Reconstructing 3D Human Avatars from Monocular Images

Von der Carl-Friedrich-Gauß-Fakultät
der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktoringenieurs (Dr.-Ing.)

genehmigte Dissertation
(kumulative Arbeit)

von

Thiemo Alldieck
geboren am 05.03.1989
in Aachen

Eingereicht am:	28.08.2019
Disputation am:	13.03.2020
1. Referent:	Prof. Dr.-Ing. Marcus Magnor
2. Referent:	Dr.-Ing. Gerard Pons-Moll

(2020)

To Philine

Abstract

Modeling 3D virtual humans has been an active field of research over the last decades. It plays a fundamental role in many applications, such as movie production, sports and medical sciences, or human-computer interaction. Early works focus on artist-driven modeling or utilize expensive scanning equipment. In contrast, our goal is the fully automatic acquisition of personalized avatars using low-cost monocular video cameras only. In this dissertation, we show fundamental advances in 3D human reconstruction from monocular images. We solve this challenging task by developing methods that effectively fuse information from multiple points in time and realistically complete reconstructions from sparse observations. Given a video or only a single photograph of a person in motion, we reconstruct, for the first time, not only his or her 3D pose but the full 3D shape including the face, hair, and clothing.

In a first scenario, we estimate 3D human poses from unconstrained video. Hereby, we leverage optical flow to enforce fluent and time-consistent motion. While a body model helps to make the problem tractable, it so far lacks hair, clothing, and personal details. In subsequent work, we reconstruct these properties, for the first time, from videos of people turning around in a so-called A-pose. Our method generalizes visual hull reconstruction to articulated motion by merging silhouette information in a canonical representation. We additionally estimate surface colors by stitching full textures from re-projected frames. Our novel semantic prior helps greatly improving the visual fidelity of the final textures. Further, we enhance surface details via multi-frame shape-from-shading. In the following work, we significantly reduce the required input for high-quality reconstruction to only about eight frames. Additionally, we speed up the reconstruction time by several magnitudes. We achieve this by combining the advantages of bottom-up Deep Learning and weakly supervised top-down optimization at test time. In the final scenario, we again simplify and accelerate the reconstruction process and further increase the level of detail in the results. We open up the input to photos of people in various camera-facing poses and enable 3D reconstruction based on only a single photograph. The key insight of this work is that single-image 3D human reconstruction can be performed by transforming 3D reconstruction into pose-independent 2D image-to-image translation in UV-space. The reconstructed results feature, for the first time, fine details like garment wrinkles, even on parts that are occluded in the input image.

In this dissertation, we explore various approaches to monocular image and video-based 3D human reconstruction. We demonstrate both straight-forward and sophisticated reconstruction methods focused on accuracy, simplicity, usability, and visual fidelity. During extensive evaluations, we give insights into important parameters, reconstruction quality, and the robustness of the methods. For the first time, our methods enable camera-based, easy-to-use self-digitization for exciting new applications like, for example, telepresence or virtual try-on for online fashion shopping.

Kurzfassung

Die 3D-Modellierung virtueller Menschen ist seit einigen Jahrzehnten Gegenstand aktiver Forschung. Sie spielt für verschiedenste Anwendungen, wie zum Beispiel in der Filmproduktion, in Sport- und Medizinwissenschaften oder bei Mensch-Computer-Interaktion eine entscheidende Rolle. Viele Arbeiten setzen auf von Designern erschaffene 3D-Modelle oder auf die Verwendung von teuren 3D-Scannern. Im Gegensatz dazu ist das Ziel dieser Arbeit, die ausschließliche Verwendung von kostengünstigen Videokameras. In dieser Dissertation zeigen wir fundamentale Entwicklungen in der 3D Rekonstruktion von Menschen aus monokularen Bilddaten. Wir lösen dieses anspruchsvolle Problem, indem wir Methoden entwickeln, die Informationen aus mehreren Zeitpunkten effektiv zusammenführen und Rekonstruktionen aus wenigen Beobachtungen realistisch vervollständigen. Aus monokularen Videos oder sogar nur einem einzelnen Bild einer Person in Bewegung, rekonstruieren wir erstmalig nicht nur dessen 3D Pose, sondern auch die 3D Körperform inklusive des Gesichtes, Haaren und Kleidung.

Zunächst beschreiben wir ein Verfahren, das aus regulären Videos menschliche 3D Posen schätzt. Durch Zuhilfenahme von optischem Fluss erzeugen wir flüssige und zeitkonsistente Bewegung. Ein statistisches Modell des menschlichen Körpers hilft hierbei das Problem besser zu beschreiben. Dieses verfügt bisher aber weder über Haare noch Kleidung noch persönliche Details der Person. Wir rekonstruieren diese Eigenschaften in einem weiteren Verfahren erstmalig aus Videos von Personen, die sich vor der Kamera drehen und eine sogenannte A-Pose einnehmen. Unsere Methode erweitert die „Visual hull“-Rekonstruktion für bewegte Objekte durch die Kombination von Silhouetteninformationen in einer kanonischen Darstellung. Zusätzlich schätzen wir das Erscheinungsbild durch Zusammenfügen einer Textur aus in den Texturraum projizierten Einzelbildern. Hierbei verbessert die Zuhilfenahme von semantischen Informationen die Qualität erheblich. Weiter verbessern wir die Oberfläche durch „Shape-from-shading“ basierend auf einer Vielzahl von Einzelbildern. Im nachfolgenden Verfahren reduzieren wir die benötigte Eingabe für hoch-qualitative Rekonstruktionen auf nur etwa acht Einzelbilder. Zusätzlich beschleunigen wir die Rekonstruktion um mehrere Größenordnungen. Dies wird durch die Kombination der Vorteile von „bottom-up“ Deep Learning und

Bild-basierter „top-down“ Optimierung zur Ausführungszeit erreicht. In einem letzten Verfahren vereinfachen und beschleunigen wir nochmals den Rekonstruktionsprozess und erhöhen noch einmal den Detailgrad der Ergebnisse. Wir erlauben beliebige Bilder von der Kamera zugewandten Personen als Eingabe und ermöglichen die 3D Rekonstruktionen aus nur einem einzelnen Foto. Die wichtigste Erkenntnis dieses Verfahrens ist, dass 3D Rekonstruktion von Personen durch Repräsentation der 3D Rekonstruktion als posenunabhängiges 2D Bildumwandlungsverfahren im Texturraum erreicht werden kann. Die rekonstruierten Ergebnisse enthalten erstmalig feine Strukturen, wie etwa Faltenwurf in der Kleidung, selbst auf Körperteilen, die der Kamera abgewandt waren.

In dieser Dissertation untersuchen wir verschiedenste Ansätze der 3D Rekonstruktion von Menschen aus monokularen Bilddaten. Wir beschreiben sowohl unkomplizierte als auch komplexere Methoden, die auf hohe Genauigkeit, Einfachheit, Nutzbarkeit oder Darstellungsqualität setzen. In umfangreichen Auswertungen untersuchen wir wichtige Parameter, die Qualität der Rekonstruktionen und die Robustheit der Methoden. Unsere Methoden erlauben erstmalig die kamerabasierte und benutzerfreundliche Digitalisierung von Menschen für spannende neue Anwendungsgebiete, wie etwa die Telepräsenz oder virtuelle Anprobe im Onlineshopping.

Acknowledgements

I want to first thank my supervisor Marcus Magnor for his steady support and for giving me the opportunity to challenge myself in an interesting and emerging research direction. He has always been trusting me and has given me great freedom in exploring the field and pursuing my own ideas. I further thank Gerard Pons-Moll for his guidance, many discussions, honest feedback, and tireless support. I am more than thankful for my inspiring stay at the Max Planck Institute for Informatics he made possible and for the ongoing close cooperation. Gerard truly made me an equal member of his group and has been an inspiring mentor. Another great thank you goes to Christian Theobalt for many interesting discussions and great support during and after my stay in Saarbrücken.

I have worked with and met many inspiring people during the last few years. Thank you to all of them. Thank you to my colleagues in Braunschweig for all the support I received and for creating such an enjoyable atmosphere. Special thanks go to JP Tauscher, Matthias Überheide, and Marc Kassubeck for numerous technical discussions and strong moral support. Thank you also to my colleagues in Saarbrücken for including me in the group, especially to Bharat Lal Bhatnagar and Verica Lazova for the fruitful collaboration.

Thank you to my family and their unlimited support. And finally, thank you to my wife Philine! You gave steady moral support during exciting and intensive years. I could always rely on you and you truly have made this dissertation possible!

Funding for this dissertation was generously provided through German Science Foundation (DFG) – project MA2555/12-1 “Umfassende digitale Modellierung von Darstellern anhand konventioneller Filmaufnahmen” / “Comprehensive Human Performance Capture from Monocular Video Footage”.

List of Publications

The following peer-reviewed and published scientific papers are part of this cumulative dissertation:

A

Thiemo Alldieck, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor:
Optical flow-based 3D human motion estimation from monocular video
In *German Conference on Pattern Recognition*. Springer, 2017, pp. 347–360.

Presented at the 39th German Conference on Pattern Recognition (GCPR), September 13-15, 2017, Basel, Switzerland.

B

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll:
Video Based Reconstruction of 3D People Models
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8387-8397.

Presented at the Conference on Computer Vision and Pattern Recognition (CVPR), June 18-23, 2018, Salt Lake City, UT, USA.

C

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll:
Detailed Human Avatars from Monocular Video
In *International Conference on 3D Vision*. IEEE, 2018, pp. 98-109.

Presented at the 6th International Conference on 3D Vision (3DV), September 5-8, 2018, Verona, Italy.

D

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll:
Learning to Reconstruct People in Clothing from a Single RGB Camera
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 1175-1186.

Presented at the Conference on Computer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long Beach, CA, USA.

E

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, Marcus Magnor:
Tex2Shape: Detailed Full Human Body Geometry from a Single Image
In *IEEE/CVF International Conference on Computer Vision*. IEEE, 2019, pp. 2293-2303.

Presented at the International Conference on Computer Vision (ICCV), October 27 - November 02, 2019, Seoul, Korea.

Thiemo Alldieck has participated in all publications as the first author. In close cooperation with the co-authors and under the supervision of the advisors, he has conducted the experiments, has written the manuscripts and has presented the work at the mentioned conferences and in additional talks.

The co-authors further contributed in accordance with the following: Bastian Wandt provided results for the comparison with his work (in Paper A). Gerard Pons-Moll provided scripts for numerical evaluation (in Paper B and Paper C). Bharat Lal Bhatnagar helped with data-processing and network training (in Paper D). Furthermore, Rudolf Martin, Juan Mateo Castrillon Cuervo, and Verica Lazova helped with data-processing and collection (in Paper B and Paper D). Some comparative results have been shared by the original authors.

All manuscripts are included in this thesis as they have been published. The format has been adjusted to match the layout of the dissertation and small editorial changes have been made. Corrections are highlighted in form of per chapter errata.

Contents

1 Introduction 1
1.1 Problem Statement . . . 2
1.2 Motivation . . . 4
1.3 Challenges . . . 6
1.4 Contributions . . . 8
1.5 Outline . . . 12

2 Related Work 13
2.1 Body Models based on Geometric Primitives . . . 13
2.2 Artist-Driven and Anatomical Models . . . 14
2.3 Data-Driven Body Models and Applications . . . 14
2.4 Free-form and Template-based Surface Reconstruction . . . 17
2.5 Deep Learning-based Human Reconstruction . . . 18

3 Methodology 21
3.1 Body Model . . . 21
3.2 Analysis-by-Synthesis . . . 23
3.2.1 Image keypoints . . . 24
3.2.2 Optical flow . . . 24
3.2.3 Image segmentation . . . 25
3.2.4 Shape-from-shading . . . 28
3.3 Deep Learning . . . 30

4 Publications 35

A Optical Flow-based 3D Human Motion Estimation from Monocular Video 37
A.1 Introduction . . . 38
A.2 Related Work . . . 40
A.3 Method . . . 41
A.3.1 Scene model . . . 42
A.3.2 Flow renderer . . . 42

A.3.3 Flow matching 43
A.3.4 Pose constraints 43
A.3.5 Silhouette coverage 44
A.3.6 Initialization 45
A.3.7 Optimization 45
A.4 Evaluation 46
A.5 Conclusions 49

B Video Based Reconstruction of 3D People Models 53
B.1 Introduction 54
B.2 Related Work 56
B.3 Method 59
B.3.1 SMPL body model with offsets 59
B.3.2 Pose reconstruction 60
B.3.3 Consensus shape 61
B.3.4 Frame refinement and texture generation 64
B.4 Experiments 65
B.4.1 Results on rendered images 65
B.4.2 Qualitative results on RGB images 67
B.4.3 Comparison with KinectCap 69
B.4.4 Surface refinement using shading 69
B.5 Discussion and Conclusion 71
B.6 Appendix: Additional Results and Implementation Details 73
B.6.1 Implementation details 73
B.6.2 Scale ambiguity 74
B.6.3 Comparison with the depth camera based approach [25] . 74
B.6.4 More results 74

C Detailed Human Avatars from Monocular Video 81
C.1 Introduction 82
C.2 Related Work 84
C.3 Method 86
C.3.1 Subdivided SMPL body model 86
C.3.2 Medium-level body shape reconstruction 88
C.3.3 Modeling fine-level surface details 89
C.3.4 Texture generation 93
C.4 Experiments 95
C.4.1 Qualitative results and comparisons 96
C.4.2 Face similarity measure 96

C.4.3 Ablation analysis . . . 97
C.4.4 User study . . . 98
C.5 Discussion and Conclusion . . . 101

D Learning to Reconstruct People in Clothing from a Single RGB Camera 103
D.1 Introduction . . . 104
D.2 Related Work . . . 106
D.3 Method . . . 108
D.3.1 Shape representation . . . 108
D.3.2 Model and data representation . . . 110
D.3.3 Loss functions . . . 111
D.3.4 Instance-specific top-down optimization . . . 113
D.4 Experiments . . . 113
D.4.1 Experimental setup . . . 114
D.4.2 Numerical evaluation . . . 117
D.4.3 Analysis of key parameters . . . 119
D.4.4 Type of supervision . . . 119
D.4.5 Qualitative results and comparisons . . . 121
D.5 Discussion and Conclusion . . . 122
D.6 Appendix: Implementation Details . . . 124
D.6.1 Body model parametrization . . . 124
D.6.2 Instance-specific top-down optimization . . . 124

E Tex2Shape: Detailed Full Human Body Geometry From a Single Image 129
E.1 Introduction . . . 130
E.2 Related Work . . . 132
E.3 Method . . . 134
E.3.1 Parametric body model . . . 134
E.3.2 UV parameterization . . . 134
E.4 Dataset Generation . . . 136
E.4.1 Scan registration . . . 137
E.4.2 Spherical harmonic lighting . . . 137
E.4.3 UV map synthetization . . . 138
E.5 Model and Training . . . 138
E.5.1 Network architectures . . . 138
E.5.2 Losses and training scheme . . . 139
E.5.3 Input partial texture map . . . 139

E.6 Experiments 140
  E.6.1 Qualitative results and comparisons 140
  E.6.2 Type of supervision 144
  E.6.3 Impact of UV mapping 145
  E.6.4 Impact of visibility 146
  E.6.5 Garment transfer 147
E.7 Conclusion 149
E.8 Appendix: Additional Results and Experiments 150
  E.8.1 Influence of illumination 150
  E.8.2 Influence of camera intrinsics 150
  E.8.3 Numerical comparison with HMR 152
  E.8.4 UV resolution ablation 152
  E.8.5 Additional qualitative results 153

5 Conclusion and Future Perspectives 157
5.1 Conclusions 157
5.2 Future Work and Applications 159

Bibliography 167

1 Introduction

Capturing and modeling the 3D human body from monocular video or photographs is a core problem in Computer Vision and Computer Graphics. For the past decades, estimating the 3D pose of a subject, encoded through the locations of distinct body parts or by joint angles, played a central role in Computer Vision and is still an active field of research. Researchers have enabled various applications in scene analysis, medical diagnostics, or human-computer interfaces. Recently, the automatic 3D reconstruction of the *full body* gained more and more attention. Hereby, one aims at reconstructing not only a 3D skeleton but the whole 3D human shape including hair, clothing, and appearance. Essentially, the goal is to create an avatar that is indistinguishable from the actual human.

