failure examples of ECON. (A-B) Failures in recovering a SMPL-X body result, *e.g.*, bent legs or wrong limb poses, cause ECON failures by extension. (C-D) Failures in normal-map estimation provide erroneous geometry to ECON to work with.



(a) Humans with challenging poses (🔍 **zoom in** to see 3D wrinkle details)

(b) Humans with loose clothing (🔍 **zoom in** to see 3D clothing details)

Figure 4.12: Qualitative results on in-the-wild images. We show 8 examples of reconstructing detailed clothed 3D humans from images with: (a) challenging poses and (b) loose clothing. For each example we show the input image along with two views (**front** and **rotated**) of the reconstructed 3D humans. ECON is robust to pose variations, generalizes well to loose clothing, and contains detailed geometry.

4.3.5 Multi-person reconstruction

Thanks to the shape completion module, ECON can deal with occlusions. Unlike other crowd body estimators [198–200, 238], ECON makes it possible to reconstruct multiple detailed "clothed" 3D humans from an image with inter-person occlusions, even though ECON has not been trained for this. Figure 4.13 shows three examples. The occluded parts, colored in red, are successfully recovered.

4.4 Discussion

Limitations: ECON takes as input an RGB image and an estimated SMPL-X body. However, recovering SMPL-X bodies (or similar models) from a single image is still an open problem, and not fully solved. Any failure in this could lead to ECON failures, such as in Fig. 4.11-A and Fig. 4.11-B. As the synthetic data [17, 70, 219] is getting sufficiently realistic, their domain gap with real data is significantly narrowed, it is predictable that such limitations will be eliminated. The reconstruction quality of ECON primarily relies on the accuracy of the predicted normal maps. Poor normal maps can result in overly close-by or even intersecting front and back surfaces, as shown in Fig. 4.11-C and Fig. 4.11-D.

Future work: Apart from addressing the above limitations, several other directions are useful for practical applications. Currently, ECON reconstructs only 3D geometry. One could additionally recover an underlying skeleton and skinning weights [116, 232], to obtain fully-animatable avatars. Moreover, generating back-view texture [28, 180, 181, 254] would result in fully-textured avatars. Disentangling clothing [2, 171, 262], hairstyle [242], or accessories [53] from the recovered geometry, would enable the simulation [61], synthesis, editing and transfer of styles [49] for these. ECON's reconstructions, together with its underneath SMPL-X body, could be useful as 3D shape prior to learn neural avatars [43, 62, 86].

In particular, ECON could be used to augment existing datasets of 2D images with 3D humans. Datasets of real clothed humans with 3D ground truth [144, 164, 179, 240, 257] are limited in size. In contrast, datasets of images without 3D ground truth are widely

available in large sizes [50, 58, 139]. We can "augment" such datasets by reconstructing detailed 3D humans from their images. We apply ECON on SHHQ [50] and recover normal maps and 3D humans; see Fig. 4.10. As ECON-like methods mature, they could produce pixel-aligned 3D humans from photos at scale, enabling the training of generative models of 3D clothed avatars with details [56, 60, 71, 156, 201, 226, 252].

Possible negative impact: As the reconstruction matures, it opens the potential for low-cost realistic avatar creation. Although such a technique benefits entertainment, film production, tele-presence and future metaverse applications, it could also facilitate deep-fake avatars. Regulations must be established to clarify the appropriate use of such technology.

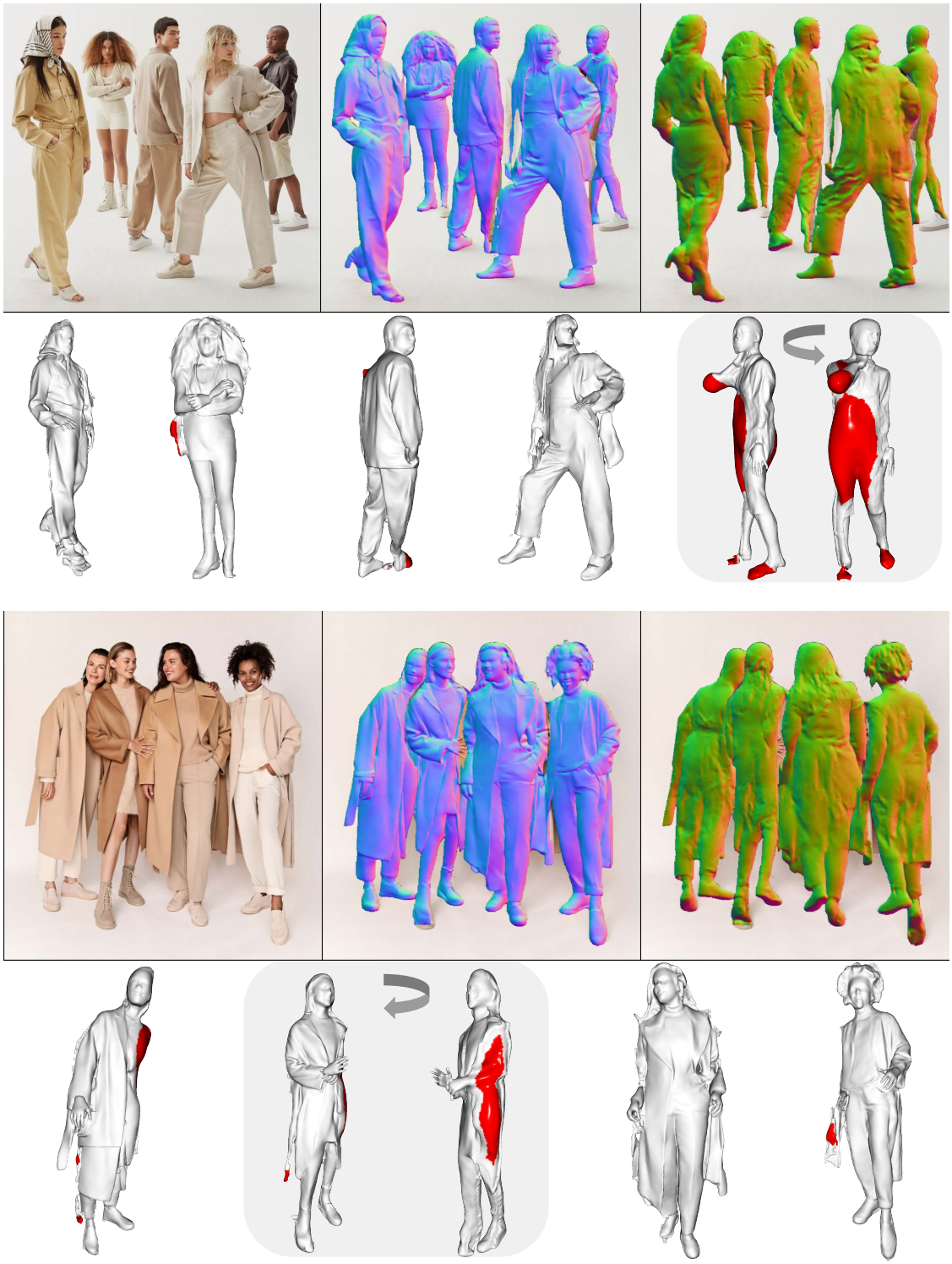


Figure 4.13: Multiple humans with occlusions. We detect multiple people and apply ECON to each separately. Although ECON is not trained on multiple people, it is robust to inter-person occlusions. We show three examples, and for each: (top) input image and the predicted front and back normal maps, (bottom) ECON's reconstruction. Red areas on the estimated mesh indicate occlusions.

5

TeCH: Text-guided Reconstruction of Lifelike Clothed Humans

Contents

5.1 Introduction

ICON (Chapter 3) and ECON (Chapter 4) primarily aim to improve the geometry quality of 3D reconstructions under challenging poses and with loose clothing. However, single-image-based reconstruction of fully clothed humans remains an ill-posed problem due to

the lack of observations in non-visible areas, which will harm the details of both geometry and texture, especially on the backsides. Efforts to predict *invisible* regions (*e.g.* back-side) based on *visible* visual cues (*e.g.* colors [7, 80, 185], normal estimates [186, 228, 229]) have proven unsuccessful, resulting in the blurry texture and smoothed-out geometry, see Fig. 5.7. As a result, inconsistencies arise when observing these reconstructions from different angles. To address this issue, introducing multi-view supervision could be a potential solution. But is it feasible given only a single input image? Here, we propose TeCH to answer this question. Unlike prior research that primarily explores the connection between visible frontal cues and non-visible regions, TeCH integrates textual information derived from the input image with a personalized Text-to-Image diffusion model, *i.e.*, DreamBooth [183], to guide the reconstruction process.

Specifically, we divide the information from the single input image into the semantic information that can be accurately described by texts and subject's distinctive and fine-detailed appearance which is not easily describable by text:

1) Describable semantic prompts, including the detailed descriptions of colors, styles of garments, hairstyles, and facial features, are *explicitly* parsed from the input image using a garment parsing model (*i.e.*, SegFormer [225]) and a pre-trained visual-language VQA model (*i.e.*, BLIP [114]).

2) Indescribable appearance information, which *implicitly* specifies the subject's distinctive appearance and fine-grained details, is embedded into a unique token "[V]", by a personalized Text-to-Image (T2I) diffusion model [183].

Based on these information sources, we optimize the 3D human using multi-view Score Distillation Sampling (SDS) [169], reconstruction losses based on the original observations, and regularization obtained from off-the-shelf normal estimators, to enhance the fidelity of the reconstructed 3D human models while preserving their original identity. To represent a high-resolution geometry at an affordable memory cost, we propose a hybrid 3D representation based on DMTet [55, 190]. This hybrid 3D representation combines an explicit tetrahedral grid to approximate the overall body, with an implicit signed distance field (SDF) and RGB field to capture fine details in geometry and texture. In a two-stage optimization process, we first optimize this tetrahedral grid, extract the geometry represented as a mesh, and then optimize the texture.

TeCH enables the reconstruction of high-fidelity 3D clothed humans with detailed full-body geometry, and intricate textures with consistent color and patterns. As a result,

it facilitates various downstream applications such as novel view rendering, character animation, and shape & texture editing. Quantitative evaluations performed on 3D clothed human datasets, covering various poses (CAPE [168]) and outfits (THuman2.0 [240]), demonstrate TeCH's superiority in reconstructing geometric details. Qualitative comparisons conducted on in-the-wild images, accompanied by a perceptual study, further confirm that TeCH surpasses SOTA methods in terms of rendering quality.

5.2 Method

Given a single image as input, TeCH aims to reconstruct a high-fidelity 3D clothed human. As depicted in Fig. 5.1, TeCH follows a two-step procedure: Firstly, a text prompt that describes the human in the input image is obtained via the human parsing model SegFormer [225] and the VQA model BLIP [114] (Sec. 5.2.1). This descriptive prompt is used to guide the generation process in DreamBooth [183], a personalized Text-to-Image diffusion model fine-tuned on augmented input images. Secondly, the 3D human, which is represented as hybrid DMTet and initialized with SMPL-X (Sec. 5.2.3), is optimized with Score Distillation Sampling (SDS) losses [169] computed from the personalized DreamBooth (Sec. 5.2.4). Note that the SDS loss has been introduced in DreamFusion [169] for the task of Text-to-3D generation of general objects, by optimizing a neural radiance field (NeRF) with gradients from a frozen diffusion model. For these preliminaries, we refer the reader to Appendix C.1.

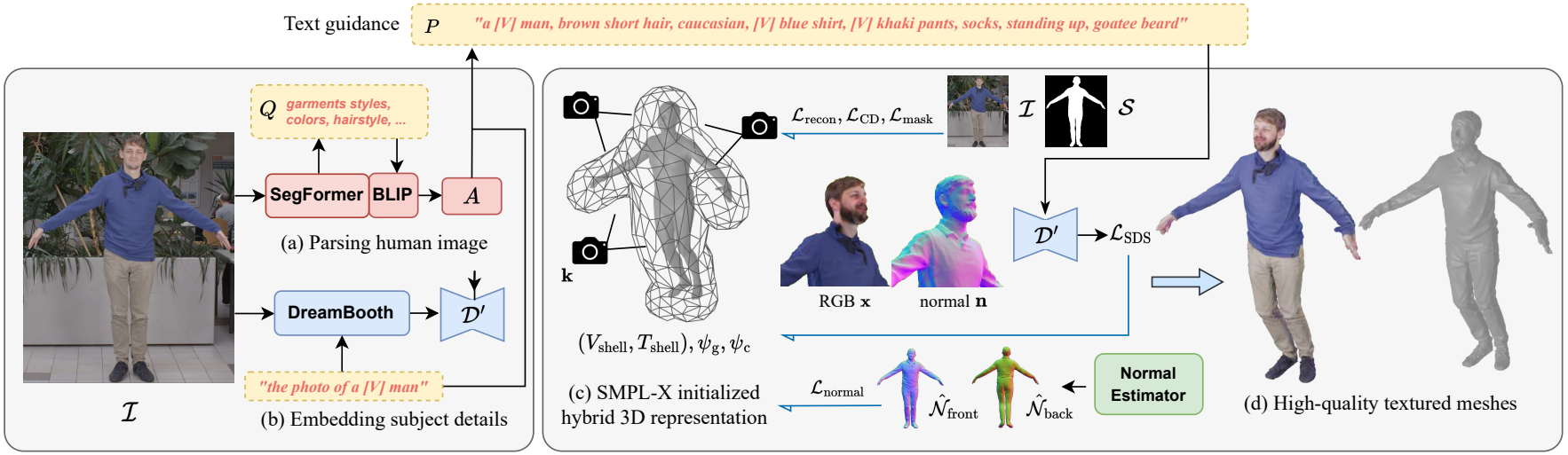
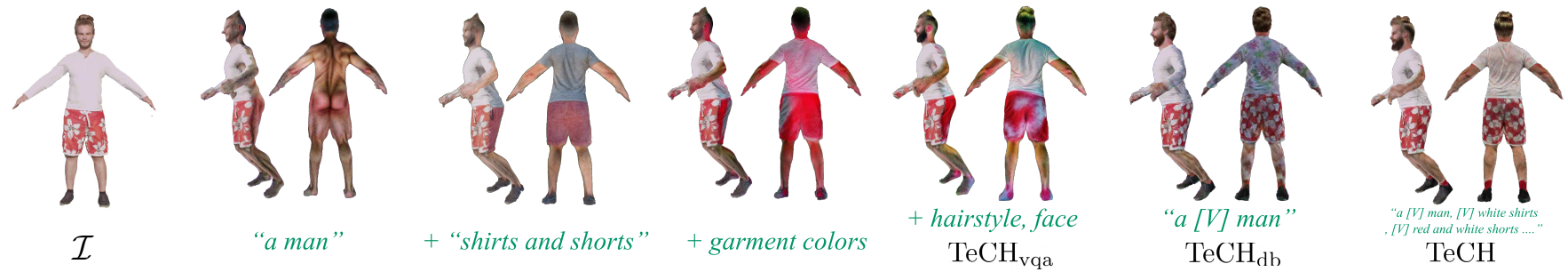


Figure 5.1: Method overview. TeCH takes an image $\mathcal{I}$ of a human as input. Text guidance is constructed through **(a)** using a garment parsing model (SegFormer) and a VQA model (BLIP) to parse the human attributes A with pre-defined problems Q, and **(b)** embedding with subject-specific appearance into DreamBooth $\mathcal{D}'$ as unique token $[V]$. Next, TeCH represents the 3D clothed human with **(c)** DMTet initialized by SMPL-X body, and optimizes both geometry and texture using $\mathcal{L}_{\text{SDS}}$ guided by prompt $P = [V] + P_{\text{VQA}}(A)$. During the optimization, $\mathcal{L}_{\text{recon}}$ is introduced to ensure input view consistency, $\mathcal{L}_{\text{CD}}$ is used to enforce the color consistency between different views, and $\mathcal{L}_{\text{normal}}$ serves as a surface regularizer. Finally, the extracted textured meshes **(d)** are ready to be used in various downstream applications.

5.2.1 Extracting Text-guidance from the Observation

Parsing human attributes. As depicted in Fig. 5.3, given the input image of a human, SegFormer [225], which is fine-tuned on the ATR dataset [127, 128], is applied to recognize each part of the garments (*e.g.* hat, skirt, pants, belt, shoes). To obtain detailed descriptions of the parsed garments, we utilize the vision-language model BLIP [114] as a VQA captioner. This model has been pre-trained on a vast collection of image-text pairs, enabling it to automatically generate descriptive prompts. Rather than using naive image captioning, we employ a series of fine-grained VQA questions $\{Q_i\}$ as input to BLIP (see Appendix C.2). These questions cover garment styles, colors, facial features, and hairstyles, with the corresponding answers denoted as $\{A_i\}$. The set of $\{A_i\}$ is inserted into a predefined template to create text prompts P_{VQA}, which serve as text-guidance to condition the text-to-image diffusion model.

Embedding subject-specific appearance. Does the text prompt P_{VQA} comprehensively capture all the visual characteristics of the subject? No, a picture is worth a thousand words. Thus, we utilize DreamBooth [183] to learn the *indescribable* visual appearance. DreamBooth is a method for "personalizing" a diffusion model through few-shot tuning (3$\sim$5 images). We perform DreamBooth's fine-tuning on a pre-trained Stable Diffusion (v1.5) as the base model. To generate the needed inputs, we augment the single input image with five different backgrounds. To prevent language drift, we assign the subject classes "`man`" or "`woman`" based on the gender determined by the VQA. After fine-tuning DreamBooth, the subject-specific distinctive appearance is encoded within a unique identifier token "$[V]$". We insert "$[V]$" into the prompt P_{VQA}, to construct the final text prompt P used by the personalized DreamBooth $\mathcal{D}'$. Figure 5.4 illustrates how these individual prompts contribute to the final appearance; additional information is provided in Sec. 5.2.2.



(a) Detailed ablation results of textual guidance

(b) Editing garment colors with subject-specific token

Figure 5.2: Impact of textual guidance. (a) Top depicts the impact of specific elements within the textual guidance, such as garment styles & colors, hairstyle, facial features, and the placement & inclusion of "$[V]$". **(b)** Bottom demonstrates that TeCH facilities text-guided garment color editing.

5.2.2 Deeper analysis of description P

In Fig. 5.2 (a), we first show the impact of individual elements within the text prompt, including garment styles & colors, hairstyle, and face, which guide the model to recover the appearance of each attribute of the clothed human. The first column shows that a basic class description alone cannot effectively guide the reconstruction process. However, in the subsequent columns, text guidance incorporating detailed descriptions of clothing proves successful in accurately reconstructing the structure of clothed humans. Furthermore, with additional information regarding colors and hairstyles, the characters reconstructed by $\mathrm{TeCH}_{\mathrm{vqa}}$ exhibit greater semantic consistency to the input view. However, merely relying on VQA descriptions is insufficient for generating a convincing appearance.

Only using the DreamBooth guidance ($\mathrm{TeCH}_{\mathrm{db}}$), helps to recover original garment patterns, which demonstrates that DreamBooth has a high-level understanding of texture patterns. However, it sometimes will *diffuse* the patterns to the entire human. By combining "$[V]$" with the VQA parsing text prompts P_{VQA}, TeCH produces remarkably realistic texture with consistent color and intricate patterns.

In Fig. 5.2 (b), we also demonstrate some text-guided garment color editing examples based on a fine-tuned DreamBooth model D' and subject-specific token "$[V]$".

5.2.3 Hybrid 3D Representation

To efficiently represent the 3D clothed human at a high resolution, we embed DMTet [55, 190] around the SMPL-X body mesh [157]. Specifically, we construct a compact tetrahedral grid $(V_{\mathrm{shell}}, T_{\mathrm{shell}})$ within an outer shell M_{shell}, shown in Fig. 5.1-(c). Compared to the DMTet cubic-based tetrahedral grid, the outer shell tetrahedral grid is more computationally efficient for high-resolution geometry modeling of a human. Using PIXIE [47], we estimate an initial body $\mathcal{M}_{\mathrm{body}}$. To create M_{shell}, a series of mesh dilation, down-sampling, and up-sampling steps are applied to the body mesh M_{body} (see details in Appendix C.3).

We use two MLP networks $\Psi_{\mathrm{g}}, \Psi_{\mathrm{c}}$ with hash encoding [153], parameterized by ψ_g and, ψ_c to learn the geometry and color separately. The geometry network Ψ_{g}

predicts the SDF value $\Psi_{\mathrm{g}}(v_i) = s(v_i; \psi_{\mathrm{g}})$ of each DMTet vertex v_i. It is initialized by fitting it to the SDF of M_{shell}:

$$\mathcal{L}_{\mathrm{init}} = \sum_{p_i \in \mathbf{P}} \left\| s(p_i; \psi_{\mathrm{g}}) - \mathrm{SDF}(p_i) \right\|_2^2, \tag{5.1}$$

where $\mathbf{P} = \{p_i \in \mathbb{R}^3\}$ is a point set randomly sampled near M_{shell}, and $\mathrm{SDF}(p_i)$ is the pre-computed pointwise SDF. Triangular meshes can be extracted from this efficient hybrid 3D representation by Marching Tetrahedra (MT) [42]:

$$M = \mathrm{MT}(V_{\mathrm{shell}}, T_{\mathrm{shell}}, s(V_{\mathrm{shell}}; \psi_{\mathrm{g}})). \tag{5.2}$$

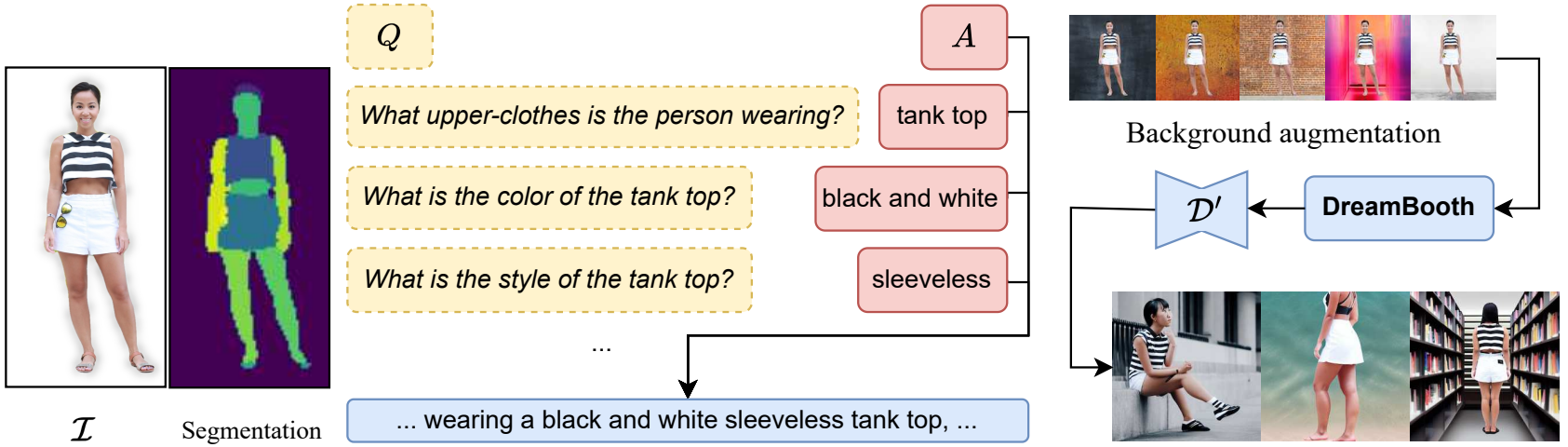
Given the camera parameters $\mathbf{k}$ (see details in Appendix C.4), the generated mesh is rendered through differentiable rasterization $\mathcal{R}$ [110], to get the back-projected 3D locations $\mathcal{P}(M, \mathbf{k})$, mask $\mathcal{M}(M, \mathbf{k})$, and rendered normal image $\mathcal{N}(M, \mathbf{k})$:

$$\mathcal{R}(M, \mathbf{k}) = (\mathcal{P}(M, \mathbf{k}), \mathcal{M}(M, \mathbf{k}), \mathcal{N}(M, \mathbf{k}))\,. \tag{5.3}$$

The albedo of each back-projected pixel is predicted by the color network $\Psi_{\mathbf{c}}$, where ψ_{c} represents the parameters:

$$\mathcal{I}'(M, \psi_{\mathrm{c}}, \mathbf{k}) = \Psi_{\mathbf{c}}(\psi_{\mathrm{c}}, \mathcal{P}(M, \mathbf{k})). \tag{5.4}$$

As detailed in Sec. 5.2.4, we optimize this 3D representation using a coarse-to-fine strategy by applying successive subdivisions on the tetrahedral grids. Specifically, a more detailed surface $M_{\mathrm{subdiv}}(\psi_{\mathrm{g}})$ can be obtained by applying volume subdivision on the surface tetrahedral grids $(V_{\mathrm{surface}}, T_{\mathrm{surface}})$ that intersect with $M(\psi_{\mathrm{g}})$. Note that the SDF values of the refined vertices are still inferred by Ψ_{g}.



(a) Descriptive prompt P_{VQA} (b) Subject-specific generation from $[V]$

Figure 5.3: Prompt construction ($P = P_{\text{VQA}} + [V]$). (a) Inquire VQA model with predefined questions on individual appearance to construct *describable* prompts P_{VQA}. (b) Fine-tuned DreamBooth with background-augmented images to embed *indescribable* subject-specific details into a unique identifier $[V]$.

Figure 5.4: The effects of text guidance. We compare the effectiveness of using only VQA descriptions (TeCH_{vqa}), only the DreamBooth identity token (TeCH_{db}), and both of them (TeCH), in the SDS optimization process.

5.2.4 Multi-stage Optimization

We adopt a multi-stage, coarse-to-fine optimization process to sequentially recover the subject's geometry and texture. In the initial stage, we utilize the tetrahedral representation to model the subject's geometry. Next, the appearance is recovered using the mesh that is extracted from the tetrahedral grid. Both stages leverage SDS-based losses using the personalized DreamBooth model, which provides multi-view supervision by sampling new camera views. **Geometry Stage:** We optimize the geometry based on a silhouette loss $\mathcal{L}_{\text{sil}}$ using the original image, a text-guided SDS loss on rendered normal images $\mathcal{L}_{\text{SDS}}^{\text{norm}}$, and geometric regularization $\mathcal{L}_{\text{reg}}$ based on predicted normals $\mathcal{L}_{\text{norm}}$ and surface smoothness $\mathcal{L}_{\text{lap}}$:

$$\begin{aligned} \mathcal{L}_{\text{geometry}} &= \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{SDS}}\mathcal{L}_{\text{SDS}}^{\text{norm}} + \mathcal{L}_{\text{reg}} \\ \mathcal{L}_{\text{reg}} &= \lambda_{\text{norm}}\mathcal{L}_{\text{norm}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}}, \end{aligned} \tag{5.5}$$

where λ represents the weights to balance the losses. During optimization of this loss, we perform a coarse-to-fine subdivision on DMTet, to robustly produce a high-resolution mesh for the clothed body. Specifically, the optimization is first performed w/o subdivision for $t_{\text{coarse}} = 5000$ iterations, and then with subdivision for $t_{\text{fine}} = 5000$ iterations.

Pixel-aligned silhouette loss. The silhouette loss [251] enforces pixel alignment with the foreground mask $\mathcal{S}$ of the input image $\mathcal{I}$ under the input camera view $\mathbf{k}$:

$$\mathcal{L}_{\text{sil}} = \|\mathcal{S} - \mathcal{M}(M, \mathbf{k})\|_2^2 + \sum_{x \in \text{Edge}(\mathcal{M}(M,\mathbf{k}))} \min_{\hat{x} \in \text{Edge}(\mathcal{S})} \|x - \hat{x}\|_1 . \tag{5.6}$$

It consists of (1) a pixel-wise L2 loss over the foreground mask $\mathcal{S}$ and the rendered silhouette $\mathcal{M}$, and (2) an edge distance loss, based on the distance of each silhouette boundary pixel $x \in \text{Edge}(\mathcal{M}(M, \mathbf{k}))$ to the nearest foreground mask boundary pixel $\hat{x} \in \text{Edge}(\mathcal{S})$.

SDS loss on normal images. Inspired by Fantasia3D [29], TeCH integrates normal renderings with the SDS loss [169]. It enables TeCH to effectively capture intricate geometric details without rendering the color image. Given the surface normals $\mathbf{n} = \mathcal{N}(M, \mathbf{k})$, $\mathcal{L}_{\text{SDS}}^{\text{norm}}$ is defined as:

$$\mathcal{L}_{\text{SDS}}^{\text{norm}} = \nabla_{\psi_{\text{g}}} \mathcal{L}_{\text{SDS}}^{\text{norm}}(\mathbf{n}, \mathbf{c}^{P_{\text{norm}}}) = \mathbb{E}_{\text{t},\varepsilon} \left[w_t \left(\hat{\varepsilon}_{\phi'}(\mathbf{z}_t^{\mathbf{n}}; \mathbf{c}^{P_{\text{norm}}}, t) - \varepsilon \right) \frac{\partial \mathbf{n}}{\partial \psi_{\text{g}}} \frac{\partial \mathbf{z}^{\mathbf{n}}}{\mathbf{n}} \right], \tag{5.7}$$

where $\mathbf{c}^{P_{\text{norm}}}$ is the text condition with an augmented prompt P_{norm}. We construct P_{norm} from P by adding an extra description "a detailed sculpture of" to better reflect the intrinsic characteristics of normal maps.

Geometric regularization. We found that relying solely on silhouette and SDS losses may lead to the generation of noisy surfaces, which is particularly evident for subjects wearing complex clothing. To address this, we leverage normal estimations as an additional constraint to regularize the surface reconstruction (see Fig. 5.5):

$$\mathcal{L}_{\text{norm}}(\hat{\mathcal{N}}_{\mathbf{k}}, \mathbf{n}) = \lambda_{\text{MSE}}^{\text{norm}} \left\| \hat{\mathcal{N}}_{\mathbf{k}} - \mathbf{n} \right\|_2^2 + \text{LPIPS}(\hat{\mathcal{N}}_{\mathbf{k}}, \mathbf{n}), \tag{5.8}$$

where $\hat{\mathcal{N}}_{\mathbf{k}}$ are the front and back normal maps *estimated* by ICON (Chapter 3). $\mathbf{n}$ are the corresponding *rendered* normal images of the 3D shape Ψ_{g}.We use a combination of LPIPS and MSE loss to enhance the similarity between $\hat{\mathcal{N}}_{\mathbf{k}}$ and $\mathbf{n}$. Furthermore, we utilize a Laplacian smoothing [8] regularizer, as $\mathcal{L}_{\text{lap}}$.

Mesh extraction. We use Marching Tetrahedra [42] to extract the mesh from the tetrahedral grid. Like ECON (Chapter 4), we register SMPL-X to this mesh, which allows us to transfer skinning weights for animation (see Fig. 5.8). In addition, we replace the hands with SMPL-X ones which effectively mitigates the artifacts introduced during reposing, which is needed in the subsequent texture generation stage.

Texture Stage: Given the triangular mesh from the geometry stage, we optimize the full texture. To recover the consistent details and color, even for self-occluded regions, we render both the input pose (M_{in}) and the A-pose (M_{A}) during optimization. The textures of M_{in} and M_{A} are modeled by Ψ_{color} in the 3D space of M_{A}. We optimize the texture from scratch with ψ_{c} randomly initialized. In Fig. 5.6, we show the effect of this multi-pose training. We utilize an occlusion-aware reconstruction loss $\mathcal{L}_{\text{recon}}$ on the input view of M_{in}, an SDS loss $\mathcal{L}_{\text{SDS}}^{\text{color}}$ with text guidance on rendered color images of both M_{in} and M_{A}, and a color consistency regularization $\mathcal{L}_{\text{CD}}$, with respective weights λ to balance the individual losses:

$$\mathcal{L}_{\text{texture}} = \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{SDS}}\mathcal{L}_{\text{SDS}}^{\text{color}} + \lambda_{\text{CD}}\mathcal{L}_{\text{CD}}. \tag{5.9}$$

Note that $\mathcal{L}_{\text{CD}}$ is only utilized after the full-body texture convergence (5000 iterations), in an additional optimization phase of 2000 iterations for enforcing color consistency.