The advent of Virtual Reality (VR) and Augmented Reality (AR) consumer hardware laid the foundation for new ways of entertainment, communication, or online shopping. For these applications, personalized and highly realistic 3D avatars are crucial. These avatars should feature all the details that form our identity and make us unique. This includes accurate body shapes, faithfully reconstructed faces, detailed clothing, and realistic hair. Reconstruction failures lead to avatars that are not being identified by others or, more importantly, to users not feeling represented by their virtual self. No less important, the acquisition process of these avatars should be fast, easy, and should not require special equipment or training. However, the classical Computer Graphics approach to 3D modeling of virtual humans still requires considerable manual effort and expert knowledge: A specially trained artist defines the 3D geometry of body and clothing, that is then rigged in order to enable animation. The avatar's 3D motion is driven by manual keyframe-based animation or marker-based motion capture. This laborious process presents an important practical barrier to the needs of the aforementioned applications. In contrast, the goal of this work is to take advantage of the omnipresence of cameras nowadays to develop automatic methods, that efficiently utilize images and video for realistic 3D avatar creation and animation.

In this dissertation, we explore the emerging topic of 3D reconstruction of human shape and pose from monocular images. We present novel approaches for reconstructing and tracking mesh-based 3D reconstructions of humans as

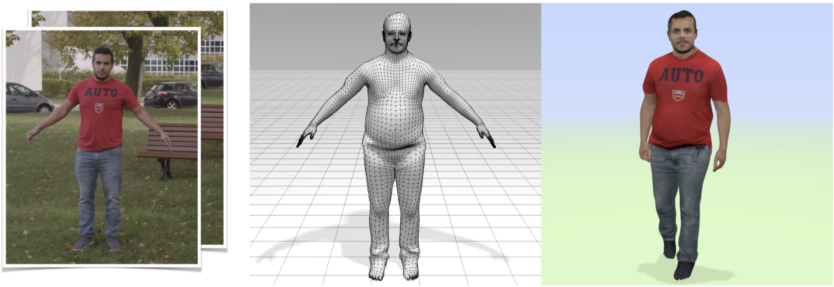
(a) (b) (c)

Figure 1.1: Given one monocular image or a short video clip of a person (a), the goal of this work is to reconstruct a detailed full-body 3D avatar (b) that can be photo-realistically animated, for example in virtual environments (c).

depicted in monocular videos and even single photographs. We show fundamental advances in the challenging task of 3D human avatar reconstruction from monocular images by developing methods that effectively fuse information from multiple points in time and realistically complete reconstructions from sparse observations. Our work enables, for the first time, easy acquisition of animatable 3D avatars for everybody and paves the way for various exciting new applications.

1.1 Problem Statement

Image-based 3D pose and shape reconstruction of humans is a wide field of research with many approaches and interpretations. Some researchers are mainly interested in the 3D skeleton and approximate body-proportions of a subject [147, 250, 26, 168, 112]. Others reconstruct the naked body shape without clothing [15, 260, 274] or focus on the garments worn by their subjects [184, 123]. Again others only focus on specific body parts alone like for example the face [23, 13, 28, 132, 231], hands [61, 210, 116, 201], or hair [141, 155]. In contrast, the goal of this dissertation is to track and model the whole human body, including hair and clothing. We are interested in reconstructing the observed subject *as detailed as possible*. Not only body and clothing geometry but also its coloring and surface structure carries important information. To this end, besides capturing the 3D shape, we aim at reconstructing the surface colors in form of texture maps, too. Finally, we want to be able to *re-use* the estimated avatars. For this purpose, the reconstructions should come in a common format that can be easily used, animated, and manipulated by other applications.

We have now defined the desired characteristics of the reconstructions. The second main aspect of this work is the capturing process and equipment. Computer Vision researchers have used a broad range of sensors and systems to capture and analyze the world. Commonly used are multi-camera set-ups, marker-aided capturing, depth sensors, or active scanners. These systems can capture 3D data to a resolution of a few millimeters. While this is undoubtedly valuable data, the systems are, unfortunately, not widely accessible. Usually, they are only found in laboratories or professional video studios. At the other extreme, standard cameras are nowadays all around us. Many of the devices we use on a daily basis have one or even multiple cameras built in and are easily accessible. Another valuable advantage of standard cameras is the unobtrusiveness and low complexity of the capturing process. While complex capturing systems often require careful and laborious calibration and set-up, or even interfere with the captured scene, cameras allow easy recording with nearly zero setup time. Additionally, they are lightweight and small and thus can be flexibly used nearly everywhere. For these purposes, as input to our algorithms, we rely in this work only on monocular image material as recorded by a standard webcam. This additionally ensures that advances of our work can be seamlessly integrated with modern devices, such as phones, tablets, or smart displays. At the same time, our work is compatible with the large amount of available legacy photo and video material.

The common pipeline and goal of all methods in this thesis is illustrated in Figure 1.1. Given a single image or multiple frames of monocular video of a person, we reconstruct a mesh-based full-body 3D virtual avatar. Additionally, we optionally reconstruct the appearance in form of a texture. The final avatar can then be animated and placed into new scenes or entire virtual environments.

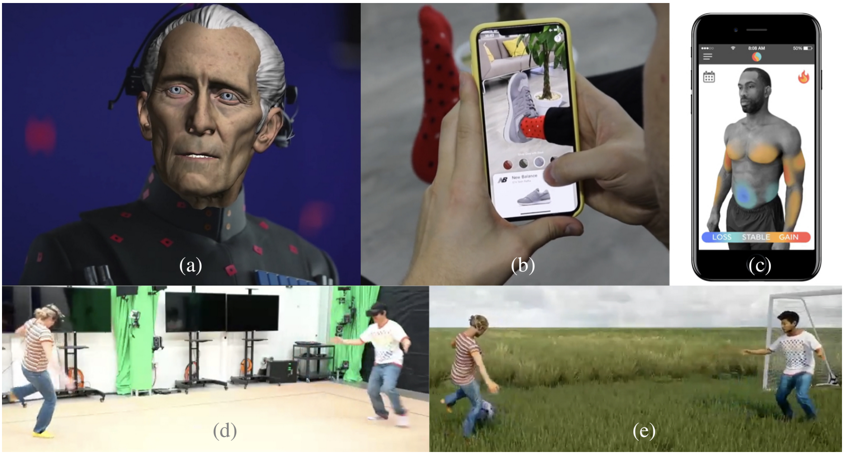

Figure 1.2: Possible use-cases for our methods: Modeling of virtual actors[1] (a), virtual try-on of clothing, shoes, and accessories[2] (b), body measurements for fitness and health[3] (c), and virtual telepresence[4] (d) and (e).

1.2 Motivation

3D virtual human avatars have been used for various tasks in the past and will potentially play a central role in many future applications. For example, in the movies industry virtual actors are commonly used in order to digitally edit and augment real-world video footage or even to produce fully computer-generated movies. To this end, producers and designers make huge efforts in order to produce highly realistic and physically-plausible virtual doubles of real-world actors. Similarly, in the games industry, developers put more and more work into realistic characters in order to produce a truly immersive gaming experience. Both industries will largely benefit from fully automatic and widely accessible reconstruction of highly realistic virtual humans.

Beyond entertainment, 3D virtual humans are potentially useful or already play an important role in many applications. Examples are human understanding for human-computer interfaces, medical diagnostics, virtual assistance, fit-

---

[1] Industrial Light & Magic / Lucasfilm via https://www.youtube.com/watch?v=OUIHzanm5Mk

[2] https://wanna.by/

[3] https://shapescale.com/

[4] Facebook Reality Labs via https://www.youtube.com/watch?v=FhiAFo9U_sM

ness and health tracking, virtual try-on in online shopping, body language interpretation and understanding, and many more. Figure 1.2 illustrates some of the examples. All these applications can benefit from more accurate reconstructions and easier acquisition. One emerging topic for 3D virtual humans are future applications in communications, as for example VR or AR telepresence. Enabling these applications is an active and emerging field of research [80, 235, 136, 170]. Telepresence applications are closely related to the problem statement of this dissertation, as they require easy 3D reconstruction and tracking of humans using low-cost sensors. Once established, these applications can significantly change our travel behavior, the way we communicate, and generally the way we live.

Highly realistic 3D human avatars and widely accessible reconstruction pipelines can make an impact in many science subjects and industries: We humans cannot *not* communicate [172], thus our visual appearance always carries rich information. From visual inspection of other human beings we are able to understand their mood, state of health, personal preferences, engagement, and much more. In this dissertation, we only lay the foundation for computers to model and understand the subtle visual cues that help us humans to understand the human body and its language. On the other hand, our scientific findings can already now directly be used by a large number of applications, for example in entertainment, fitness and health, or online shopping.

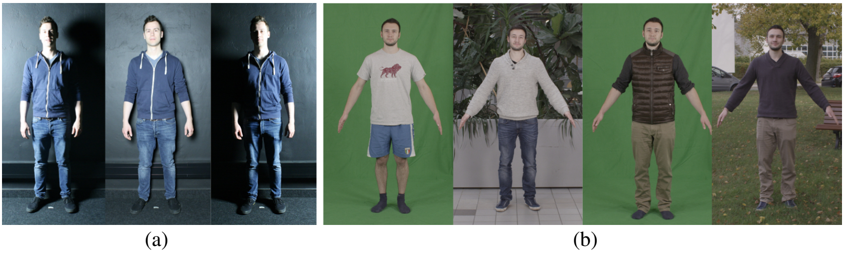

Figure 1.3: The appearance variation of a human is one of the main challenges in 3D human reconstruction. The same person might look very different in varying lighting conditions (a), in front of different backgrounds, or while wearing different sets of garments (b), even when other parameters, e.g. pose and camera, remain fixed.

1.3 Challenges

Humans are extremely good at understanding and predicting the human body and its language. From just a 2D photograph of people we can tell their 3D body pose, 3D body shape and approximate height, we understand their facial expressions and the performed actions, their intentions, and we even can tell how they might look on unseen parts. We are able to do all this because we have rich experience about how humans look like and how they move and behave.

While humans are able to process monocular video and photos with ease, the same remains a challenging task for computers and algorithms. Relevant information is often encoded, noisy, or ambiguous. For example, images lack direct depth information. Depth is only encoded indirectly through perspective, shading, and semantics. However, this information is much harder to understand and process than direct depth values. Another challenge originates from the image formation itself: distance to the camera, the actual size of an object, and the focal length of the camera all affect the projected size of an object in an image. This implies, for example, that multiple 3D skeletons all project into the same 2D skeleton. Conversely, this means that the true 3D pose often cannot be recovered from its 2D projection. Even with given intrinsic parameters of the camera, the true bone lengths and height of a person can only be approximated. While this ambiguity is one fundamental problem of 2D imaging, many more challenges exist: Unknown lens-distortion or recording parameters may prevent accurate measurements. Perspective distortion and foreshortening effects have to be handled by the algorithms. Further, images only describe the scene from one single view-point. Consequently, crucial information is often missing due to occlusion or self-occlusion. When working with videos, occluded scene content may be revealed, but connecting information from multiple time instances to a joint scene model is a non-trivial task.

While information retrieval from images is already challenging, there is usually also much additional information in an image that is not directly relevant for the task. Often the background of the scene must be ignored, and shadows and reflections may fool or impede the algorithms. Further, sensor noise, compression artifacts, or the appearance of new objects can erroneously be interpreted as a relevant signal.

Finally, the object of interest may change its appearance and shape over time. This is especially challenging when working with videos, as differing information from multiple frames have to be aggregated. Humans are particularly challenging, as we come in various shapes and appearances. Humans may look completely different at different points in time due to changes in pose, changed illumination conditions, through altered camera settings, or different wrinkle patterns in clothing, changed hair, and much more. When looking at images from longer time spans, humans may even have changed their clothing, their hairstyle, or have gained or lost weight. See Figure 1.3 for some examples on how the appearance of a human can vary.

When reconstructing humans, one presumably faces another challenge: The Uncanny Valley [153]. The Uncanny Valley is a theory by Mori et al. describing the relationship between the degree of realism of an artificial human and the emotional response to it. The *valley* denotes a dip in the curve of familiarity with artificial humans plotted against their human likeness. Very human-like robots or avatars seem to cause a response of uncanniness or revulsion. While the Uncanny Valley is a hypothesis, some studies provide empirical evidence [154].

Naturally, we can not tackle all of the listed challenges. To this end, we provide algorithms that work in more or less constrained settings. However, we take particular care to constrain the set-ups not too much, so that our algorithms can be reproduced and data acquisition is as easy as possible. Since our focus lies on the detailed acquisition of 3D shape, we constrain our setup to images of people in A-poses or standing poses, which is practical for many applications. The following section give an overview over the contributions of this work, how the described challenges are approached and partly overcome, and which methods, tools, and concepts have been used in order to achieve this.

1.4 Contributions

This dissertation describes advances in 3D human pose and shape estimation from monocular images. Each of the following chapters corresponds to one publication and describes specific advances in this field. All described methods have the input and output modalities in common: Input are monocular videos or photos; output are animatable 3D meshes describing the apparent shape, pose, or motion of a human depicted in the input material. Solving this joint task – creating 3D reconstructions of humans from monocular images – summarizes the overall contribution of this thesis: By only relying on regular video or even photos, our work democratizes the digitization of humans. For the first time, it eliminates the need for specialized equipment. We enable automatic reconstruction of detailed shapes and widespread usage of virtual humans in emerging technologies.

Our work explores different approaches to 3D human pose and shape reconstruction. We show advantages of optimization-based and learning-based approaches, study different forms of data representation and supervision losses, and discuss the robustness and limitations of the individual methods. In the following, the main contributions of each publication are briefly summarized:

[Paper A] Optical flow-based 3D human motion estimation from monocular video: A 3D representation of an actor in a video sequence needs to match the sequence both in shape and in motion. While most previous works focused on identifying 3D poses individually per frame, this work presents a method to estimate the 3D motion of an approximate 3D body shape that matches the *apparent motion* of the video sequence. By minimizing the difference between calculated and synthesized optical flow, we are able to reconstruct fluent 3D motion of up to 100 frames after initializing on a single frame.

[Paper B] Video Based Reconstruction of 3D People Models: While 3D pose and motion estimation became more and more popular concurrent to our work, monocular 3D human shape estimation was still limited to estimating parameters of a parametric body model. The paper corresponding to this chapter has presented the first method to estimate the full 3D shape of a clothed human from video. From a video depicting a person such that he or she is visible from all sides, we aggregate silhouette information from all frames into a single frame of reference. To this end, we *unpose* the silhouette cone in each frame, allowing for efficient 3D shape estimation independently of pose. Extensive experiments demonstrate a reconstruction accuracy of 4.5mm and robustness of the method to noisy 3D pose estimates.

[Paper C] Detailed Human Avatars from Monocular Video: A convincing digital avatar of a human should comprise all the unique properties of this person. In this work, we add many of those properties to avatars that have been calculated using the method from Paper B. Specifically, we improve the reconstructed faces by relying on detected 2D facial landmarks and add clothing wrinkles and fine structured details to the shapes based on multi-frame shape-from-shading. Finally, we introduce a novel texture stitching strategy that leverages a semantic prior and stitch a high-detailed texture that adds important appearance information to the meshes. In a user study, we show that *details matter* and the additional reconstruction steps undoubtedly pay off.

[Paper D] Learning to Reconstruct People in Clothing from a Single RGB Camera: In order to make 3D human reconstruction widely available, the process has to be fully automatic, robust, and fast. Building on recent advances in geometric Deep Learning, we present a learning-based model that enables robust 3D shape reconstruction of clothed humans from only a small number of frames. The presented model combines advantages from both learning and optimization-based methods: A reconstruction predicted by a single forward-pass through a neural network can be refined for a few seconds via weak *render and compare* supervision using the same model at test-time. We further present an extensive analysis of key parameters and demonstrate that the model can partly be trained with weak supervision.

[Paper E] Tex2Shape: Detailed Full Human Body Geometry from a Single Image: In the last publication, we further reduce the input data. From only a single photograph of a person, we reconstruct an avatar that compromises fine details such as hair and garment wrinkles even on *occluded parts*. To this end, we train a conditional Generative Adversarial Network that effectively translates incomplete texture maps into normal and displacement maps. These maps add the desired level of detail to a smooth parametric body model. The key insight of the work is to transform the pose-dependent and unaligned reconstruction problem into a pose-independent and aligned image-to-image translation problem by encoding the input image in UV space. Despite being trained purely with synthetic data, the model generalizes well to real-world photographs, laying the foundation for wide-spread 3D reconstruction of people for various applications.

Table 1.1 summarizes inputs and outputs of the proposed methods. Each of the papers describes individual advances and focuses on different tasks and problems in 3D human reconstruction from images. Together, the publications have impacted the emerging field of human digitization from images, and enabled, for the first time, to create detailed 3D avatars from monocular images of subjects in motion.

	Input		Output	
Paper	**Format**	**Allowed Poses**	**Pose Reconstruction**	**Shape Personalization**
A	video	any	✓	✗
B	video	only A-poses	✓	medium
C	video	only A-poses	✓	high
D	1 - 8 images	only A-poses	✓	medium
E	single image	camera facing	pose-invariant	high

Table 1.1: Input, output, and properties of the proposed methods. See Section 1.4 for detailed descriptions and individualcontributions.

1.5 Outline

The remainder of this dissertation is organized as follows: In the following chapter, an in-depth review of the topic of 3D human modeling and reconstruction from images, depth-sensors, and 3D data is given. The chapter outlines the development in the field from geometric primitives to complex data-driven models as well as model-free approaches. Finally, it discusses recent advances based on Deep Learning techniques. In Chapter 3, we discuss the different concepts that have been used in the publications forming this dissertation. We give an overview of the core methods and explain crucial tools and algorithms in more detail. The Papers A to E contain the above introduced publications and form the core contribution of this dissertation. Chapter 5 concludes this dissertation with a discussion about the achieved results and an extensive discussion about possible directions for future work.

2 Related Work

3D human body pose and shape modeling and reconstruction has changed dramatically over the past few years and recently received more and more attention. Starting from models based on geometric primitives, researchers have developed more and more complex models of the human body learned from large scan-datasets of real humans. These models again have been successfully deployed for various applications, such as pose tracking, video editing, or statistical analyses. The advent of Deep Learning resulted in a paradigm shift not only for this specific topic, but generally in the fields of Computer Vision, Computer Graphics, and many more. It accelerated the progress in these fields and enabled many new applications, while at the same time reduced the amount of needed input to often only a single image.

This chapter gives a systematic overview of the topic of 3D human body pose and shape geometry modeling and reconstruction. We illustrate its origins, how the topic has developed before and concurrent to this dissertation, as well as its most recent advances.

2.1 Body Models based on Geometric Primitives

Researchers have understood very early how their methodology can benefit from a model of the human body. Early works model the human body in form of geometric primitives, such as the pioneering mathematical model by Hanavan Jr [82]. In this work, a personalized body model is constructed from 15 simple 3D polygonal shapes. Even simpler 2D models have been constructed and successfully applied to human gait analysis [169, 91, 165]. 3D human pose estimation and tracking have been the driving force to develop more and more complex 3D models of the human body and its kinematic chain [149, 199, 69, 212]. Finally, also the human shape was taken into account, introducing the first full parametric yet completely synthetic body models [233, 178, 219, 220].

2.2 Artist-Driven and Anatomical Models

At the same time, the Computer Graphics community introduced the first digital actors and began to revolutionize the movie industry [142]. Similar to the models of the Computer Vision community, these characters have been constructed from geometric shapes and an implanted skeleton used for animation. Seeking more realism, researchers soon developed layered models of bones, muscles, and skin that are artist-driven [38] or anatomically inspired [205, 157]. However, these models are difficult to build and simulation requires time-expensive calculation. To this end, the use of *skinning* techniques became popular [129, 114, 115]. A skinning function defines how the surface of a model bends and moves according to the movement of an implanted skeleton. This technique is also employed in state-of-the-art data-driven models, that are introduced in the following.

2.3 Data-Driven Body Models and Applications

To represent the human body more realistically, models learned from data of real humans have been developed (see Figure 2.1). These models describe the shape variations of the naked body without hair or clothing. In the process of learning such models, typically a template mesh is deformed to match 3D data of a large number of subjects in various poses and of different body shapes. Then a statistical formulation is found that minimizes the error between low-dimensional, parametrization-based predictions and the alignments. A similar concept has already been used in the pioneering work by Kakadiaris and Metaxas [110], where a body model is constructed from three orthogonal views. Allen et al. [10] learn a rich model of the human shape from laser scans of 250 subjects. Later the model has been updated to also modeling pose-dependent shape deformation [11]. Both models operate in global model space which means that they directly output global vertex positions. With SCAPE [14], a popular parametric body model has been presented, that operates on mesh triangle level. Pose and shape deformation components of the model are applied separately to each face, which is then rigidly transformed to match the pose. This formulation simplifies the mathematical formulation into a rigid and a non-rigid component, which supports the learning process. Based on SCAPE, BlendSCAPE [90] is inspired by skinning functions and deforms each triangle based on a linear combination of multiple influencing parts. Another variant that incorporates correlations between body shape and pose has been introduced by Hasler et al. [84]. The DYNA model [182] extends SCAPE with dynamic soft-tissue deformation based on the performed motion. However, because the mesh triangles are transformed independently and form no watertight mesh, all SCAPE variants depend on a least-squares solver to connect the triangles to a smooth and coherent surface. This drawback prevents the models from being used in standard graphics pipelines, which has been addressed by the following works.

The SMPL body model [138] is an accurate parametric body model learned from thousands of scans from real people. For posing, it transforms a template mesh using standard linear blend skinning, thus it requires no post-optimization and is compatible with standard graphics pipelines. A key insight is that pose-dependent deformations can be linearly regressed from the pose rotation matrices. SMPL is more accurate and more straight-forward to use than SCAPE and thus is heavily used in various research problems. We, too, use SMPL in this work as a template and prior for reconstructing poses and shapes of clothed people. Like SCAPE, SMPL has been extended for soft-tissue deformations. The DMPL model, a data-driven extension to SMPL is included in the original paper. Kim et al. [118] present a layered combined data-driven and physics-based model.

Despite its popularity, SMPL comes with some drawbacks. By design it models the body only at a coarse scale. Neither facial expressions nor finger movement are covered by the original model. To this end, multiple works focus on extending SMPL and adding missing functionality. The first work along this line has been SMPL+H [201], a SMPL model with an incorporated hand model. Joo et al. [109] propose Frank, a model stitched together from three different models. They use SMPL without the pose-dependent deformations for the body, an artist-rigged hand model, and a generative PCA face model learned from the FaceWarehouse dataset [35]. However, the components are learned individually and thus the model lacks realism. To this end, Pavlakos et al. [174] present SMPL-X. SMPL-X extends SMPL with articulated hands and an expressive face. In contrast to Frank [109], the model is learned in a unified fashion. Finally, several special-case models exist. ClothCap [184] presents the first SMPL with clothing but does not model pose-depended deformations. Hesse et al. [88] retrain SMPL for infants based on RGB-D captures, and Zuffi et al. [290] introduce a SMPL-like model of animals that has been learned from 3D scans of toy figures.

Besides models of the whole body, a large number of 3D parametric body part models exist. These models include models of the face [23, 13, 175, 28, 98,

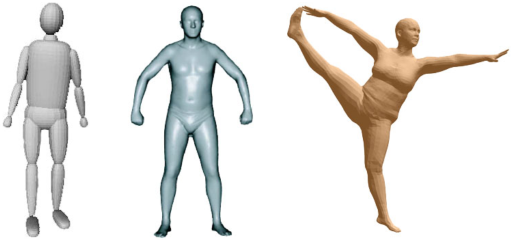
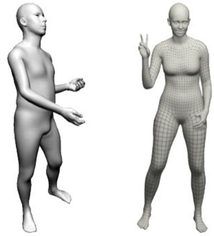

Figure 2.1: Body models with varying degree of realism. From left to right: Superquadric model [219], SCAPE [14], SMPL [138], Frank [109], and SMPL-X [174].

70, 231], the head [47, 132], hand models [210, 116, 201], models of the whole arm [159], and even of the ear [48]. In the following, we will review methods that utilize parametric models of the whole body.

Parametric body models have been heavily used to reconstruct and encode 3D pose [179]. In early works, researchers formulate analysis-by-synthesis problems to recover the 3D pose from multiple views [17], depth data [254], or single images [213, 85]. For this purpose, posed 3D human shapes are reconstructed that project into the image silhouettes or match with the input data. Similarly, we present a work where we minimize silhouette and additional optical-flow differences to recover time-consistent 3D poses [5] (Paper A). In an alternative strategy, the reprojection error of 3D joint locations is minimized. First, these 2D joint landmarks have been manually clicked [76]. Later the process was automated [26, 124, 95]. The automation was made possible by the advent of human landmark detection networks [177, 101, 36, 12]. Another streamline of works uses a set of inertial measurement units (IMUs) attached to the subjects body alone [245, 246, 96] or in combination with images [180, 247] to reconstruct 3D motion. The advent of Deep Learning accelerated the advances in model-based 3D human motion estimation from images. We review this field separately in Section 2.5.

Besides the pose properties of parametric body models, also the shape components have been utilized in the literature. For example, the SCAPE model has been used to recover the *naked* body shape of people from photographs in regular clothing [15, 54]. The SMPL model has been used with 4D scanner data to recover the body shape of the subject under clothing [274]. Pons-Moll et al. jointly estimate garments as a separate clothing layer [184]. A similar system has recently been introduced also for depth data [227]. The methods by Guo et al. [79] and Chen et al. [39] recover the clothed and naked shapes from a single image but require manual initialization of pose and clothing parameters. Fully automatic acquisition of the full shape including tight [25] and loose clothing [275, 226] has been presented, too. However, these works require RGB-D data. Our works [7, 6] (Paper B, Paper C) have been the first to present 3D human shape and clothing reconstruction from monocular video in which the subject is allowed to move. Similarly to the works by Zhang et al. [274] or Pons-Moll et al. [184], we extend SMPL with a deformation field for modeling clothing and hair. In contrast to these works we use a single RGB camera as input.

Due to their shape variation, parametric body models have also been successfully deployed in other science fields, for example, to study body-related clothing preferences [204] or self-perception in anorexia nervosa [152]. Finally, body models have been used for image and video editing, for example, to change the body shape of subjects in images [280] or videos [106], to augment actors with new clothing [198], or to “wake-up” subjects in photos and artwork to display them in VR or AR [255].

2.4 Free-form and Template-based Surface Reconstruction

While body models are rich priors for human body shape reconstruction problems, they also limit the shape space. All shapes that do not share the human topology cannot be well approximated using a body model. To this end, researchers have developed free-form and template-based reconstruction methods, which we review in the following.

Even before body models have been available, researchers have used body templates. These typically have been artist-made, rigged meshes that represent a single person. For personalization, these templates have been non-rigidly deformed to match image silhouettes in multi-view set-ups [89, 1]. These early methods share in large parts the methodology of those using parametric models. However, they cannot benefit from the low-dimensional shape space. Nevertheless, these methods enabled for the first time multi-view body pose and shape reconstruction [234, 50] and even free-viewpoint video of human actors [37]. Later, the artist-made templates have been replaced with laser-scans of the subjects [51, 243, 66], enabling detailed reconstructions and also complex clothing like skirts and dresses. Also, temporal surface deformation tracking has been enabled for detailed free-viewpoint video [223, 45, 127]. In an alternative strategy, the methods by Rhodin et al. [192] and Robertini et al. [194] leverage a flexible sum of Gaussians body model [224] to reconstruct human motion and shape. Also related, general frameworks for 3D shape tracking based on volumetric shape representations have been presented [94, 4].

While all these methods require multi-view input, methods utilizing a single depth sensor for shape reconstruction have been developed, too [46, 131, 273, 209]. These methods, however, do not allow for free movement but require the subject to carefully take the same pose at different angles to the camera or hold the pose while the camera is moving around the subject. Subtle pose changes are then compensated by non-rigid alignment of the point clouds. For easier acquisition, Tong et al. [237] propose to use a turntable. Later, the restriction of static poses has been removed by utilizing multiple depth-sensors [57, 170]. Live performance capture using a small number of depth-sensors was made possible. Finally, Newcombe et al. [163] introduce a real-time method to dynamically fuse the incoming depths stream of a single RGB-D camera into a canonical model. The model is warped to match the latest frame, enabling single sensor live performance capture. Based on this idea, methods enabling for example volumetric non-rigid reconstruction [100] or less-controlled motion [218] have been presented.

Xu et al. [264] present for the first time monocular performance capture including surface deformation, what made the use of depth-sensors obsolete. A pre-captured template of the actor is tracked and deformed based on 2D and 3D human landmark detection and image silhouettes. Following the proposed methodology, Habermann et al. [80] present the first real-time human performance capture based on a single view RGB video-stream only.

2.5 Deep Learning-based Human Reconstruction

Deep Learning techniques like CNNs have accelerated advances in Computer Vision in general, and advances in human pose and shape reconstruction in particular. Numerous learning-based works on 2D and 3D landmark detectors or reconstruction and tracking of specific body parts exist in the literature. In the following, we will review image-based methods that reconstruct the full human body.

In early works, methods that reconstruct the shape in the space of a parametric body model have been presented [53, 55]. These methods use only a single silhouette image but are restricted to a small set of poses. In the following, more flexible works have been presented that reconstruct 3D pose and shape from single images by integrating the SMPL body model into a network architecture. Different works leverage either color images [239, 112], color images plus segmentation [173], or body part segmentation [168]. Other works have focused on the temporal aspect and successfully reconstruct temporal-coherent 3D human motion [113]. While these approaches reliably recover the 3D human pose from in-the-wild images, the reconstructions tend to feature average body shapes. The reason for this is, that the methods heavily rely on the body model statistics and return shape regressed from bone lengths. For more exact reconstructions that better align with the images, methods perform mesh fitting after network inference [78]. This fitting step also allows to additionally reconstruct face and hand motion [261]. Our work [8] (Paper D) has been the first to reconstruct the human shape beyond the parametrization of a body model from a small number of frames. We, too, refine our results via optimization at test time. Similarly, Zhu et al. [286] perform a multi-step approach. They first find an initial SMPL pose and shape parametrization, then repose the mesh based on silhouettes, and finally, leverage shading to refine the surface beyond the shape parametrization.

Recently, the question of the best 3D human shape representation in the context of CNNs gained more and more attention (see Figure 2.2). BodyNet [241] was the first work to directly regress a volumetric representation of the human body from a single image. A similar approach has been introduced by Jackson et al. [105], demonstrating a higher level of detail. More recently, synthesizing novel silhouette views to represent the 3D shape of the person, before reconstructing the final 3D volume has been proposed [156]. Zheng et al. [278] refine results from volumetric regression via a shading-based normal refinement network to alleviate the limited spacial resolution of volumetric approaches. In a different direction, Kolotouros et al. [121] propose to directly regress vertices and optionally infer body model parametrization from there. Other works regress and represent vertices in the UV space [267] or similarly as geometry images [187]. In our work [9] (Paper E), we use the UV space to reconstruct detailed human shape to a wrinkle-level independently of the 3D pose. In contrast to concurrent work, our results feature details even on the unseen back-side of the person.

In recent work, 3D shapes have been encoded as implicit functions like volumetric occupancy fields or signed distance functions [171, 148, 42, 150, 263]. The first works deploying this idea in 3D human shape and pose reconstruction use spare multi-view setups [97, 72]. Saito et al. [202] use this form of representation for single-view human shape and texture reconstruction. The main

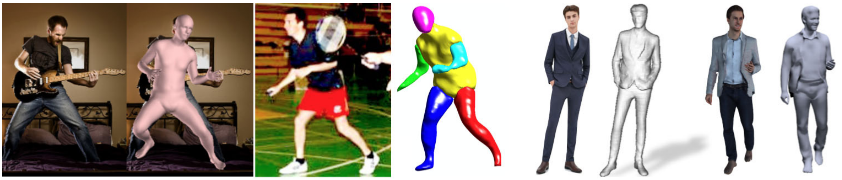

Figure 2.2: Learning-based human shape reconstruction using different forms of 3D shape representation. From left to right: SMPL model-based [112], voxel-based [241], using implicit functions [202], and by augmenting SMPL in UV-space [9] (Paper E).

idea of this work is to sample the occupancy field along pixel-aligned projection rays, which favors local details.

Finally and very recently, methods with or without coarse explicit 3D representation have been presented. In the work by Shysheya et al. [211] the appearance of a subject is learned as per-part textures of the SMPL body model. Given a 3D pose and a view-point, the parts are used to synthesize an image of the subject utilizing a subject-specific neural renderer. Other recent works present first ideas to encode complex scenes in coarse voxel grids [217, 137] or as feature point clouds [3]. A learned renderer allows synthesizing images of the scenes from novel viewpoints, featuring view-dependent surface effects or thin structures and semi-transparent materials like human hair or smoke. While this is an exciting avenue to explore, artifacts are still prominent and in contrast to mesh-based solutions, compatibility with existing rendering pipelines is not given.