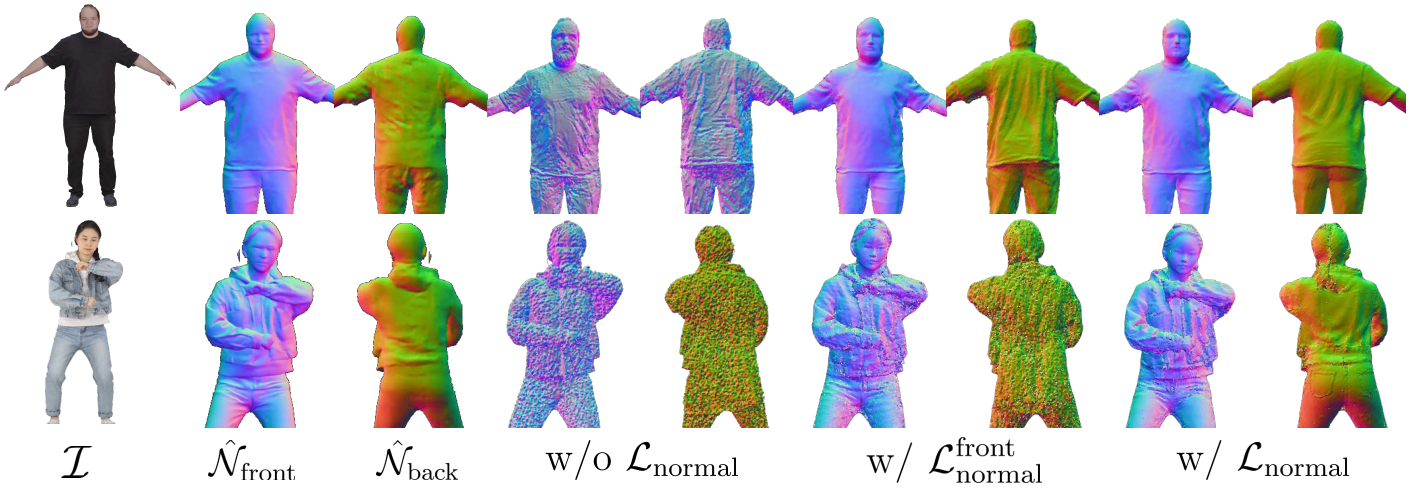


Figure 5.5: The effects of normal regularization. $\mathcal{L}_{\text{norm}}$ regularizes the surface with predicted normal images $\hat{\mathcal{N}}_{\text{front}}, \hat{\mathcal{N}}_{\text{back}}$.

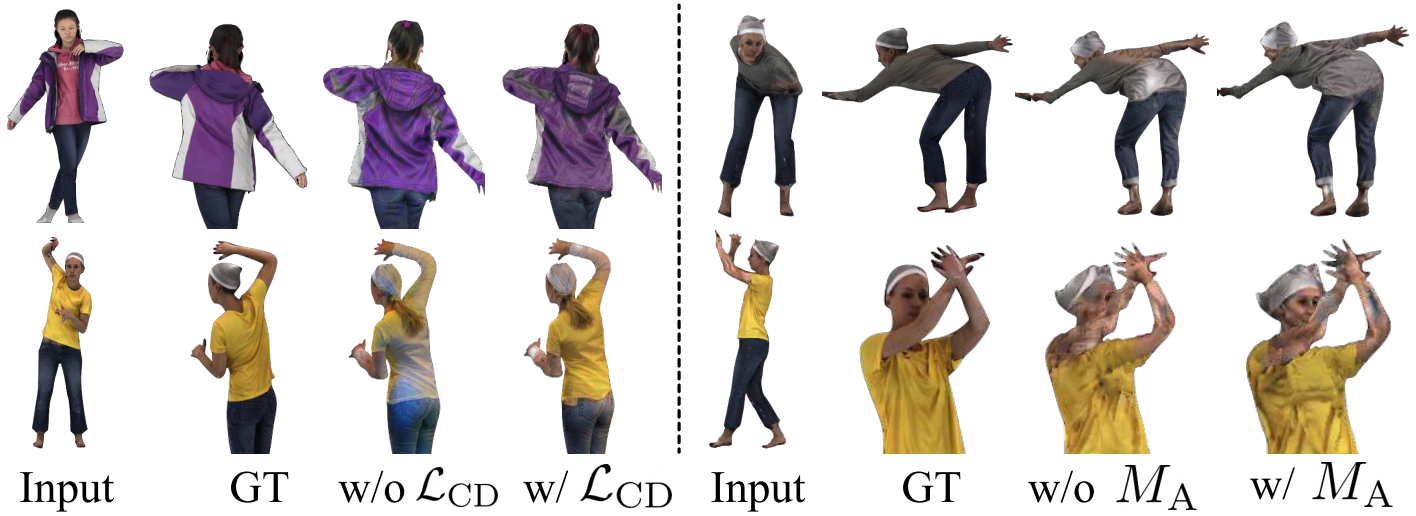


Figure 5.6: The effects of color consistency loss $\mathcal{L}_{\text{CD}}$ and multi-pose training (M_{A}) for texture optimization. $\mathcal{L}_{\text{CD}}$ corrects the over-saturated back-side color generated by SDS, while M_{A} improves the texture quality under self-occlusion or extreme poses. Note that, the "w/o M_{A}" refers to training only with input pose M_{in}.

Occlusion-aware reconstruction loss. We apply an input view reconstruction loss $\mathcal{L}_{\text{recon}}$ to minimize the difference between the input image $\mathcal{I}$ and the rendered image $\mathcal{I}'(M, \psi_{\text{c}}, \mathbf{k}_{\mathcal{I}})$. Self-occluded areas may lead to incorrect texture due to geometry misalignment, thus, an occlusion-aware mask m_{occ} is introduced:

$$\mathcal{L}_{\text{recon}} = m_{\text{occ}}(\lambda_{\text{MSE}} \left\| \mathcal{I} - \mathcal{I}'(M, \psi_{\text{c}}, \mathbf{k}_{\mathcal{I}}) \right\|_2^2 + \text{LPIPS}(\mathcal{I}, \mathcal{I}'(M, \psi_{\text{c}}, \mathbf{k}_{\mathcal{I}}))), \tag{5.10}$$

where $\mathbf{k}_{\mathcal{I}}$ denotes the input view camera, and λ_{MSE} is a weight to balance the two loss terms.

SDS loss on color images. To recover the full-body texture, including unseen regions, we update ψ_{c} via SDS loss $\mathcal{L}_{\text{SDS}}^{\text{color}}$ with text guidance. This loss is calculated based on random-view color renderings $\mathbf{x} = \mathcal{I}'(\psi_{\text{g}}, \psi_{\text{c}}, \mathbf{k})$, and DreamBooth $\mathcal{D}'$ parameterized by ϕ' and guided by text prompt P.

$$\mathcal{L}_{\text{SDS}}^{\text{color}} = \nabla_{\psi_{\text{c}}} \mathcal{L}_{\text{SDS}}^{\text{color}}(\mathbf{x}, \mathbf{c}^P) = \mathbb{E}_{\text{t},\varepsilon} \left[w_t \left(\hat{\varepsilon}_{\phi'}(\mathbf{z}_t^{\mathbf{x}}; \mathbf{c}^P, t) - \varepsilon \right) \frac{\partial \mathbf{x}}{\partial \psi_{\text{c}}} \frac{\partial \mathbf{z}^{\mathbf{x}}}{\mathbf{x}} \right], \tag{5.11}$$

where $\mathbf{k}$ is the camera pose, and $\mathbf{c}^P$ is the text embedding of P.

Chamfer-based color consistency loss. As mentioned in DreamFusion [169], the SDS loss may result in over-saturated colors, which will cause a noticeable color disparity between visible and invisible regions. To mitigate this, we incorporate a color consistency loss that measures the disparity between the color distributions of the real and rendered images using a Chamfer Distance (CD) by treating the pixels as points within the RGB color space:

$$\mathcal{L}_{\text{CD}} = \sum_{x \in \mathbf{F}_{\mathbf{x}}} \min_{y \in \mathbf{F}_{\mathcal{I}}} ||x - y||_2^2 + \sum_{y \in \mathbf{F}_{\mathcal{I}}} \min_{x \in \mathbf{F}_{\mathbf{x}}} ||x - y||_2^2, \tag{5.12}$$

where $\mathbf{F}_{\mathbf{x}}$ and $\mathbf{F}_{\mathcal{I}}$ respectively represent the foreground pixels of the novel-view albedo rendering $\mathbf{x}$, and the input view $\mathcal{I}$. The improvement using $\mathcal{L}_{\text{CD}}$ is shown in Fig. 5.6.

Camera sampling during optimization: To optimize the 3D shape and texture using multi-view renderings, cameras are all-around sampled in a way that ensures comprehensive coverage of the entire body by adjusting various parameters, see more details at Appendix C.4. To mitigate the occurrence of mirrored appearance artifacts (*i.e.*, Janus-head), we incorporate view-aware prompts ("`front/side/back/overhead view`") w.r.t. the viewing angle in the diffusion-based generation process, whose effectiveness has been demonstrated in DreamBooth [169]. To improve facial details, we

also sample cameras positioned around the face, together with the additional prompt "`face of`" (see Appendix C.4).

5.3 Experiments

We compare TeCH with state-of-the-art image-based 3D clothed human reconstruction methods, including body-agnostic methods, such as PIFu [185], PIFuHD [186] and PHORHUM [7], as well as methods that utilize a SMPL-(X) [140, 165] body prior, such as PaMIR [256], ICON (Chapter 3) and ECON (Chapter 4). For a fair comparison, all methods (*i.e.*, PIFu, PaMIR, ICON, ECON) utilize the same normal estimator from ICON. Official PIFu, PaMIR, and PHORHUM are used to evaluate the quality of texture. For ECON, we use ECON_{EX}, due to its superior performance on both "OOD poses" and "OOD outfits" cases, as reported in Chapter 4. Note that PHORHUM uses a different camera model which is incompatible with TeCH's testing data. Thus, we use PHORHUM only for qualitative comparisons. Implementation details of the network structure and optimization settings can be found in Appendix C.5.

Method	3D Metrics						2D Image Quality Metrics					
	CAPE			THuman2.0			CAPE			THuman2.0		
	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o SMPL-X body prior												
PIFu	1.9683	1.6236	0.0623	1.9305	1.8031	0.0802	27.0994	0.9362	0.0987	23.5068	0.9296	0.1083
PIFuHD	3.2018	2.9930	0.0758	2.4613	2.3605	0.0924	-	-	-	-	-	-
w/ SMPL-X body prior												
PaMIR	1.3756	1.1852	0.0526	1.2979	**1.2188**	0.0676	27.7279	0.9456	0.0904	22.5466	0.9266	0.1082
ICON	0.8689	0.8397	0.0360	**1.1382**	1.2285	0.0623	-	-	-	-	-	-
ECON	0.9186	0.9227	0.0330	1.2585	1.4184	**0.0612**	-	-	-	-	-	-
TeCH (Ours)	**0.7416**	**0.6962**	**0.0306**	1.2364	1.2715	0.0642	**28.3601**	**0.9490**	**0.0639**	**25.2107**	**0.9363**	**0.0835**

Table 5.1: Quantitative evaluation with SOTA methods. TeCH surpasses the SOTA methods w.r.t. both 3D metrics (unit of Chamfer and P2S is *cm*) and 2D metrics. This demonstrates its superior performance in accurately reconstructing clothed human geometry with intricate details, as well as producing high-quality textures with consistent appearance. The best results are marked with "**bold**", and the second-best with "underline".

5.3.1 Models and Datasets

Off-the-shelf models: TeCH relies on multiple off-the-shelf pre-trained models and does not need any additional training data. Specifically, we use the stable-diffusion-v1.5 (runwayml) as Text-to-Image (T2I) diffusion model, which is trained on LAION-5B, the VQA model BLIP [114] pre-trained on 129M images from multiple datasets [25, 108, 134, 155, 159, 188] and fine-tuned on VQA2.0 [59], SegFormer[i] [225] pretrained from [20, 37, 41, 259] and fine-tuned on ATR[127], PIXIE [47] trained on human images from multiple datasets [36, 134, 163, 222, 264], and the normal predictor of ICON (Chapter 3) trained on AGORA [164].

Datasets for evaluation. Based on the high-fidelity 3D textured scans from CAPE [144] and THuman2.0 [240], we perform quantitative evaluations.We follow ICON (Chapter 3) to analyze the robustness of reconstructions under both simple and complex poses (150 scans from CAPE). An additional 150 THuman2.0 scans are included, which comprise 100 subjects that were manually selected to represent a diverse range of clothing styles (*e.g.*, open jackets, long coats, garments with intricate patterns, *etc.*), and 50 randomly sampled subjects. The images are rendered at a resolution of 512×512. For qualitative comparison, we selected the SHHQ dataset [50] due to its wide range of textures, outfits, and poses. From this dataset, we randomly sampled 90 images with official mask annotations.

5.3.2 Quantitative Comparison

We quantitatively evaluate the quality of geometry and appearance, using the **Chamfer** (bidirectional point-to-surface) and **P2S** (1-directional point-to-surface) distance. Additionally, we report the L2 **Normal** error between normal images rendered from both meshes, to measure the consistency and fineness of local surface details, by rotating the camera by $\{0°, 90°, 180°, 270°\}$ w.r.t. to the input view. To evaluate the quality of the texture, we report 2D image quality metrics, on the multi-view colored images rendered in the same way as the normal images, including **PSNR** (Peak Signal-to-Noise Ratio), **SSIM** (Structural Similarity) and **LPIPS** (learned perceptual image path similarity).

[i] matei-dorian/segformer-b5-finetuned-human-parsing

Preference (%, ↑)	PIFu	PaMIR	PHORHUM	ICON	ECON
Geometry	88.6	87.0	81.7	97.94	90.48
Colored Rendering	95.1	93.7	93.0	-	-

Table 5.2: Perceptual study. The percentages of participants preference for TeCH compared to other baselines are reported. Most participants preferred TeCH in both geometry and colored rendering (texture).

As shown in Tab. 5.1, TeCH demonstrates superior performance across all 2D metrics and 3D metrics on CAPE. This reveals that TeCH can accurately reconstruct both geometry and texture, even for subjects with challenging poses (CAPE) or loose clothing (THuman2.0). However, on THuman2.0, it achieves comparable reconstruction accuracy to prior-based methods. This can be attributed to the fact that the hallucinated back-side may differ from the ground truth while still appearing realistic.

5.3.3 Perceptual Evaluation

We conducted a perceptual study using 90 randomly sampled in-the-wild images from the SHHQ dataset [50]. Participants were shown videos showcasing rotating 3D humans reconstructed by TeCH, as well as the baselines (PaMIR [256], PIFu [185], ICON (Chapter 3), ECON (Chapter 4) and PHORHUM [7]). They were asked to choose the more realistic and consistent result based on the input image. We gathered a total of 3,150 pairwise comparisons from 63 participants, uniformly covering 90 SHHQ subjects. The results in Tab. 5.2 show that TeCH is preferred, both, in terms of geometry and texture. As illustrated in Fig. 5.7 and Appendix C.6, unlike other methods that reconstruct overly smooth surfaces and blurry textures, TeCH shows better generalizability, featuring diverse clothing styles and gestures, even for unseen regions.

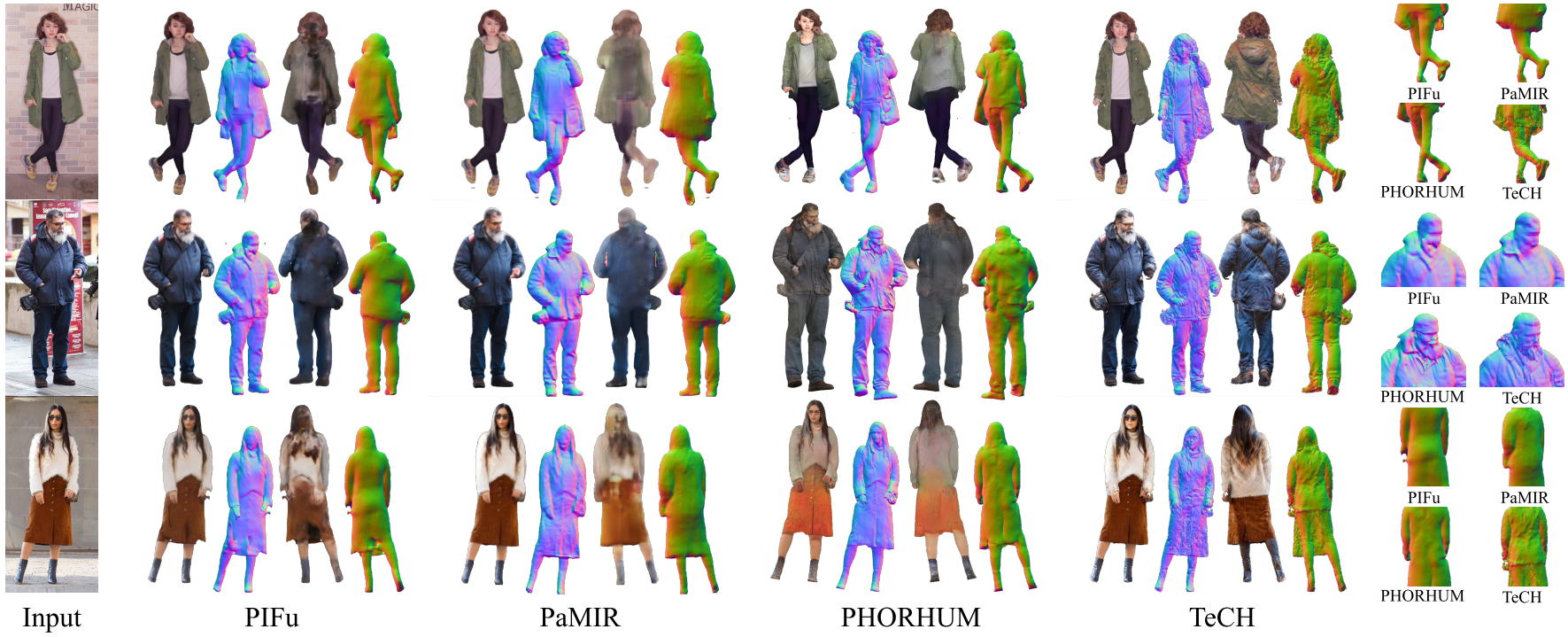


Figure 5.7: Qualitative comparison on SHHQ images. TeCH generalizes well on in-the-wild images with diverse clothing styles and textures. It successfully recovers the overall structure of the clothed body with text guidance, and generates realistic full-body texture which is consistent with the colored pattern and the material of the clothes. 🔍 **Zoom in** to see the geometric details.

	Experiment settings						3D Metrics			2D Image Quality Metrics		
	VQA	DreamBooth	$\mathcal{L}_{\text{norm}}$	$\mathcal{L}_{\text{CD}}$	M_{A}	multi-stage	Chamfer ↓	P2S ↓	Normal ↓	PSNR ↑	SSIM ↑	LPIPS ↓
TeCH	✓	✓	✓	✓	✓	✓	**0.9794**	0.9779	0.0466	26.7565	**0.9428**	**0.0741**
	✓	✗	✓	✓	✓	✓	0.9959	1.0192	**0.0454**	26.2078	0.9405	0.0813
A.	✗	✓	✓	✓	✓	✓	1.0032	1.0218	0.0470	**26.9602**	**0.9428**	0.0785
	✓	✓	✓	✓	✓	✗	0.9957	0.9963	0.0468	26.0465	0.9395	0.0775
B.	✓	✓	✗	✗	✗	✗	1.0882	**0.9203**	0.0870	-	-	-
C.	✓	✓	✓	✓	✗	✓	-	-	-	26.6500	0.9427	0.0746
	✓	✓	✓	✗	✓	✓	-	-	-	26.6506	0.9425	0.0786

Table 5.3: Ablation study. We quantitatively ablate each component. The best results are marked with "**bold**", and the second-best with "underline". All the factors are grouped w.r.t. to their influence: A. text guidance, B. geometry only, C. texture only.

5.3.4 Ablation Studies

To assess the effectiveness of key designs in TeCH, we perform ablation studies on a 10% subset of the test set, consisting of 15 subjects from THuman2.0 and 15 from CAPE. The detailed analysis of these results is as follows:

Text guidance. Figure 5.4 shows that VQA prompts help to recover the overall structure of clothing, while DreamBooth enhances the fine details of the texture pattern. Combining both text guidance sources yields the best results. This is also confirmed by the metrics in Tab. 5.3-A. A detailed analysis of individual descriptive texts (*e.g.*, garments, hairstyles, *etc.*) is shown in Fig. 5.2.

Geometric regularization. As shown in Fig. 5.5, using only $\mathcal{L}_{\text{SDS}}^{\text{norm}}$ to optimize the geometry will produce noisy artifacts, particularity noticeable in loose clothes. The significant increase in "Normal" error shown in Tab. 5.3-B echos this. This issue can be mitigated by incorporating $\mathcal{L}_{\text{norm}}$ at the beginning of the optimization.

Consistent texture recovery. The results presented in Fig. 5.6 demonstrate that $\mathcal{L}_{\text{CD}}$ notably enhances color consistency between the frontal and back sides, and "multi-pose" training (M_{A}) improves texture quality when dealing with self-occlusion scenarios. This improvement is further supported by Tab. 5.3-C, across all 2D image quality metrics.

Multi-stage optimization. As shown in Tab. 5.3-A, compared to our decoupled two-stage optimization, the joint optimization results in a performance drop across both 3D and 2D metrics. This may be attributed to the entanglement of the gradients from the geometry and texture branches during optimization. Notably, in the separate texture stage, a colored image is rendered from the extracted mesh, saving 20% of the run time compared to joint optimization, which involves rendering from the DMTet mesh.

5.4 Applications

5.4.1 Avatar animation & Editing

Following the geometry optimization phase, TeCH aligns the clothed body mesh with the SMPL-X model, enabling us to animate the reconstructed avatar with SMPL-X motions [146], as shown in Fig. 5.8.

Figure 5.8: Animating TeCH with SMPL-X motions.

The text-guided texture generation feature also allows us to edit the texture of the generated avatars. Figure 5.9 illustrates stylization results with different painting styles, like "`pop art, pixel art, van gogh`". The resulting texture features the desired styles and preserves the inherent appearance traits of the original character.

5.5 Discussion

Limitations. Despite achieving impressive results on diverse datasets, some failure cases still exist, see Fig. 5.10: **A.** TeCH occasionally fails for extremely loose clothing, this may relate to the SMPL-X-based initialization. **B.** Texture patterns can sometimes be misaligned, like the tattoos not matching the input image observations. **C.** TeCH relies on robust SMPL-X pose estimation, which is still an unsolved problem, especially for challenging poses.

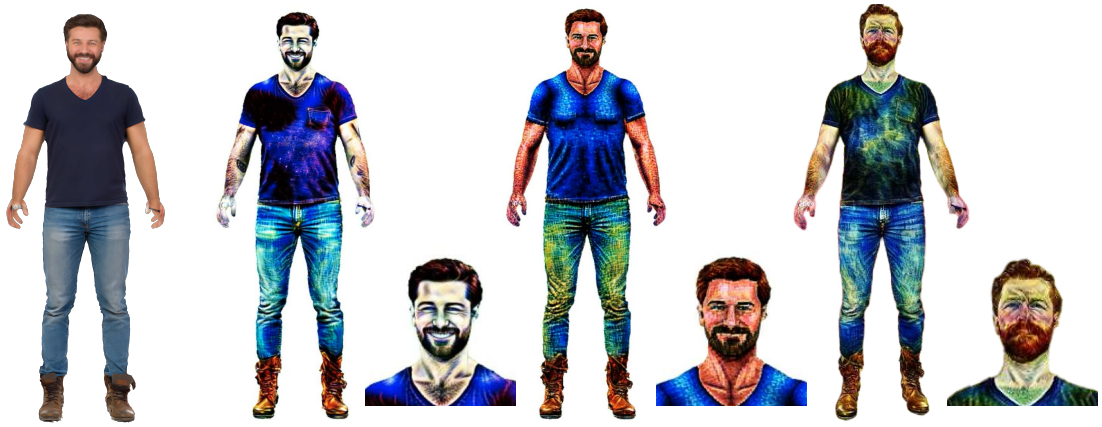


Figure 5.9: Text-guided stylization.

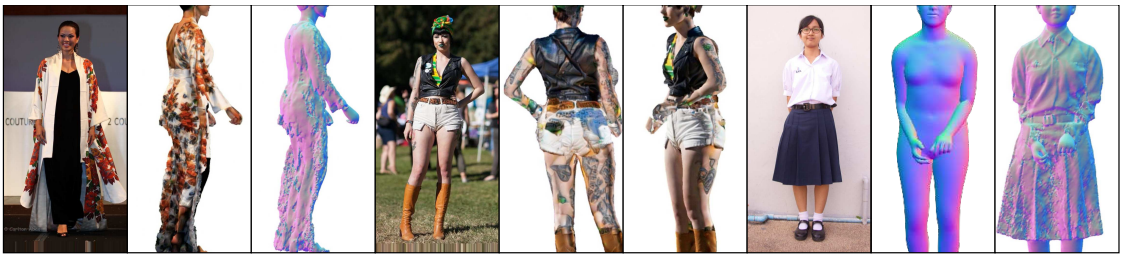


Figure 5.10: TeCH's failure case. TeCH might exhibit noise for extremely loose clothing, or hallucinated tattoos. And the pose error from HPS, like PIXIE [47], will propagate to final output.

Efficiency. For each subject, training DreamBooth takes 20 min, DMTet SMPL-X initialization takes 20 min, the geometry stage (coarse-50 min, fine-50 min), mesh post-processing takes 10 min (re-meshing, SMPL-X registration, hand replacement), the texture stage takes 140 min, resulting in 4.5 hours in total on a V100 GPU, which remains time-consuming. Improving efficiency is crucial to facilitate broader applications.

Future work. Controllable T2I models [91, 152, 254] and parameter-efficient fine-tuning techniques (PEFT) [74, 137, 172] could help to improve the controllability of the generation process and efficiency of DreamBooth training. Improved tetrahedral representation (FlexiCubes [191] and G-Shell [138]), and 3D Gaussian Splatting [98, 241], could be suitable alternatives or complements to DMTet, especially to model open-surfaces with adaptive topology. Furthermore, the generation of compositional assets, such as haircuts [193], accessories [53], and decoupled outfits [48, 49, 101, 102, 120, 263], is also valuable to explore. We leave these for future research.

6

PUZZLEAVATAR: ASSEMBLING 3D AVATARS FROM PERSONAL ALBUMS

Contents

6.1 Introduction

In all chaos there is a cosmos, in all disorder a secret order.

Carl Jung

Advances in text-guided digital human synthesis open the door to 3D avatar creation with arbitrary skin tones, clothing styles, hairstyles and accessories. While these advances have demonstrated great potential by generating iconic figures (such as Superman or Bruce Lee) and editing specific human features (such as wavy hair or full beards),

the problem of crafting one's *personalized* full-body avatar is relatively unexplored. Imagine that you are given a personal "outfit of the day" (OOTD) photo album in casual snapshots: strolling through a park, crouching to tie a shoelace, seated at a café, *etc*. These snapshots, capturing full-body actions, upper-body poses and close-up selfies with diverse backgrounds, lighting and camera settings, form a rich photo collection. Notably, this collection is relatively "unconstrained", that is, its only constraint is having a consistent identity, outfit, hairstyle and accessories, while every other factor can vary arbitrarily, see Fig. 6.1. Can we effectively construct from this album a personalized 3D avatar that vividly characterizes the user's clothes, physique, and facial details? In this work, we investigate this novel task, which we call "**Album2Human**", that transforms everyday album collections into textured 3D humans.

Compared to work that reconstructs general 3D scenes from photos with varying lighting conditions, cropping ratio, background and camera settings [147, 197], **Album2Human** is more challenging due to the additional factor of varying body articulation. On the other hand, **Album2Human** drastically differs from prior work [5, 166, 208] that creates personalized avatars from images captured in laboratory settings [32, 82, 144, 189, 227, 240, 257], in which full human bodies in limited body poses are captured using well-calibrated and synchronized cameras with controlled lighting and simple backgrounds; see Fig. 6.2.

While it is possible to create avatars from monocular (image or video) input as shown by some methods [63, 229, 235], such methods perform poorly for unusual body poses, motion blur, and occlusions, because they rely on accurate human and camera pose estimation from full-body shots. Instead, we bypass pose estimation, and follow the new paradigm of "reconstruction as conditional generation", as recently demonstrated for Text-to-Image (T2I) generation [57, 78, 220, 234, 253]. Specifically, these works cast reconstruction from partial observations as "inpainting" unobserved regions through foundational-model priors, while imposing cross-view consistency. We adapt existing T2I work [9] to learn **subject-specific** priors from a **personal OOTD** image collection, by finetuning T2I models on such images to capture identity, pieces of clothing, accessories, and hairstyle into unique and inter-exchangeable tokens, and extracting 3D geometry and texture with Score Distillation Sampling (SDS) based techniques [169]. Metaphorically, PuzzleAvatar swallows relatively "unstructured" data and digests this into a "structured library"; that is, *"seeking order in chaos, finding harmony in turmoil."*

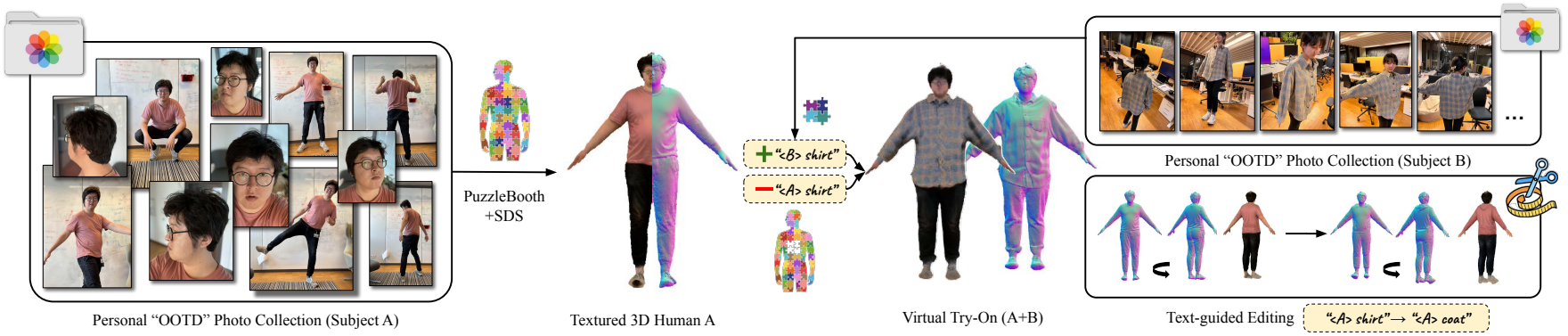


Figure 6.1: PuzzleAvatar reconstructs avatars from peronal album. It takes as input a set of "OOTD" (Outfit Of The Day) personal photos with unconstrained body poses, camera poses, framing, lighting and backgrounds, albeit with a consistent outfit and hairstyle. All these consistent factors are learned as separate unique tokens `<asset X>` in a compositional manner, like pieces of a puzzle. PuzzleAvatar allows one to easily inter-change tokens for downstream tasks, such as for customizing avatars and performing virtual try-on while preserving identity.

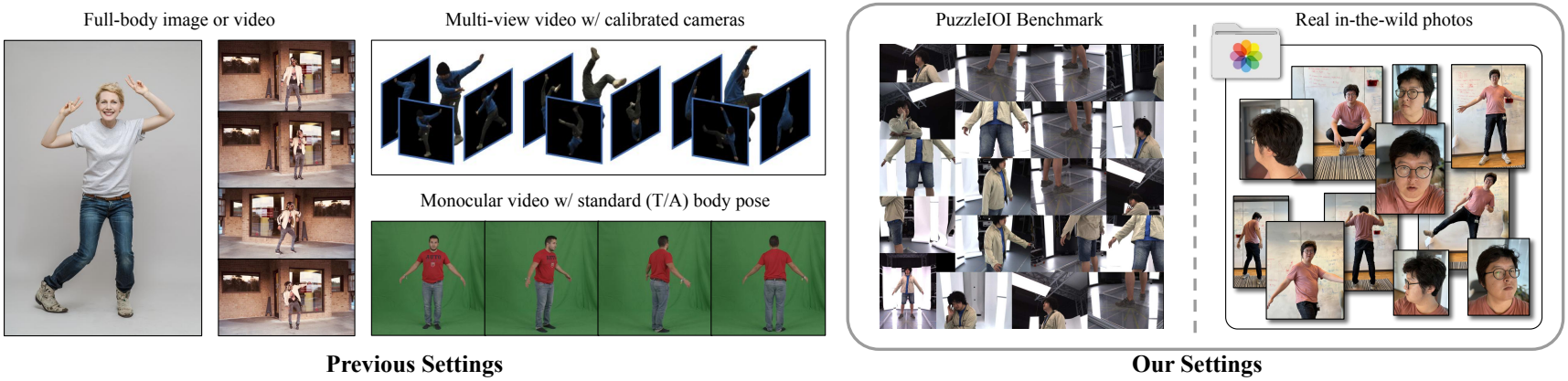


Figure 6.2: Image settings for avatar creation. Past work (left) requires images with full-body visibility, known cameras, or simple human poses. PuzzleAvatar operates on in-the-wild photos (right); it assumes a consistent outfit, hairstyle and accessories, but deals with unconstrained human poses, camera settings, lighting and background. The PuzzleIOI dataset contains multi-view images with challenging crops paired with 3D ground truth.

The insight of treating T2I models as personalized priors enables us to not only avoid *explicit* per-pixel correspondences to a canonical human space, but also to build avatars in a compositional manner. To this end, given a photo collection of a person, various assets are extracted via an open-vocabulary segmentation method [178], such as the face, garments, accessories, and hairstyles. Each of these assets is labeled by a unique token as "`<asset X>`". We exploit these token-asset pairs, to finetune a pre-trained T2I model, so that it learns to generate "personalized" assets given a respective token. Based on this personalized T2I model, we produce a 3D human avatar via Score Distillation Sampling (SDS) given a descriptive and compositional text prompt, *e.g.*, "`a DSLR photo of a man, with <asset1> face, wearing <asset0> shirt, ...`" (see Fig. 6.1). Here, each unique asset is like a puzzle piece, characterizing the identity, hairstyle and dressing style of the person. In a sense, the learned tokens are used as puzzle pieces to assemble avatars, guided by text prompts. Thus, we call it **"PuzzleAvatar"**.

Since there exists no benchmark for this new **Album2Human** task, we collect a new dataset, called PuzzleIOI, of 41 subjects in a total of roughly 1k configurations (outfits, accessories, hairstyles). The evaluation metrics include both *3D reconstruction errors* (*e.g.*, Chamfer distances, P2S distances) between reconstructed shapes and ground-truth 3D scans, as well as *2D image similarity measures* (*e.g.*, PSNR, SSIM) between rendered multi-view images of the reconstructed surface and ground-truth textured scans. PuzzleAvatar is compatible with different types of diffusion models. We ablate various diffusion models on PuzzleIOI, including single-view Stable Diffusion [181] and multi-view MVDream [192]. Moreover, we evaluate the contribution of each model component both qualitatively and quantitatively with an in-depth ablation analysis (Sec. 6.3.4).

In summary, here we make the following main contributions:

Task: We introduce a novel task, called "Album2Human", for reconstructing a 3D avatar from a personal photo album with a consistent outfit, hairstyle and accessories, but unconstrained human pose, camera settings, framing, lighting and background.

Benchmark: For evaluation of this novel task, we collect a new dataset, called PuzzleIOI, with challenging cropped images and paired 3D ground truth. This facilitates quantitatively evaluating methods on both 3D reconstruction and view-synthesis quality.

Methodology: PuzzleAvatar follows the fresh paradigm of "reconstruction as conditional generation", that is, it performs implicit human canonicalization using a personalized T2I model to bypass explicit pose estimation, or re-projection pixel losses.

Analysis: We conduct detailed evaluation and ablation studies to analyze the effectiveness and scalability of PuzzleAvatar and each of its components, shedding light on potential future directions.

Downstream applications: We show that PuzzleAvatar's highly-modular tokens and text guidance facilitates downstream tasks through two examples: character editing and virtual try-on.

PuzzleAvatar is a step towards personalizing 3D avatars. Please check out more qualitative results, demos, code, and PuzzleIOI dataset at puzzleavatar.is.tue.mpg.de

6.2 Method

Given an image collection $\{\mathcal{I}_1, \mathcal{I}_2, \ldots \mathcal{I}_N\}$ of a person, we build a 3D avatar that captures the person's shape ψ_g and appearance ψ_c. Notably, personal daily-life photos are unconstrained (see Fig. 6.2) as humans (1) appear in diverse poses and scales, (2) are often occluded or largely truncated, and (3) are captured from unknown viewpoints in diverse backgrounds. Thus, camera calibration and pose canonicalization for these photos are extremely challenging, making direct reconstruction of human avatars difficult.

PuzzleAvatar's key insight is to circumvent estimating human body poses and cameras, and, instead, to perform implicit human canonicalization via a foundation vision-language model (*e.g.*, Stable Diffusion [181]). PuzzleAvatar is summarized visually in Fig. 6.3, and has two main stages. Specifically, we first "decompose" photos into multiple assets (*e.g.*, garments, accessories, faces, hair), all of which are linked with unique learned tokens by a personalized T2I model, PuzzleBooth (Sec. 6.2.1), that is G_{puzzle} in Fig. 6.3. Then, we "compose" these multiple assets into a 3D full-body representation ψ_g, ψ_c via Score Distillation Sampling (SDS) (Sec. 6.2.2).

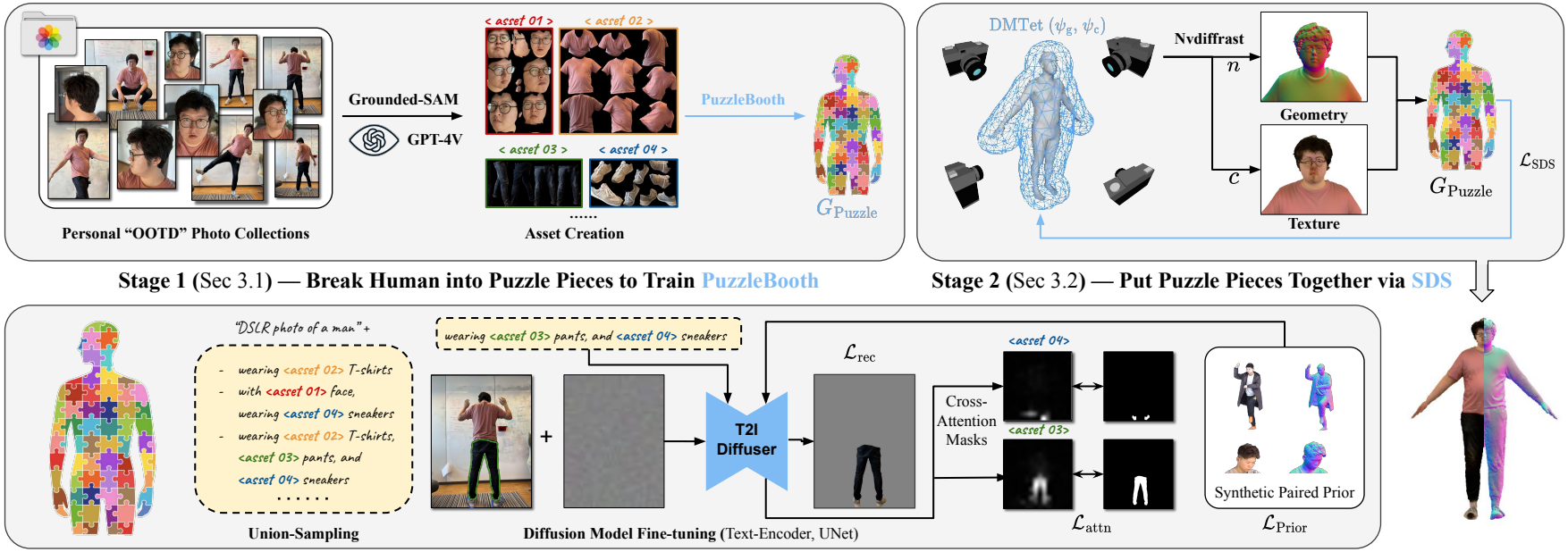


Figure 6.3: Overview of PuzzleAvatar. The upper figure shows the two main stages: (1) *PuzzleBooth* (Sec. 6.2.1), where the unconstrained photo collections are captioned and segmented to create personalized puzzle pieces, for training PuzzleBooth, G_{puzzle}, and (2) *Create-3D-Avatar* (Sec. 6.2.2), where the T-posed textured tetrahedral body mesh is optimized using a multi-view SDS loss $\mathcal{L}_{\text{SDS}}$. The bottom figure illustrates the training details of PuzzleBooth; the Text-Encoder and the UNet of T2I Diffuser (*i.e.*, Stable Diffusion) are fine-tuned using the masked diffusion loss, $\mathcal{L}_{\text{rec}}$ (Eq. (6.1)), cross-attention loss, $\mathcal{L}_{\text{attn}}$ (Eq. (6.2)), and prior preservation loss, $\mathcal{L}_{\text{prior}}$ (Eq. (6.3)). Components marked in light blue are trainable or optimizable.

6.2.1 PuzzleBooth – Personalize Puzzle Pieces

PuzzleAvatar's first step is to segment subject images into multiple assets representing different human parts such as trousers, shoes, and hairstyles. While one could build each asset individually, we adapt the "Break-A-Scene" [9] method, which shows that jointly learning multiple concepts significantly boosts performance, possibly because this facilitates global reasoning when multiple regions are simultaneously generated. Such a strategy is even more beneficial in PuzzleAvatar's setting since human-related concepts, such as face and hair, are harder to learn as their properties are strongly correlated compared to clearly distinct objects in the setting of "Break-A-Scene."

Asset Creation: All images are segmented into multiple assets V_k, each of which is associated with a segmentation mask $\mathcal{M}_k$, a dedicated learnable token $[v_k]$, and its textual name $[c_k]$, such as "pants" or "skirt." In addition, we also obtain a coarse view direction d for each image. All such information is obtained automatically by Grounded-SAM [178] and GPT-4V [158]. Specifically, we query GPT-4V with an image to directly get the property of each asset $[c_k]$ and coarse view direction d. Then, given the full list of queried asset names $\{[c_k]\}_{k=1}^{K}$, Grounded-SAM outputs segmentation masks if they are present. Please refer to Appendix D.1 for the full prompt template.

Two-Stage Personalization: We finetune the pretrained text-to-image diffusion model [181, 192] so that it adapts to the new assets. Following "Break-A-Scene" [9], we optimize the "text" part, *i.e.*, the text embedding of asset token $[v_j]$, and the "visual" part, *i.e.*, the weights of the diffusion model, in two stages: In the first stage, only text embeddings of the asset tokens $[v_k]$ are optimized with a large learning rate. In the second stage, both the "text" and "visual" part are optimized with a small learning rate. This strategy effectively prevents guidance collapse [54] between newly introduced tokens $[v_k]$ and existing asset names $[c_k]$, or, equivalently, preserves the compositionality of visual concepts.

During training, we randomly select, for every image $\mathcal{I}$, a subset of assets that appear in the image and train the model on the union set of these selected assets. This union sampling strategy, originally introduced in [9], is crucial for effective asset disentanglement. Specifically, the *mask union* is done via a pixel-wise union operation $\mathcal{M}_{\cup} = \cup_{i=1}^{j}\mathcal{M}_i$, while the *image union* applies the union mask on the

image, $\mathcal{I}_\cup = \mathcal{I} \odot \mathcal{M}_\cup$. The union text prompt $p_\cup$ is constructed by concatenating selected assets, *i.e.* "`a high-resolution DSLR colored image of a man/woman with` [v_1] [c_1], ..., [v_2] [c_2], `and wearing` [v_3] [c_3], ..., [v_j] [c_j], [d] `view`".

Losses: In both optimization stages, the model is trained to encourage concept separation while still retaining its generalization capability. To do so, the model is optimized with three loss terms: a Masked Diffusion Loss, $\mathcal{L}_{\text{rec}}$, Cross-Attention Loss, $\mathcal{L}_{\text{attn}}$, and Prior Preservation Loss, $\mathcal{L}_{\text{prior}}$. The overall training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \mathcal{L}_{\text{prior}}$ where $\lambda_{\text{attn}} = 0.01$.

The *Masked Diffusion Loss* encourages fidelity in replicating each concept via a pixel-wise reconstruction within the segmented mask:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{z,\varepsilon\sim\mathcal{N}(0,1),t}\Big[\|[\varepsilon - \varepsilon_\theta(z_t, t, p_\cup)] \odot \mathcal{M}_\cup\|_2^2\Big], \tag{6.1}$$

where $\mathcal{M}_\cup$ is the union mask, and $\varepsilon_\theta(z_t, t, p_\cup)$ is the denoised output at diffusion step t given the union prompt, $p_\cup$.

To disentangle different learned assets, we use a *Cross-Attention Loss* [9] to encourage each of the newly-added tokens to be exclusively associated with only the target asset:

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{z,j,t}\Big[\|\mathcal{CA}_\theta(v_j, z_t) - \mathcal{M}_j\|_2^2\Big], \tag{6.2}$$

where $\mathcal{CA}_\theta(v_j, z_t)$ is the cross-attention map in the diffusion U-Net between the newly-added token, [v_j], and the visual feature, z_t.

Lastly, we apply a *Prior Preservation Loss* [183] to retain the generalization capability of the vanilla T2I model — Stable Diffusion (SD-2.1). The model is trained to reconstruct images with general concepts when the special tokens are removed from prompts. General human images come from two sources: (1) Generated images, $\mathcal{I}^{\text{pr}}_{\text{gen}}$, come from SD. (2) Synthetic color-normal pairs (see Fig. 6.4), $\mathcal{I}^{\text{pr}}_{\text{syn}}$, rendered from multiple views, come from THuman2.0 [240]. The latter is to improve the geometric quality and color-normal consistency [76]. Instead of applying the prior preservation loss to individual concepts separately, we find it beneficial to compute the loss on the entire human images:

$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{z^{\text{pr}},\varepsilon\sim\mathcal{N}(0,1),t}\Big[\|[\varepsilon - \varepsilon_\theta(z^{\text{pr}}_t, t, p^*_\cup)]\|_2^2\Big] \tag{6.3}$$

where $p^*_\cup$ is the text prompt without special tokens.

*Prompt (GPT-4V): "a high-resolution DSLR **colored image** / **detailed sculpture of (the headshot of) a woman,** with oval **face**, eyes with visible epicanthic folds, and medium length, straight, dark brown **haircut**, wearing loose-fitting, teal-colored with long sleeves **shirt**, wide-legged, dark gray or black **pants** and black, ankle-high **boots**"*

Figure 6.4: Color-Normal Synthetic Prior. The corresponding descriptions are generated via GPT-4V [158], where starts with "`a high-resolution DSLR colored image`", while that of the normal image starts with "`a detailed sculpture of`" The zoomed-in head images are generated by appending "`the headshot of`".

6.2.2 PuzzleAvatar – Put Puzzle Pieces Together

With the fine-tuned diffusion model customized for all provided assets, we are able to distill a descriptive 3D avatar via SDS.

Score Distillation Sampling (SDS): A pretrained diffusion model over images $D(\mathrm{z})$ captures the data distribution $\log p(\mathrm{z}_\psi)$. SDS [169] is a technique that guides some parameterization of images $\mathrm{z}(\psi)$ (raw pixels, neural networks, etc.) to align with conditions (*e.g.*, text, landmark, *etc.*). The core idea is to approximate the parameter gradient $\nabla_\psi \mathcal{L}$ as a weighted reconstruction residual. As the vanilla method suffers from color oversaturation, we use an improved SDS – Noise-Free Distillation Sampling (NFDS) [94]. This modifies the guidance from a single reconstruction residual into two composed residual terms δ_C and δ_D. Specifically, by denoting the derived gradient of a network ψ from NFSD as $\nabla \mathcal{L}_{\text{NFDS}}[\mathrm{x}, \psi]$:

$$\nabla_\psi \mathcal{L}_{\text{NFDS}}[\mathrm{z}, \psi] = w(t)(\delta_D + s\delta_C)\frac{\partial \mathrm{z}}{\partial \psi}, \quad \text{where} \tag{6.4}$$

$$\begin{aligned} \delta_{\mathrm{C}}(z_t, p, t) &= \varepsilon_\theta(z_t; p, t) - \varepsilon_\theta(z_t; \varnothing, t), \\ \delta_{\mathrm{D}}(z_t, t) &= \begin{cases} \varepsilon_\theta(z_t; \varnothing, t), & \text{if } t \leq 200 \\ \varepsilon_\theta(z_t; \varnothing, t) - \varepsilon_\theta(z_t; p^{\text{neg}}, t), & \text{otherwise,} \end{cases} \end{aligned} \tag{6.5}$$

In our case, z is the (latent of) diffusion output (human images or normals) and ψ denotes the 3D avatar representation (both ψ_g, ψ_c), s is the guidance scale, and we follow NFDS and set $s = 7.5$.

Representation and Initialziation: The 3D human avatar is parameterized with DMTet [55, 190], a flexible tetrahedron-based 3D neural representation. The geometry, ψ_g, and appearance, ψ_c, are optimizable, and can be differentially rendered into normal, n, and colored images, c. The geometry ψ_g is first initialized to an A-posed SMPL-X body [165].

Optimization: We use the full-text description of the human p^{all} as a guiding prompt. It is a concatenation of text prompts from all assets *i.e.*, $(v_i, c_i), \ldots, (v_K, c_K)$. We optimize geometry and color separately in two optimization stages, both using Noise-Free-Score Distillation (NFSD). In the first stage, the avatar's geometry is guided in the surface normal space, $\nabla\mathcal{L}^{\text{norm}} \equiv \nabla\mathcal{L}_{\text{NFDS}}[\mathrm{n}, \psi_g]$. We additionally prepend "`a detailed sculpture of`" to the full-text to indicate the guidance space. In the second stage, its appearance is guided by $\nabla\mathcal{L}^{\text{color}} \equiv \nabla\mathcal{L}_{\text{NFDS}}[\mathrm{c}, \psi_c]$. The camera settings for multi-view SDS are in Appendix D.2.

6.3 Experiments

It has been a long-standing challenge in the field of "Text-to-3D" (including "Text-to-Avatar") to *quantitatively* benchmark new algorithms. Existing benchmarks are typically neither reliable nor objective because they sample 3D avatars from a relatively small collection of prompts and evaluate the quality of these avatars through perceptual studies with a limited number of participants.

While PuzzleAvatar adopts "Text-to-3D" techniques, its goal is to reconstruct avatars from photos of a specific person in a specific outfit, rather than to randomly generate avatars. As a result, a natural and reliable way to benchmark PuzzleAvatar is to exploit a 4D scanner (synced with IOI color cameras[i]) to capture ground-truth 3D shape and appearance, and to measure the reconstruction error between the reconstructed and ground-truth shape and appearance. We thus build a dataset, called PuzzleIOI (Sec. 6.3.1), on which we evaluate PuzzleAvatar, and the ablation of its components. Some examples of PuzzleIOI are shown in Fig. 6.2 (Our Settings).

[i] https://www.ioindustries.com/cameras

Dataset	#Views	#ID	#Outfits	#Actions	SMPL-X	Scan	Text	Texture
ActorsHQ [82]	160	8	8	52	✓	✓	✗	✓
MVHumanNet [227]	48	4500	9000	500	✓	✗	✓	✓
HuMMan [21]	10	1000	1000	500	✗	✗	✗	✓
DNA-Rendering [32]	60	500	1500	1187	✗	✗	✓	✗
THuman2.0 [240]	–	200	500	–	✗	✗	✗	✓
CAPE [144]	–	15	8	600	✓	✓	✗	✗
BUFF [245]	–	5	2	3	✓	✓	✗	✓
PuzzleIOI	22	41	933	40	✓	✓	✓	✓

Table 6.1: Datasets related to PuzzleIOI. "–" means image captures are unavailable. "Scan" is A-posed, and "SMPL-X" is its respective SMPL-X fits.

6.3.1 PuzzleIOI Dataset

We create PuzzleIOI (see statistics in Tab. 6.1) to simulate real-world album photos of humans, which: (1) cover a wide range of human identities (**#ID** column in Tab. 6.1) and daily outfits (**#Outfits**), (2) span numerous views (**#Views**) to mimic real-world captures (*e.g.*, occlusion, out-of-frame cropping), and (3) include text descriptions (**Text**), and ground-truth textured A-posed scans (**Scan, Texture**) and their SMPL-X fits (**SMPL-X**) for shape initialization purposes.

A-Pose SMPL-X & Scan: Almost all "Text-to-Avatar" methods [23, 76, 106, 130, 241] use an A-pose body for shape initialization due to its minimal self-occlusions. Thus, we adhere to this empirical setting in PuzzleIOI. For each subject **(ID+Outfit)**, we capture a ground-truth A-posed 3D scan and fit a SMPL-X model to it, as in AGORA [164].

Multiple Views: To simulate the diversity and imperfections of real-world photos, for each subject **(ID+outfit)** we randomly sample 120 photos from the multi-view human action sequence (approx. 760 frames / subject) captured by 22 cameras; see Fig. 6.2. The captured images are segmented and shuffled to build the training dataset for PuzzleBooth (Sec. 6.2.1).

Text Description: Similar to how image captioning is done in Sec. 6.2.1, here we randomly select two frontal full-body images and use GPT-4V to query the asset names and corresponding descriptions of visible assets. We use the position of the ground truth camera to categorize the photos into 4 view groups `{front, back, side, overhead}` in PuzzleIOI, while we use GPT-4V to automatically label viewpoints from in-the-wild images.

6.3.2 2D and 3D Metrics

We conduct quantitative evaluation on the PuzzleIOI dataset (Sec. 6.3.1). To evaluate the quality of **shape reconstruction** we report three metrics: (1) **Chamfer distance** (bidirectional point-to-surface, *cm* as unit), (2) **P2S distance** (1-directional point-to-surface, *cm* as unit) distance, and (3) **L2 error for Normal maps** rendered for four views ($\{0°, 90°, 180°, 270°\}$) to capture local surface details.

To evaluate the quality of **appearance reconstruction**, we render multi-view color images as above, and report three image-quality metrics: **PSNR** (Peak Signal-to-Noise Ratio), **SSIM** (Structural Similarity) and **LPIPS** (Learned Perceptual Image Path Similarity).

6.3.3 Benchmark

PuzzleAvatar is a general framework, compatible with different diffusion models. In Tab. 6.2 we benchmark variants of PuzzleAvatar with two different backbones: (1) vanilla Stable Diffusion [181], *i.e.*, SD-2.1 [ii], and (2) MVDream [192] [iii] fine-tuned from vanilla SD using multi-view images rendered from Objaverse [40]. The shared basic pipeline for PuzzleAvatar, the SOTA image-to-3D methods TeCH [78] and MVDreamBooth [192] is: 1) first to finetune these backbones with subject images and 2) later to extract avatars with text-guided SDS optimization.

Quantitative Evaluation. Tab. 6.2 shows that PuzzleAvatar is on par with TeCH on 3D metrics, while outperforming it on all 2D metrics. Note that, to enhance shape quality, TeCH employs multiple supervision signals and regluarization terms, including normal maps predicted from the input image via ECON (Chapter 4), silhouette masks produced by SegFormer [225] and a Laplacian regularizer. In terms of texture quality, TeCH uses an RGB-based chamfer loss to minimize color shift between the input image and the backside texture, while its front-side texture is computed by back-projecting the input image. In contrast, PuzzleAvatar achieves on-par 3D accuracy and better texture quality ***without*** any of these auxiliary losses, regularizers, or pixel back-projection.

[ii] huggingface.co/stabilityai/stable-diffusion-2-1-base
[iii] huggingface.co/ashawkey/mvdream-sd2.1-diffusers

PuzzleAvatar outperforms MVDreamBooth on texture quality by a large margin (PSNR +10.09%, LPIPS -8.79%), and on geometry quality (measured by Chamfer and P2S), while showing comparative performance with the baselines on normal consistency. The key difference of PuzzleAvatar, compared to MVDreamBooth and TeCH, is its puzzle-wise training strategy. Without this, 2D diffusion models fine-tuned on human photos with complex poses and cropping might produce completely flawed 3D humans, with low-quality (even full black) textures or overly smooth shapes; see Fig. 6.8.

Qualitative Evaluation. As depicted in Figs. 6.7 and 6.8, PuzzleAvatar has various advantages over TeCH: (1) *Enhanced front-back consistency*, because PuzzleAvatar treats all views with ID-consistent generation, while TeCH introduces inconsistency between the front view created by reconstruction and the back view created by imagination. (2) *Reduced non-human artifacts*, PuzzleAvatar bypasses the dependence on numerous off-the-shelf estimators used in TeCH, for which non-human artifacts arise when segmentation or normal map estimation fails. (3) *Improved geometry-texture disentanglement*, where PuzzleAvatar excels in separating shirt stripes compared to TeCH. This is mainly due to the failed normal map estimated from the input image (see Fig. 6.8, rightmost normal estimate), which relies on often incorrectly estimated normal maps from the input image. Notably, the comparison with MVDreamBooth highlights PuzzleAvatar's proficiency in producing intricate geometric details and textures. We also compare with AvatarBooth [244], which addresses the similar problem. Since its code and trained models have not been released yet, we test PuzzleAvatar on the same photo collections used by AvatarBooth, and show the results in Fig. 6.11.

Method	Backbone	3D Metrics (Shape)			2D Metrics (Color)		
		Chamfer ↓	P2S ↓	Normal ↓	PSNR↑	SSIM↑	LPIPS↓
TeCH†	SD-2.1-base	1.646	**1.590**	**0.076**	23.635	0.919	0.065
PuzzleAvatar	SD-2.1-base	**1.617** -1.76%	1.613 +1.45%	0.077 +1.32%	**24.687** +4.45%	**0.930** +1.20%	**0.062** -4.62%
MVDreamBooth†	MVDream	1.705	1.835	0.100	19.401	0.909	0.091
PuzzleAvatar	MVDream	1.697 -0.47%	1.811 -1.31%	0.101 +1.00%	21.361 +10.09%	0.906 -0.33%	0.083 -8.79%

Table 6.2: Evaluation on full PuzzleIOI (933 OOTD). † means using SMPL-X fits of ground-truth scans to initialize DMTet and factor out pose error (unlike the vanilla TeCH [78], which estimates SMPL-X using PIXIE [47]). The best results are marked with "**bold**". "Ratio%" is the relative performance drop, while "ratio%" is the relative performance gain, w.r.t. the competitors, *i.e.* TeCH and MVDreamBooth [192].

Group	Method	3D Metrics (Shape)			2D Metrics (Color)		
		Chamfer ↓	P2S ↓	Normal ↓	PSNR↑	SSIM↑	LPIPS↓
	TeCH†	1.600	1.541	0.073	23.665	0.919	0.065
	PuzzleAvatar	1.589	1.570	0.075	24.718	**0.931**	**0.061**
A.	w/ detailed GPT-4V description	1.604 +0.9%	1.607 +2.4%	0.079 +5.3%	24.208 -2.1%	0.929 -0.2%	0.062 +1.6%
	w/o view prompt	1.641 +3.3%	1.653 +5.3%	0.082 +9.3%	23.929 -3.2%	0.928 -0.3%	0.064 +4.9%
	w/o NFSD (vanilla SDS)	1.624 +2.2%	1.604 +2.2%	0.072 -4.0%	20.441 -17.3%	0.923 -0.9%	0.071 +16.4%
B.	w/o synthetic normal+color	2.194 **+38.1%**	2.493 **+58.8%**	0.130 **+73.3%**	21.940 -11.2%	0.912 -2.0%	0.078 **+27.9%**
	w/o synthetic normal	2.089 **+31.5%**	2.335 **+48.7%**	0.123 **+64.0%**	23.684 -4.2%	0.919 -1.3%	0.074 **+21.3%**
	w/o synthetic color	1.680 +5.7%	1.687 +7.5%	0.084 +12.0%	23.813 -3.7%	0.927 -0.4%	0.063 +3.3%
C.	multi-subject training (5 subjects / model)	1.809 +13.8%	1.560 -0.6%	0.080 +6.7%	**24.990** +1.1%	0.929 -0.2%	0.062 +1.6%
	w/o full-body images	1.603 +0.9%	1.580 +0.6%	0.073 -2.7%	23.703 -4.1%	**0.931** 0.0%	0.062 +1.6%
	50% training data	1.590 +0.1%	1.569 -0.1%	0.074 -1.3%	24.095 -2.5%	0.930 -0.1%	**0.061** 0.0%
	10% training data	**1.583** -0.4%	**1.531** -2.5%	**0.069** -8.0%	23.477 -5.0%	0.928 -0.3%	0.062 +1.6%

Table 6.3: Ablation study on subset of PuzzleIOI (120 OOTD). The best results are marked with "**bold**", the second best results are marked with and underline. The "ratio%" is the relative performance drop, and "ratio%" is the relative performance gain, w.r.t. PuzzleAvatar, where the drop larger than 20% are marked with "**bold**". Group-A summarizes the *failed attempts*, Group-B justifies the *key components*, and Grounp-C analyses the *scalability* of PuzzleAvatar.

6.3.4 Ablations

Common Practices: In Tab. 6.3-B, we analyze the effect of common practices that have been shown to be beneficial for general scenes, including view-specific prompt [183], NFSD over vanilla SDS [94], and prior preservation loss [76, 183]. The performance gain confirms that the problems of PuzzleAvatar could also be mitigated with these practices. Some qualitative comparisons are shown in Figs. 6.9 and 6.10. PuzzleAvatar's ablation results show the effectiveness of PuzzleIOI metrics in measuring the performance of different methods in our setting, and also help us answer the following questions.

Does the view prompt [d] helps the reconstruction? Yes. This is a common practice of numerous existing works [29, 78, 130, 169], and has not yet been quantitatively justified. As detailed in Tab. 6.3 (B. w/o view prompt), the normal error increases by +9.3%. Apart from view prompts captioned by LLM, there is still room to grow with improved representations for cameras, such as the camera pose embedding used in LGM [203] and Cameras-as-Rays [250].

Does NSFD outperforms vanilla SDS? Yes. For fair comparison, we set the guidance scale $s = 7.5$ for both NSFD and vanilla SDS. As detailed in Tab. 6.3 (B. w/o NFSD), compared with NFSD (Noise-Free Score Distillation [94]), vanilla SDS degrades the geometry quality a bit by +2.2%, while considerably degrading the texture quality (PSNR +17.3%, LPIPS +16.4%), as the SDS often crashes, leading to full-gray/yellow textures.

Does the synthetic human prior helps? Yes, and it significantly improves the reconstruction quality, in both the geometry (chamfer error -38.1%, P2S error -58.8%, Normal error -73.3%), and texture (PSNR +11.2%, LPIPS -27.9%). And synthetic normals appear to contribute more than synthetic RGB (chamfer error -31.5% vs. -5.7%, LPIPS -21.3% vs. -3.3%). Introducing photorealistic synthetic data during fine-tuning proves beneficial, and the performance boost from color-normal pairs surpasses that from only using single mode (color/normal) of data, such as chamfer (+38.1% > +31.5% + +5.7%) and LPIPS (+27.9% > +21.3% + +3.3%), see Fig. 6.5. We attribute such **"1+1>2 effect"** to the enhanced geometry-texture alignment, which benefits from such pairwise training. Please see Fig. 6.9 for more qualitative ablation results.

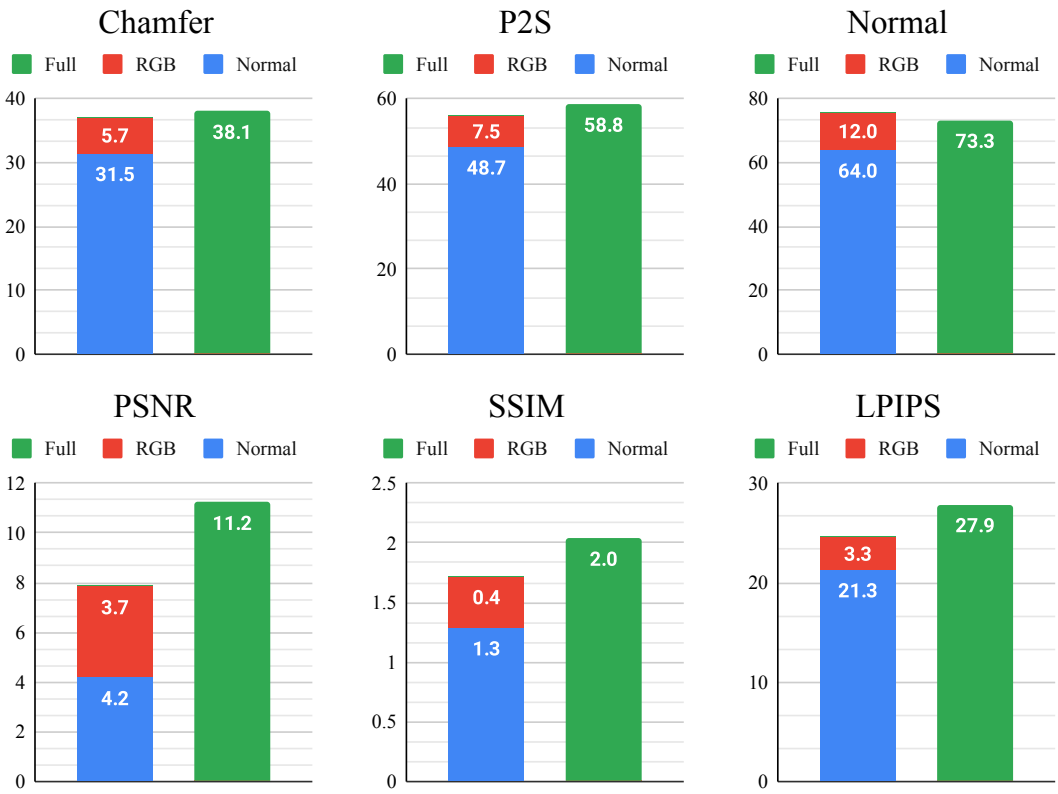
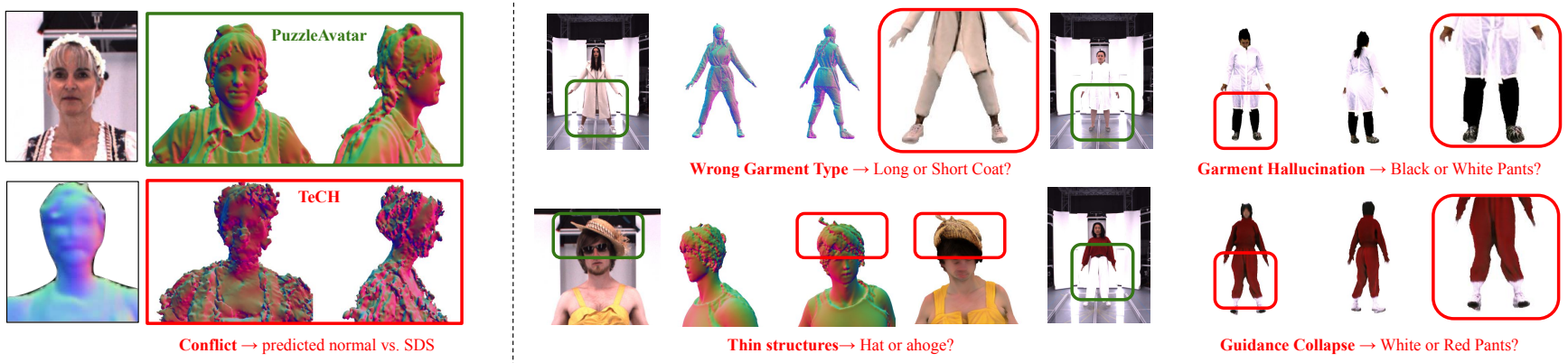


Figure 6.5: "1+1>2 Effect" of Synthetic Priors. All the numbers refer to the performance gain (%), where **Full** means training with color-normal pairs, and **RGB** and **Normal** means training with a single modality.