3 Methodology

In the following, we will outline some of the basic tools, techniques, and principles, that have been used in the publications that form the core of this dissertation. In the last few years, the advent of Deep Learning significantly changed the methodology of Computer Vision research. This also reflects in the different approaches to 3D human shape and pose reconstruction in this work, as outlined in the following. All presented approaches have, however, one thing in common: To make the problems tractable, we leverage a parametric body model. Parametric body models are statistical models of the variation of the human body shapes and poses. See Section 2.3 for an introduction. These models help by regularizing the search space and reducing the dimensionality of tasks related to the human body. In other words, they provide a template as an approximate solution that can be further refined by relying on its parametrization alone or as a regularization prior. We now introduce the body model and the different methods and concepts that have been utilized in this dissertation.

3.1 Body Model

In this work, we utilize the SMPL body model [138], presented by Loper et al. in 2015. SMPL is designed as a function $M(\cdot) \in \mathbb{R}^{N \times 3}$ that maps pose $\boldsymbol{\theta} \in \mathbb{R}^{3K}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ parameters to a mesh of $N = 6890$ vertices. To form a watertight mesh, the vertices are connected to $F = 13776$ faces. The pose is determined through $K = 23$ skeleton joints parametrized by $\boldsymbol{\theta}$ in axis-angle representation. The SMPL model has been learned from scans of real people. It can, therefore, produce realistic body shapes and pose-depended shape deformations. SMPL exists in three variants: A male, a female, and a neutral version, covering only male, only female, or all subjects respectively.

SMPL produces a posed mesh by performing the following steps: To create realistic body shapes, a template mesh $\mathbf{T} \in \mathbb{R}^{N \times 3}$ is deformed with shape deformation offsets $B_s(\boldsymbol{\beta}) \in \mathbb{R}^{N \times 3}$ (Figure 3.1 (b)). The offsets are based on a low-dimensional basis of the principal components of the body shape distribution among the SMPL subjects. The shape parametrization $\boldsymbol{\beta}$ forms a vector of linear shape coefficients of the shape space. Additionally, a linear regressor

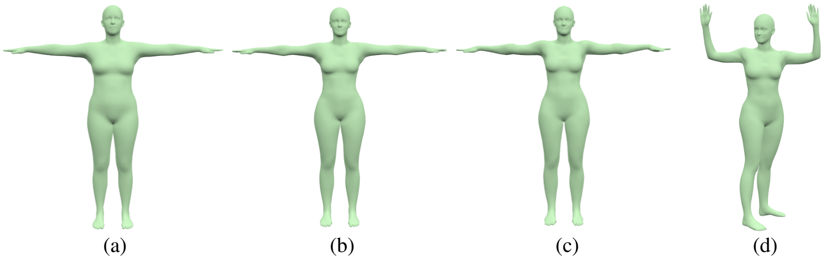

Figure 3.1: Setting pose and shape of the SMPL model: From a template (a) the new shape (b) is formed. Then pose-dependent offsets are applied (c). Finally, the pose is set via blend skinning (d).

determines the positions of the skeleton joints $J(\boldsymbol{\beta}) \in \mathbb{R}^{K\times 3}$. Next, pose-dependent deformations $B_p(\boldsymbol{\theta}) \in \mathbb{R}^{N\times 3}$ are applied on the reshaped template (Figure 3.1 (c)). $B_p(\cdot)$ is a learned linear function parametrized with the desired pose $\boldsymbol{\theta}$. It accounts for muscle and soft-tissue deformations as well as skinning artifacts potentially introduced in the last step. Finally, the mesh is posed using standard linear blend skinning $W(\cdot) \in \mathbb{R}^{N\times 3}$ with blend weights $\mathbf{W} \in \mathbb{R}^{N\times K}$ (Figure 3.1 (d)). The final equation reads as:

$$M(\boldsymbol{\beta},\boldsymbol{\theta}) = W(T(\boldsymbol{\beta},\boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{3.1}$$

$$T(\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}). \tag{3.2}$$

SMPL only covers naked subjects and its shape parametrization does not allow for detailed personalization. For this reason, we augment the standard formulation with additional details in large parts of this work. We add additional per-vertex offsets $\mathbf{D} \in \mathbb{R}^{3\times N}$ to the function [182, 274, 184]. SMPL+D, SMPL extended with offsets $\mathbf{D}$, is formed as follows:

$$M(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}) = W(T(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{3.3}$$

$$T(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D}. \tag{3.4}$$

Additionally, we augment SMPL using UV mapping. In Papers B, C, and D we apply textures to the mesh. In Paper E we augment its surface using normal and displacement-maps. UV mapping [24] unfolds the body surface onto a 2D image such that a given pixel corresponds to a 3D point on the body surface. The mapping is defined over the faces such that every face consisting of three 3D vertices has a counterpart consisting of three 2D UV-coordinates. Hereby, U and V denote the 2 axes of the image. The mapping of points inside a face is determined via barycentric interpolation of neighboring coordinates. The 2D image can be used to augment the 3D surface. A texture defines a color per surface point. Similarly, a normal-map stores a surface normal that can add or enhance visual details through shading. A 3D displacement-map actually displaces the surface point in the given direction. Hence, it can be used to create a highly detailed surface without changing the resolution of the underlying mesh. Some tasks, however, require a higher mesh resolution. It can be derived by subdividing the SMPL base mesh. Hereby, a new vertex is placed on the center of each side of a triangular face. The old face is removed and four new faces are created by connecting subsets of the six vertices. This processes can be repeated. Please see Paper C for details.

As described beforehand, the body model can be used as a template, as a prior, or as a representation for methods that work on 3D human body shapes and poses. We will now elaborate on how we utilize the SMPL body model in our work. More importantly, we introduce the general methods and principles we have consulted to provide effective solutions to 3D human pose and shape reconstruction from monocular images.

3.2 Analysis-by-Synthesis

Paper A presents an approach to time-consistent 3D pose estimation from video. The principal idea of this work is that 3D pose is encoded in the 2D vector field of optical flow. To recover the 3D pose changes from the 2D optical flow field, we follow the *analysis-by-synthesis* or *inverse graphics* approach. In analysis-by-synthesis, one aims at recreating the apparent scene using a rich synthetic scene model. In our case, we minimize the difference between the observed and synthesized optical flow. We synthesize optical flow by rendering different pose parametrization of the SMPL body model using a specialized renderer. Similarly, in papers Paper B and Paper C we estimate the 3D body shape of a human by comparing its rendered silhouette with observed silhouettes. Generally speaking, in *analysis-by-synthesis* we define one or multiple objective functions, that are optimized with respect to our scene model. In our case, the scene model is the SMPL model plus possible additional components, for example, an image formation function. The objective functions measure the similarity between the synthesized and the observed images. Typically, we choose to recreate abstractions of the images, e.g. segmentation, optical flow, or keypoints, rather than images itself. These abstractions or features have far less variation in appearance and thus are easier to synthesize. In the following, we shortly introduce different analysis-by-synthesis techniques that we have utilized in the different works of this dissertation.

3.2.1 Image keypoints

The easiest of the mentioned abstractions are image keypoints. Keypoints are 2D locations of image observations often with a specific semantic. For example, for our setting keypoints can be facial landmarks or skeleton joint locations. First, for every image keypoint $\boldsymbol{k}_i \in \mathbb{R}^2$ one finds a corresponding point $\boldsymbol{l}_i \in \mathbb{R}^3$ in the scene model. Then, during optimization one aims at finding a scene description such that every $\boldsymbol{l}_i$ projects onto $\boldsymbol{k}_i$ under a given projection $\pi(\cdot)$:

$$\sum_i ||\pi(\mathbf{R}\boldsymbol{l}_i + \boldsymbol{t}) - \boldsymbol{k}_i|| = 0. \tag{3.5}$$

$\mathbf{R}$ and $\boldsymbol{t}$ are rotation and translation parameters in an exemplary scene model. As mentioned earlier, in the problem settings of this work, the scene is described by the SMPL model. Global rotation and translation are generally applied, too. 3D points corresponding to image keypoints are regressed from the surface of the body model by a linear combination of a set of vertices, for example through barycentric interpolation.

3.2.2 Optical flow

Optical flow [71] is the perception of motion by our visual sense. For two images it can be described as a 2D vector field that matches all points in the first image to their apparent counterpart in the second image. For the first time, optical flow between two images was computed by Horn and Schunck [93]. The described method makes two assumptions: First, the *brightness constancy* constraint assumes that the intensity of a pixel at position $[x, y]$ in an image $\mathbf{I}$ at time-step t remains constant after displacement:

$$\mathbf{I}_{x,y,t} = \mathbf{I}_{x+\Delta x, y+\Delta y, t+\Delta t}. \tag{3.6}$$

Second, it is assumed that all motion is small, i.e. less than a pixel wide. The latter assumption was later replaced in extensions using image pyramids [140]. Hereby, larger motion can be estimated but local smoothness is assumed. Based on these constraints, one can optimize for the beforehand described 2D vector fields.

Although calculated in the image plane, optical flow contains 3D information. Optical flow can be interpreted as the projection of 3D scene flow [242]. Assuming the presence of optical flow in the sequence, large parts of the observed optical flow are caused by relative movement between object and camera. Optical flow contains information about boundaries of rigid structures. On the other hand, unique appearance effects such as texture and shading are removed. To this end, optical flow is a well-suited abstraction for analysis-by-synthesis problems. In Paper A, we extract 3D poses from an image sequence by minimizing the difference between computed and synthesized optical flow. For synthesizing optical flow, we have developed a differential flow renderer that renders the projected scene flow between two parameterizations of the SMPL model. By relying on optical flow, we enforce small differences between subsequent images and therefore produce time-consistent and fluent motion.

3.2.3 Image segmentation

Image segmentation is a well-established scene abstraction that is heavily used in analysis-by-synthesis problems. In image segmentation, each pixel is represented by a certain label. One has to differentiate between binary and multi-part segmentation. Binary segmentation usually differentiates between foreground and background. Foreground and background are hereby defined task-specific. Often all moving objects belong to the foreground and all static objects belong to the background. In Paper B and Paper C we are only interested in the person, hence we define foreground as person and the rest of the scene as background. In Paper D, we utilize multi-part segmentation. In multi-part segmentation, each label corresponds to a certain object type or instance. In our case, the segmentation differentiates between classes of garments, certain body parts, and background.

During the optimization of an analysis-by-synthesis problem based on image segmentation, one simply minimizes the difference between the predicted silhouette and the observed silhouette. For the binary case, this reads as:

$$\min_{\mathbf{R},\boldsymbol{t}} |G(\mathbf{R},\boldsymbol{t}) - \mathbf{S}| \tag{3.7}$$

$$G(\mathbf{R},\boldsymbol{t}) = R_c(F(\mathbf{R},\boldsymbol{t})), \tag{3.8}$$

where $\mathbf{S}$ is an observed segmentation image, $F(\cdot)$ is an exemplary scene function, and $R(\cdot)$ is a binary image formation function under camera c. While this formulation has the expected minimum, it can be slow and problematic to optimize. The reason for this is that the optimization might get stuck in local minima and gradients only describe a one pixel-wide neighborhood. Hence, usually more advanced formulations are used. One straight-forward way of solving this issue is to formulate the problem at the same time on different resolutions of the images. This way, the optimization can take larger steps and eventually passes local minima. In Paper A and Paper B we follow a slightly different approach: For each point in one silhouette, we minimize the difference to the closest point in the other silhouette. This process is called Chamfer matching (see Figure 3.2). The idea is that the predicted silhouette should not exceed the observed silhouette, while at the same time cover it as much as possible. The distances of all points in silhouette $\mathbf{M}$ to the nearest points in the silhouette $\mathbf{N}$ can easily be computed by multiplying silhouette $\mathbf{M}$ with the distance transform $C(\cdot)$ of silhouette $\mathbf{N}$. The distance transform of a binary image contains for every pixel its distance to the closest non-zero pixel. The Chamfer matching objective sums the errors over all image pixels p:

$$\min_{\mathbf{R},\boldsymbol{t}} \sum_p G_p(\mathbf{R},\boldsymbol{t}) \cdot C_p(\mathbf{S}) + \mathbf{S}_p \cdot C_p(G(\mathbf{R},\boldsymbol{t})). \tag{3.9}$$

Unfortunately, $C(\cdot)$ is not differentiable. To make the objective differentiable, we slightly change it to:

$$\min_{\mathbf{R},\boldsymbol{t}} \sum_p G_p(\mathbf{R},\boldsymbol{t}) \cdot C_p(\mathbf{S}) + (1 - G_p(\mathbf{R},\boldsymbol{t})) \cdot C_p(\mathbf{1} - \mathbf{S}). \tag{3.10}$$

While the minimum of the objective stays the same, we don't have to calculate the distance transform of the predicted silhouette.

While silhouette overlap and Chamfer matching are calculated in the image plane, they actually supervise a 3D problem: The depicted scene is three-dimensional, hence every pixel in an image corresponds to a virtual ray shooting from the camera into the 3D scene (see Figure 3.3). Given the camera intrinsics are known, we can supervise silhouette matching directly in 3D by computing point to line distances. The key contribution of Paper B is based on this observation: Instead of minimizing the per-frame 2D silhouette overlap, we minimize 3D point to line distances in a joint canonical representation. To this end, we find correspondences $(\boldsymbol{v}, \mathbf{r}) \in \mathcal{M}$ between each SMPL vertex $\boldsymbol{v}$ and a 3D silhouette ray $\mathbf{r}$ in every frame. Then, we *unpose* the silhouette rays based on the estimated 3D pose of the subject. Unposing generalizes visual hull for

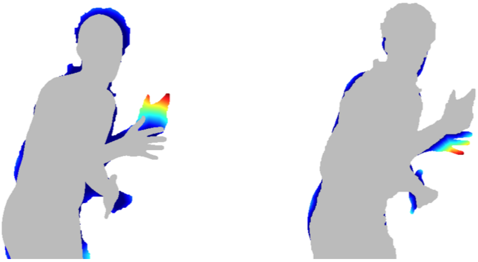

Figure 3.2: Pose reconstruction using Chamfer matching: The predicted silhouette should not exceed the observed silhouette (right), while at the same time cover it as much as possible (left). Red means large error and blue means small error compared to the corresponding silhouette (grey).

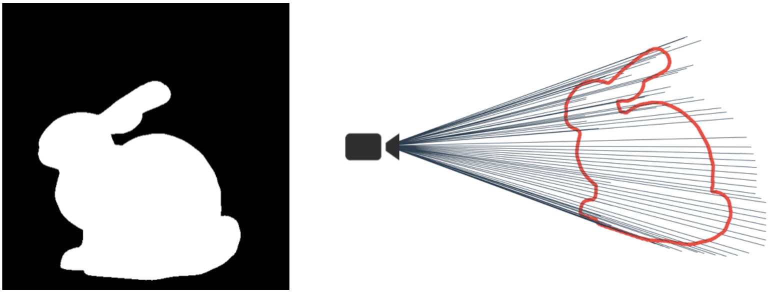

Figure 3.3: Every pixel in an image corresponds to a virtual ray shooting from the camera into the scene. Silhouette pixels in an image (left) form a silhouette-ray cone (right) that limits possible 3D object positions and dimensions.

articulated motion. Please see Paper B for details. Having silhouette ray and vertex correspondences from all frames a canonical representation, we now can minimize point to line distances to recover the underlying shape. The point to line distances can be efficiently computed by expressing the rays using Plucker coordinates:

$$\mathbf{r} = \boldsymbol{r}_m, \boldsymbol{r}_n. \tag{3.11}$$

Hereby $\boldsymbol{r}_n$ corresponds to the direction of a line passing through point $\boldsymbol{p}$ and $\boldsymbol{r}_m = \boldsymbol{p} \times \boldsymbol{r}_n$ is referred to as *moment vector*. A point $\boldsymbol{q}$ lies on a line if and only if $\boldsymbol{q} \times \boldsymbol{r}_n = \boldsymbol{r}_m$. When $\boldsymbol{r}_n$ is a unit vector, the norm of the moment gives the distance from the origin to the line. This means, given a set of correspondences $(\boldsymbol{v}, \mathbf{r}) \in \mathcal{M}$, we can minimize

$$\min \sum_{(\boldsymbol{v},\mathbf{r}) \in \mathcal{M}} ||\boldsymbol{v} \times \boldsymbol{r}_n - \boldsymbol{r}_m||. \tag{3.12}$$

to recover a 3D shape that maximizes the silhouette overlap. In Paper B, all silhouette rays are transformed into a joint canonical representation. Therefore, we can jointly optimize for a consensus shape that maximizes the silhouette overlap in all frames. Additionally, we do not have to differentiate through the SMPL blend shape-based posing and an image formation function, what makes the optimization very efficient and less memory-intensive.