(a) **TeCH** may create non-human noisy artifacts

(b) **PuzzleAvatar** creates realistic 3D humans, **but struggles from** …

Figure 6.6: Failure Cases. Non-human artifacts mainly cause errors in TeCH (see a), whereas errors in PuzzleAvatar stem from hallucination and flawed DMTet modeling of thin structures. For the right-top case, the black pants showing through the white coat, while realistic, deviates from the original input. As a result of this hallucination, the failures of PuzzleAvatar are distinct from ground-truth, but not completely catastrophic (see b).

Can token [v_i] encode the identity and features of assets? Yes. As shown in Tab. 6.3 (A. w/ detailed GPT-4V description), both shape and color quality slightly decrease when too-detailed descriptions are used in the prompt, such as "`wearing sleeveless <asset1> t-shirts, and fitted <asset2> jeans`", instead of "`wearing <asset1> t-shirts, and <asset2> jeans`". Surprisingly, more detailed prompts can introduce bias, conflicting with the original identity and harming performance; see Fig. 6.10.

Does PuzzleAvatar work without using any full-body shots? Yes, but with some performance drop. Excluding the full-body shots (*i.e.*, complete images), slightly decreases the quality of both geometry and texture (Chamfer +0.9% and PSNR -4.1%; Tab. 6.3, C. w/o full-body images). Nevertheless, it is unsurprising to find that PuzzleAvatar without training on full-body images still outperforms the best TeCH setting (better texture plus on-par geometry quality).

How much data does PuzzleAvatar need? With just a fraction of the training data (10%), PuzzleAvatar can already achieve satisfactory reconstruction performance. As the number of training images increases, the view synthesis performance initially keeps improving in both texture and geometry quality (shown in Tab. 6.3, C. 50% / 10% training data) but interestingly starts to deteriorate in geometry quality. We hypothesize that training PuzzleBooth using more RGB images could impair the qualiy of SDS gradients in the space of normal maps, thus degrading the geometry optimized via SDS. We find some empirical evidence supporting this hypothesis in Tab. 6.3 (B. without synthetic normal), where the absence of normal priors leads to a notable decline in geometry quality compared to texture (P2S +48.7% vs. SSIM -1.3%).

Does PuzzleAvatar support multi-subject training? Yes. In fact, and perhaps surprisingly, multi-subject training even slightly improves reconstruction quality (Tab. 6.3-C). This demonstrates the power of Stable Diffusion to process and integrate numerous human identities simultaneously, and the robustness of this puzzle-based training strategy in learning disentangled human identities.

6.4 Applications

The compositionality of PuzzleAvatar through its tokens and text prompts supports diverse applications like Virtual Try-On and text-guided avatar editing, as shown in Fig. 6.1. Moreover, the A-Posed output can simplify the rigging and skinning process. With the underlying SMPL-X parametric body, the 3D output could be easily animated with SMPL-X motion data, like AMASS [146] and AIST++ [123], as the common practice in [80, 229, 256].

6.5 Discussion

Limitations & Future Work. Since PuzzleAvatar builds on PuzzleBooth and Score Distillation Sampling (SDS), while using no re-projection terms, some hallucination is inevitable. As Fig. 6.6 shows, PuzzleAvatar may incorrectly hallucinate garment texture or types, and suffer from description contamination, a common issue in T2I models. Despite being trained with synthetic paired data, PuzzleAvatar sometimes struggles to perfectly disentangle shape and color, leading to baked-in texture. Additionally, preserving facial identity is challenging without high-resolution headshot selfies in the training data. Potential solutions for better identity preservation may include enhancing segmented faces with super-resolution techniques [215], conducting personalized restoration [26], or incorporating face ID embeddings [212]. PuzzleAvatar's main bottleneck currently is its computational complexity, as it takes roughly 4 hours to train PuzzleBooth and perform SDS-based optimization. This is impractical for certain applications. In the future, we will explore better training-free strategies [125, 205] and sampling methods for diffusion models [143, 196]. And the compositional 3D could be achieved through non-watertight or multi-layer representations [49, 101, 138, 195].

Multi-subject training with PuzzleAvatar seems promising. Thus, it might be feasible to extend PuzzleAvatar to decentralized training settings. By fine-tuning a shared T2I model through federated learning [129], users across the globe could upload their personal albums to build a global "style set" of really diverse clothing, accessories, and hairstyles, for customizing avatars.

Indirect vs. Direct Reconstruction. We acknowledge that our "SDS-based person-specific generation loss" (*indirect reconstruction*) are less sensitive to fine-grained geometric misalignment (*e.g.*, specific wrinkles in clothing, state of hair, or facial expression) than traditional "re-projection loss" (*direct reconstruction*). PuzzleAvatar leans more towards semantic-aligned rather than pixel-aligned reconstruction. This explains why the front-side rendering of TeCH always looks more pixel-aligned with the original input than PuzzleAvatar, as shown in Figs. 6.7 and 6.8. However, "re-projection loss" necessitate precise estimating camera, body pose or geometric maps (*i.e.*, depth, normal), which is challenging in our unconstrained setting where both the human and camera move freely against random backgrounds. Thus, the pixel-aligned scheme is not scalable enough in the case of diverse unstructured inputs. Finally, incorrect estimates of *direct reconstruction* cause non-human artifacts in TeCH (Fig. 6.6-a), whereas errors of PuzzleAvatar (*indirect reconstruction*) mainly stem from model hallucination, as shown in Fig. 6.6-b, where the reconstructed shapes look realistic but vary slightly in identity.

Potential Negative Effect. As discussed in Sec. 6.3.4, the performance of PuzzleAvatar relies heavily on existing public/commercial synthetic datasets and therefore may inherit their gender, racial and age biases. One may address such an issue by curating balanced datasets from real-world images (with off-the-shelf methods to estimate normals [11, 186, 228, 229]) or by "simply" building better synthetic datasets.

Contributions to the Community. PuzzleAvatar paves the way for reconstructing detailed articulated humans from personal, natural photo collections – introducing the new "Album2Human" task. Meanwhile, PuzzleIOI offers a new benchmark that facilitates *objective* evaluation for various tasks, including but not limited to model customization, model personalization and distillation sampling. We believe that our new task, Album2Human, together with our new benchmark, PuzzleIOI, could push the boundary of the field of AI-Generated Content (AIGC). Furthermore, PuzzleAvatar offers a simple yet scalable reconstruction system, with which users may ignore the technical details of reconstruction parameters. More importantly, we believe that PuzzleAvatar demonstrates a new and practical paradigm for *"puzzle-assembled clothed human reconstruction"* that produces a 3D avatar from everyday photos in a more scalable and constraint-free manner than *"pixel-aligned clothed human reconstruction"* [185].

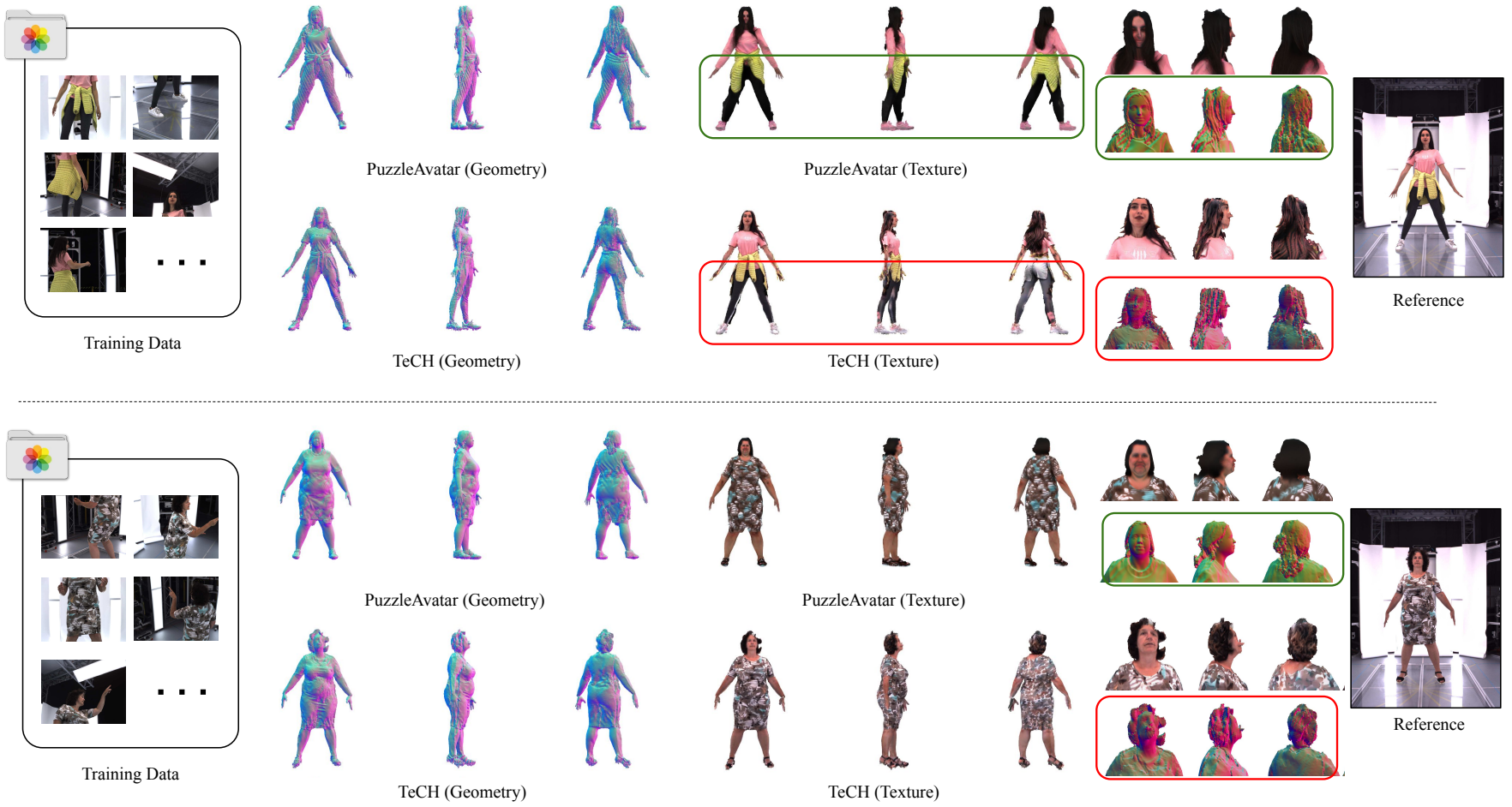


Figure 6.7: Qualitative Results. We compare PuzzleAvatar, TeCH on randomly sampled subjects. PuzzleAvatar offers various advantages over TeCH: (1) Enhanced front-back consistency. (2) Reduced non-human artifacts. (3) Improved geometry-texture disentanglement. **Zoom in** to see more 3D and color details.

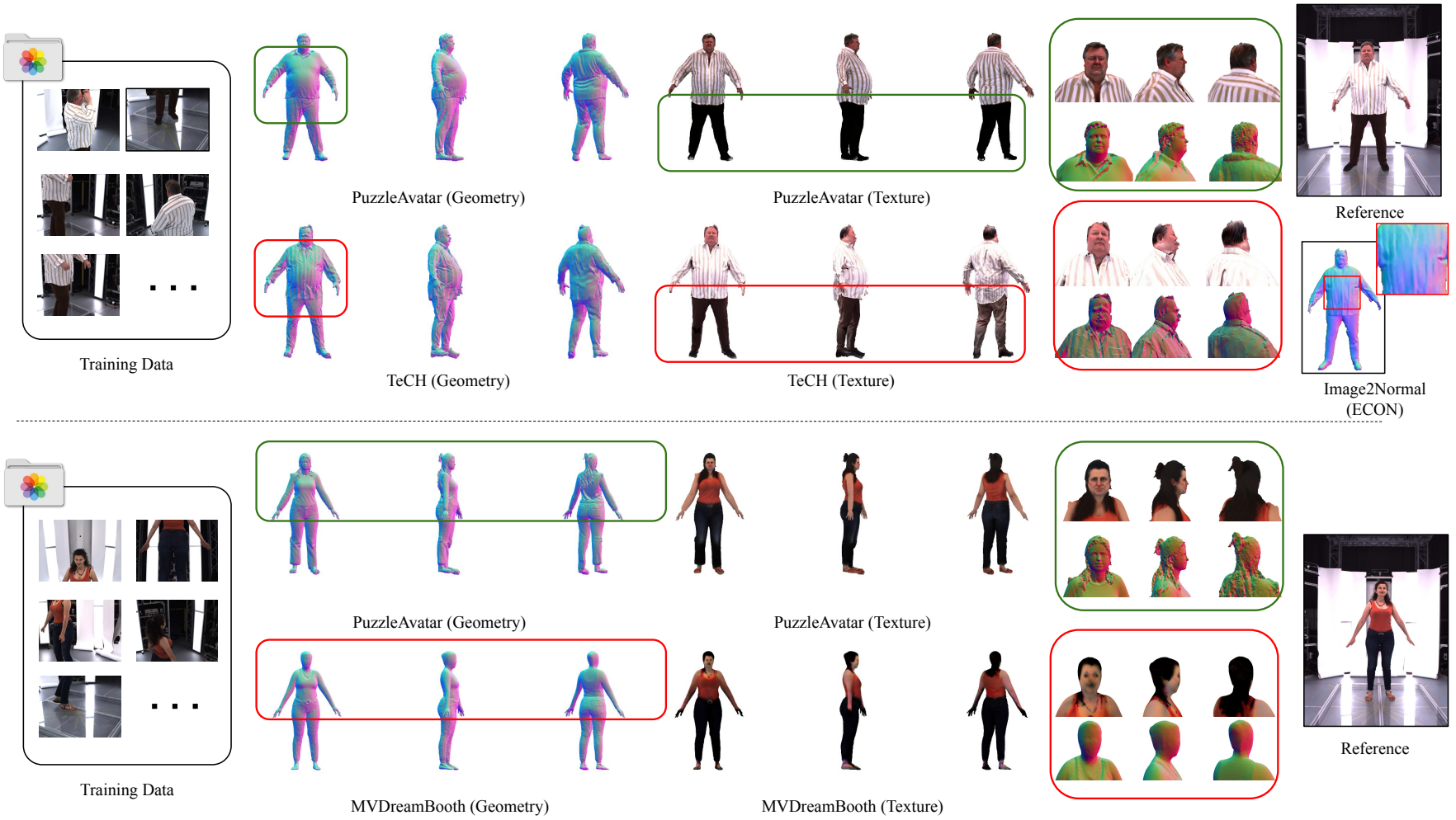


Figure 6.8: Qualitative Results. We compare PuzzleAvatar, TeCH and MVDreamBooth on randomly sampled subjects. PuzzleAvatar offers various advantages over TeCH: (1) Enhanced front-back consistency. (2) Reduced non-human artifacts. (3) Improved geometry-texture disentanglement. At the bottom, MVDreamBooth highlights PuzzleAvatar's proficiency in producing intricate geometric details and textures. **Zoom in** to see more 3D and color details.

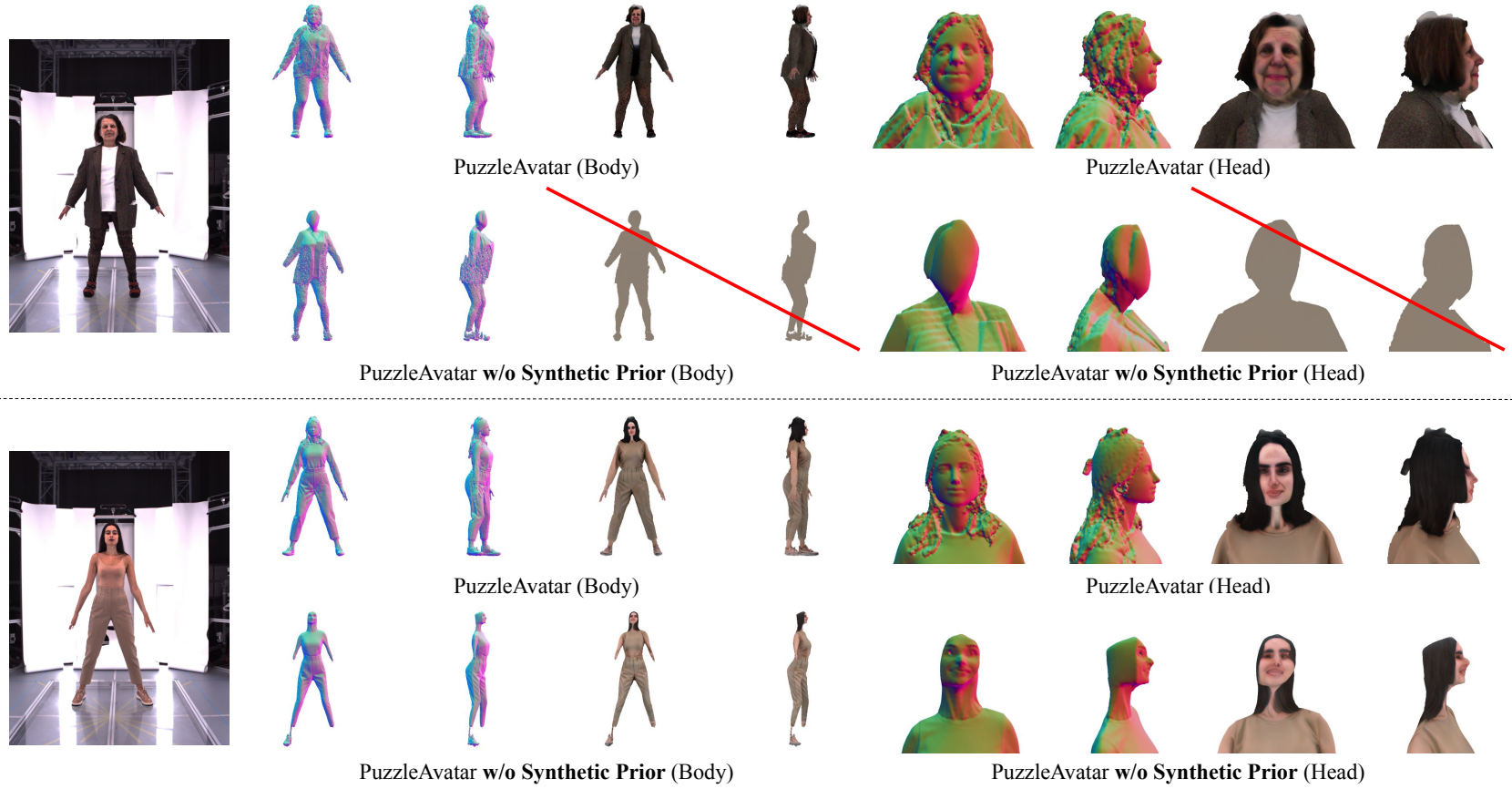


Figure 6.9: How Synthetic Prior Helps? See Fig. 6.5 for more in-depth analysis.

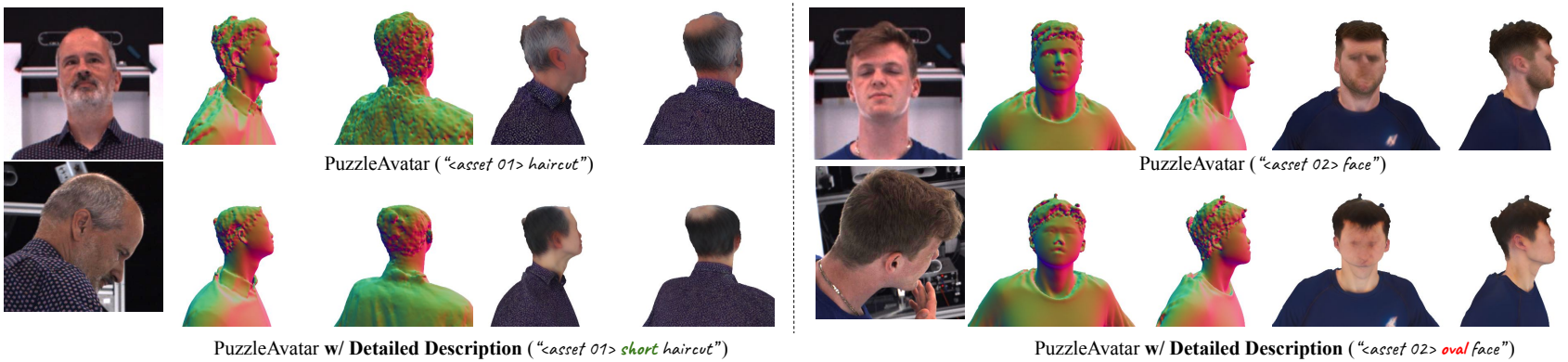


Figure 6.10: Detailed vs. Plain Prompt. Token <asset X> suffices to maintain the appearance of assets. Elaborate prompts introduce bias.

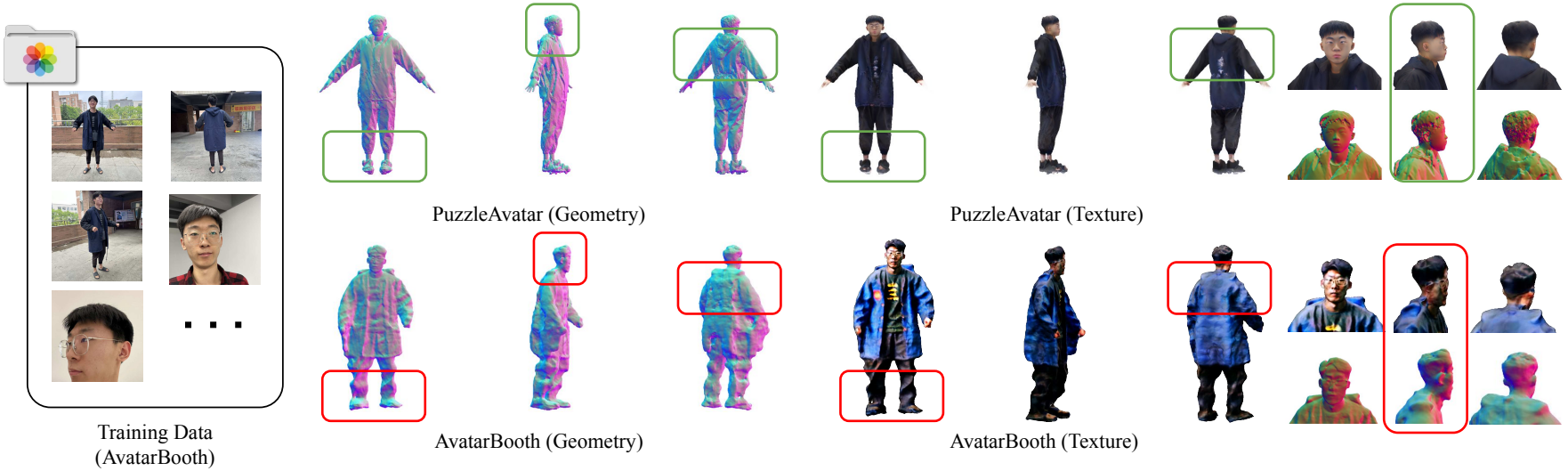


Figure 6.11: AvatarBooth [244] vs. PuzzleAvatar AvatarBooth introduces a similar task, but overlooks the compositionality of garments and utilizes two separate DreamBooths (Head, Body) along with ControlNet, making it more complex and less scalable than PuzzleAvatar.

7

CONCLUSION

Contents

7.1 Future Works

At the end of each chapter, we have outlined the technical limitations and potential solutions for ICON (Sec. 3.5), ECON (Sec. 4.4), TeCH (Sec. 5.5), and PuzzleAvatar (Sec. 6.5). Here, we highlight two long-term directions worth pursuing.

1) Lifelong Multimodal Personalization: Despite the daily emergence of GenAI techniques claiming to "democratize" 2D/3D content generation and free designers and artists, the created avatars remain limited to well-known celebrities (*e.g.*, Barack Obama, Bruce Lee), popular Disney cartoons, or Marvel Cinematic characters (*e.g.*, Spider-Man, Elsa), rather than representing the general public. This contradicts the principle that *"all humans are created equal."* These recognizable 3D celebrities, with consistent identities and attire, can be easily extracted from foundational models, like diffusion-based generators, due to their extensive exposure during training. However, ordinary individuals differ in this regard, as they frequently alter their outfits, and only a few photographs are taken of each specific outfit. However, from a lifelong perspective, a significant amount of personal data could be collected or captured over time. Is there a way to use this "lifelong personal data" to build a personalized avatar that evolves with you and even acts as your personal assistant, helping with daily affairs? This avatar would not only look like you,

sound like you, and behave like you, but also think like you — the ultimate form of digital twin agent. PuzzleAvatar (Chapter 6) is our first attempt at this vision, but it is clear that the road ahead is long and challenging. All kinds of modalities (*e.g.*, audio, text, image, video, *etc.*) should definitely be considered under a unified framework.

2) Compositionality and Controllability: From the modeling view, "One-piece" reconstructed or generated digital humans are incompatible with the mainstream simulation engines (*e.g.*, PhysX, Clo3D, MAYA, Houdini). In this context, different components (*e.g.*, garments, hair, accessories) need to be separated after the reconstruction process [118], or generated in a compositional manner in the first place [101]. Additionally, there has been ongoing debate about the best 3D representation: free-form implicit, regularized explicit, NeRF [150], or Gaussian Splatting [98]. It depends. For instance, strand-based suits hair [65], watertight meshes fit articulated bodies [140], open surfaces work for garments [138], neural-based Gaussian Splats or radiance fields are ideal for smoke, and particles model liquids. Instead, the key question should be — how can we efficiently render all these types together [49] and handle collisions between them?

Regarding rendering, if we view the vision foundation models, like Stable Diffusion [181], as a powerful renderer, then Zero-1-to-3 [136] focuses on camera control in the generation process. Moreover, it is equally important to disentangle other rendering factors such as camera, albedo, lighting, and material to enable physically based rendering and controllable editing. This is the generation perspective. From an estimation standpoint, since Marigold [97] has shown that a diffusion-based generator can be repurposed as a depth estimator, could all the above intrinsic factors be accurately estimated from a single RGB image? The field has made some efforts in this direction, such as StableNormal [237], but it remains uncharted territory, and the story is far from over. Finally, compositionality exists not only within the single human subject, but also in the interaction between humans and the physical world. Is it possible to generate a 3D avatar that holds an iPhone or sits on a chair in a compositional manner [100]? Although there is still much work ahead, we should think big and eagerly anticipate even more incredible advances in this field.

7.2 Summary

Compared to text and images, the digitization of the 3D world is underdeveloped due to the costly and time-consuming process of capturing or synthesizing. Specifically, the challenge of creating 3D digital humans with photorealistic appearances, lifelike movements, and faithful personalities in a scalable manner remains unresolved.

This Ph.D. thesis has advanced the field of image-based human reconstruction, transitioning from the "Image-to-Human" framework to a more general "Album-to-Human" setting. This progression includes the development of the algorithms (ICON, ECON, TeCH, PuzzleAvatar) and a new PuzzleIOI benchmark, providing new perspectives on challenges such as out-of-distribution pose robustness, generalization for loose clothing, reconstruction of non-visible sides, and handling unconstrained inputs.

Despite these advancements, several limitations remain. Achieving real-time efficiency and narrowing the quality gap between our reconstructions and studio-captured images, especially for facial features, are ongoing challenges. Additionally, issues such as ineffective shape-color disentanglement and dependency on off-the-shelf estimators (*e.g.*, pose, camera, normal) persist.

While this thesis primarily focuses on image-based human personalization, future research should explore personalization across various modalities (*e.g.*, video, audio, text, *etc.*), adopt lifelong learning approaches, and improve the compositionality and controllability of generation and reconstruction processes. The ultimate ambition of this research is to develop lifelike digital twins for everyone — models that accurately reflect physiological structures and personalities. Such advancements would facilitate human-centric simulations, human-robot interactions, humanoid robots, virtual reality, mixed reality, visual effects, and virtual teleportation, broadening the scope of future AI applications, and ultimately beneficial to all of humanity.

Appendices

A

ICON: Implicit Clothed humans Obtained from Normals

Contents

We provide more details for the method and experiments of ICON, as well as more quantitative and qualitative results, as an extension of Sec. 3.2, Sec. 3.3 and Sec. 3.4 of the main paper of ICON.

A.1 Method & Experiment Details

A.1.1 Dataset

Dataset size: We evaluate the performance of ICON and SOTA methods for a varying training-dataset size (Fig. 3.5 and Tab. A.6). For this, we first combine AGORA [164] ($3,109$ scans) and THuman [257] (600 scans) to get $3,709$ scans in total. This new dataset is 8x times larger than the 450 RenderPeople ("450-Rp") scans used in [185,

186]. Then, we sample this "8x dataset" to create smaller variations, for $1/8$x, $1/4$x, $1/2$x, 1x, and 8x the size of "450-Rp".

Dataset splits: For the "8x dataset", we split the $3,109$ AGORA scans into a new training set ($3,034$ scans), validation set (25 scans) and test set (50 scans). Among these, $1,847$ come from RenderPeople [179] (see Fig. A.2a), 622 from AXYZ [10], 242 from Humanalloy [81], 398 from 3DPeople [1], and we sample only 600 scans from THuman (see Fig. A.2b), due to its high pose repeatability and limited identity variants (see Tab. 3.1), with the "select-cluster" scheme described below. These scans, as well as their SMPL-X fits, are rendered after every 10 degrees rotation around the yaw axis, to totally generate $(3109\ \texttt{AGORA} + 600\ \texttt{THuman} + 150\ \texttt{CAPE}) \times 36 = 138,924$ samples.

Dataset distribution via "select-cluster" scheme: To create a training set with a rich pose distribution, we need to select scans from various datasets with poses different from AGORA. Following SMPLify [18], we first fit a Gaussian Mixture Model (GMM) with 8 components to all AGORA poses, and **select** 2K THuman scans with low likelihood. Then, we apply M-Medoids ($\texttt{n\_cluster} = 50$) on these selections for **clustering**, and randomly pick 12 scans per cluster, collecting $50 \times 12 = 600$ THuman scans in total; see Fig. A.2b. This is also used to split CAPE into "CAPE-FP" (Fig. A.2c) and "CAPE-NFP" (Fig. A.2d), corresponding to scans with poses similar (in-distribution poses) and dissimilar (out-of-distribution poses) to AGORA ones, respectively.

Perturbed SMPL: To perturb SMPL's pose and shape parameters, random noise is added to θ and β by:

$$\begin{aligned} \theta &+= s_\theta * \mu, \\ \beta &+= s_\beta * \mu, \end{aligned} \tag{A.1}$$

where $\mu \in [-1, 1]$, $s_\theta = 0.15$ and $s_\beta = 0.5$. These are set empirically to mimic the misalignment error typically caused by off-the-shell HPS during testing.

Discussion on simulated data: The wide and loose clothing in CLOTH3D++ [14, 145] demonstrates strong dynamics, which would complement commonly used datasets of commercial scans. Yet, the domain gap between CLOTH3D++ and real images is still large. Moreover, it is unclear how to train an implicit function from multi-layer non-watertight meshes. Consequently, we leave it for future research.