3.2.4 Shape-from-shading

Shape-from-shading [92, 276] is a classical Computer Vision technique developed by Horn in the 1970s. The motivation behind shape-from-shading is that the shading of a 3D object is a strong cue for its 3D shape. Shading refers to the different levels of darkness of an originally uniformly colored surface caused by the illumination and the shape and properties of the surface. For shape-from-shading, we assume the shading of a pixel at position $[x, y]$ in a brightness image $\mathbf{B}$ only depends on the normal $\boldsymbol{n}$ of the surface point projecting into the pixel and the scene reflectance map P:

$$\mathbf{B}_{x,y} = P(\boldsymbol{n}_{x,y}) \tag{3.13}$$

Further, we assume a Lambertian reflectance model. This means shading forms from the dot product of the light direction and the surface normal:

$$\mathbf{B}_{x,y} = \cos(\boldsymbol{l}, \boldsymbol{n}_{x,y}) = \frac{\boldsymbol{l}}{|\boldsymbol{l}|} \cdot \frac{\boldsymbol{n}_{x,y}}{|\boldsymbol{n}_{x,y}|}. \tag{3.14}$$

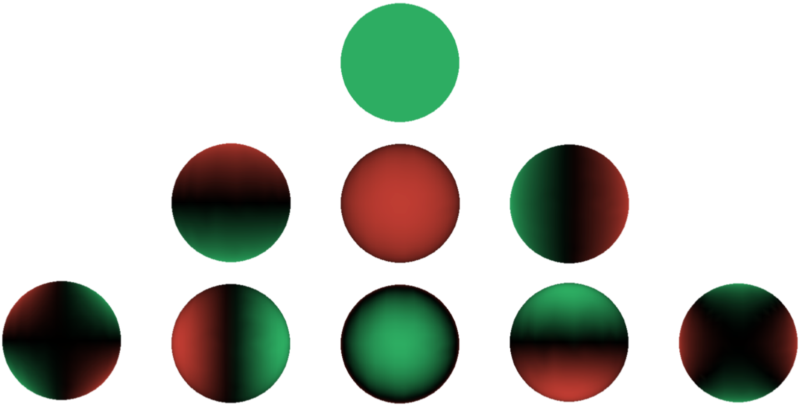

Figure 3.4: The first nine spherical harmonics visualized on the unit sphere. Positive values are green and negative values are red.

Using this simple illumination model, we can recover surface normals. Then, we can optimize for a 3D shape that explains the estimated normals.

In Paper C, we refine the estimated SMPL surface using shape-from-shading. Similar to the silhouette ray-based shape estimation step, we estimate the refinement in the canonical representation. This way, we can fuse noisy estimates from many frames and solve for the whole shape refinement jointly. Further, we use a more advanced illumination model. Instead of modeling the illumination using a single light direction $\boldsymbol{l}$, we utilize spherical harmonic lighting. Spherical harmonics are orthogonal basis functions defined over the surface of the sphere (Figure 3.4). Spherical harmonic lighting uses the first nine spherical harmonics to describe the directions from where light is shining into the scene [189]. By using spherical harmonic lighting, we can describe realistic illumination conditions with the low dimensional vector $\boldsymbol{c} \in \mathbb{R}^9$. We further use spherical harmonic lighting in Paper E for synthesizing realistic images of humans. To create realistic illumination conditions, we convert images of a light probe dataset into diffuse spherical harmonics coefficients $\boldsymbol{c}$. In contrast to Paper C, Paper E does not explicitly utilize shape-from-shading. Instead, we train a neural network to reconstruct the 3D shape of subjects. In the following, we give an introduction to Deep Learning and describe learning-based methods which are used in the different works of this dissertation.

3.3 Deep Learning

Learning-based methods follow a fundamentally different approach than optimization-based methods that build on the analysis-by-synthesis methodology. Learning-based methods find a parameterization for a complex function, a so-called neural network, such that it produces the desired output. This is achieved by tuning the parameters with the help of a large number of input-output pairs. The process of finding these parameters is called training the network. Recently, especially convolutional neural networks (CNNs) have revolutionized the field of Computer Vision. Their performance on various tasks beats classical methods often by far. Further, they provide solutions to new tasks where classical methods find their limits. While analysis-by-synthesis can play a role during the training of neural networks, their functioning is very different. Neural networks and similar learning architectures extract high-dimensional features from input variables, for example, from images. These features are then mapped to the task-specific result space. In the following we give a basic introduction to neural networks and CNNs. Systematic introductions to the Deep Learning methodology, mathematical backgrounds, and specific models are given by Goodfellow et al. [75] and Bishop [22].

A basic neural network is composed of several small functions, the so-called *neurons*, that are organized in layers. Each neuron receives an input vector $\boldsymbol{x}$. It computes a weighted average based on learnable weights $\boldsymbol{w}$ and adds a learnable bias b:

$$y = \boldsymbol{w}^\top \boldsymbol{x} + b. \tag{3.15}$$

Finally, the result is passed through a non-linear function $h(\cdot)$:

$$a = h(y). \tag{3.16}$$

The function $h(\cdot)$ is called *activation*. Typical activation functions are Rectified Linear Units (ReLU) or the Sigmoid function. To compose a complete layer l, we stack $i \in \{1, \dots, I\}$ neurons of the form:

$$y_i^{[l]} = \boldsymbol{w}_i^{[l]\top} \boldsymbol{a}^{[l-i]} + b_i^{[l]} \tag{3.17}$$

$$a_i^{[l]} = h^{[l]}(y_i^{[l]}). \tag{3.18}$$

The neurons of a layer l can be vectorized to:

$$\boldsymbol{y}^{[l]} = \mathbf{W}^{[l]} \boldsymbol{a}^{[l-1]} + \boldsymbol{b}^{[l]}. \tag{3.19}$$

$$\boldsymbol{a}^{[l]} = h^{[l]}(\boldsymbol{y}^{[l]}) \tag{3.20}$$

A set of layers forms a basic neural network. The number of neurons per layer and the types of activations can be freely chosen. A network composed of many layers is called a *deep* neural network. Different combinations of different layer sizes and activations, referred to as *architectures*, determine the complexity of the network and therefore its computational capabilities.

To train the neural network so that it computes the desired output, one optimizes over a *loss function* with respect to the network parameters; its weights $\mathbf{W}$ and biases $\boldsymbol{b}$. The loss function measures the distance of the computed output from the desired output. Given a specific loss value, we can compute partial derivatives with respect to the parameters of the neural network. Finally, we can update the parameters to decrease the loss value. During training, many examples are computed and the network parameters are adjusted to minimize the overall loss value.

In Computer Vision, CNNs, a variant of neural networks, became popular. Instead of computing a weighted average over the whole input vector $\boldsymbol{x}$, one applies a *convolution* over a local neighborhood. In other words, each element y is dependent on only a subset of $\boldsymbol{x}$. A convolution of f and g is defined as:

$$(f * g)[n] = \sum_{m=-K}^{K} f[m]g[n-m]. \tag{3.21}$$

Hereby, f is referred to as *kernel*. In a CNN, a kernel consists of learnable weights and is shared across the input to its layer. This has three main benefits: First, the number of network parameters is comparably low, even for large inputs. Second, the dimensionality of the input vector can vary. And finally, the network is (approximate) invariant to translations in the inputs.

In this work, we utilize Deep Learning methods in two ways: First, we use pre-trained models to calculate the beforehand mentioned abstractions including, but not limited to, foreground segmentation, semantic segmentation, reflectance and shading separation, and keypoint localization. Second, we provide algorithms that make use of Deep Learning at their core. In Paper D and Paper E, we perform the change from optimization-based methods to learning-based models. This allows us to significantly reduce the number of input frames, while at the same time robustify our methods. The questioning for methods utilizing Deep Learning is, however, different from the classical methods. Instead of designing a specific algorithm for the problem, we focus more on what data is needed and how the problem and data can efficiently be represented. Most importantly, Deep Learning methods require large amounts of data. For solving a task using Deep Learning, one has to find the right data and formulate matching supervision losses. This problem again is closely related to optimization. Methods from optimization can be exploited to make use of certain data. For example, often data is only annotated in the image domain. However, we can still supervise a three-dimensional problem, by utilizing ideas from analysis-by-synthesis to construct loss functions in the image domain. Using noisy, limited, or imprecise sources is referred to as *weak supervision*. In contrast, *full supervision* refers to comparing the results directly with ground truth data.

The problems of finding data and matching supervision losses are tackled in the works of Paper D and Paper E. Hereby, the SMPL model again serves as a base template and approximate solution. We incorporate the model as fixed algorithmic layers into our neural networks or utilize its parametrization to represent our data. This way we effectively reduce the dimensionality of the output space and align training data in a common parametrization.

In the following chapters we present our published works and additional material on human pose and shape reconstruction.

Publications

A Optical Flow-based 3D Human Motion Estimation from Monocular Video

Thiemo Alldieck[1], Marc Kassubeck[1], Bastian Wandt[2],
Bodo Rosenhahn[2], and Marcus Magnor[1]

[1] Computer Graphics Lab, TU Braunschweig

[2] Institut für Informationsverarbeitung, Leibniz Universität Hannover

Abstract

This paper presents a method to estimate 3D human pose and body shape from monocular videos. While recent approaches infer the 3D pose from silhouettes and landmarks, we exploit properties of optical flow to temporally constrain the reconstructed motion. We estimate human motion by minimizing the difference between computed flow fields and the output of our novel flow renderer. By just using a single semi-automatic initialization step, we are able to reconstruct monocular sequences without joint annotation. Our test scenarios demonstrate that optical flow effectively regularizes the under-constrained problem of human shape and motion estimation from monocular video.

Published in *German Conference on Pattern Recognition*. Springer, 2017, pp. 347–360.

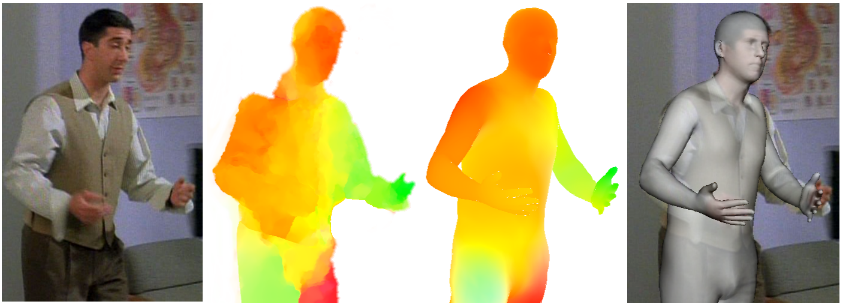

Figure A.1: Following our main idea we compute the optical flow between two consecutive frames and match it to an optical flow field estimated by our proposed optical flow renderer. From left to right: input frame, color-coded observed flow, estimated flow, resulting pose.

A.1 Introduction

Human pose estimation from video sequences has been an active field of research over the past decades with various applications such as surveillance, medical diagnostics or human-computer interfaces [151]. One branch of human pose estimation is referred to as *articulated motion parsing* [289], which defines the combination of monocular pose estimation and motion tracking in uncontrolled environments. We present a new approach to temporally coherent human shape and motion estimation in uncontrolled monocular video sequences. Our work follows the *generative* strategy, where both pose and shape parameters of a 3D body model are found to match the input image through analysis-by-synthesis [143].

The 3D pose of a human figure is highly ambiguous when inferred from only a 2D image. Common generative approaches [76, 85, 41] try to find human poses that are a good match to given silhouettes. However, human silhouettes can often be explained by multiple poses [76]. Existing methods for landmark-based 3D human motion estimation from monocular images [188, 215, 2, 281, 282] can find a pose per frame independently. Although 3D reconstructions from both approaches look very convincing on single images, they can result in significant jumps in position and joint angles between two successive frames. This creates highly unrealistic 3D reconstructions in the temporal domain. Temporal consistency of tracked landmarks is only considered by few researchers [191, 249, 250].

In our work we exploit the properties of the optical flow in the sequence to not only enforce temporal coherence but also resolve the pose ambiguities of purely silhouette-based or landmark-based approaches. We develop a motion tracker based on our novel optical flow renderer. Optical flow has proven to improve 2D tracking while also sharing much of the properties of range data [200]. By exploiting properties of the optical flow we construct a robust and stable 3D human motion tracker working on monocular image sequences.

The main idea behind our work is that the optical flow between two consecutive frames largely depends on the change of the human pose between them. Following this idea, we propose an energy minimization problem that infers those model parameters that minimize the distance between observed and rendered flow for two input frames (Fig. A.1). Additional energy terms are derived based on typical constraints of the human body, namely joint angle limits, limb interpenetration and continuous motion. For stable tracking, silhouette coverage is enforced.

We evaluate the proposed method using two well known datasets. We analyze the performance of our approach qualitatively and evaluate its 3D and 2D precision quantitatively. In the first experiment, 3D joint positions are compared against ground truth of the HumanEva-I dataset [214] and results of two recently published methods [26, 250]. The second evaluation compares projected 2D joint positions against ground truth of the VideoPose 2.0 dataset [203] featuring camera movement and rapid gesticulation. We compare our results against a recent deep-learning-based method for joint localization [177]. Results demonstrate the strengths and potential of the proposed method.

Summarizing, our contributions are:

- We develop a novel optical flow renderer for analysis-by-synthesis.
- We propose a complete pipeline for 3D reconstruction of human poses from monocular image sequences, that is independent of previous annotations of joints. It only uses a single semi-automatic initialization step.
- Optical flow is exploited to retrieve 3D information and achieve temporal coherence, instead of solely relying on silhouette information.

A.2 Related Work

Human pose estimation is a broad and active field of research. Here, we focus on model-based approaches and work that exploits optical flow during pose estimation.

Human pose from images. 3D human pose estimation is often based on the use of a body model. Human body representations exist in 2D and 3D. Many of the following methods utilize the 3D human body model SCAPE [14]. SCAPE is a deformable mesh model learned from body scans. Pose and shape of the model are parametrized by a set of body part rotations and low dimensional shape deformations. In recent work the SMPL model, a more accurate blend shape model compatible with existing rendering engines, has been presented by Loper et al. [138].

A variety of approaches to 3D pose estimation have been presented using various cues including shape from shading, silhouettes and edges. Due to the highly ill-posed and under-constrained nature of the problem these methods often require user interaction e.g. through manual annotation of body joints on the image. Guan et al. [76] have been the first to present a detailed method to recover human pose together with an accurate shape estimate from single images. Based on manual initialization, parameters of the SCAPE model are optimized exploiting edge overlap and shading. The work is based on [17], a method that recovers the 3D pose from silhouettes from 3-4 calibrated cameras. Similar methods requiring multi-view input have been presented, e.g. [16, 213, 192, 60]. Hasler et al. [85] fit their own statistical body model [84] into monocular image silhouettes with the help of sparse annotations. Chen et al. [41] infer 3D poses based on learned shape priors. In recent work, Bogo et al. [26] present the first method to extract both pose and shape from a single image fully automatically. 2D joint locations are found using the CNN-based approach DeepCut [177], then projected joints of the SMPL model are fitted against the 2D locations. In contrast to our work no consistency with the image silhouette or temporal coherency is taken into consideration.

Pose reconstruction for image based rendering. 3D human pose estimation can serve as a preliminary step for image based rendering techniques. In early work Carranza et al. [37] have been the first to present free-viewpoint video using model-based reconstruction of human motion using the subject's silhouette in multiple camera views. Zhou et al. [280] and Jain et al. [106] present updates to model-based pose estimation for subsequent reshaping of humans in images and videos respectively. Rogge et al. [198] fit a 3D model for automatic cloth exchange in videos. All methods utilize various cues, none of them uses optical flow for motion estimation.

Optical flow based methods. Previous work has exploited optical flow for different purposes. Sapp et al. [203] and Fragkiadaki et al. [63] use optical flow for segmentation as a preliminary step for pose estimation. Both exploit the rigid structure revealing property of optical flow, rather than information about motion. Fablet and Black [62] use optical flow to learn motion models for automatic detection of human motion. Efros et al. [58] categorize human motion viewed from a distance by building an optical flow-based motion descriptor. Both methods label motion without revealing the underlying movement pattern. In recent work, Romero et al. [200] present a method for 2D human pose estimation using optical flow only. They detect body parts by porting the random forest approach used by the Microsoft Kinect to use optical flow. Brox et al. [32] have shown that optical flow can be used for 3D pose tracking of rigid objects. They propose the use for objects *modeled as kinematic chains*. They argue that optical flow provides point correspondences inside the object contour which can help to identify a pose where silhouettes are ambiguous. Inspired by the above mentioned characteristics, we investigate the extent to which optical flow can be used for 3D human motion estimation from monocular video.