A.1.2 Refining SMPL

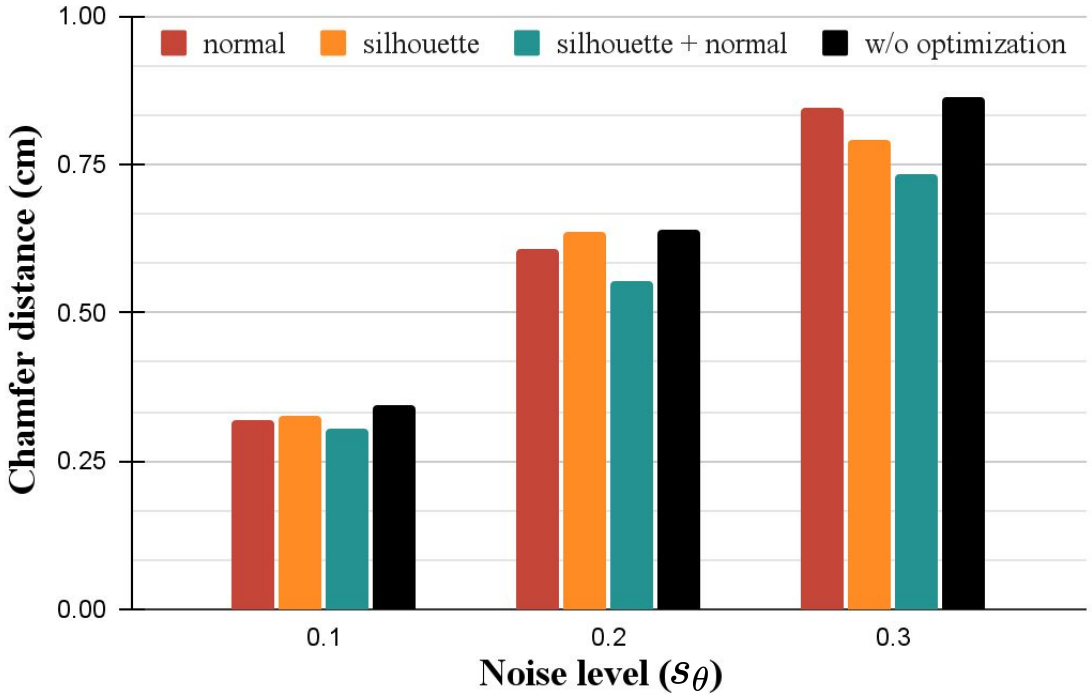


Figure A.1: SMPL refinement error with different losses and noise levels. SMPL refinement error (y-axis) with different losses (see colors) and noise levels, s_θ, of pose parameters (x-axis).

To statistically analyze the necessity of $\mathcal{L}_{\text{N\_diff}}$ and $\mathcal{L}_{\text{S\_diff}}$ in Eq. (3.4), we do a sanity check on AGORA's validation set. Initialized with different pose noise, s_θ (Eq. (A.1)), we optimize the $\{\theta, \beta, t\}$ parameters of the perturbed SMPL by minimizing the difference between rendered SMPL-body normal maps and ground-truth clothed-body normal maps for 2K iterations. As Fig. A.1 shows, $\mathcal{L}_{\text{N\_diff}} + \mathcal{L}_{\text{S\_diff}}$ always leads to the smallest error under any noise level, measured by the Chamfer distance between the optimized perturbed SMPL mesh and the ground-truth SMPL mesh.

	Methods	SMPL-X condition.	AGORA-50			CAPE-FP			CAPE-NFP			CAPE		
			Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓	Chamfer ↓	P2S ↓	Normal ↓
	ICON	✓	1.583	1.987	0.079	**1.364**	**1.403**	0.080	**1.444**	**1.453**	0.083	**1.417**	**1.436**	0.082
D	SMPL-X perturbed	✓	1.984	2.471	0.098	1.488	1.531	0.095	1.493	1.534	0.098	1.491	1.533	0.097
D	$\text{ICON}_{\text{enc(I,N)}}$	✓	1.569	**1.784**	**0.073**	1.379	1.498	**0.070**	1.600	1.580	**0.078**	1.526	1.553	**0.075**
D	$\text{ICON}_{\text{enc(N)}}$	✓	**1.564**	1.854	0.074	1.368	1.484	0.071	1.526	1.524	**0.078**	1.473	1.511	0.076
D	$\text{ICON}_{\text{N}^\dagger}$	✓	1.575	2.016	0.077	1.376	1.496	0.076	1.458	1.569	0.080	1.431	1.545	0.079

Table A.1: Quantitative errors (cm) for several ICON variants conditioned on perturbed SMPL-X fits ($s_\theta = 0.15, s_\beta = 0.5$).

A.1.3 Perceptual study

Reconstruction on in-the-wild images: We perform a perceptual study to evaluate the perceived realism of the reconstructed clothed 3D humans from in-the-wild images. ICON is compared against 3 methods, PIFu [185], PIFuHD [186], and PaMIR [256]. We create a benchmark of 200 unseen images downloaded from the internet, and apply all the methods on this test set. All the reconstruction results are evaluated on Amazon Mechanical Turk (AMT), where each participant is shown pairs of reconstructions from ICON and one of the baselines, see Fig. A.3. Each reconstruction result is rendered in four views: front, right, back and left. Participants are asked to choose the reconstructed 3D shape that better represents the human in the given color image. Each participant is given 100 samples to evaluate. To teach participants, and to filter out the ones that do not understand the task, we set up 1 tutorial sample, followed by 10 warm-up samples, and then the evaluation samples along with catch trial samples inserted every 10 evaluation samples. Each catch trial sample shows a color image along with either (1) the reconstruction of a baseline method for this image and the ground-truth scan that was rendered to create this image, or (2) the reconstruction of a baseline method for this image and the reconstruction for a different image (false positive), see Fig. A.3c. Only participants that pass 70% out of 10 catch trials are considered. This leads to 28 valid participants out of 36 ones. Results are reported in Tab. 3.3.

Normal map prediction: To evaluate the effect of the body prior for normal map prediction on in-the-wild images, we conduct a perceptual study against prediction without the body prior. We use AMT, and show participants a color image along with a pair of predicted normal maps from two methods. Participants are asked to pick the normal map that better represents the human in the image. Front- and back-side normal maps are evaluated separately. See Fig. A.4 for some samples. We set up 2 tutorial samples, 10 warm-up samples, 100 evaluation samples and 10 catch trials for each subject. The catch trials lead to 20 valid subjects out of 24 participants. We report the statistical results in Tab. A.2. A chi-squared test is performed with a null hypothesis that the body prior does not have any influence. We show some results in Fig. A.5, where all participants unanimously prefer one method over the other. While results of both methods look generally similar on front-side normal maps, using the body prior usually leads to better back-side normal maps.

	w/ SMPL prior	w/o SMPL prior	P-value
Preference (front)	47.3%	52.7%	8.77e-2
Preference (back)	52.9%	47.1%	6.66e-2

Table A.2: Perceptual study on normal prediction.

	w/ global encoder	pixel dims	point dims	total dims
PIFu*	✓	12	1	13
PaMIR*	✓	6	7	13
$\text{ICON}_{\text{enc(I,N)}}$	✓	6	7	13
$\text{ICON}_{\text{enc(N)}}$	✓	6	7	13
ICON	✗	0	7	7

Table A.3: Feature dimensions for various approaches. "pixel dims" and "point dims" denote the feature dimensions encoded from pixels (image/normal maps) and 3D body prior, respectively.

A.1.4 Implementation details

Network architecture: ICON's body-guided normal prediction network uses the same architecture as PIFuHD [186], originally proposed in [89], and consisting of residual blocks with 4 down-sampling layers. The image encoder for PIFu*, PaMIR*, and ICON_{enc} is a stacked hourglass [154] with 2 stacks, modified according to [83]. Tab. A.3 lists feature dimensions for various methods; "total dims" is the neuron number for the first MLP layer (input). The number of neurons in each MLP layer is: 13 (7 for ICON), 512, 256, 128, and 1, with skip connections at the 3rd, 4th, and 5th layers.

Training details: For training $\mathcal{G}^{\text{N}}$ we do not use THuman due to its low-quality texture (see Tab. 3.1). On the contrary, $\mathcal{IF}$ is trained on both AGORA and THuman. The front-side and back-side normal prediction networks are trained individually with batch size of 12 under the objective function defined in Eq. (3.3), where we set $\lambda_{\text{VGG}} = 5.0$. We use the ADAM optimizer with a learning rate of 1.0×10^{-4} until convergence at 80 epochs.

Test-time details: During inference, to iteratively refine SMPL and the predicted clothed-body normal maps, we perform 50 iterations (each iteration takes $\sim$ 460 ms on a Quadro RTX 5000 GPU) and set $\lambda_{\text{N}} = 2.0$ in Eq. (3.4). We conduct an experiment to show the influence of the number of iterations (#iterations) on accuracy, see Tab. A.4. The resolution of the queried occupancy space is 256^3. We use `rembg`[i] to segment the humans

[i] https://github.com/danielgatis/rembg

# iterations (460ms/it)	0	10	50
Chamfer ↓	1.417	1.413	**1.339**
P2S ↓	1.436	1.515	**1.378**
Normal ↓	0.082	0.077	**0.074**

Table A.4: ICON errors w.r.t. iterations

Receptive field	139	271	403
Chamfer ↓	1.418	1.478	**1.366**
P2S ↓	1.236	1.320	**1.214**
Normal ↓	0.083	0.084	**0.078**

Table A.5: PaMIR's receptive field

in in-the-wild images, and use `Kaolin`[ii] to compute per-point the signed distance, $\mathcal{F}_{\text{s}}$, and barycentric surface normal, $\mathcal{F}_{\text{n}}^{\text{b}}$.

Discussion on receptive field size: As Tab. A.5 shows, simply reducing the size of receptive field of PaMIR does not lead to better performance. This shows that our informative 3D features as in Eq. (3.6) and normal maps $\widehat{\mathcal{N}}^{\text{c}}$ also play important roles for robust reconstruction. A more sophisticated design of smaller receptive field may lead to better performance and we would leave it for future research.

A.2 More Quantitative Results

Table A.1 compares several ICON variants conditioned on perturbed SMPL-X meshes. For the plot of Fig. 3.5 of the main paper (reconstruction error w.r.t. training-data size), extended quantitative results are shown in Tab. A.6.

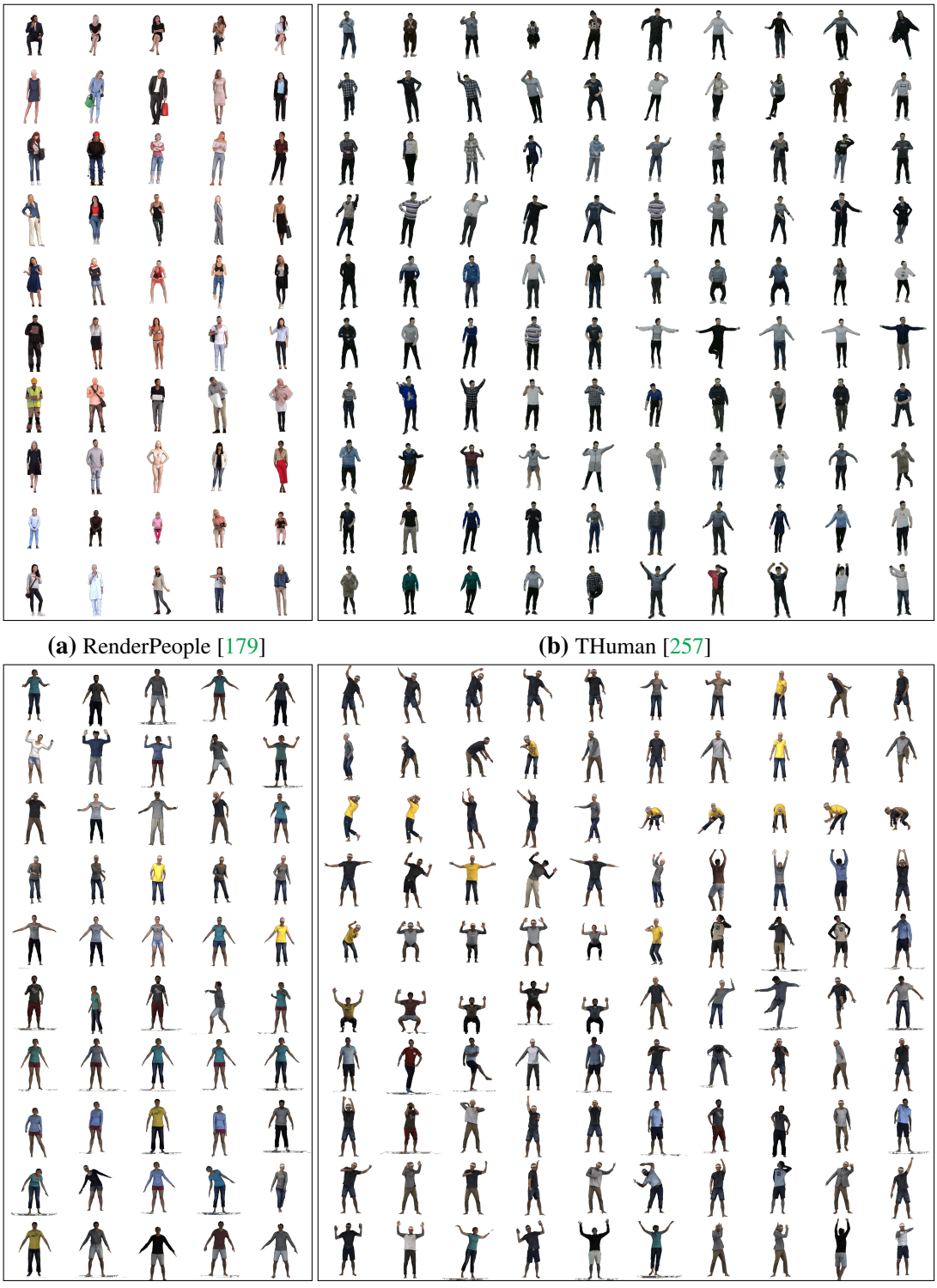
Training set scale		1/8x	1/4x	1/2x	1x	8x
PIFu*	Chamfer ↓	3.339	2.968	2.932	2.682	1.760
	P2S ↓	3.280	2.859	2.812	2.658	1.547
PaMIR*	Chamfer ↓	2.024	1.780	1.479	1.350	1.095
	P2S ↓	1.791	1.778	1.662	1.283	1.131
ICON	Chamfer ↓	1.336	1.266	1.219	1.142	**1.036**
	P2S ↓	1.286	1.235	1.184	1.065	**1.063**

Table A.6: Reconstruction error (cm) w.r.t. training-data size. "Training set scale" is defined as the ratio w.r.t. the 450 scans used in [185, 186]. The "8x" setting is all 3,709 scans of AGORA [164] and THuman [257]. Results outperform ground-truth SMPL-X, which has 1.158 cm and 1.125 cm for Chamfer and P2S in Tab. 3.2.

[ii] https://github.com/NVIDIAGameWorks/kaolin

A.3 More Qualitative Results

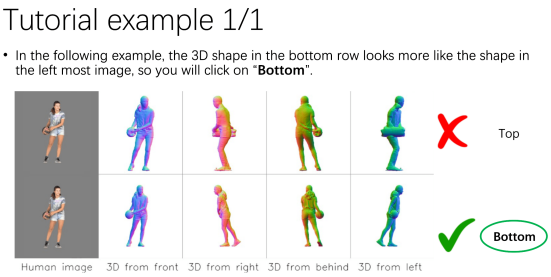
Figures A.6 to A.8 show reconstructions for in-the-wild images, rendered from four different view points; normals are color coded. Figure A.9 shows reconstructions for images with out-of-frame cropping. Figure A.10 shows additional representative failures. The videos on homepage *icon.is.tue.mpg.de* shows animation examples created with ICON and SCANimate.



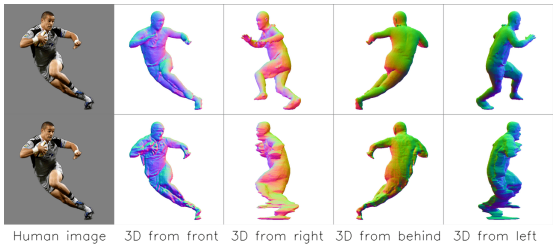
(a) RenderPeople [179] **(b)** THuman [257]

(c) "CAPE-FP" [144] **(d)** "CAPE-NFP" [144]

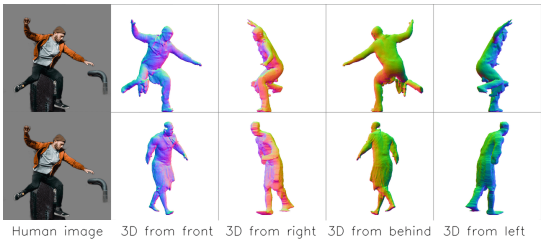
Figure A.2: Representative poses for different datasets.

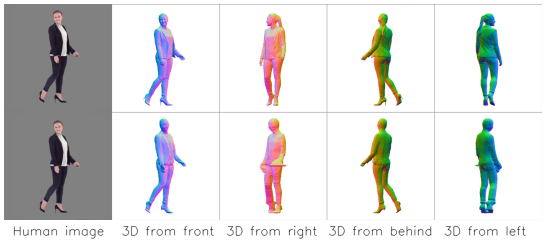


(a) A tutorial sample.



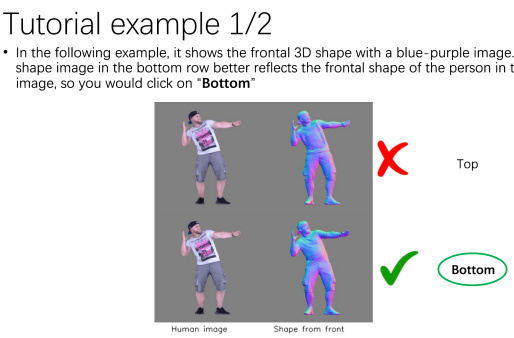
(b) An evaluation sample.

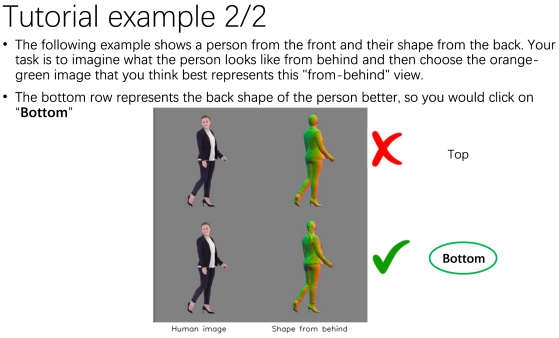




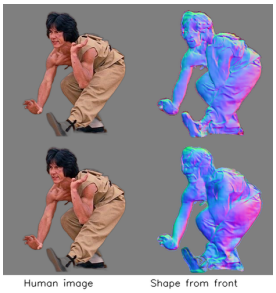
(c) Two samples of catch trials. Left: result from this image (top) vs from another image (bottom). Right: ground-truth (top) vs reconstruction mesh (bottom).

Figure A.3: Some samples in the perceptual study to evaluate **reconstructions** on in-the-wild images.

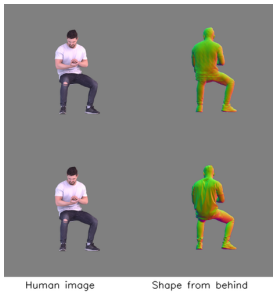




(a) The two tutorial samples.

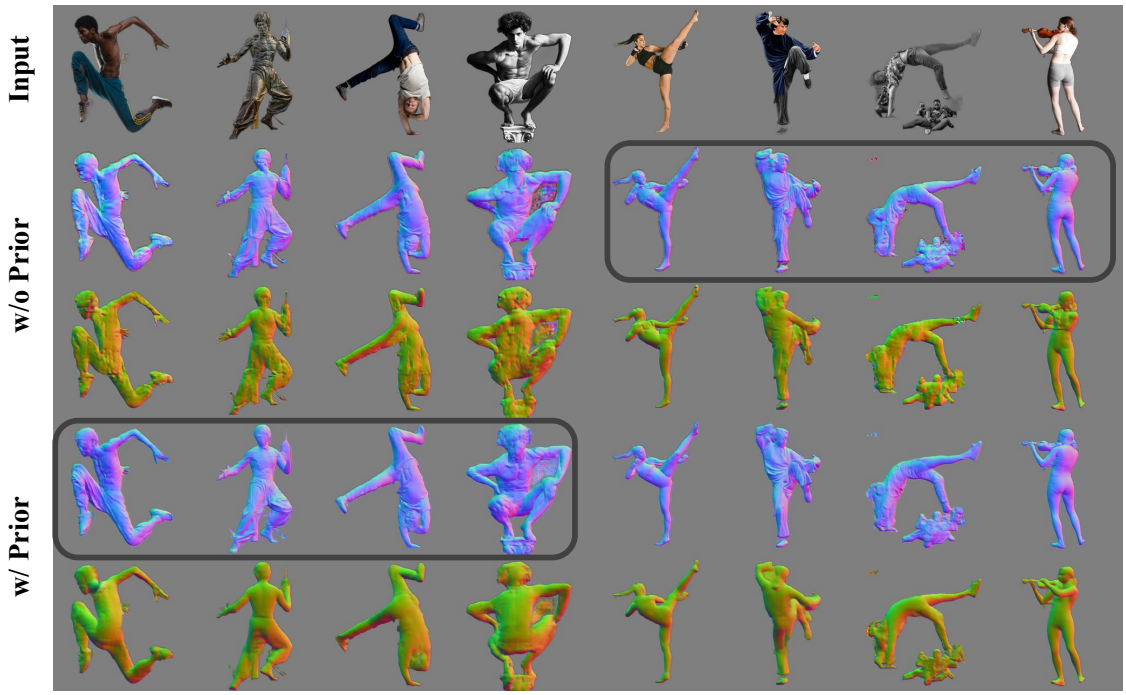


(b) Two evaluation samples.

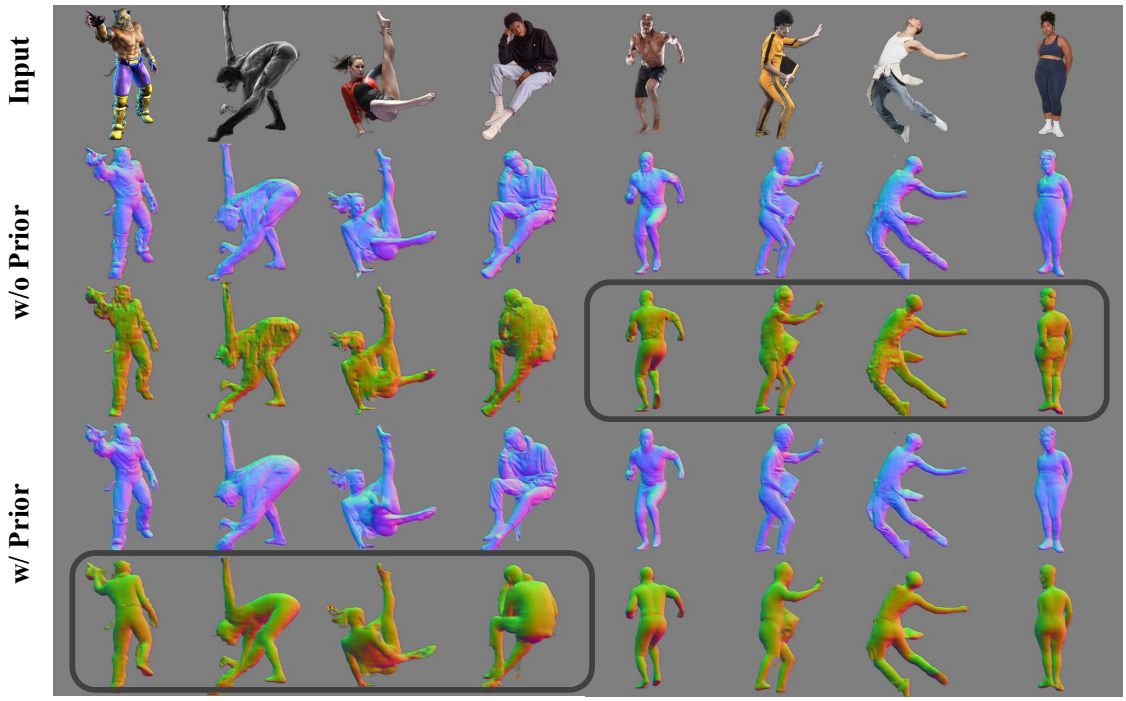


(c) Two catch trial samples.

Figure A.4: Some samples in the perceptual study to evaluate the effect of the **body prior** for **normal prediction** on in-the-wild images.



(a) Examples of perceptual preference on **front** normal maps. Unanimously preferred results are in black boxes. The back normal maps are for reference.



(b) Examples of perceptual preference on **back** normal maps. Unanimously preferred results are in black boxes. The front normal maps are for reference.

Figure A.5: Qualitative results to evaluate the effect of body prior for normal prediction on in-the-wild images.

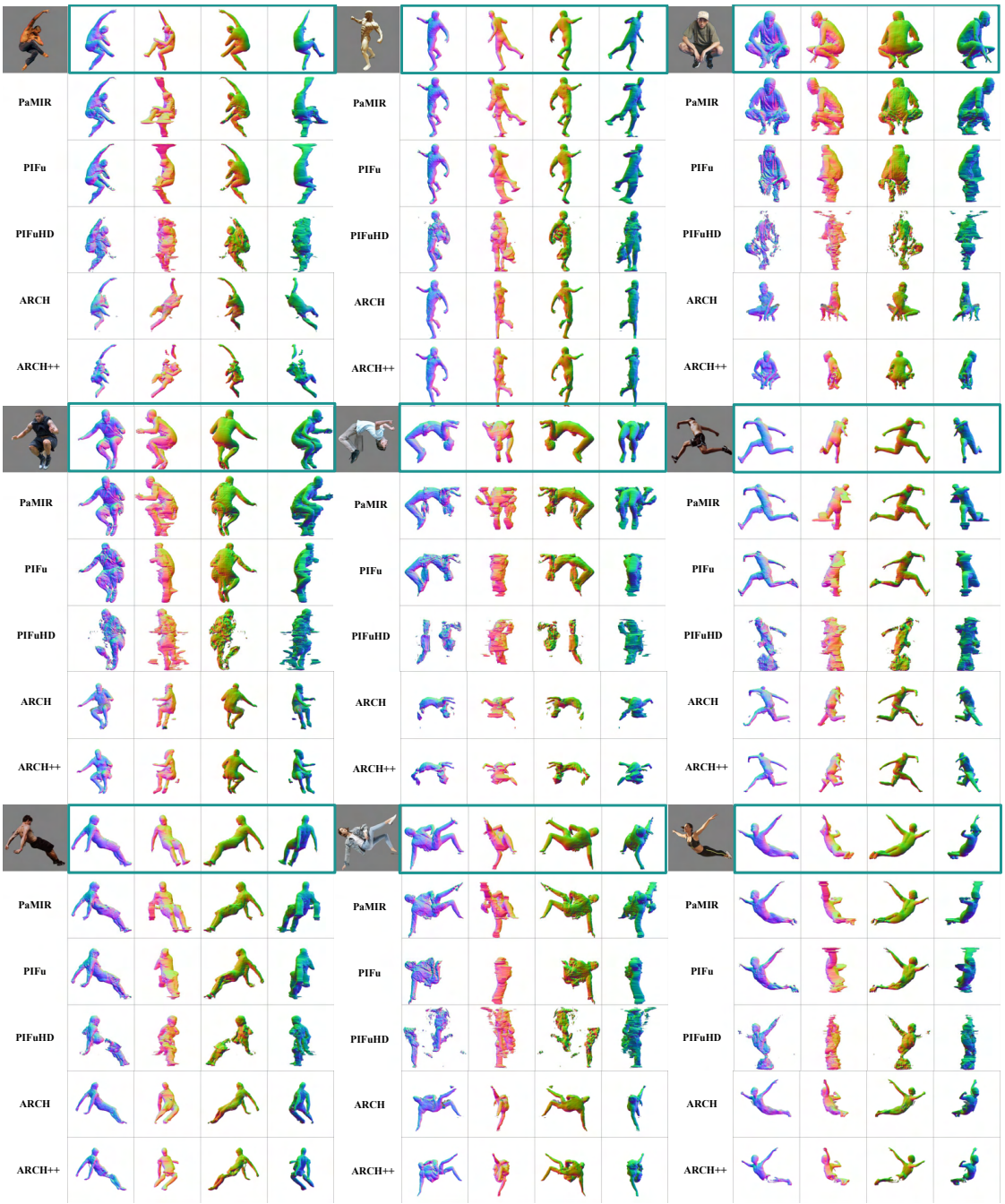


Figure A.6: Qualitative comparison of reconstruction for ICON vs SOTA. Four view points are shown per result.

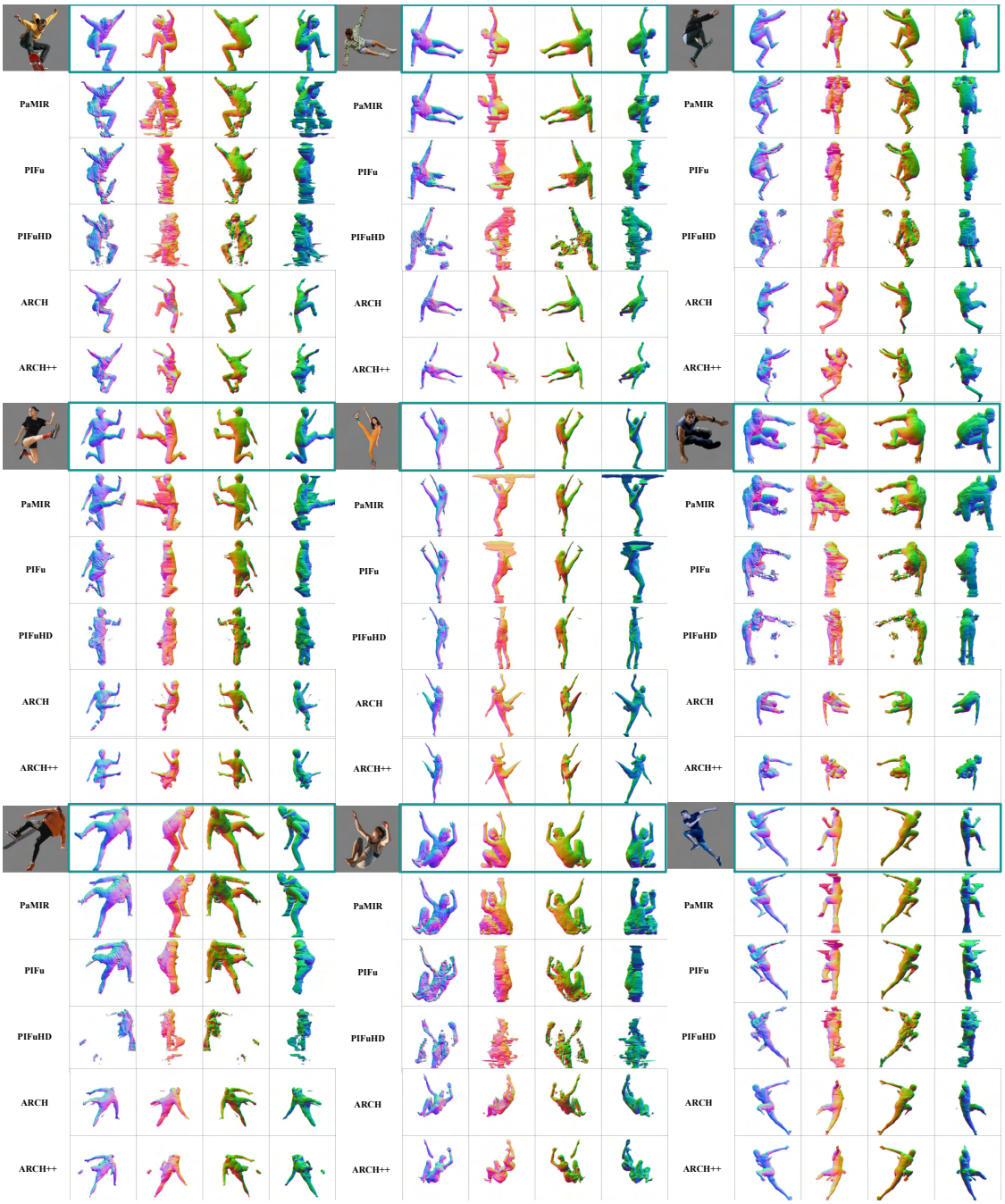


Figure A.7: Qualitative comparison of reconstruction for ICON vs SOTA. Four view points are shown per result.

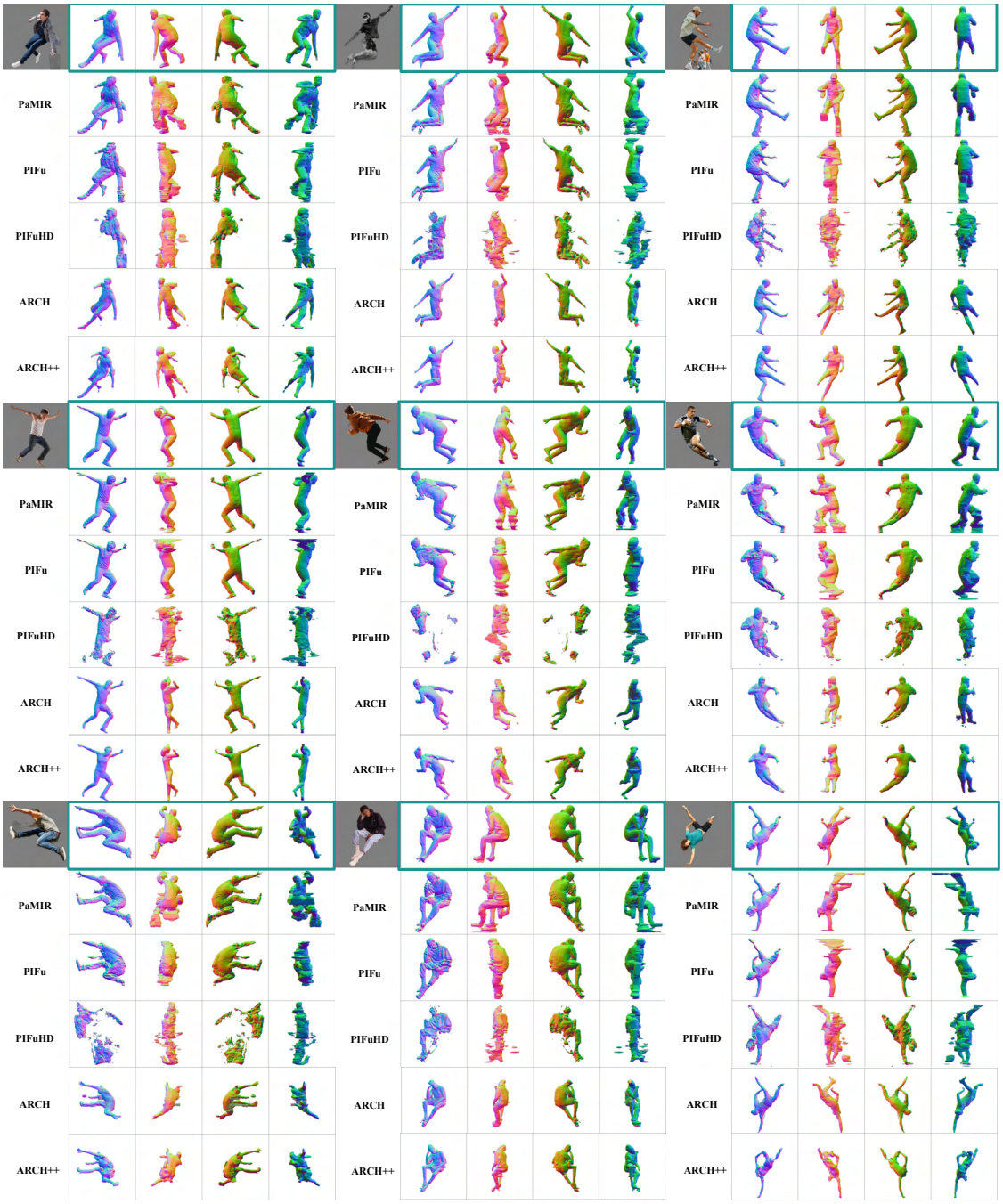


Figure A.8: Qualitative comparison of reconstruction for ICON vs SOTA. Four view points are shown per result.

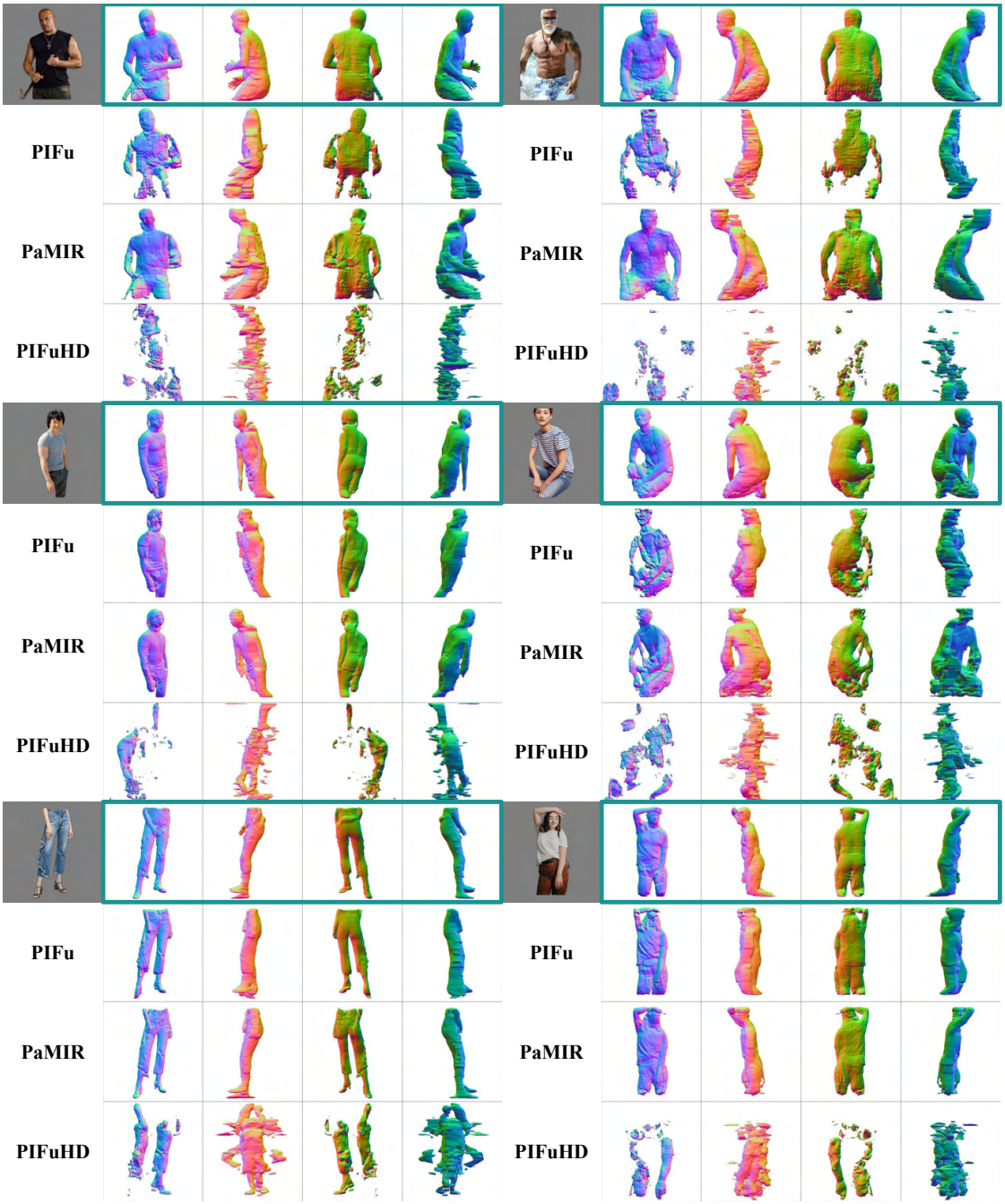


Figure A.9: Qualitative comparison (ICON vs SOTA) on images with out-of-frame cropping.

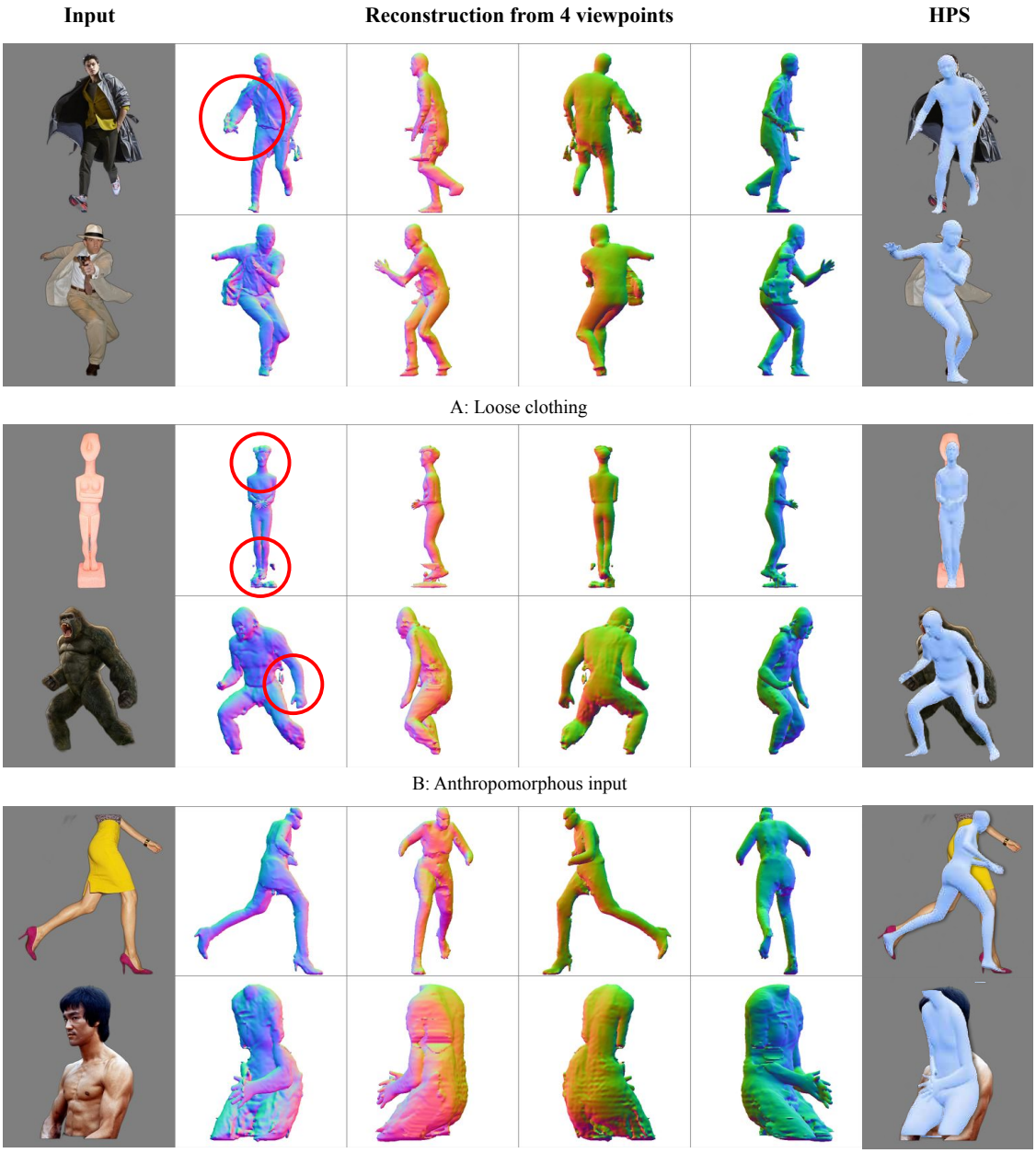


Figure A.10: More failure cases of ICON.

B

ECON: Explicit Clothed humans Optimized via Normal integration

Contents

In the following, we provide more details and discussion on normal prediction, d-BiNI and IF-Nets+, as well as more qualitative results in the perceptual study, as an extension of Sec. 4.2 and Sec. 4.3 of the main paper. We also explore future applications.

B.1 Implementation details

B.1.1 Normal map prediction

We set the loss weights $\lambda_{\text{J\_diff}}$, $\lambda_{\text{N\_diff}}$, and $\lambda_{\text{S\_diff}}$ in Eq. (4.1) to 5.0, 1.0, and 1.0 respectively. However, if the overlap ratio between clothing and body mask is smaller than 0.5, it means humans are dressed with loose clothing. In this situation we trust the 2D joints more and increase the $\lambda_{\text{J\_diff}} = 50.0$. Similarly, when the overlap between body mask inside the clothing mask and full body mask is smaller than 0.98, occlusion happens. In such cases we set $\lambda_{\text{S\_diff}} = 0.0$ to avoid limb self-intersection after pose refinement.

During inference, following ICON (Chapter 3), we iteratively refine SMPL-X and clothed-body normals for 50 iterations (1.10 iter/s on Quadro RTX 5000 GPU). We use *rembg*[i] plus *Mask R-CNN (ResNet50-FPN-V2)* [66] for multi-person segmentation, *Mediapipe* [142] to estimate full-body landmarks, *Open3D* for poisson surface reconstruction [95], and *MonoPort* [121, 122] for fast implicit surface query, and *PyTorch3D* [176] for marching cubes.

B.1.2 d-BiNI

Optimization details: To better present the optimization details, we first write the d-BiNI objective function in a matrix form. Figure 4.3 shows the four inputs to d-BiNI. We vectorize the front and back clothed and prior depth maps $\{\widehat{\mathcal{Z}}^{\mathrm{c}}_{\mathrm{F}}, \widehat{\mathcal{Z}}^{\mathrm{c}}_{\mathrm{B}}, \mathcal{Z}^{\mathrm{b}}_{\mathrm{F}}, \mathcal{Z}^{\mathrm{b}}_{\mathrm{B}}\}$ within Ω_{n} as $\{\widehat{\mathbf{z}_{\mathrm{F}}}, \widehat{\mathbf{z}_{\mathrm{B}}}, \mathbf{z}_{\mathrm{F}}, \mathbf{z}_{\mathrm{B}}\}$; all vectors are of length $|\Omega_{\mathrm{n}}|$. d-BiNI then jointly solves for the front and back clothed depth $\widehat{\mathbf{z}_{\mathrm{F}}}$ and $\widehat{\mathbf{z}_{\mathrm{B}}}$ by minimizing the objective function consisting of the five terms:

$$\begin{aligned}\mathcal{L}(\widehat{\mathbf{z}_{\mathrm{F}}}, \widehat{\mathbf{z}_{\mathrm{B}}}) = &(\mathbf{A}_{\mathrm{F}}\widehat{\mathbf{z}_{\mathrm{F}}} - \mathbf{b}_{\mathrm{F}})^{\top}\mathbf{W}_{\mathrm{F}}(\mathbf{A}_{\mathrm{F}}\widehat{\mathbf{z}_{\mathrm{F}}} - \mathbf{b}_{\mathrm{F}}) \\ &+(\mathbf{A}_{\mathrm{B}}\widehat{\mathbf{z}_{\mathrm{B}}} - \mathbf{b}_{\mathrm{B}})^{\top}\mathbf{W}_{\mathrm{B}}(\mathbf{A}_{\mathrm{B}}\widehat{\mathbf{z}_{\mathrm{B}}} - \mathbf{b}_{\mathrm{B}}) \\ &+\lambda_{\mathrm{d}}(\widehat{\mathbf{z}_{\mathrm{F}}} - \mathbf{z}_{\mathrm{F}})^{\top}\mathbf{M}(\widehat{\mathbf{z}_{\mathrm{F}}} - \mathbf{z}_{\mathrm{F}}) \\ &+\lambda_{\mathrm{d}}(\widehat{\mathbf{z}_{\mathrm{B}}} - \mathbf{z}_{\mathrm{B}})^{\top}\mathbf{M}(\widehat{\mathbf{z}_{\mathrm{B}}} - \mathbf{z}_{\mathrm{B}}) \\ &+\lambda_{\mathrm{s}}(\widehat{\mathbf{z}_{\mathrm{F}}} - \widehat{\mathbf{z}_{\mathrm{B}}})^{\top}\mathbf{S}(\widehat{\mathbf{z}_{\mathrm{F}}} - \widehat{\mathbf{z}_{\mathrm{B}}}).\end{aligned} \tag{B.1}$$

Here, $\mathbf{A}_{\mathrm{F}} \in \mathbb{R}^{4|\Omega_{\mathrm{n}}|\times|\Omega_{\mathrm{n}}|}$ and $\mathbf{b}_{\mathrm{F}} \in \mathbb{R}^{4|\Omega_{\mathrm{n}}|}$ are constructed from the front normal map following Eq. (21) of BiNI [22]; $\mathbf{A}_{\mathrm{B}}$ and $\mathbf{b}_{\mathrm{B}}$ are from the back normal map. $\mathbf{W}_{\mathrm{F}}$ and $\mathbf{W}_{\mathrm{B}} \in \mathbb{R}^{4|\Omega_{\mathrm{n}}|\times 4|\Omega_{\mathrm{n}}|}$ are bilateral weight matrices for front and back depth maps, respectively; both are constructed following Eq. (22) of BiNI [22] and depend on the unknown depth. $\mathbf{M}$ and $\mathbf{S}$ are $|\Omega_{\mathrm{n}}|\times|\Omega_{\mathrm{n}}|$ diagonal matrices whose diagonal entries indicate the pixels with depth priors and located at the silhouette, respectively. Specifically, the i-th diagonal entry m_i of $\mathbf{M}$ is

$$m_i = \begin{cases} 1, & \text{if } i\text{-th entry of } \widehat{\mathbf{z}_{\mathrm{F}}} \text{ in } \Omega_{\mathrm{z}} \\ 0, & \text{otherwise} \end{cases}, \tag{B.2}$$

[i] https://github.com/danielgatis/rembg

while the i-th diagonal entry s_i of $\mathbf{S}$ is

$$s_i = \begin{cases} 1, & \text{if } i\text{-th entry of } \widehat{\mathbf{z}_\mathrm{F}} \text{ in } \partial\Omega_\mathrm{n} \\ 0, & \text{otherwise} \end{cases} . \tag{B.3}$$

Stacking $\widehat{\mathbf{z}_\mathrm{F}}$ and $\widehat{\mathbf{z}_\mathrm{B}}$ as $\widehat{\mathbf{z}} = \begin{bmatrix} \widehat{\mathbf{z}_\mathrm{F}} \\ \widehat{\mathbf{z}_\mathrm{B}} \end{bmatrix}$, Eq. (B.1) then reads

$$\begin{aligned} \mathcal{L}(\widehat{\mathbf{z}}) = (\mathbf{A}\widehat{\mathbf{z}} - \mathbf{b})^\top \mathbf{W}(\mathbf{A}\widehat{\mathbf{z}} - \mathbf{b}) + \\ \lambda_\mathrm{d}(\widehat{\mathbf{z}} - \mathbf{z})^\top \widetilde{\mathbf{M}}(\widehat{\mathbf{z}} - \mathbf{z}) + \lambda_\mathrm{s}\widehat{\mathbf{z}}^\top \widetilde{\mathbf{S}}\widehat{\mathbf{z}}, \end{aligned} \tag{B.4}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_\mathrm{F} & \\ & \mathbf{A}_\mathrm{B} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{b}_\mathrm{F} \\ \mathbf{b}_\mathrm{B} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}_\mathrm{F} & \\ & \mathbf{W}_\mathrm{B} \end{bmatrix},$$
$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_\mathrm{F} \\ \mathbf{z}_\mathrm{B} \end{bmatrix}, \quad \widetilde{\mathbf{M}} = \begin{bmatrix} \mathbf{M} & \\ & \mathbf{M} \end{bmatrix}, \quad \widetilde{\mathbf{S}} = \begin{bmatrix} \mathbf{S} & -\mathbf{S} \\ -\mathbf{S} & \mathbf{S} \end{bmatrix}.$$

To minimize Eq. (B.4), we perform an iterative optimization similar to BiNI [22]. At each iteration, we first fix the weights $\mathbf{W}$ and jointly solve for the front and back depth $\widehat{\mathbf{z}}$, then compute the new weights from the updated depth. When $\mathbf{W}$ is fixed and treated as a constant matrix, solving for the depth becomes a convex least-squares problem. The necessary condition for the global optimum is obtained by equating the gradient of Eq. (B.4) to $\mathbf{0}$:

$$(\mathbf{A}^\top \mathbf{W}\mathbf{A} + \lambda_\mathrm{d}\widetilde{\mathbf{M}} + \lambda_\mathrm{s}\widetilde{\mathbf{S}})\widehat{\mathbf{z}} = \mathbf{A}^\top \mathbf{W}\mathbf{b} + \lambda_\mathrm{d}\widetilde{\mathbf{M}}\mathbf{z}. \tag{B.5}$$

Equation (B.5) is a large-scale sparse linear system with a symmetric positive definite coefficient matrix. We solve Eq. (B.5) using a CUDA-accelerated sparse conjugate gradient solver with a Jacobi preconditioner [ii].

B.1.3 IF-Nets+

Network structure: As Fig. B.1 shows, similar to IF-Nets [33], IF-Nets+ applies multi-scale voxel 3D CNN encoding on voxelized d-BiNI and the SMPL-X surface, namely $\mathcal{F}_1^{\text{d-BiNI}}$ and $\mathcal{F}_1^{\text{SMPL-X}}$, generating multi-scale deep feature grids to account for both local

[ii] https://docs.cupy.dev/en/stable/reference/generated/cupyx.scipy.sparse.linalg.cg.html

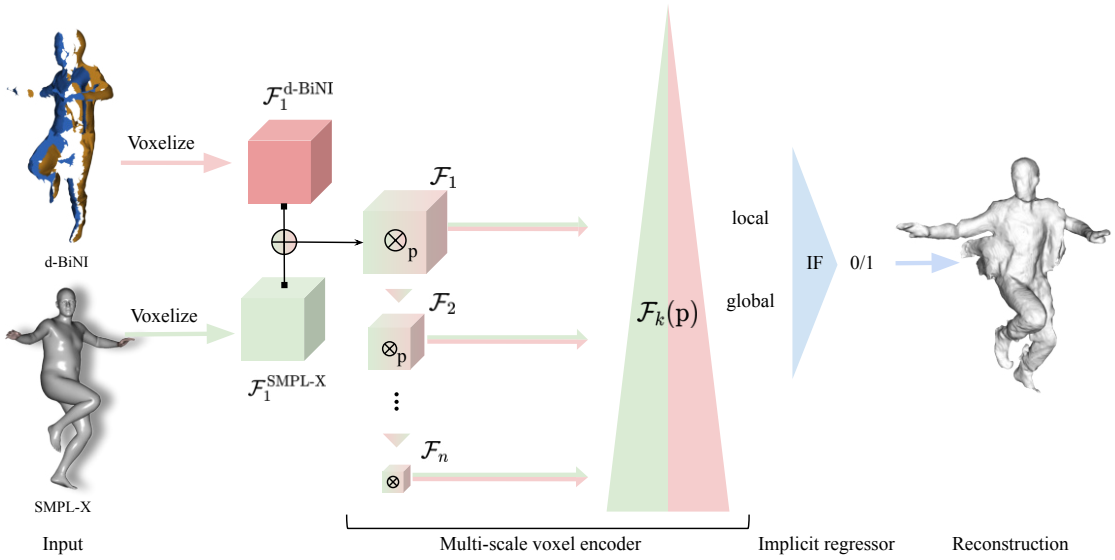


Figure B.1: Overview of IF-Nets+

and global information, $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n, \mathcal{F}_k \in \mathbb{R}^{K \times K \times K \times C_k}, n = 6$. These deep features are with decreasing resolution $K = \frac{N}{2^{k-1}}, N = 256$ and variable dimension channels $C = \{32, 32, 64, 128, 128, 128\}$. All these features are then fed into an implicit function regressor, parameterized by a Multi-Layer Perceptron (MLP), to predict the occupancy value of point P. This MLP regressor is trained with BCE loss.

Training setting: IF-Nets and IF-Nets+ share the same training setting. The voxelization resolution for both SMPL-X and d-BiNI surfaces is 256^3. We use RMSprop as an optimizer, with a learning rate $1e^{-4}$, and weight decay by a factor of 0.1 every 10 epochs. These networks are trained on an NVIDIA A100 for 20 epochs with a batch size of 48. Following ICON (Chapter 3), we sampled 10000 points with the mixture of cube-uniform sampling and surface-around sampling, with standard deviation of 5cm.

Dataset details: We augment THuman [240] by (1) rotating the scans every 10 degrees around the yaw axis, to generate $525 \times 36 = 18900$ samples in total, and (2) randomly selecting a rectangle region from the d-BiNI depth maps, and erasing its pixels [258]. In particular, the erasing operation is being performed with $p = 0.8$ probability, the range of aspect ratio of erased area is between 0.3 and 3.3, and its range of proportion are $\{0.01, 0.05, 0.2\}$.

Speed analysis of ECON vs. ICON: d-BiNI takes 6.2 secs (150 iterations). For ECON_{IF}, the IF-Nets+ plus Marching cubes takes 2.6 secs (for 256^3 resolution), and the Poisson

step takes 10.7 secs (level=10). For a single image, ECON_{IF} takes 112 secs, and ECON_{EX} takes 97 secs. ICON, which shares the same SMPL-X fitting (w/ landmarks), takes 78 secs, and w/ cloth-refinement (50 iterations) it takes 115 secs.

B.2 Qualitative results

Figure B.2 shows PaMIR's results on the same photos in Fig. 4.12. Figures B.3 to B.5 show more comparisons used in our perceptual study, containing the results on in-the-wild images with challenging poses, loose clothing, and standard fashion poses, respectively. For each image, we display the results obtained by ECON, PaMIR [256], ICON (Chapter 3), and PIFuHD [186]. In each row, we show normal maps rendered in $\{0°, 90°, 180°, 270°\}$ views. The videos on homepage econ.is.tue.mpg.de shows more reconstructions with a rotating virtual camera.

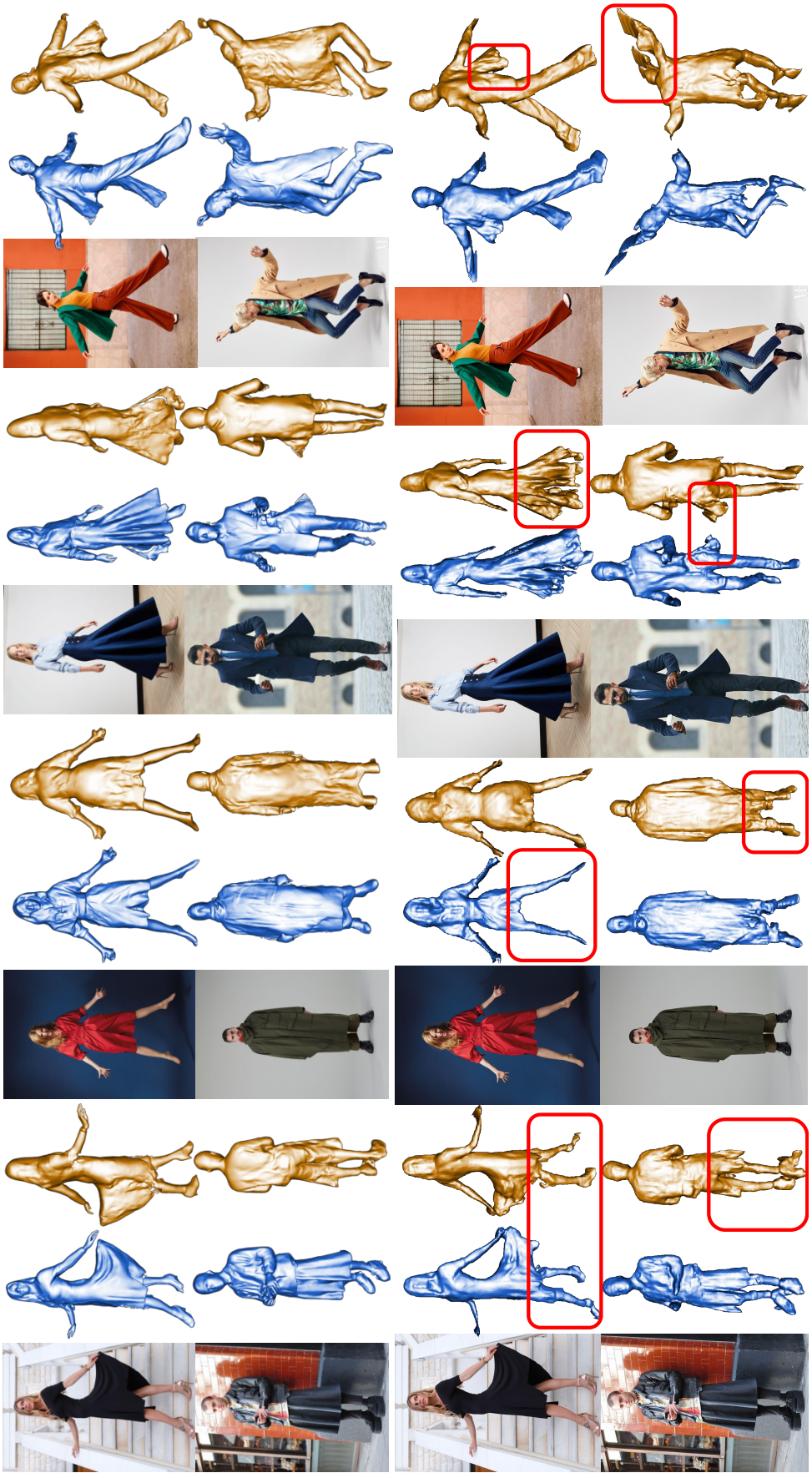


Figure B.2: ECON (Top) vs. **PaMIR** (Bottom) on loose clothes; **Q Zoom in** to see front/back 3D details.

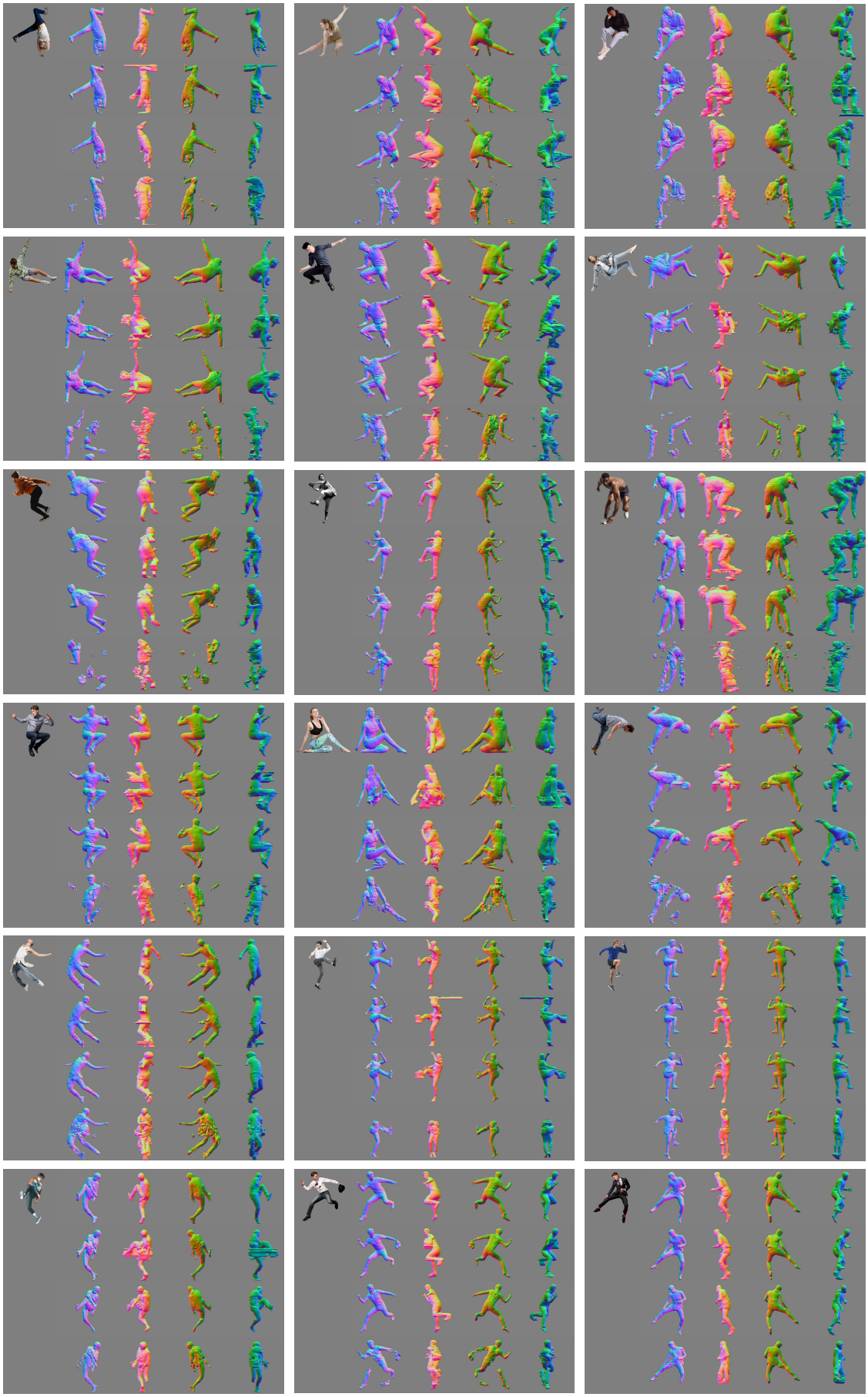


Figure B.3: Results on in-the-wild images with challenging poses: For each example, the format is as follows: **Top** → **bottom:** ECON, PaMIR [256], ICON (Chapter 3), and PIFuHD [186]. **Left** → **right:** Virtual camera rotated by $\{0°, 90°, 180°, 270°\}$. 🔍 **Zoom in** to see 3D details.