A.3 Method

Optical flow [71] is the perception of motion by our visual sense. For two successive video frames, it is described as a 2D vector field that matches a point in the first frame to the displaced point in the following frame [93]. Although calculated in the image plane, optical flow contains 3D information, as it can be interpreted as the projection of 3D scene flow [242]. Assuming the presence of optical flow in the sequence (i.e. all observed surfaces are diffuse, opaque and textured), the entire observed optical flow is caused by relative movement between object and camera. Besides the motion of individual body parts, optical flow contains information about boundaries of rigid structures and is an abstraction layer to the input images. Unique appearance effects such as texture and shading are removed [62, 200]. We argue that these features make optical flow highly suitable for generative optimization problems.

The presented method estimates pose parameters (i.e. joint angles), global position, and rotation of a human model (Sec. A.3.1) frame by frame. The procedure only requires a single semi-automatic initialization step (Sec. A.3.6) and then runs automatically. The parameters for each frame are inferred by minimizing the difference between the observed and rendered flow (Sec. A.3.3) from our flow renderer (Sec. A.3.2). A set of energy functions based on pose constraints (Sec. A.3.4) and silhouettes (Sec. A.3.5) is defined to regularize the solution to meaningful poses and to make the method more robust.

A.3.1 Scene model

In this work, we use the human body model SMPL [138]. The model can be reshaped using 10 shape parameters $\boldsymbol{\beta}$. For different poses, 72 pose parameters $\boldsymbol{\theta}$ can be set, including global orientation. $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ produce realistic vertex transformations and cover a large range of body shapes and poses. We define $(\gamma, \boldsymbol{\beta}, \boldsymbol{\theta}_i, \boldsymbol{\sigma}_i)$ as the model state at time step i, with global translation vector $\boldsymbol{\sigma}$ and gender γ. Here, for simplicity we assume that the camera positions and rotations as well as its focal lengths are known and static. It is however not required that the cameras of the actual scene are fixed, as the body model can rotate and move around the camera (cf. Sec. A.4).

A.3.2 Flow renderer

The core of the presented method is our differential flow renderer built upon OpenDR [139], a powerful open source framework for analysis-by-synthesis. The rendered flow image depends on the vertex locations determined by the virtual human model's pose parameters $\boldsymbol{\theta}$ and its translation $\boldsymbol{\sigma}$. To be able to render the flow *in situ*, we calculate the flow from frame i to $i-1$, referred to as backward flow. With this approach each pixel, and more importantly, each vertex location contains the information where it came from rather than were it went and can be rendered in place. The calculation of the flow is achieved as follows: The first step calculates the displacement of all vertices between two frames i and j in the image plane. Then the flow per pixel is calculated through barycentric interpolation of the neighboring vertices. Visibility and barycentric coordinates are calculated through the standard OpenGL rendering pipeline.

The core feature of the utilized rendering framework OpenDR is the differentiability of the rendering pipeline. To benefit from that property, our renderer estimates the partial derivatives of each flow vector with respect to each projected vertex position.

A.3.3 Flow matching

Having a flow renderer available, we can formulate the pose estimation as an optimization problem. The cost function E_f over all pixels p is defined as follows:

$$E_f = \sum_p ||F_\mathrm{o}(i, i-1, p) - F_\mathrm{r}(i, i-1, p)||^2 \tag{A.1}$$

where F_r refers to the *rendered* and F_o to the *observed* flow field calculated on the input frames i and $i-1$. The objective drives the optimization in such way that the rendered flow is similar to the observed flow (Fig. A.1). As proposed in [139], we evaluate E_f not over the flow field but over its Gaussian pyramid in order to perform a more global search.

For this work we use the method by Xu et al. [262] to calculate the observed optical flow field. The method has its strength in the ability to calculate large displacements while at the same time preserving motion details and handling occlusions. The definition of the objective shows that the performance of the optical flow estimation is crucial to the overall performance of the presented method. To compensate for inaccuracies of the flow estimation and to lower the accumulated error over time, we do not rely exclusively on the flow for pose estimation, but employ additional constraints as well (Sec. A.3.4 and Sec. A.3.5).

A.3.4 Pose constraints

SMPL does not define bounds for deformation. We introduce soft boundaries to constrain the joint angles in form of a cost function for pose estimation:

$$E_b = ||\max(e^{\boldsymbol{\theta}_\mathrm{min} - \boldsymbol{\theta}_i} - 1, 0) + \max(e^{\boldsymbol{\theta}_i - \boldsymbol{\theta}_\mathrm{max}} - 1, 0)||^2 \tag{A.2}$$

where $\boldsymbol{\theta}_\mathrm{min}$ and $\boldsymbol{\theta}_\mathrm{max}$ are empirical lower and upper boundaries and e and $\max$ are applied component-wise.

Furthermore, we introduce extended Kalman filtering per joint and linear Kalman filtering for translation. In addition to temporal smoothness, the Kalman filters are used to predict an *a priori* pose for the next frame before optimization, which significantly speeds up computation time.

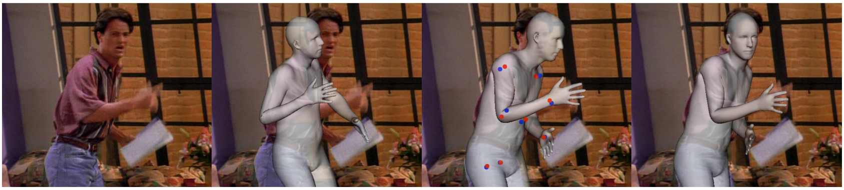

Figure A.2: Method initialization. Observed image, manual pose initialization, first optimization based on joint positions (red: model joints; blue: manually marked joints), final result including silhouette coverage and optical flow based correction.

During optimization the extremities of the model may intersect with other body parts. To prevent this, we integrate the interpenetration error term E_{sp} from [26]. The error term is defined over a capsule approximation of the body model. By using an error term interpenetration is not strictly prohibited but penalized.

A.3.5 Silhouette coverage

Pose estimation based on flow similarity requires that the rendered human model accurately covers the subject in the input image. Only body parts that cover the correct counterpart in the image can be moved correctly based on flow. To address inaccuracies caused by flow calculation, we introduce boundary matching.

We use the method presented by Bălan et al. [17] and adapt it to make it differentiable (cf. Sec. A.3.7). A cost function measures how well the model fits the image silhouette S_I by penalizing non-overlapping pixels by the shortest distance to the model silhouette S_M. For this purpose Chamfer distance maps C_I for the image silhouette and C_M for the model are calculated. The cost function is defined as:

$$E_c = \sum_p ||aS_{M_i}(p)C_I(p) + (1-a)S_I(p)C_{M_i}(p)||^2 \tag{A.3}$$

where a weighs $S_{M_i}C_I$ stronger as image silhouettes are wider to enforce the model to reside within in the image silhouette than to completely cover it. To be able to compute derivatives, we approximate C_M by calculating the shortest distance of each pixel to the model capsule approximation used for E_{sp}. The distance at p is the shortest distance among all distances to each capsule. To lower computation time, we calculate only a grid of values and interpolate in between.

A.3.6 Initialization

For the initialization of the presented method two manual steps are required. First the user sets the joints of the body model to a pose that roughly matches the observed pose. It is sufficient that only the main joints such as shoulder, elbow, hip and knee are manipulated. In a second step the user marks joint locations of hips, knees, ankles, shoulders, elbows and wrists in the first frame. If the position of a joint cannot be seen or estimated it may be skipped. From this point no further user input is needed.

The initialization is then performed in three steps (Fig. A.2). The first step minimizes the distance between the marked joints and their model counterparts projected to the image plane, while keeping E_{sp} and E_b low. We optimize over translation $\boldsymbol{\sigma}$, pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$. To guide the process we regularize both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ with objectives that penalize high differences to the manually set pose and the mean shape. In the second step we include the silhouette coverage objective E_c. Finally, we optimize the estimated pose for temporal consistency. We initialize the second frame with the intermediate initialization result and optimize on the flow field afterwards. While optimizing E_f we still allow updates for $\boldsymbol{\theta}_0$ and $\boldsymbol{\sigma}_0$.

A.3.7 Optimization

After initialization we now iteratively find each pose using the defined objectives. The final objective function is a weighted sum of the energy terms of the previous sections:

$$\min_{\boldsymbol{\sigma},\boldsymbol{\theta}}(\lambda_f E_f + \lambda_c E_c + \lambda_b E_b + \lambda_{sp} E_{sp} + \lambda_M E_M) \tag{A.4}$$

with scalar weights λ. E_M regularizes the current state with respect to the last state

$$E_M = ||\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}||^2 + ||\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_{i-1}||^2. \tag{A.5}$$

Each frame is initialized with the Kalman prediction as described in Sec. A.3.4.

For the optimization we use the OpenDR toolbox [139]. It allows for automatic differentiation of most partially differentiable functions. Therefore we

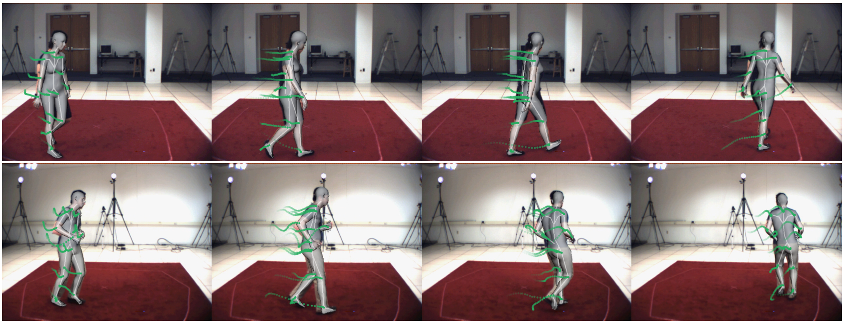

Figure A.3: Resultant poses of frames 30 to 120 of the HumanEva-I test sets. Green traces show the history of evaluated joints.

can avoid the laborious and inaccurate task of calculating finite differences. All our energy terms are designed to be fully or partially differentiable. Using this auto-differentiation we are able to optimize Eq. (A.4) efficiently.

A.4 Evaluation

We evaluate the 3D and 2D pose accuracy of the presented method using two publicly available datasets: HumanEva-I [214] and VideoPose2.0 [203]. Ground truth is available for both datasets. We compare our results in both tests, 3D and 2D, against state-of-the-art methods [26, 250, 177]. Foreground masks needed for our method have been hand-annotated using an open-source tool for image annotation[1].

HumanEva-I. The HumanEva-I datasets features different actions performed by 4 subjects filmed under laboratory conditions. We reconstruct 130 frames of the sets *Walking C1* by subject 1 and *Jog C2* by subject 2 without reinitialization. The camera focal length is known. We do not adjust our method for the dataset except setting the λ weights. Fig. A.3 shows a qualitative analysis. The green plots show the history of the joints used for evaluation. The traces demonstrate clearly the temporal coherence of the presented method. The low visual error in the last frames demonstrates that the presented method is robust over time.

---

[1] https://bitbucket.org/aauvap/multimodal-pixel-annotator

	Walking C1		Jog C2	
	local	global	local	global
Bogo et al. [26]	6.6	17.4	7.5	10.4
Wandt et al. [250]	5.7	34.0	**6.3**	38.0
Our method	**5.5**	**7.6**	7.9	**9.9**

Table A.1: Mean 3D joint error in cm for local per frame Procrustes alignment and global per sequence alignment.

	Chandler	Ross	Rachel
DeepCut [177]	25.3	**10.5**	32.8
Our method	**23.3**	21.9	**15.9**

Table A.2: Mean 2D joint error (shoulders, elbows, and wrists) in pixels.

We compare our method against the state-of-the-art methods of Bogo et al. [26] and Wandt et al. [250]. We use [26] without the linear pose regressor learned for the HumanEva sequences, which is missing in the publicly available source code. Frames that could not be reconstructed because of undetected joints have been excluded for evaluation. The 3D reconstruction of [250] is initialized with the same DeepCut [177] results as used for [26]. We measure the precision of the methods by calculating the *3D positioning error* as introduced by [215]. It calculates the mean euclidean distance of 13 reconstructed 3D joint locations to ground truth locations from MoCap data. Beforehand, optimal linear alignment of the results of all methods is achieved by Procrustes analysis. In order to demonstrate the global approach of our method, we follow two strategies here: First we measure the joint error after performing Procrustes alignment per frame. Afterwards we calculate a per sequence alignment over all joint locations in all frames and measure the resulting mean error. Table A.1 shows the result of all tests.

The results show that our method performs best in three of four test scenarios. In contrast to [26] and [250], our method does not require prior knowledge about the performed motion or training of plausible poses. The better performance of our method can be explained by the temporal coherent formulation using optical flow. This strength is especially noticeable in the global analysis. The method of [26] takes no temporal consistency into consideration, which results in jumps of joint locations between two frames and unresolved pose ambiguities (cf. Fig. A.6). Note that some frames cannot be reconstructed due

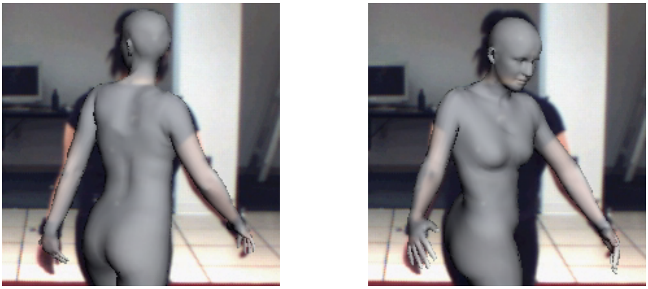

Figure A.4: Frame 120 of *Walking C1* in comparison to reconstruction with E_f set to zero.

to the joint detector failing to find a feasible skeleton. The algorithm of [250] also estimates the camera trajectory. A slightly wrongly estimated person size results in a global offset of the camera path and causes a larger global error. In order to demonstrate, that our method resolves ambiguities successfully, we conduct the experiment again with E_f set to zero. The resultant motion does no longer resemble the performed action (Fig. A.4) and the positioning error raises significantly to 9.8 and 15.9 for local and global analysis of *Walking C1* and 14.5 and 22.3 for *Jog C2* respectively.

VideoPose2.0. After evaluation with fixed camera and under laboratory conditions, we test our method under a more challenging setting. The second evaluation consists of three clips of the VideoPose2.0 dataset. We choose the "full-frame, every frame" (720 × 540px) variant in order to face camera movement. Ground truth is given in form of projected 2D location of shoulders, elbows, and wrists for every other frame. The camera focal length has been estimated.

We evaluate our method in 2D by comparison against DeepCut [177], the same method that has been used before as input for the 3D reconstruction methods. Table A.2 shows the mean euclidean distance to ground truth 2D joint locations. We use the first detected person by DeepCut and exclude several undetected joints from its evaluation. For our method, we project the reconstructed 3D joint locations to the image plane. The mixed performance of [177] is due to problems of the CNN with background objects. In order to enable fair comparison, we hand filter the results of [177] to foreground detections only and exclude several undetected joints. The comparison shows that our method produces similar precision while providing much more information. However,

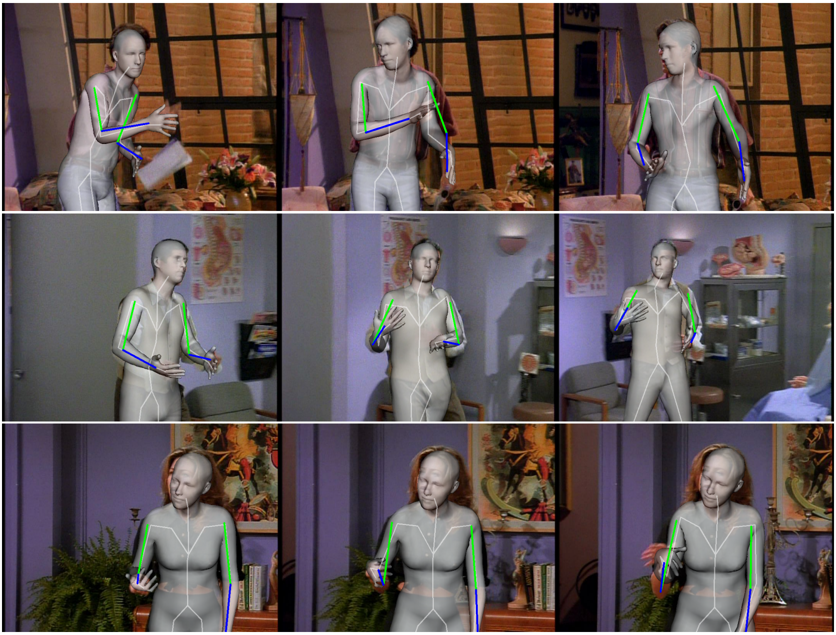

Figure A.5: Resultant poses of frames 1, 21 and 41 of the VideoPose2.0 sets (Chandler, Ross, Rachel) with ground truth arm locations (green and blue).

the increasing performance of CNN-based methods suggests that our method can benefit from semantic scene information for reinitialization in future work.