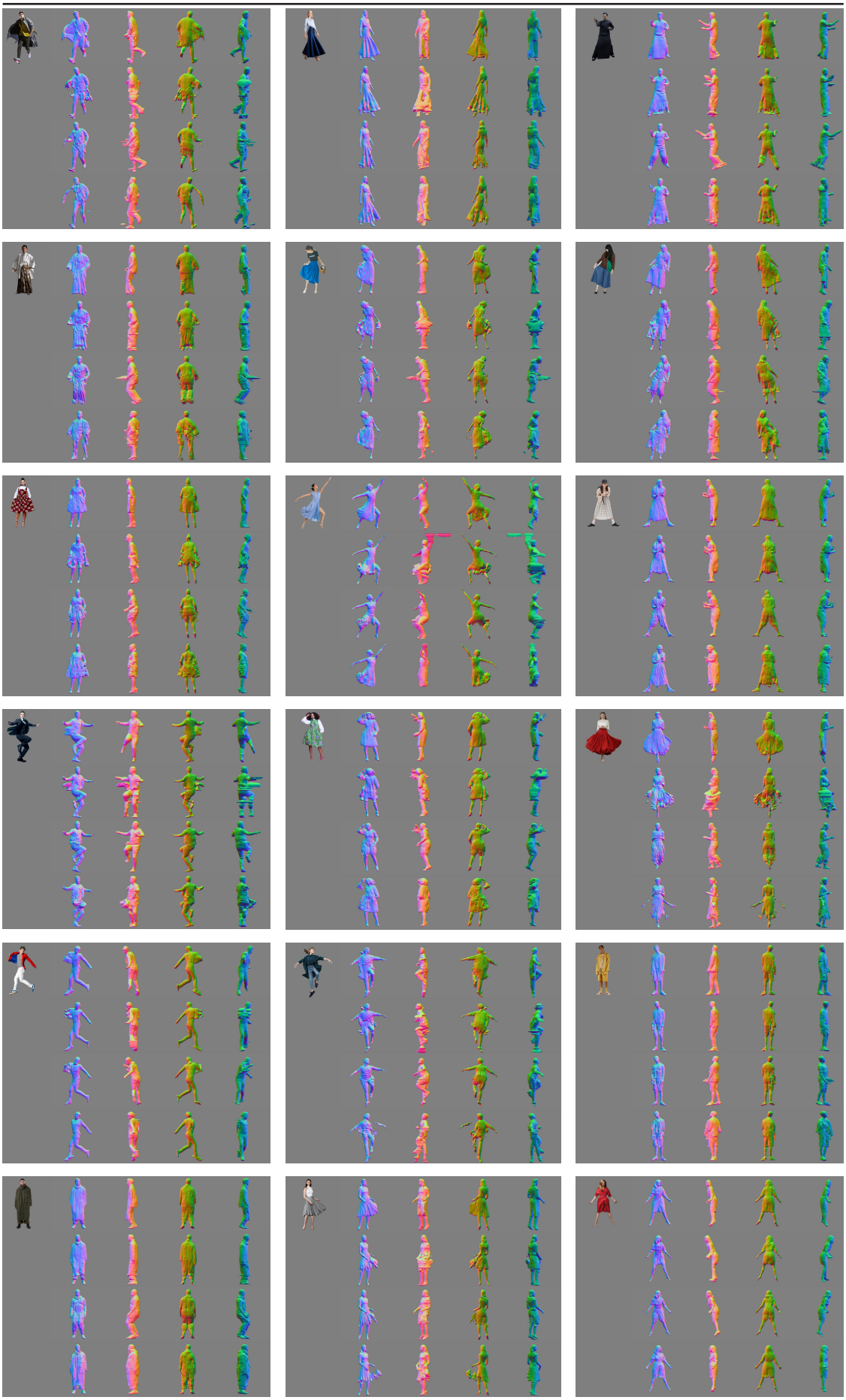


Figure B.4: Results on in-the-wild images with loose clothing: For each example, the format is as follows: **Top** → **bottom:** ECON, PaMIR [256], ICON (Chapter 3), and PIFuHD [186]. **Left** → **right:** Virtual camera rotated by $\{0°, 90°, 180°, 270°\}$. 🔍 **Zoom in** to see 3D details.

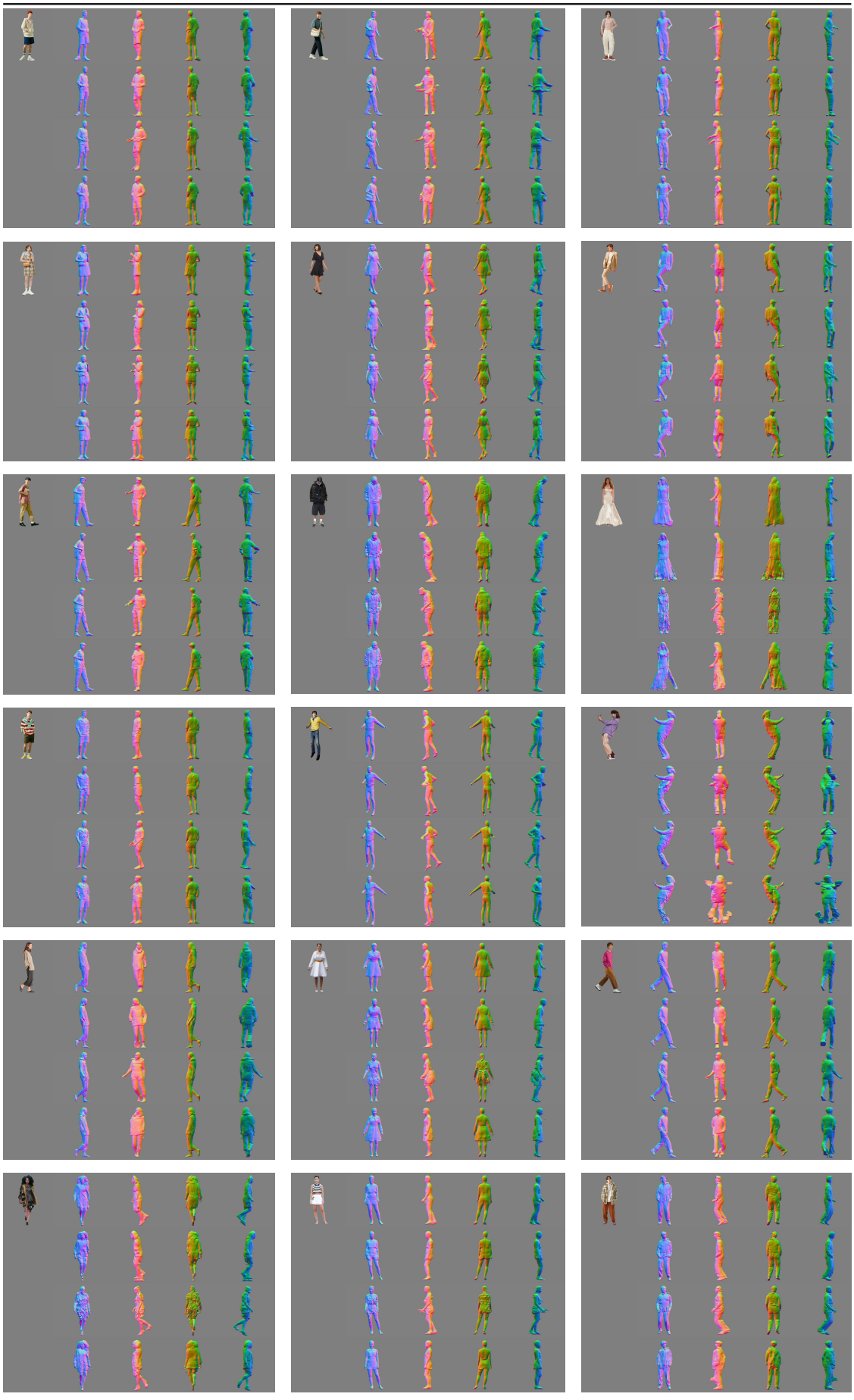


Figure B.5: Results on in-the-wild fashion images: For each example the format is as follows: **Top** → **bottom:** ECON, PaMIR [256], ICON (Chapter 3), and PIFuHD [186]. **Left** → **right:** Virtual camera rotated by $\{0°, 90°, 180°, 270°\}$. 🔍 **Zoom in** to see 3D details.

C

TeCH: Text-guided Reconstruction of Lifelike Clothed Humans

Contents

We provide the preliminaries (Appendix C.1) of TeCH. We list the VQA questions P_{VQA} (Appendix C.2), and provide a more in-depth analysis of the description prompt P (Sec. 5.2.2). Additional implementation details to construct the outer shell around SMPL-X (Appendix C.3), as well as details on the camera sampling strategy (Appendix C.4) are given. Implementation details of the network structure and optimization settings are presented in Appendix C.5. The limitations, efficiency of training and testing, future works, and broader impact are discussed in Sec. 5.5. Further, two applications enabled by TeCH are shown (Sec. 5.4), and additional qualitative results based on the benchmark datasets (CAPE, THuman2.0) and in-the-wild photos used in the perceptual studies are presented in Figs. C.1 to C.3.

C.1 Preliminaries

DreamBooth. Pretrained text-to-Image diffusion models [175, 181, 184] lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. To enable subject-driven image generation, DreamBooth [183] *personalizes* the pretrained diffusion model through few-shot tuning.

Specifically, for a pre-trained image diffusion model $\hat{\mathbf{x}}_\phi$, the model takes an initial noise $\varepsilon \sim \mathcal{N}(0,1)$, and a text embedding $\mathbf{c} = \Gamma(P)$, generated by the text encoder Γ and a text prompt P, to produce an image $\mathbf{x}_{\text{gen}} = \hat{\mathbf{x}}_\phi(\varepsilon, \mathbf{c})$. DreamBooth uses 3$\sim$5 images of the same subject to fine-tune the diffusion model using MSE denoising losses:

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\mathbf{c},\varepsilon,\varepsilon',t} = \Big[w_t \left\| \hat{\mathbf{x}}_\phi(\alpha_t \mathbf{x}_{\text{gt}} + \sigma_t \varepsilon, \mathbf{c}) - \mathbf{x}_{\text{gt}} \right\|_2^2 \\ + \lambda w_{t'} \left\| \hat{\mathbf{x}}_\phi(\alpha_{t'} \mathbf{x}_{\text{prior}} + \sigma_{t'} \varepsilon', \mathbf{c}_{\text{prior}}) - \mathbf{x}_{\text{prior}} \right\|_2^2 \Big].\end{aligned} \tag{C.1}$$

Where $\mathbf{x}_{\text{gt}}$ represents ground-truth images, and $\mathbf{c}$ is the embedding of a text prompt with a rare token as the unique identifier, and α_t, σ_t, w_t controls the noise schedule and sample quality of the diffusion process at time $t \sim \mathcal{U}([0,1])$. The second term is the prior-preservation loss weighted by λ, which is supervised by self-generated images $\mathbf{x}_{\text{prior}}$ conditioned with the class-specific embedding $\mathbf{c}_{\text{prior}} = \Gamma(\text{“}\texttt{a man/woman}\text{”})$. This loss mitigates the phenomenon of language drift, where the model collapses into a single mode by associating the class name with a particular instance, thus augmenting the output diversity.

Score Distillation Sampling (SDS). DreamFusion [169] introduces Score Distillation Sampling (SDS) loss, to perform Text-to-3D synthesis by using a pretrained 2D Text-to-Image diffusion model ϕ. Instead of sampling in pixel space, SDS optimizes over the 3D volume, which is parameterized with θ, with the differential renderer g, so the generated image $\mathbf{x} = g(\theta)$ closely resembles a sample from the frozen diffusion model. Here is the gradient of $\mathcal{L}_{\text{SDS}}$:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) = \mathbb{E}_{\text{t},\varepsilon} \left[w_t \left(\hat{\varepsilon}_\phi(\mathbf{z}_t^{\mathbf{x}}; \mathbf{c}, t) - \varepsilon \right) \frac{\partial \mathbf{x}}{\partial \theta} \frac{\partial \mathbf{z}^{\mathbf{x}}}{\partial \mathbf{x}} \right], \tag{C.2}$$

where $\hat{\varepsilon}_\phi(\mathbf{z}_t^{\mathbf{x}}; \mathbf{c}, t)$ denotes the noise prediction of the diffusion model with condition $\mathbf{c}$ and latent $\mathbf{z}_t^{\mathbf{x}}$ of the generated image $\mathbf{x}$. Such SDS-guided optimization is performed with

random camera poses to improve the multi-view consistency. In contrast to DreamFusion, the 3D shape here is parameterized with an improved DMTet instead of NeRF.

Deep Marching Tetrahedra (DMTet). DMTet [55, 190] is a hybrid 3D representation designed for high-resolution 3D shape synthesis and reconstruction. It incorporates the advantages of both explicit and implicit representations, by learning Signed Distance Field (SDF) values on the vertices of a deformable tetrahedral grid. For a given DMTet, represented as (V_T, T), where V_T are the vertices in the tetrahedral grid T, comprising K tetrahedrons $T_k \in T$, with $k \in \{1, \ldots, K\}$. Each tetrahedron is defined by four vertices $\{v_k^1, v_k^2, v_k^3, v_k^4\}$. The objective of the model is firstly to estimate the SDF value $s(v_i)$ for each vertex, then to iteratively refine the surface and subdivide the tetrahedral grid by predicting the position offsets Δv_i and SDF residual values $\Delta s(v_i)$. A triangular mesh can be extracted through Marching Tetrahedra [42]. As noted by Magic3D [132], DMTet offers two advantages over NeRF, **fast-optimization** and **high-resolution**. It achieves this by efficiently rasterizing a triangular mesh into high-resolution image patches using a differentiable renderer [110], enabling interaction with pre-trained high-resolution latent diffusion models, such as eDiff-I [12], and Stable Diffusion [181].

C.2 VQA Questions

To construct the descriptive prompt P_{VQA}, we designed a series of questions to parse clothed human attributes. First, we use BLIP [114] and a series of general questions Q_{general} to parse genders, facial appearance, hair colors, hairstyles, facial hairs, and body poses. Secondly, we use SegFormer [225] to parse human garments, consisting of 10 categories `{hat, sunglasses, upper-clothes, skirt, pants, dress, belt, shoes, bag, scarf}`, denoted as G, and use another group of questions Q_{garments} to parse the attribute of each garment $g \in G$. All the questions are listed in Tab. C.1.

Empirically, we found that the BLIP [114] VQA model tends to use $1 \sim 3$ words to answer these questions, so we simply concatenate all the answers and remove repeated words to construct P_{VQA}. Note that for the CAPE dataset, we add the dataset-specific description "`hairnet`" to the guidance, as it is difficult to be recognized by BLIP.

Groups	Quetions
Q_{general}	Is this person a man or a woman?
	What is this person wearing?
	What is the hair color of this person?
	What is the hairstyle of this person?
	Describe the facial appearance of this person.
	Does this person have facial hair?
	How is the facial hair of this person?
	Describe the pose of this person.
Q_{garments}	Is this person wearing g?
	What g is the person wearing? $\rightarrow d$
	What is the color / style of the d + g?

Table C.1: Predefined questions for parsing clothed human attributes. g is the segmentation category of a part of the garments, and d is the recognized garment category from the answer to the second question in Q_{garments}.

C.3 Construction of the Outer SMPL-X Shell

To construct a compact tetrahedral grid $(V_{\text{shell}}, T_{\text{shell}})$, we calculate a coarse outer shell M_{shell} from SMPL-X estimated body mesh M_{body}. Specifically, we dilate M_{body} with an offset of $\Delta M_{\text{body}} = 0.1$ and simplify the mesh by reducing triangle numbers by $r_{\text{decimate}} = 90\%$ using quadric decimation [73]. Then, we generate the tetrahedral grid $(V_{\text{shell}}, T_{\text{shell}})$ of this outer shell by TetGen [64] with a maximum volume size of 5×10^{-8}.

C.4 Camera Sampling

To ensure full coverage of the entire body and the human face, during optimization process, we sample virtual camera poses into two groups: 1) $\mathbf{K}_{\text{body}}$ cameras with a field of view (FOV) covering the full body or the main body parts, and 2) zoom-in cameras $\mathbf{K}_{\text{face}}$ focusing the face region.

The ratio $\mathcal{P}_{\text{body}}$ determines the probability of sampling $\mathbf{k} \in \mathbf{K}_{\text{body}}$, while the height h_{body}, radius r_{body}, elevation angle ϕ_{body}, and azimuth ranges θ_{body} are adjusted relative to the SMPL-X body scale. Empirically, we set $\mathcal{P}_{\text{body}} = 0.7$, $h_{\text{body}} = (-0.4, 0.4)$, $r_{\text{body}} = (0.7, 1.3)$, $\theta_{\text{body}} = [-180°, 180°)$, $\phi_{\text{body}} = \{0°\}$, with the M_{body} proportionally

scaled to a unit space with xyz coordinates in the range $[-0.5, 0.5]$. To mitigate the occurrence of mirrored appearance artifacts (*i.e.*, Janus-head), we incorporate view-aware prompts, "`front/side/back/overhead view`", w.r.t. the viewing angle during generation process, whose effectiveness has been demonstrated in DreamBooth [169].

To enhance facial details, we sample additional virtual cameras positioned around the face $\mathbf{k} \in \mathbf{K}_{\text{face}}$, together with the additional prompt "`face of`". With a probability of $\mathcal{P}_{\text{face}} = 1 - \mathcal{P}_{\text{body}} = 0.3$, the sampling parameters include the view target c_{face}, radius range r_{face}, rotation range θ_{face}, and azimuth range ϕ_{face}. Empirically, we set c_{face} to the 3D position of SMPL-X head keypoint, $r_{\text{face}} = [0.3, 0.4]$, $\theta_{\text{face}} = [-90°, 90°]$ and $\phi_{\text{face}} = \{0°\}$.

C.5 Implementation Details

C.5.1 Network Structure

We use two networks Ψ_{g} and Ψ_{c} to predict the SDF for geometry modeling and to predict the RGB value for albedo texture modeling, respectively. For Ψ_{g}, we use a 2-layer MLP network with a hidden dimension of 32 and a hash positional encoding with a maximum resolution of 1028 and 16 resolution levels. During the forward process, we use coordinates of V_{shell} in the normalized unit space, the vertices of the tetrahedral grid as the input of Ψ_{g} to query SDF value for each vertex.

For Ψ_{c}, we use a similar network with 1-layer MLP and a hash positional encoding with a maximum resolution of 2048. We model the albedo texture in the canonical A-pose 3D space. Specifically, for the post-processed result mesh $M_{\text{in}} = (V_{\text{in}}, F)$, we register the model with SMPL-X, and repose it with the standard A-pose $M_{\text{A}} = (V_{\text{A}}, F)$. During rendering, if a target pixel is projected onto a triangle $(v^i_{\text{in}}, v^j_{\text{in}}, v^j_{\text{in}})$, where$(i, j, k) \in F$ of the M_{in}. We query the pixel color with its corresponding 3d position in the A-pose space, calculated by interpolation of the triangle $(v^i_{\text{A}}, v^j_{\text{A}}, v^j_{\text{A}})$. Additionally, we use two 2-layer MLP $\Psi^{\text{g}}_{\text{bg}}, \Psi^{\text{c}}_{\text{bg}}$ conditioned by camera $\mathbf{k}$ to learn adaptive 3D background colors for both normal map rendering $\mathcal{N}(M, \mathbf{k})$ and color rendering $\mathcal{I}'(M, \psi_{\text{c}}, \mathbf{k})$.

C.5.2 Optimization Details

In both stages of TeCH's multi-stage optimization pipeline, we use an Adam optimizer with a base learning rate of $\eta = 1 \times 10^{-}3$, and weight decay of $\lambda_{\text{WD}} = 5 \times 10^{-4}$. We utilize super-sampling with a factor of $r_{SS} = 4$ to render normal maps and colored images, which aids in antialiasing during optimization.

Geometry-stage optimization. We optimize Ψ_{g} in a coarse-to-fine manner, with $t_{\text{coarse}} = 5000$ steps w/o mesh subdivision and $t_{\text{fine}} = 5000$ steps w/ mesh subdivision. We use a loss weight setting of $\lambda_{\text{sil}} = 1 \times 10^4$, $\lambda_{\text{SDS}} = 1$, $\lambda_{\text{lap}} = 1 \times 10^4$, and a base loss weight $\lambda_{\text{norm}}^{\text{base}} = 1 \times 10^4$. For λ_{norm}, to ensure robust convergence of the geometry, we start with a higher value of λ_{norm} during each stage and gradually decrease it using a two-round cosine annealing, where $\lambda_{\text{norm}}(t)$ is the weight of $\mathcal{L}_{\text{norm}}$ at the t-th iteration:

$$\lambda_{\text{norm}}(t) = \begin{cases} 0.5\lambda_{\text{norm}}^{\text{base}} \left(1 + \cos\left(\frac{t}{t_{\text{coarse}}}\pi\right)\right) & \text{if } t < t_{\text{coarse}} \\ 0.5\lambda_{\text{norm}}^{\text{base}} \left(1 + \cos\left(\frac{t - t_{\text{coarse}}}{t_{\text{fine}}}\pi\right)\right) & \text{if } t \geq t_{\text{coarse}} \end{cases}, \tag{C.3}$$

Texture-stage optimization. We optimize Ψ_{c} for $t_{\text{texture}} = 7000$ steps, with $\lambda_{\text{recon}} = 2 \times 10^4$ and $\lambda_{\text{SDS}} = 1$. Besides, we set $\lambda_{\text{CD}} = 0$ at the beginning of the training, and $\lambda_{\text{CD}} = 1 \times 10^6$ at the last $t_{\text{CD}} = 2000$ iterations to enforce color consistency.

C.6 More Qualitative Results

In addition to Fig. 5.7, we show more qualitative comparisons between TeCH and other baselines (PIFu [185], PIFuHD [186], PaMIR [256], PHORHUM [7], ICON (Chapter 3), ECON (Chapter 4)) on CAPE, THuman2.0, and SHHQ [50] images (Figs. C.1 to C.3 of Sup. Mat.), by visualizing multi-view surface normals, color renderings, and zoomed-in details. For subjects in CAPE and THuman2.0, TeCH precisely recover the human shape and generate high-quality details of garments and facial features, regardless of hard poses, complex texture, loose clothing, or self-occlusion. Furthermore, Fig. C.3 demonstrates the strong generalizability of TeCH on in-the-wild images, more rotating 3D humans are provided in videos on homepage *huangyangyi.github.io/TeCH*.

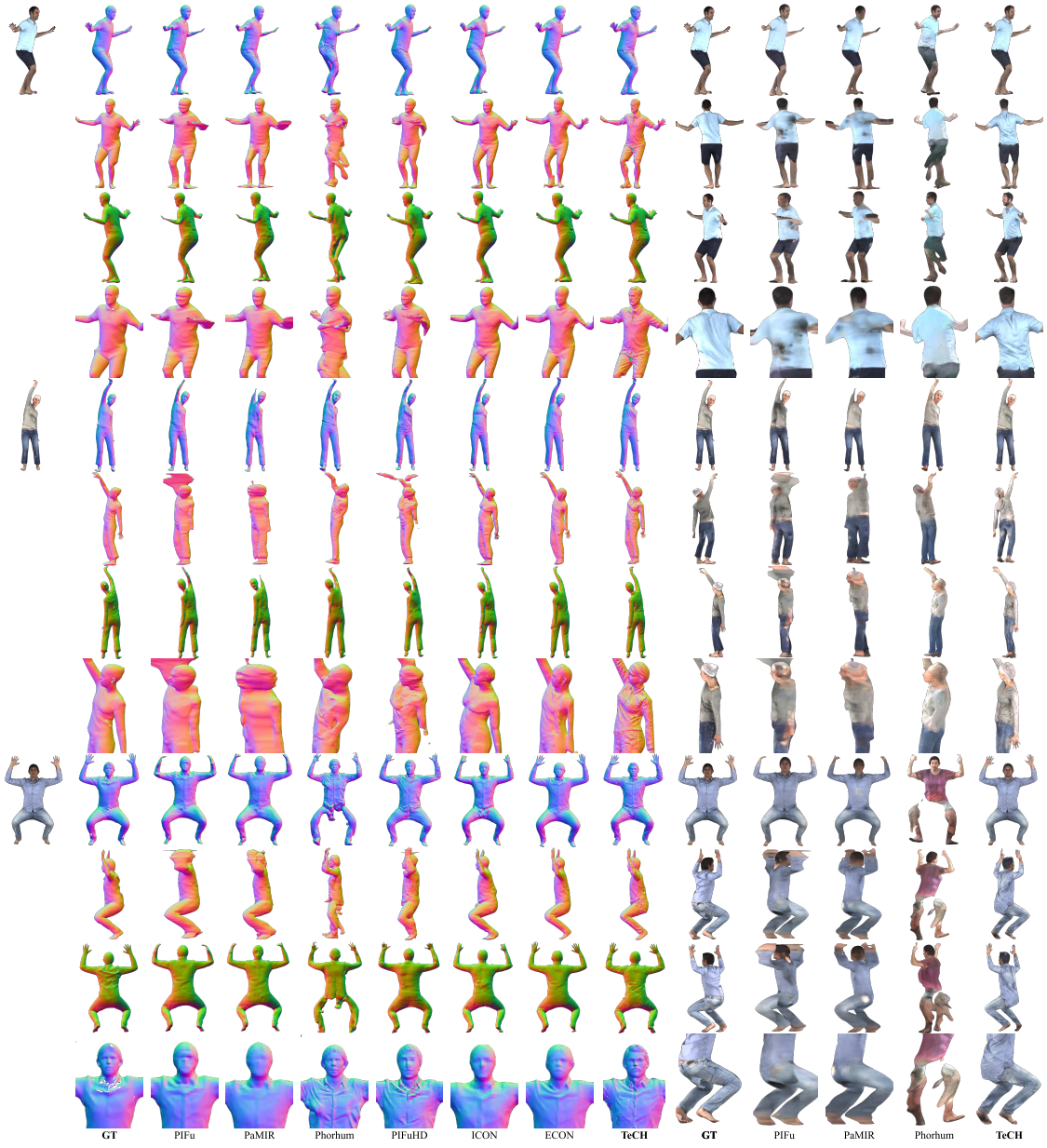


Figure C.1: Qualitative comparison on CAPE. TeCH performs better on subjects with challenging poses.

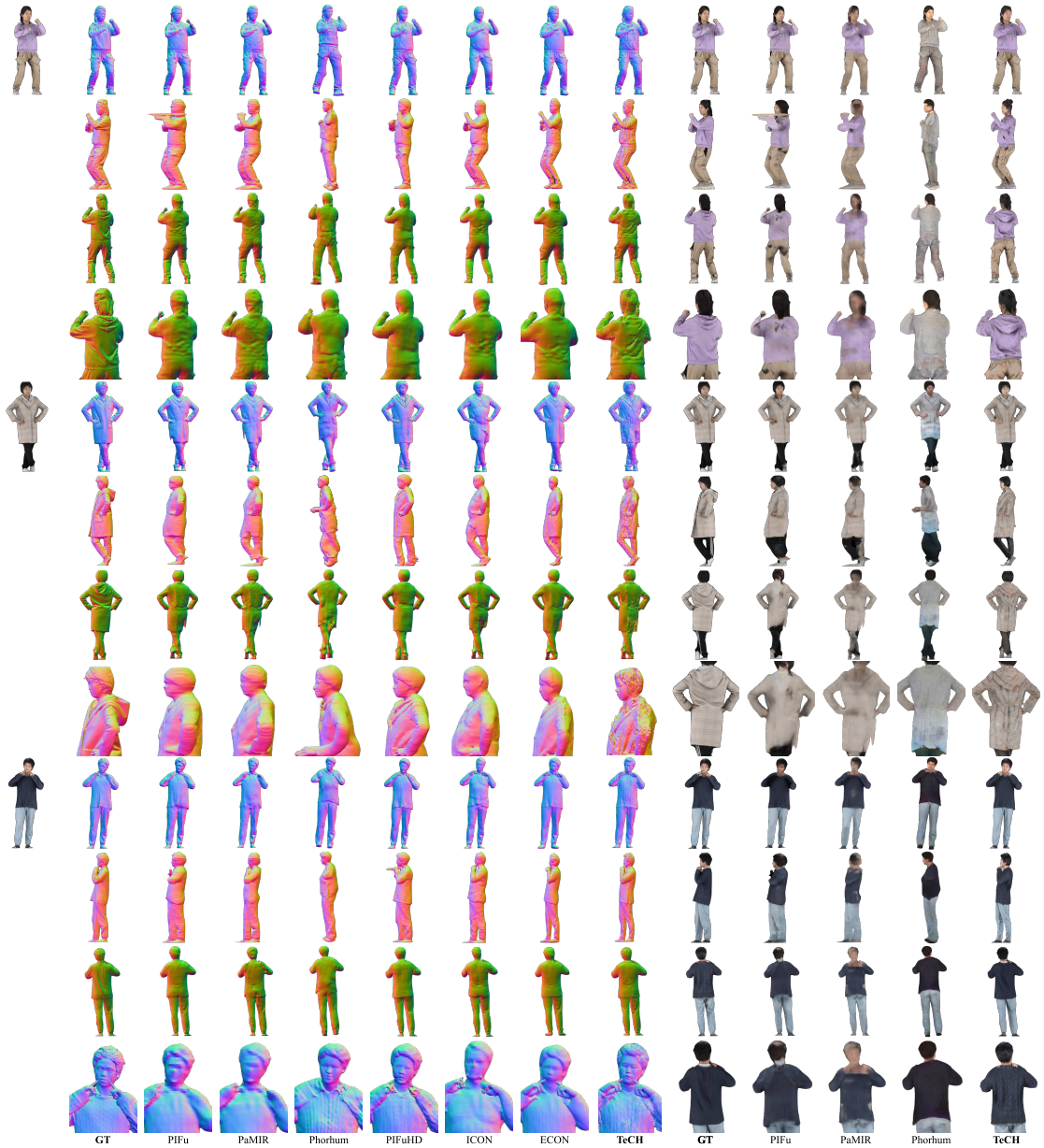


Figure C.2: Qualitative comparison on THuman2.0. TeCH performs better regardless of hard pose, complex texture, or loose clothing.

Figure C.3: Qualitative comparison on SHHQ images. TeCH generalizes well on in-the-wild images with diverse clothing styles and textures. It successfully recovers the overall structure of the clothed body with text guidance, and generates realistic full-body texture which is consistent with the colored pattern and the material of the clothes. 🔍 **Zoom in** to see the geometric details.

D

PUZZLEAVATAR: ASSEMBLING 3D AVATARS FROM PERSONAL ALBUMS

Contents

D.1 GPT-4V Prompt for PuzzleBooth

Queried Prompt: *"Analyze the provided images, each featuring an individual. Identify and describe the individual's gender, facial features (excluding hair), haircut, and specific clothing items such as shirts, hats, pants, shoes, dresses, skirts, scarves, etc. Return the results in a dictionary format with keys for "gender", "face", "haircut", and each type of clothing. The corresponding value should provide 1-3 adjective or noun words, which describe the topological or geometric features, such as length (*e.g.*, short, long, midi, mini, knee-length, floor length, ankle-length, hip-length, calf-length), shape (*e.g.*, oval, round, square, heart-shaped, diamond-shaped, rectangular, voluminous, razor-cut, tousled, layered, messy), tightness (*e.g.*, tight, snug, fitted, skin-tight, loose, tight-fitting, clingy), style (*e.g.*, modern, casual, sporty, classic, formal, vintage, bohemian, avant-garde), or haircut types (*e.g.*, long, short, wavy, straight, curly, bald, medium-length, pony tail, bun, plaits, beard, sideburns, dreadlocks, goatee), without referencing color or texture pattern. Exclude accessories and don't include any clothing item in the description of another. Omit any keys for which the clothing item does not appear or the description*

is empty. The response should be a dictionary only, without any additional sentences, explanations, or markdowns syntax (like json)"

Negative Prompt: *"unrealistic, blurry, low quality, out of focus, ugly, low contrast, dull, dark, low-resolution, gloomy, shadow, worst quality, jpeg artifacts, poorly drawn, dehydrated, noisy, poorly drawn, bad proportions, bad anatomy, bad lighting, bad composition, bad framing, fused fingers, noisy, many people, duplicate characters"*

D.2 Camera setting

To familiarize the diffusion model with the camera positions sampled during SDS optimization, we rendered the synthetic color-normal image pairs in the exact same manner as the SDS sampling strategy. This rendered data will be used in preserving synthetic human prior ($\mathcal{L}_{\text{prior}}$), while training the 2D generator G_{puzzle}.