A.5 Conclusions

We have presented a new method for estimating 3D human motion from monocular video footage. The approach utilizes optical flow to recover human motion over time from a single initialization frame. For this purpose a novel flow renderer has been developed that enables direct interpretation of optical flow. The rich human body model SMPL provides the description of estimated human motion. Different test cases have shown applicability and robustness of the approach.

The presented method is dependent on realistic flow fields and good segmentation. It finds its natural limitations in the typical limits of optical flow

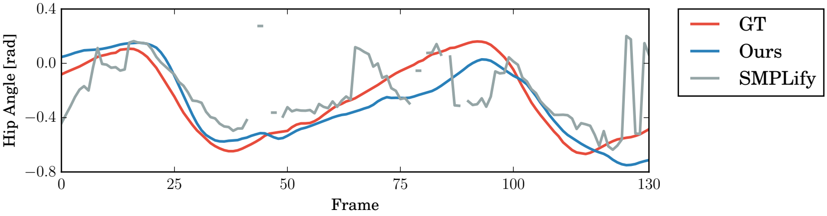

Figure A.6: Temporal behavior of the left hip angle of our method for *Walking C1* in comparison against ground truth (GT) and Bogo et al. (SMPLify) [26].

estimation. Improvements in optical flow estimation, especially multi-frame optical flow, can help to further improve our method. Although our temporal coherent formulation allows for a good occlusion handling, large occlusions and reappearances can still lead to tracking errors.

Our work is focused on automatic estimation of human motion from monocular video. In future work we plan to further automatize our method. The method might benefit from recent developments in semantic segmentation [167] and joint angle priors [2]. Building upon the presented framework, the next steps are texturing of the model and geometry refinement, enabling new video editing and virtual reality applications.

B Video Based Reconstruction of 3D People Models

Thiemo Alldieck[1], Marcus Magnor[1], Weipeng Xu[2],
Christian Theobalt[2], and Gerard Pons-Moll[2]

[1] Computer Graphics Lab, TU Braunschweig

[2] Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

This paper describes a method to obtain accurate 3D body models and texture of arbitrary people from a single, monocular video in which a person is moving. Based on a parametric body model, we present a robust processing pipeline to infer 3D model shapes including clothed people with 4.5mm reconstruction accuracy. At the core of our approach is the transformation of dynamic body pose into a canonical frame of reference. Our main contribution is a method to transform the silhouette cones corresponding to dynamic human silhouettes to obtain a visual hull in a common reference frame. This enables efficient estimation of a consensus 3D shape, texture and implanted animation skeleton based on a large number of frames. Results on 4 different datasets demonstrate the effectiveness of our approach to produce accurate 3D models. Requiring only an RGB camera, our method enables everyone to create their own fully animatable digital double, e.g., for social VR applications or virtual try-on for online fashion shopping.

Published in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8387-8397.

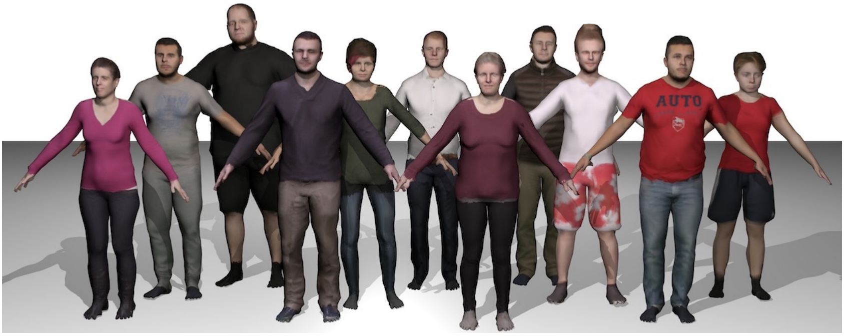

Figure B.1: Our technique allows to extract for the first time accurate 3D human body models, including hair and clothing, from a single video sequence of the person moving in front of the camera such that the person is seen from all sides.

B.1 Introduction

A personalized realistic and animatable 3D model of a human is required for many applications, including virtual and augmented reality, human tracking for surveillance, gaming, or biometrics. This model should comprise the person-specific static geometry of the body, hair and clothing, alongside a coherent surface texture.

One way to capture such models is to use expensive active scanners. But size and cost of such scanners prevent their use in consumer applications. Alternatively, multi-view passive reconstruction from a dense set of static body pose images can be used [65, 161]. However, it is hard for people to stand still for a long time, and so this process is time-consuming and error-prone. Also, consumer RGB-D cameras can be used to scan 3D body models [131], but these specialized sensors are not as widely available as video. Further, all these methods merely reconstruct surface shape and texture, but no rigged animation skeleton inside. All aforementioned applications would benefit from the ability to automatically reconstruct a personalized movable avatar from monocular RGB video.
Despite remarkable progress in reconstructing 3D body models [25, 254, 275] or free-form surface [287, 163, 170, 57] from depth data, 3D reconstruction of humans in clothing from monocular video (without a pre-recorded scan of the person) has not been addressed before. In this work, we estimate the shape of people in clothing from a single video in which the person moves. Some

methods infer shape parameters of a parametric body model from a single image [26, 55, 15, 85, 280, 106], but the reconstruction is limited to the parametric space and can not capture personalized shape detail and clothing geometry.

To estimate geometry from a video sequence, we could jointly optimize a single free-form shape constrained by a body model to fit a set of F images. Unfortunately, this requires to optimize F poses at once and more importantly it requires storing F models in memory during optimization which makes it computationally expensive and unpractical.

The key idea of our approach is to generalize visual hull methods [145] to monocular videos of people in motion. Standard visual hull methods capture a static shape from multiple views. Every camera ray through a silhouette point in the image casts a constraint on the 3D body shape. To make visual hulls work for monocular video of a moving person it is necessary to "undo" the human motion and bring it to a canonical frame of reference. In this work, the geometry of people (in wide or tight clothing) is represented as a deviation from the SMPL parametric body model [138] of naked people in a canonical T-pose; this model also features a pose-dependent non-rigid surface skinning. We first estimate an initial body shape and 3D pose at each frame by fitting the SMPL model to 2D detections similar to [124, 26]. Given such fits, we associate every silhouette point in every frame to a 3D point in the body model. We then transform every projection ray according to the inverse deformation model of its corresponding 3D model point; we call this operation unposing (Fig. B.3). After unposing the rays for all frames we obtain a visual hull that constrains the body shape in a canonical T-pose. We then jointly optimize body shape parameters and free-form vertex displacements to minimize the distance between 3D model points and unposed rays. This allows us to efficiently optimize a single displacement surface on top of SMPL constrained to fit all frames at once, which requires storing only one model in memory (Fig. B.2). Our technique allows for the first time extracting accurate 3D human body models, including hair and clothing, from a single video sequence of the person moving in front of the camera such that the person is seen from all sides.
Our results on several 3D datasets show that our method can reconstruct 3D human shape to a remarkable accuracy of 4.5 mm (even higher 3.1 mm with ground truth poses) despite monocular depth ambiguities. We provide our dataset and source code of our method for research purposes[1].

---

[1] https://graphics.tu-bs.de/people-snapshot

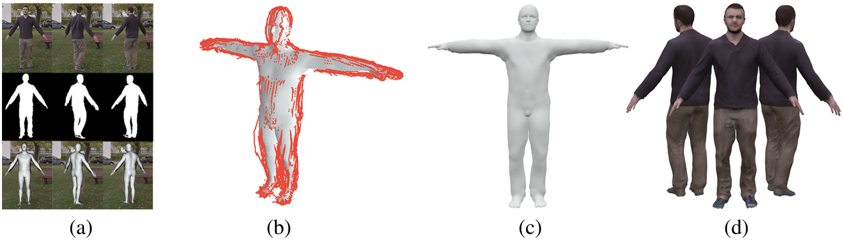
(a) (b) (c) (d)

Figure B.2: Overview of our method. The input to our method is an image sequence with corresponding segmentations. We first calculate poses using the SMPL model (a). Then we unpose silhouette camera rays (unposed silhouettes depicted in red) (b) and optimize for the subjects shape in the canonical T-pose (c). Finally, we are able to calculate a texture and generate a personalized blend shape model (d).

B.2 Related Work

Shape reconstruction of humans in clothing can be classified according to two criteria: (1) the type of sensor used and (2) the kind of template prior used for reconstruction. *Free-form* methods typically use multi-view cameras, depth cameras or fusion of sensors and reconstruct surface geometry quite accurately without using a strong prior on the shape. In more unconstrained and ambiguous settings, such as in the monocular case, a parametric body model helps to constrain the problem significantly. Here we review free-form and model-based methods and focus on methods for monocular images.

Free-form methods reconstruct the moving geometry by deforming a mesh [37, 51, 34] or using a volumetric representation of shape [94, 4]. The advantage of these methods is that they allow reconstruction of general dynamic shapes provided that a template surface is available initially. While flexible, such approaches require high-quality *multi-view* input data which makes them impractical for many applications. Only one approach showed reconstruction of human pose and deforming cloth geometry from monocular video using a pre-captured shape template [264]. Using a *depth camera*, systems like KinectFusion [104, 162] allow reconstruction of 3D rigid scenes and also appearance models [279] by incrementally fusing geometry in a canonical frame. A number of methods adapt KinectFusion for human body scanning [209, 131, 273, 46]. The problem is that these methods require separate shots at different time instances. The person thus needs to stand still while the camera is turned around, or subtle pose changes need to be explicitly compensated. The approach in [163] generalized KinectFusion to non-rigid objects. The approach performs non-rigid registration between the incoming depth frames and a concurrently updated, initially incomplete, template. While general, such template-free approaches [162, 100, 218] are limited to slow and careful motions. One way to make fusion and tracking more robust is by using multiple kinects [57, 170] or multi-view [223, 127, 45]; such methods achieve impressive reconstructions but do not register all frames to the same template and focus on different applications such as streaming or remote rendering for telepresence, e.g., in the holoportation project [170]. Pre-scanning the object or person to be tracked [287, 51] reduces the problem to tracking the non-rigid deformations. Some works are in-between free-form and model-based methods. In [66, 243] they pre-scan a template and insert a skeleton and in [271] they use a skeleton to regularize dynamic fusion. Our work is also related to the seminal work of [43, 44] where they align visual hulls over time to improve shape estimation. In the articulated case, they need to segment and track every body part separately and then merge the information together in a coarse voxel model; more importantly, they need multi-view input. In [117] they compensate for small motions of captured objects by de-blurring occupancy images but no results are shown for moving humans. In [285] they reconstruct the shape of clothed humans in outdoor environments from RGB video, requiring the subject to stand still. All these works use either multi-view systems, depth cameras or do not handle moving humans. In contrast, we use a single RGB video of a moving person, which makes the problem significantly harder as geometry can not be directly unwarped as it is done in depth fusion papers.

Model-based. Several works leverage a parametric body model for human pose and shape estimation from images [179]. Early models in computer vision were based on simple primitives [149, 69, 178, 212]. Recent ones are learned from thousands of scans of real people and encode pose, and shape deformations [14, 84, 138, 288, 182]. Some works reconstruct the body shape from *depth data* sequences [254, 86, 268, 275, 25] exploiting the temporal information. Typically, a single shape and multiple poses are optimized to exploit the temporal information. Using *multi-view* some works have shown performance capture outdoors [192, 194] by leveraging a sum of Gaussians body model [224] or using a pre-computed template [270]. A number of works are restricted to estimating the shape parameters of a body model [15, 76] from multiple views or single images with manually clicked points; silhouettes shading cues and color have been used for inference. Some works fit a body model to images using manual intervention [280, 106, 198] with the goal of image manipulation.

Shape and clothing from a single image is recovered in [79, 39] but the user needs to click points in the image and select the clothing types from a database. In [122] they obtain shape from contour drawings. The advance in 2D pose detection [253, 36, 102] has made 3D pose and shape estimation possible in challenging scenarios. In [26, 124] they fit a 3D body model [138] to 2D detections; since only model parameters are optimized and these methods heavily rely on 2D detections, results tend to be close to the shape space mean. In [5] they add a silhouette term to reduce this effect.

Shape Under Clothing. The aforementioned methods ignore clothing or treat it as noise, but a number of works explicitly reason about clothing. Typically, these methods incorporate constraints such as the body should lie inside the clothing silhouette. In [15] they estimate body shape under clothing by optimizing model parameters for a set of images of the same person in different clothing. In [260, 265] they exploit temporal sequences of scans to estimate shape under clothing. Results are usually restricted to the (naked) model space. In [274] they estimate detailed shape under clothing from scan sequences by optimizing a free-form surface constrained by a body model. The approach in [184] jointly captures clothing geometry and body shape using separate meshes but requires 3D scan sequences as input. DoubleFusion [226] reconstructs clothing geometry and inner body shape from a single depth camera in real time.

Learning based. Only very few works predict human shape from images using learning methods since images annotated with ground truth shape, pose and clothing geometry are hardly available. A few exceptions are the approach of [55] that predicts shape from silhouettes using a neural network and [49] that predicts garment geometry from a single image. Predictions in [55] are restricted to model shape space and tend to look over-smooth; only garments seen in the dataset can be recovered in [49]. Recent works leverage 2D annotations to train networks for the task of 3D pose estimation [146, 185, 283, 225, 236, 197]. Such works typically predict a stick figure or bone skeleton only, and can not estimate body shape or clothing.

B.3 Method

Given a single monocular RGB video depicting a moving person, our goal is to generate a personalized 3D model of the subject, which consists of the shape of body, hair and clothing, a personalized texture map, and an underlying skeleton rigged to the surface. Non-rigid surface deformations in new poses are thus entirely skeleton-driven. Our method consists of 3 steps: 1) *pose reconstruction* (Sec. B.3.2) 2) *consensus shape estimation* (Sec. B.3.3) and 3) *frame refinement and texture map generation* (Sec. B.3.4). Our main contribution is step 2), the consensus shape estimation; step 1) builds on previous work and step 3) to obtain texture and time-varying details is optional.

In order to estimate the consensus shape of the subject, we first calculate the 3D pose in each frame (Sec. B.3.2). We extend the method of [26] to make it more robust and enforce better temporal coherence and silhouette overlap. In the second step, the *consensus shape* is calculated as detailed in Sec. B.3.3. The consensus shape is efficiently optimized to maximally explain the silhouettes at each frame instance. Due to time-varying cloth deformations the posed consensus shape might be slightly misaligned with the frame silhouettes. Hence, in order to compute texture and capture time-varying details, in step 3) deviations from the consensus shape are optimized per frame in a sliding window approach (Sec. B.3.4). Given the refined frame-wise shapes we can compute the texture map. Our method relies on a foreground segmentation of the images. Therefore, we adopt the CNN based video segmentation method of [33] and train it with 3-4 manual segmentations per sequence. In order to counter ambiguities in monocular 3D human shape reconstruction, we use the SMPL body model [138] as starting point. In the following, we briefly explain how we adapt original SMPL body model for our problem formulation.

B.3.1 SMPL body model with offsets

SMPL is a parameterized model of naked humans that takes 72 pose and 10 shape parameters and returns a triangulated mesh with $N = 6890$ vertices. The shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ deformations are applied to a base template $\mathbf{T}$, which in the original SMPL model corresponds to the statistical mean shape in the training scans $\mathbf{T}_\mu$:

$$M(\boldsymbol{\beta},\boldsymbol{\theta}) = W(T(\boldsymbol{\beta},\boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{B.1}$$

$$T(\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{T}_\mu + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) \tag{B.2}$$

where W is a linear blend-skinning function applied to a rest pose $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ based on the skeleton joints $J(\boldsymbol{\beta})$ and after pose-dependent deformations $B_p(\boldsymbol{\theta})$ and shape dependent deformations $B_s(\boldsymbol{\beta})$ are applied. Shape-dependent deformations $B_s(\boldsymbol{\beta})$ model subject identity. However the Principal Component shape space of SMPL was learned from scans of naked humans, so clothing and other personal surface detail cannot be modeled. In order to personalize the SMPL model, we simply add a set of auxiliary variables or offsets $\mathbf{D} \in \mathbb{R}^{3N}$ from the template:

$$T(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D}) = \mathbf{T}_\mu + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D} \tag{B.3}$$

Such offsets $\mathbf{D}$ allow us to deform the model to better explain details and clothing. Offsets are optimized in step 2.