To ensure complete coverage of the entire body and face, we sample virtual camera poses around the full body and zoom in on the face region. To reduce the occurrence of mirrored appearance artifacts (e.g., Janus-head), we incorporated view-aware prompts (*i.e.*, "`front/side/back/overhead view`"), regarding the viewing angle during the generation process. The effectiveness of this approach has been demonstrated in DreamFusion [169].

To ensure full coverage of the entire body and the human face, we sample virtual camera poses into two groups: 1) $\mathbf{K}_{\text{body}}$ cameras with a field of view (FOV) covering the full body or the main body parts, and 2) zoom-in cameras $\mathbf{K}_{\text{face}}$ focusing the face region.

The ratio $\mathcal{P}_{\text{body}}$ determines the probability of sampling $\mathbf{k} \in \mathbf{K}_{\text{body}}$, while the height h_{body}, radius r_{body}, elevation angle ϕ_{body}, and azimuth ranges θ_{body} are adjusted relative to the SMPL-X body scale. Empirically, we set $\mathcal{P}_{\text{body}} = 0.5$, $h_{\text{body}} = [-0.4, 0.4]$, $r_{\text{body}} = (0.7, 1.3)$, $\theta_{\text{body}} = [60°, 120°]$, $\phi_{\text{body}} = [0°, 360°]$, with the M_{body} proportionally scaled to a $[-0.5, 0.5]$ unit space.

To enhance facial details, we sample additional virtual cameras positioned around the face $\mathbf{k} \in \mathbf{K}_{\text{face}}$, together with the additional prompt "`face of`". With a probability of $\mathcal{P}_{\text{face}} = 1 - \mathcal{P}_{\text{body}} = 0.5$, the sampling parameters include the view target c_{face}, radius range r_{face}, rotation range θ_{face}, and azimuth range ϕ_{face}. Empirically, we set c_{face} to

the 3D position of SMPL-X head keypoint, $r_{\text{face}} = [0.3, 0.4]$, $\theta_{\text{face}} = [90°, 90°]$ and $\phi_{\text{face}} = [-90°, 90°]$.

Regarding the synthetic data, we use all the subjects (525 textured scans) in THuman2.0. For each subject, we render 8 full-body views and 8 head views, as shown in Fig. 6.4, and query their descriptive prompts via GPT-4V [158]. This gives us $525 \times 8 \times 2 = 8400$ color-normal pairs in total.

REFERENCES

[1] *3DPeople*. 3dpeople.com. 2018.

[2] Alakh Aggarwal, Jikai Wang, Steven Hogue, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. "Layered-Garment Net: Generating Multiple Implicit Garment Layers from a Single Image". In: *Asian Conference on Computer Vision (ACCV)*. 2022.

[3] Thiemo Alldieck, Marcus A. 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.

[4] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. "Detailed Human Avatars from Monocular Video". In: *International Conference on 3D Vision (3DV)*. 2018, pp. 98–109.

[5] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. "Video Based Reconstruction of 3D People Models". In: *Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 8387–8397.

[6] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus A. Magnor. "Tex2Shape: Detailed Full Human Body Geometry From a Single Image". In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2293–2303.

[7] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. "Photorealistic monocular 3d reconstruction of humans wearing clothing". In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[8] Rie Ando and Tong Zhang. "Learning on graph with Laplacian regularization". In: *Conference on Neural Information Processing Systems (NeurIPS)* (2006).

[9] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. "Break-A-Scene: Extracting Multiple Concepts from a Single Image". In: *SIGGRAPH Asia 2023 Conference Papers*. SA '23. 2023.

[10] AXYZ. secure.axyz-design.com. 2018.

[11] Gwangbin Bae and Andrew J. Davison. "Rethinking Inductive Biases for Surface Normal Estimation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[12] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. "eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers". In: *arXiv:2211.01324* (2022).

[13] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. "Generative neural articulated radiance fields". In: *Conference on Neural Information Processing Systems (NeurIPS)* (2022).

[14] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. “CLOTH3D: Clothed 3D Humans”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[15] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. “Multi-Garment Net: Learning to Dress 3D People From Images”. In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 5419–5429.

[16] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. “Nope-nerf: Optimising neural radiance field with no pose prior”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[17] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. “BEDLAM: A Synthetic Dataset of 3D Human Bodies Exhibiting Detailed Lifelike Animated Motion”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[18] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. 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)*. Vol. 9909. 2016, pp. 561–578.

[19] Aljaz Bozic, Pablo R. Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. “Neural Deformation Graphs for Globally-consistent Non-rigid Reconstruction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 1450–1459.

[20] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. “Coco-stuff: Thing and stuff classes in context”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2018.

[21] 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: *European Conference on Computer Vision (ECCV)*. 2022.

[22] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. “Bilateral Normal Integration”. In: *European Conference on Computer Vision (ECCV)*. 2022.

[23] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. “DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models”. In: *Computer Vision and Pattern Recognition (CVPR)* (2024).

[24] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. “GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models”. In: *International Conference on Computer Vision (ICCV)*. 2023.

[25] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. “Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021.

[26] Pradyumna Chari, Sizhuo Ma, Daniil Ostashev, Achuta Kadambi, Gurunandan Krishnan, Jian Wang, and Kfir Aberman. “Personalized Restoration via Dual-Pivot Tuning”. In: *arXiv preprint arXiv:2312.17234* (2023).

[27] Bowei Chen, Brian Curless, Ira Kemelmacher-Shlizerman, and Steve Seitz. “Total Selfie: Generating Full-Body Selfies”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[28] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. “Text2Tex: Text-driven Texture Synthesis via Diffusion Models”. In: *arXiv:2303.11396* (2023).

[29] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. “Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation”. In: *International Conference on Computer Vision (ICCV)*. 2023.

[30] 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.

[31] Zhiqin Chen and Hao Zhang. “Learning Implicit Fields for Generative Shape Modeling”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5939–5948.

[32] Wei Cheng, Ruixiang Chen, Wanqi Yin, Siming Fan, Keyu Chen, Honglin He, Huiwen Luo, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, and Kwan-Yee Lin. “DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering”. In: *International Conference on Computer Vision (ICCV)*. 2023.

[33] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. “Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020.

[34] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. “Neural Unsigned Distance Fields for Implicit Function Learning”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020.

[35] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. “Accurate 3D Body Shape Regression via Linguistic Attributes and Anthropometric Measurements”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[36] 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)*. 2020.

[37] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. “The cityscapes dataset”. In: *CVPR Workshop on the Future of Datasets in Vision*. sn. 2015.

[38] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. “Smplicit: Topology-aware generative model for clothed people”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021.

[39] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. “Structured 3D Features for Reconstructing Relightable and Animatable Avatars”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[40] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. “Objaverse: A universe of annotated 3d objects”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *Computer Vision and Pattern Recognition (CVPR)*. Ieee. 2009.

[42] Akio Doi and Akio Koide. “An efficient method of triangulating equi-valued surfaces by using tetrahedral cells”. In: *IEICE TRANSACTIONS on Information and Systems* 74.1 (1991), pp. 214–224.

[43] Junting Dong, Qi Fang, Yudong Guo, Sida Peng, Qing Shuai, Xiaowei Zhou, and Hujun Bao. “TotalSelfScan: Learning Full-body Avatars from Self-Portrait Videos of Faces, Hands, and Bodies”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[44] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. “AG3D: Learning to Generate 3D Avatars from 2D Image Collections”. In: *International Conference on Computer Vision (ICCV)*. 2023.

[45] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. “PINA: Learning a Personalized Implicit Neural Avatar from a Single RGB-D Video Sequence”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[46] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. “AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time”. In: *TPAMI* (2022).

[47] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. “Collaborative Regression of Expressive Bodies using Moderation”. In: *International Conference on 3D Vision (3DV)*. 2021, pp. 792–804.

[48] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. “Learning Disentangled Avatars with Hybrid 3D Representations”. In: *arXiv* (2023).

[49] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. “Capturing and Animation of Body and Clothing from Monocular Video”. In: *SIGGRAPH Asia 2022 Conference Papers*. SA ’22. 2022.

[50] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. “StyleGAN-Human: A Data-Centric Odyssey of Human Generation”. In: *European Conference on Computer Vision (ECCV)* (2022).

[51] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. “COLMAP-Free 3D Gaussian Splatting”. In: *Computer Vision and Pattern Recognition (CVPR)* (2024).

[52] Valentin Gabeur, Jean-Sébastien 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.

[53] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. "DART: Articulated Hand Model with Diverse Accessories and Rich Textures". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022.

[54] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. "GraphDreamer: Compositional 3D Scene Synthesis from Scene Graphs". In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[55] Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. "Learning deformable tetrahedral meshes for 3d reconstruction". In: *Conference on Neural Information Processing Systems (NeurIPS)* (2020).

[56] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. "GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[57] Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, and Long Quan. "ConTex-Human: Free-View Rendering of Human from a Single Image with Texture-Consistent Synthesis". In: *arXiv preprint arXiv:2311.17123* (2023).

[58] Yuying Ge, Ruimao Zhang, Lingyun Wu, Xiaogang Wang, Xiaoou Tang, and Ping Luo. "A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images". In: *Computer Vision and Pattern Recognition (CVPR)*. 2019.

[59] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[60] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. "Stylepeople: A generative model of fullbody human avatars". In: *Computer Vision and Pattern Recognition (CVPR)*. 2021.

[61] Artur Grigorev, Bernhard Thomaszewski, Michael J Black, and Otmar Hilliges. "HOOD: Hierarchical Graphs for Generalized Modelling of Clothing Dynamics". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[62] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. "Vid2Avatar: 3D Avatar Reconstruction from Videos in the Wild via Self-supervised Scene Decomposition". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[63] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. "DeepCap: Monocular Human Performance Capture Using Weak Supervision". In: *Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2020.

[64] Si Hang. "TetGen, a Delaunay-based quality tetrahedral mesh generator". In: *ACM Trans. Math. Softw* (2015).

[65] Chengan He, Xin Sun, Zhixin Shu, Fujun Luan, Sören Pirk, Jorge Alejandro Amador Herrera, Dominik L Michels, Tuanfeng Y Wang, Meng Zhang, Holly Rushmeier, and Yi Zhou. "Perm: A Parametric Representation for Multi-Style 3D Hair Modeling". In: *arXiv preprint arXiv:2407.19451* (2024).

[66] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn". In: *International Conference on Computer Vision (ICCV)*. 2017.

[67] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. "Geo-PIFu: Geometry and Pixel Aligned Implicit Functions for Single-view Human Reconstruction". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2020.

[68] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. "ARCH++: Animation-Ready Clothed Human Reconstruction Revisited". In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 11046–11056.

[69] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. *MagicMan: Generative Novel View Synthesis of Humans with 3D-Aware Diffusion and Iterative Refinement*. 2024. arXiv: 2408.14211 [cs.CV].

[70] Charlie Hewitt, Tadas Baltrušaitis, Erroll Wood, Lohit Petikam, Louis Florentin, and Hanz Cuevas Velasquez. "Procedural Humans for Computer Vision". In: *arXiv:2301.01161* (2023).

[71] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. "EVA3D: Compositional 3D Human Generation from 2D Image Collections". In: *International Conference on Learning Representations (ICLR)*. 2023.

[72] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. "Avatarclip: Zero-shot text-driven generation and animation of 3d avatars". In: *Transactions on Graphics (TOG)* (2022).

[73] Hugues Hoppe. "New quadric metric for simplifying meshes with appearance attributes". In: *Proceedings Visualization'99 (Cat. No. 99CB37067)*. IEEE. 1999.

[74] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models". In: *International Conference on Learning Representations (ICLR)* (2022).

[75] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. "SHERF: Generalizable Human NeRF from a Single Image". In: *International Conference on Computer Vision (ICCV)*. 2023.

[76] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. "HumanNorm: Learning normal diffusion model for high-quality and realistic 3d human generation". In: *Computer Vision and Pattern Recognition (CVPR)* (2024).

[77] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. "One-shot Implicit Animatable Avatars with Model-based Priors". In: *International Conference on Computer Vision (ICCV)*. 2023.

[78] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. "TeCH: Text-guided Reconstruction of Lifelike Clothed Humans". In: *International Conference on 3D Vision (3DV)*. 2024.

[79] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. "DreamWaltz: Make a Scene with Complex 3D Animatable Avatars". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[80] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. "ARCH: Animatable Reconstruction of Clothed Humans". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020.

[81] *HumanAlloy*. humanalloy.com. 2018.

[82] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. "HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion". In: *Transactions on Graphics (TOG)* (2023).

[83] Aaron S. Jackson, Chris Manafas, and Stefan Roth Georgios Tzimiropoulos. "3D Human Body Reconstruction from a Single Image via Volumetric Regression". In: *European Conference on Computer Vision Workshops (ECCVw)*. Vol. 11132. 2018, pp. 64–77.

[84] 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.

[85] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. "Zero-Shot Text-Guided Object Generation with Dream Fields". In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[86] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. "SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video". In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[87] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. "BCNet: Learning Body and Cloth Shape from a Single Image". In: *European Conference on Computer Vision (ECCV)*. 2020, pp. 18–35.

[88] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. "AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control". In: *International Conference on Computer Vision (ICCV)*. 2023.

[89] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *European Conference on Computer Vision (ECCV)*. Vol. 9906. 2016, pp. 694–711.

[90] 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.

[91] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. "HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation". In: *International Conference on Computer Vision (ICCV)*. 2023.

[92] 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.

[93] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of StyleGAN". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020.

[94] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. "Noise-free Score Distillation". In: *International Conference on Learning Representations (ICLR)*. 2024.

[95] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. "Poisson surface reconstruction". In: *Symposium on Geometry Processing (SGP)*. 2006.

[96] Michael Kazhdan and Hugues Hoppe. "Screened poisson surface reconstruction". In: *Transactions on Graphics (TOG)* (2013).

[97] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. "Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[98] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". In: *Transactions on Graphics (TOG)* 42.4 (2023). URL: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[99] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. "Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models". In: *International Conference on Computer Vision (ICCV)*. 2023.

[100] Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. "Beyond the Contact: Discovering Comprehensive Affordance for 3D Objects from Pre-trained 2D Diffusion Models". In: *European Conference on Computer Vision (ECCV)*. 2024.

[101] Taeksoo Kim, Byungjun Kim, Shunsuke Saito, and Hanbyul Joo. *GALA: Generating Animatable Layered Assets from a Single Scan*. 2024.

[102] Taeksoo Kim, Shunsuke Saito, and Hanbyul Joo. "NCHO: Unsupervised Learning for Neural 3D Composition of Humans and Objects". In: *International Conference on Computer Vision (ICCV)*. 2023.

[103] 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.

[104] 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.

[105] 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.

[106] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. "DreamHuman: Animatable 3D Avatars from Text". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[107] 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.

[108] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations". In: *International Journal of Computer Vision (IJCV)* (2017).

[109] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. "Multi-Concept Customization of Text-to-Image Diffusion". In: *Computer Vision and Pattern Recognition (CVPR)* (2023).

[110] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. "Modular Primitives for High-Performance Differentiable Rendering". In: *Transactions on Graphics (TOG)* (2020).

[111] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. "360-Degree Textures of People in Clothing from a Single Image". In: *International Conference on 3D Vision (3DV)*. 2019, pp. 643–653.

[112] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. "NIKI: Neural Inverse Kinematics with Invertible Neural Networks for 3D Human Pose and Shape Estimation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[113] 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.

[114] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In: *International Conference on Machine Learning (ICML)*. PMLR. 2022.

[115] Lei Li, Zhizheng Liu, Weining Ren, Liudi Yang, Fangjinhua Wang, Marc Pollefeys, and Songyou Peng. "3D Textured Shape Recovery with Learned Geometric Priors". In: *arXiv:2209.03254* (2022).

[116] Peizhuo Li, Kfir Aberman, Rana Hanocka, Libin Liu, Olga Sorkine-Hornung, and Baoquan Chen. "Learning Skeletal Articulations with Neural Blend Shapes". In: *Transactions on Graphics (TOG)* (2021).

[117] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. "PSHuman: Photorealistic Single-view Human Reconstruction using Cross-Scale Diffusion". In: *arXiv preprint arXiv:2409.10141* (2024).

[118] Ren Li, Corentin Dumery, Benoit Guillard, and Pascal Fua. “Garment Recovery with Shape and Deformation Priors”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[119] Ren Li, Benoit Guillard, Edoardo Remelli, and Pascal Fua. “DIG: Draping Implicit Garment over the Human Body”. In: *Asian Conference on Computer Vision (ACCV)*. 2022.

[120] Ren Li, Benoit Guillard, Edoardo Remelli, and Pascal Fua. “ISP: Multi-Layered Garment Draping with Implicit Sewing Patterns”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[121] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. “Volumetric human teleportation”. In: *ACM SIGGRAPH 2020 Real-Time Live*. 2020.

[122] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. “Monocular real-time volumetric performance capture”. In: *European Conference on Computer Vision (ECCV)*. 2020.

[123] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. “Ai choreographer: Music conditioned 3d dance generation with aist++”. In: *International Conference on Computer Vision (ICCV)*. 2021.

[124] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. “Robust 3D Self-portraits in Seconds”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 1341–1350.

[125] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. “PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[126] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. “CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation”. In: *European Conference on Computer Vision (ECCV)*. Springer. 2022.

[127] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. “Deep Human Parsing with Active Template Regression”. In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2015).

[128] 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.

[129] Jiang Liangze and Tao Lin. “Test-Time Robust Personalization for Federated Learning”. In: *International Conference on Learning Representations (ICLR)*. 2023.

[130] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. “TADA! Text to Animatable Digital Avatars”. In: *International Conference on 3D Vision (3DV)*. 2024.

[131] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. “High-Fidelity Clothed Avatar Reconstruction from a Single Image”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[132] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. "Magic3D: High-Resolution Text-to-3D Content Creation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[133] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. "Barf: Bundle-adjusting neural radiance fields". In: *International Conference on Computer Vision (ICCV)*. 2021.

[134] 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.

[135] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund T, Zexiang Xu, and Hao Su. "One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization". In: *Conference on Neural Information Processing Systems (NeurIPS)* (2023).

[136] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. "Zero-1-to-3: Zero-shot One Image to 3D Object". In: *International Conference on Computer Vision (ICCV)*. 2023.

[137] Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. "Parameter-Efficient Orthogonal Finetuning via Butterfly Factorization". In: *International Conference on Learning Representations (ICLR)* (2024).

[138] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J. Black, and Bernhard Schölkopf. "Ghost on The Shell: An Expressive Representation of General 3D Shapes". In: *International Conference on Learning Representations (ICLR)* (2024).

[139] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations". In: *Computer Vision and Pattern Recognition (CVPR)*. 2016.

[140] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "SMPL: A Skinned Multi-Person Linear Model". In: *Transactions on Graphics (TOG)* 34.6 (2015), 248:1–248:16.

[141] William E. Lorensen and Harvey E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm". In: *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* 21.4 (1987), pp. 163–169.

[142] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. "Mediapipe: A framework for building perception pipelines". In: *arXiv:1906.08172* (2019).

[143] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. *Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference*. 2023. arXiv: 2310.04378.

[144] 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.

[145] Meysam Madadi, Hugo Bertiche, Wafa Bouzouita, Isabelle Guyon, and Sergio Escalera. “Learning Cloth Dynamics: 3D + Texture Garment Reconstruction Benchmark”. In: *Proceedings of the NeurIPS 2020 Competition and Demonstration Track, PMLR*. 2021.

[146] 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.

[147] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. “Nerf in the wild: Neural radiance fields for unconstrained photo collections”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021.

[148] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. “Occupancy Networks: Learning 3D Reconstruction in Function Space”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4460–4470.

[149] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. “Progressively optimized local radiance fields for robust view synthesis”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[150] 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.

[151] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. “3D Clothed Human Reconstruction in the Wild”. In: *European Conference on Computer Vision (ECCV)*. 2022.

[152] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. “T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models”. In: *Conference on Artificial Intelligence (AAAI)* (2023).

[153] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. “Instant neural graphics primitives with a multiresolution hash encoding”. In: *ACM Transactions on Graphics (ToG)* (2022).

[154] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked Hourglass Networks for Human Pose Estimation”. In: *European Conference on Computer Vision (ECCV)*. Vol. 9912. 2016, pp. 483–499.

[155] Edwin G. Ng, Bo Pang, Piyush Kumar Sharma, and Radu Soricut. “Understanding Guided Image Captioning Performance across Domains”. In: *Conference on Computational Natural Language Learning*. 2020.

[156] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. “Unsupervised learning of efficient geometry-aware neural articulated representations”. In: *European Conference on Computer Vision (ECCV)*. Springer. 2022.

[157] Hayato Onizuka, Zehra Haiyrci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-Ichiro Taniguchi. "TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell". In: *Computer Vision and Pattern Recognition (CVPR)*. 2020.

[158] OpenAI. *GPT-4V(ision) system card*. 2023.

[159] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. "Im2text: Describing images using 1 million captioned photographs". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2011.

[160] 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.

[161] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela Dai. "SPAMs: Structured Implicit Parametric Models". In: *Computer Vision and Pattern Recognition (CVPR)* (2022).

[162] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 165–174.

[163] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: *British Machine Vision Conference (BMVC)*. 2015.

[164] 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.

[165] 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.

[166] Sida Peng, Chen Geng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. "Implicit Neural Representations with Structured Latent Codes for Human Body Modeling". In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023).

[167] PIFuHD code on GitHub. https://github.com/facebookresearch/pifuhd. 2020.

[168] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. "ClothCap: seamless 4D clothing capture and retargeting". In: *Transactions on Graphics (TOG)* 36.4 (2017), 73:1–73:15.

[169] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. "DreamFusion: Text-to-3d using 2d diffusion". In: *International Conference on Learning Representations (ICLR)*. 2023.

[170] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. "Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors". In: *International Conference on Learning Representations (ICLR)*. 2024.

[171] Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. "REC-MV: REconstructing 3D Dynamic Cloth from Monucular Videos". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[172] Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. "Controlling Text-to-Image Diffusion by Orthogonal Finetuning". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[173] Yvain Quéau, Jean-Denis Durou, and Jean-François Aujol. "Normal integration: A survey". In: *Journal of Mathematical Imaging and Vision* (2018).

[174] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *International Conference on Machine Learning (ICML)*. PMLR, 2021.

[175] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-Shot Text-to-Image Generation". In: *International Conference on Machine Learning (ICML)*. 2021.

[176] 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).

[177] Rembg: A tool to remove images background. https://github.com/danielgatis/rembg. 2022.

[178] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. *Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks*. 2024. arXiv: 2401.14159 [cs.CV].

[179] *RenderPeople*. renderpeople.com. 2018.

[180] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. "TEXTure: Text-Guided Texturing of 3D Shapes". In: *arXiv:2302.01721* (2023).

[181] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[182] 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.

[183] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[184] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[185] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. "PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization". In: *International Conference on Computer Vision (ICCV)*. 2019, pp. 2304–2314.

[186] Shunsuke Saito, Tomas Simon, Jason M. 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.

[187] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. "SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks". In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 2886–2897.

[188] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. "LAION-5B: An open large-scale dataset for training next generation image-text models". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022.

[189] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. "X-Avatar: Expressive Human Avatars". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[190] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis". In: *Conference on Neural Information Processing Systems (NeurIPS)* (2021).

[191] Tianchang Shen, Jacob Munkberg, Jon Hasselgren, Kangxue Yin, Zian Wang, Wenzheng Chen, Zan Gojcic, Sanja Fidler, Nicholas Sharp, and Jun Gao. "Flexible Isosurface Extraction for Gradient-Based Mesh Optimization". In: *Transactions on Graphics (TOG)* 42.4 (2023). URL: https://doi.org/10.1145/3592430.

[192] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. "MVDream: Multi-view Diffusion for 3D Generation". In: *International Conference on Learning Representations (ICLR)* (2024).

[193] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. "Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction". In: *International Conference on Computer Vision (ICCV)*. 2023.

[194] 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.

[195] Sanghyun Son, Matheus Gadelha, Yang Zhou, Zexiang Xu, Ming C. Lin, and Yi Zhou. *DMesh: A Differentiable Representation for General Meshes*. 2024. arXiv: 2404.13445 [cs.CV].

[196] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. "Consistency models". In: *International Conference on Machine Learning (ICML)*. 2023.

[197] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. "Neural 3D Reconstruction in the Wild". In: *SIGGRAPH Conference Proceedings*. 2022.

[198] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. "Monocular, One-stage, Regression of Multiple 3D People". In: *International Conference on Computer Vision (ICCV)*. 2021.

[199] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. "TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[200] 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.

[201] Jiang Suyi, Jiang Haoran, Wang Ziyu, Luo Haimin, Chen Wenzheng, and Xu Lan. "HumanGen: Generating Human Radiance Fields with Explicit Priors". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[202] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lemptisky. "DINAR: Diffusion Inpainting of Neural Textures for One-Shot Human Avatars". In: *International Conference on Computer Vision (ICCV)*. 2023.

[203] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. "LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation". In: *arXiv preprint arXiv:2402.05054* (2024).

[204] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. "A Neural Network for Detailed Human Depth Estimation From a Single Image". In: *International Conference on Computer Vision (ICCV)*. 2019.

[205] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. "ConsiStory: Training-Free Consistent Text-to-Image Generation". In: *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. 2024.

[206] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. "3D Human Pose Estimation via Intuitive Physics". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[207] Twindom. twindom.com. 2018.

[208] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. "Dynamic shape capture using multi-view photometric stereo". In: *ACM SIGGRAPH Asia 2009 Papers*. 2009.

[209] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[210] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. "Disentangled Clothed Avatar Generation from Text Descriptions". In: *arXiv preprint arXiv:2312.05295* (2023).

[211] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. "PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction". In: *International Conference on Learning Representations (ICLR)*. 2024.

[212] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. "InstantID: Zero-shot Identity-Preserving Generation in Seconds". In: *arXiv preprint arXiv:2401.07519* (2024).

[213] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Revaud Jerome. "DUSt3R: Geometric 3D Vision Made Easy". In: *Computer Vision and Pattern Recognition (CVPR)* (2024).

[214] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. "Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[215] Xintao Wang, Liangbin Xie, Ke Yu, Kelvin C.K. Chan, Chen Change Loy, and Chao Dong. *BasicSR: Open Source Image and Video Restoration Toolbox*. 2022.

[216] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. "Image inpainting via generative multi-column convolutional neural networks". In: *Conference on Neural Information Processing Systems (NeurIPS)* (2018).

[217] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. "NeRF–: Neural radiance fields without known camera parameters". In: *arXiv preprint arXiv:2102.07064* (2021).

[218] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. "Novel View Synthesis with Diffusion Models (3DiM)". In: *International Conference on Learning Representations (ICLR)*. 2023.

[219] Andrew Weitz, Lina Colucci, Sidney Primas, and Brinnae Bent. "InfiniteForm: A synthetic, minimal bias dataset for fitness applications". In: *arXiv:2110.01330* (2021).

[220] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. "ReconFusion: 3D Reconstruction with Diffusion Priors". In: 2024.

[221] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. “Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction”. In: *British Machine Vision Conference (BMVC)*. 2022.

[222] 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.

[223] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. “MonoClothCap: Towards Temporally Coherent Clothing Capture from Monocular RGB Video”. In: *International Conference on 3D Vision (3DV)*. 2020, pp. 322–332.

[224] Yongqin Xiang, Julian Chibane, Bharat Lal Bhatnagar, Bernt Schiele, Zeynep Akata, and Gerard Pons-Moll. “Any-Shot GIN: Generalizing Implicit Networks for Reconstructing Novel Classes”. In: *International Conference on 3D Vision (3DV)*. 2022.

[225] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021.

[226] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, and Xiaoguang Han. “Get3DHuman: Lifting StyleGAN-Human into a 3D Generative Model using Pixel-aligned Reconstruction Priors”. In: *International Conference on Computer Vision (ICCV)*. 2023.

[227] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. “MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[228] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. “ECON: Explicit Clothed humans Optimized via Normal integration”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[229] 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.

[230] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. “PuzzleAvatar: Assembling 3D Avatars from Personal Albums”. In: *Transactions on Graphics (TOG)* (2024).

[231] 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. 6183–6192.

[232] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Landreth, and Karan Singh. “RigNet: Neural Rigging for Articulated Characters”. In: *Transactions on Graphics (TOG)* (2020).

[233] Yuxuan Xue, Bharat Lal Bhatnagar, Riccardo Marin, Nikolaos Sarafianos, Yuanlu Xu, Gerard Pons-Moll, and Tony Tung. "NSF: Neural Surface Fields for Human Modeling from Monocular Depth". In: *International Conference on Computer Vision (ICCV)*. 2023.

[234] Xihe Yang, Xingyu Chen, Daiheng Gao, Shaohui Wang, Xiaoguang Han, and Baoyuan Wang. "HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images". In: *Computer Vision and Pattern Recognition (CVPR)*. 2024.

[235] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. "D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field". In: *International Conference on Computer Vision (ICCV)*. 2023.

[236] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. "S3: Neural Shape, Skeleton, and Skinning Fields for 3D Human Modeling". In: *Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 13284–13293.

[237] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. "StableNormal: Reducing Diffusion Variance for Stable and Sharp Normal". In: *Transactions on Graphics (TOG)* (2024).

[238] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. "Decoupling Human and Camera Motion from Videos in the Wild". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[239] 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.

[240] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. In: *Computer Vision and Pattern Recognition (CVPR)*. 2021.

[241] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. "GAvatar: Animatable 3D Gaussian Avatars with Implicit Mesh Learning". In: *arXiv preprint arXiv:2312.11461* (2023).

[242] Zheng Yujian, Jin Zirong, Li Moran, Huang Haibin, Ma Chongyang, Cui Shuguang, and Han Xiaoguang. "HairStep: Transfer Synthetic to Real Using Strand and Depth Maps for Single-View 3D Hair Modeling". In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[243] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. "Point-based modeling of human clothing". In: *International Conference on Computer Vision (ICCV)*. 2021.

[244] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. "AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation". In: *arXiv:2306.09864* (2023).

[245] Chao Zhang, Sergi Pujades, Michael J. 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.

[246] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J. Black. "TECA: Text-Guided Generation and Editing of Compositional 3D Avatars". In: *International Conference on 3D Vision (3DV)*. 2024.

[247] 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: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023).

[248] 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.

[249] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. "AvatarVerse: High-quality & Stable 3D Avatar Creation from Text and Pose". In: *Conference on Artificial Intelligence (AAAI)* (2023).

[250] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. "Cameras as Rays: Pose Estimation via Ray Diffusion". In: *International Conference on Learning Representations (ICLR)*. 2024.

[251] 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)*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Lecture Notes in Computer Science. Springer International Publishing, 2020.

[252] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. "AvatarGen: A 3D Generative Model for Animatable Human Avatars". In: *European Conference on Computer Vision Workshops (ECCVw)*. 2022.

[253] Jingbo Zhang, Xiaoyu Li, Qi Zhang, Yanpei Cao, Ying Shan, and Jing Liao. "HumanRef: Single Image to 3D Human Generation via Reference-Guided Diffusion". In: *arXiv preprint arXiv:2311.16961* (2023).

[254] Lvmin Zhang and Maneesh Agrawala. "Adding Conditional Control to Text-to-Image Diffusion Models". In: *International Conference on Computer Vision (ICCV)*. 2023.

[255] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. "DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras". In: *International Conference on Computer Vision (ICCV)*. 2021, pp. 6239–6249.

[256] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. "PaMIR: Parametric Model-conditioned Implicit Representation for image-based human reconstruction". In: *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2021).

[257] 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.

[258] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. “Random erasing data augmentation”. In: *Conference on Artificial Intelligence (AAAI)*. 2020.

[259] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. “Scene parsing through ade20k dataset”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2017.

[260] Zhizhuo Zhou and Shubham Tulsiani. “SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2023.

[261] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. “Detailed Human Shape Estimation From a Single Image by Hierarchical Mesh Deformation”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4491–4500.

[262] Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. “Registering explicit to implicit: Towards high-fidelity garment mesh reconstruction from single images”. In: *Computer Vision and Pattern Recognition (CVPR)*. 2022.

[263] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. “Drivable 3D Gaussian Avatars”. In: *arXiv preprint arXiv:2311.08581* (2023).

[264] 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.