B.3.2 Pose reconstruction

The approach in [26] optimizes SMPL model parameters to fit a set of 2D joint detections in the image. As with any monocular method, scale is an inherent ambiguity. To mitigate this effect, we take inspiration from [192] and extend [26] such that it jointly considers $P = 5$ frames and optimizes a single shape and $P = 5$ poses. Note that optimizing many more frames would become computationally very expensive and many models would have to be simultaneously stored in memory. Our experiments reveal that even when optimizing over $P = 5$ poses the scale ambiguity prevails. The reason is that pose differences induce additional 3D ambiguities which cannot be uniquely decoupled from global size, even on multiple frames [228, 220, 181]. Hence, if the height of the person is known, we incorporate it as constraint during optimization. If height is not known the shape reconstructions of our method are still accurate up to a scale factor (height estimation is roughly off by 2-5 cm). The output of initialization are SMPL model shape parameters $\boldsymbol{\beta}_0$ that we keep fixed during subsequent frame-wise pose estimation. In order to estimate 3D pose more reliably, we extend [26] by incorporating a silhouette term:

$$E_{\text{silh}}(\boldsymbol{\theta}) = G(\mathbf{w}_{\text{o}}\mathbf{I}_{rn}(\boldsymbol{\theta})\mathbf{C} + \mathbf{w}_{\text{i}}(\mathbf{1} - \mathbf{I}_{rn}(\boldsymbol{\theta}))\bar{\mathbf{C}}) \tag{B.4}$$

with the silhouette image of the rendered model $\mathbf{I}_{rn}(\boldsymbol{\theta})$, distance transform of observed image mask $\mathbf{C}$ and its inverse $\bar{\mathbf{C}}$, weights $\mathbf{w}$. To be robust to local minima we optimize at 4 different levels of a Gaussian pyramid G. We further update the method to use state of the art 2D joint detections [36, 253] and a single-modal A-pose prior. We train the prior from SMPL poses fitted against

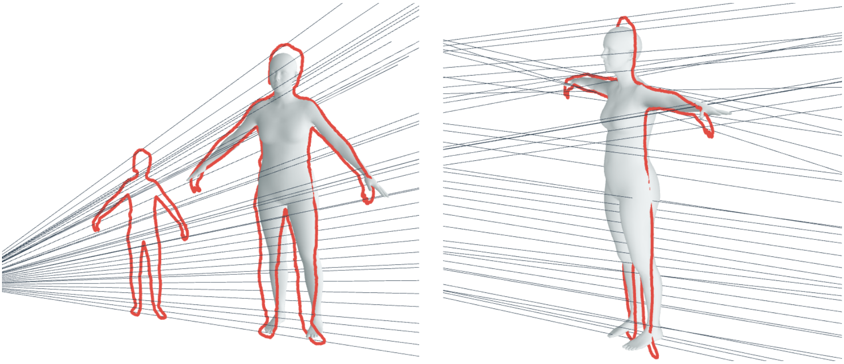

Figure B.3: The camera rays that form the image silhouette (left) are getting unposed into the canonical T-pose (right). This allows efficient shape optimization on a single model for multiple frames.

body scans of people in A-pose. Further, we enforce a temporal smoothness and initialize the pose in a new frame with the estimated pose $\boldsymbol{\theta}$ in the previous frame. If the objective error gets too large, we re-initialize the tracker by setting the pose to zero. While optimization in batches of frames would be beneficial it slows down computation and we have not found significant differences in pose accuracy. The output of this step is a set of poses $\{\boldsymbol{\theta}_p\}_{p=1}^{F}$ for the F frames in the sequence.

B.3.3 Consensus shape

Given the set of estimated poses we could jointly optimize a single refined shape matching all original F poses, which would yield a complex, non-convex optimization problem. Instead, we merge all the information into an unposed canonical frame, where refinement is computationally easier. At every frame a silhouette places a new constraint on the body shape; specifically, the set of rays going from the camera to the silhouette points define a constraint cone, see Fig. B.3. Since the person is moving, the pose is changing. Our key idea is to *unpose* the cone defined by the projection rays using the estimated poses. Effectively, we invert the SMPL function for every ray. In SMPL, every vertex $\boldsymbol{v}$ deforms according to the following equation:

$$\boldsymbol{v}'_i = \sum_{k=1}^{K} w_{k,i} G_k(\boldsymbol{\theta}, J(\boldsymbol{\beta}))(\boldsymbol{v}_i + b_{s,i}(\boldsymbol{\beta}) + b_{P,i}(\boldsymbol{\theta})) \tag{B.5}$$

where G_k is the global transformation of joint k and $b_{s,i}(\boldsymbol{\beta}) \in \mathbb{R}$ and $b_{P,i}(\boldsymbol{\theta})$ are elements of $B_s(\boldsymbol{\beta})$ and $B_p(\boldsymbol{\theta})$ corresponding to $i-th$ vertex. For every ray $\mathbf{r}$ we find its closest 3D model point. From Eq. (B.5) it follows that the inverse transformation applied to a ray $\mathbf{r}$ corresponding to model point $\boldsymbol{v}'_i$ is

$$\mathbf{r} = \left(\sum_{k=1}^{K} w_{k,i} G_k(\boldsymbol{\theta}, J(\boldsymbol{\beta})) \right)^{-1} \mathbf{r}' - b_{P,i}(\boldsymbol{\theta}). \tag{B.6}$$

Doing this for every ray effectively unposes the silhouette cone and places constraints on a canonical T-pose, see Fig. B.3. Unposing removes blend-shape calculations from the optimization problem and significantly reduces the memory foot-print of the method. Without unposing the vertex operations and the respective Jacobians would have to be computed for every frame at *every update* of the shape. Given the set of unposed rays for F silhouettes (we use $F = 120$ in all experiments), we formulate an optimization in the canonical frame

$$E_{\text{cons}} = E_{\text{data}} + w_{\text{lp}} E_{\text{lp}} + w_{\text{var}} E_{\text{var}} + w_{\text{sym}} E_{\text{sym}} \tag{B.7}$$

and minimize it with respect to shape parameters $\boldsymbol{\beta}$ of a template model and the vertex offsets $\mathbf{D}$ defined in Eq. B.3. The objective E_{cons} consists of a data term E_{data} and three regularization terms $E_{\text{lp}}, E_{\text{var}}, E_{\text{sym}}$ with weights w_* that balance its influence.

Data Term measures the distance between vertices and rays. Point to line distances can be efficiently computed expressing rays using Plucker coordinates $(\mathbf{r} = \boldsymbol{r}_m, \boldsymbol{r}_n)$. Given a set of correspondences $(\boldsymbol{v}_i, \mathbf{r}) \in \mathcal{M}$ the data term equals

$$E_{\text{data}} = \sum_{(\boldsymbol{v},\mathbf{r}) \in \mathcal{M}} \rho(\boldsymbol{v} \times \boldsymbol{r}_n - \boldsymbol{r}_m) \tag{B.8}$$

where ρ is the Geman-McClure robust cost function, here applied to the point to line distance. Since the canonical pose parameters are all zero ($\boldsymbol{\theta} = \mathbf{0}$) it follows from Eq. B.3 that vertex positions are a function of shape parameters and offsets $\boldsymbol{v}(\boldsymbol{\beta}_0, \mathbf{D}) = T_i(\boldsymbol{\beta}_0, \mathbf{D}) = (\boldsymbol{v}_{\mu,i} + b_{s,i}(\boldsymbol{\beta}_0) + \mathbf{d}_i)$, where $\mathbf{d}_i \in \mathbb{R}^3$ is the offset in $\mathbf{D}$ corresponding to vertex $\boldsymbol{v}_i$. In our notation, we remove the dependency on parameters for clarity. The remaining terms regularize the optimization.

Laplacian Term. We enforce smooth deformation by adding the Laplacian mesh regularizer [222]:

$$E_{\text{lp}} = \sum_{i=1}^{N} \tau_{l,i} ||L(\boldsymbol{v}_i) - \delta_i||^2 \tag{B.9}$$

where $\delta = L(\boldsymbol{v}(\boldsymbol{\beta}_0, \mathbf{0}))$ and L is the Laplace operator. The term forces the Laplacian of the optimized mesh to be similar to the Laplacian of the mesh at initialization (where offsets $\mathbf{D} = \mathbf{0}$).

Body Model Term. We penalize deviations of the reconstructed free-form vertices $\boldsymbol{v}(\boldsymbol{\beta}_0, \mathbf{D})$ from vertices explained by the SMPL model $\boldsymbol{v}(\boldsymbol{\beta}, \mathbf{0})$:

$$E_{\text{var}} = \sum_{i=1}^{N} \tau_{v,i} ||\boldsymbol{v}_i(\boldsymbol{\beta}_0, \mathbf{D}) - \boldsymbol{v}_i(\boldsymbol{\beta}, \mathbf{0})||^2 \tag{B.10}$$

Symmetry Term. Humans are usually axially symmetrical with respect to the Y-axis. Since the body model is nearly symmetric, we add a constraint on the offsets alone that enforces a symmetrical shape:

$$E_{\text{sym}} = \sum_{(i,j) \in \mathcal{S}} \tau_{s,i,j} \left|\left| [-1, 1, 1]^T \cdot \mathbf{d}_i - \mathbf{d}_j \right|\right|^2 \tag{B.11}$$

where $\mathcal{S}$ contains all pairs of Y-symmetric vertices. We phrase this as a soft-constraint to allow potential asymmetries in clothing wrinkles and body shapes. Since the refined consensus shape still has the mesh topology of SMPL, we can apply the pose-based deformation space of SMPL to simulate surface deformation in new skeleton poses.

Implementation Details. Body regions that are typically unclothed or where silhouettes are noisy (face, ears, hands, and feet) are more regularized towards the body model using per-vertex weights $\boldsymbol{\tau}$. We optimize E_{cons} using a "dog-leg" trust region method using the chumpy auto-differentiation framework. We alternate minimizing E_{cons} with respect to model parameters and offsets and finding point to line correspondences. We also re-initialize E_{lp}, E_{var}, E_{sym}. More implementation details and runtime metrics are given in the supplementary material.

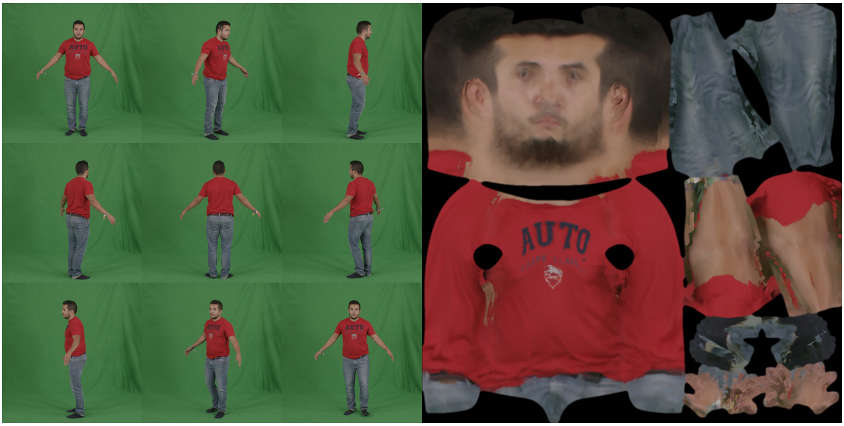

Figure B.4: We back-project the image color from several frames to all visible vertices to generate a full texture map.

B.3.4 Frame refinement and texture generation

After calculating a *global* shape for the given sequence, we aim to capture the temporal variations. We adapt the energy in Eq. B.7 to process frames sequentially. The optimization is initialized with the preceding frame and regularized with neighboring frames:

$$E_{\text{ref},j} = \sum_{j=f-m}^{f+m} \psi_j E_{\text{data},j} + w_{\text{var}} E_{\text{var},j} + w_{\text{lp}} E_{\text{lp},j} + w_{\text{last}} E_{\text{last},j} \tag{B.12}$$

where $\psi_j = 1$ for $j = k$ and $\psi_j = w_{\text{neigh}} < 1$ for neighboring frames. Hence, w_{neigh} defines the influence of neighboring frames and E_{last} regularizes the reconstruction to the result of the preceding frame. To create the texture, we warp our estimated canonical model back to each frame, back-project the image color to all visible vertices, and finally generate a texture image by calculating the median of the most orthogonal texels from all views. An example of keyframes we use for texture mapping and the resulting texture image is shown in Fig. B.4.

B.4 Experiments

We study the effectiveness of our method, qualitatively and quantitatively, in different scenarios. For quantitative evaluation, we used two publicly available datasets consisting of 3D scan sequences of humans in motion: with minimal clothing (MC) (DynamicFAUST [27]) and with clothing (BUFF [274]). Since these datasets were recorded without RGB sensors we simply render images of the scans using a virtual camera and use them as input. In order to evaluate our method on more varied clothing and backgrounds, we captured a new test dataset (People-Snapshot dataset), and present qualitative results. To the best of our knowledge, our method is the first approach that enables detailed human body model reconstruction in clothing from a single monocular RGB video without requiring a pre-scanned template or manually clicked points. Thus, there exist no methods with the same setting as ours. Hence, we provide a quantitative comparison to the state-of-the-art RGB-D based approach KinectCap [25] on their dataset. The image sequences and ground truth scans were provided by the authors of [25]. While reconstruction from monocular videos is much harder than from depth videos, a comparison is still informative. In all experiments, the method's parameters are set to two constant values, one set for clothed and one set for people in MC, which are empirically determined.

B.4.1 Results on rendered images

We take all 9 sequences of 5 different subjects in the BUFF dataset and all 9 sequences of 9 subjects from the DynamicFaust dataset performing "Hip" movements, featuring strong fabric movement or soft tissue dynamics respectively. Each dynamic sequence consists of 300-800 frames. To simulate the subject rotating in front of a camera, we create a virtual camera at 2.5 meters away from the 3D scans of the subject. We rotate the camera in a circle around the person moving one time per sequence. The foreground masks are easily obtained from the alpha channel of the rendered images. For BUFF we render images with real dynamic textures; for DynamicFAUST since textures are not available we rendered shaded models.

In Fig. B.6, we show some examples of our reconstruction results on image sequences rendered from BUFF and DynamicFAUST scans. The complete results of all 9 sequences are provided in the supplementary material. To be able to quantitatively evaluate the reconstruction quality, we adjust the pose and scale of our reconstruction to match the ground truth body scans following [274, 25].

D-FAUST

Subject ID	*full method*	*GT poses*
50002	5.13 ±6.43	3.92 ±4.49
50004	4.36 ±4.67	2.95 ±3.11
50009	3.72 ±3.76	2.56 ±2.50
50020	3.32 ±3.04	2.27 ±2.06
50021	4.45 ±4.05	3.00 ±2.66
50022	5.71 ±5.78	2.96 ±2.97
50025	4.84 ±4.75	2.92 ±2.94
50026	4.56 ±4.83	2.62 ±2.48
50027	3.89 ±3.57	2.55 ±2.33

BUFF

Subject ID		*full method*	*GT poses*
t-shirt, long pants	00005	5.07 ±5.74	3.80 ±4.13
	00032	4.84 ±5.25	3.37 ±3.59
	00096	5.57 ±6.54	4.35 ±4.66
	00114	4.22 ±5.12	3.14 ±2.99
	03223	4.85 ±4.80	2.87 ±2.58
soccer outfit	00005	5.35 ±6.67	3.82 ±3.67
	00032	7.95 ±8.62	3.04 ±3.39
	00114	4.97 ±5.81	3.01 ±2.80
	03223	5.49 ±5.71	3.21 ±3.28

KinectCap

Subject ID		*Subject ID*	
00009	4.07 ±4.20	02909	3.94 ±4.80
00043	4.30 ±4.39	03122	3.21 ±2.85
00059	3.87 ±3.96	03123	3.68 ±3.22
00114	4.85 ±4.93	03124	3.67 ±3.31
00118	3.79 ±3.80	03126	4.89 ±6.12

Table B.1: Numerical evaluation on 3 different datasets with ground truth 3D shapes. On D-FAUST and BUFF we rendered the ground truth scans on a virtual camera (see text), KinectCap already included images. We report for every subject the average surface to surface distance (see text). On BUFF, D-FAUST and KinectCap we achieve mean average errors of 5.37mm, 4.44mm, 3.97mm respectively. As expected best results are obtained using ground truth poses. Perhaps surprisingly, the results (3.40 mm for BUFF, 2.86 for D-FAUST) do not differ much from the average errors of the full pipeline. This demonstrates that our approach is robust to inaccuracies in 3D pose estimation.

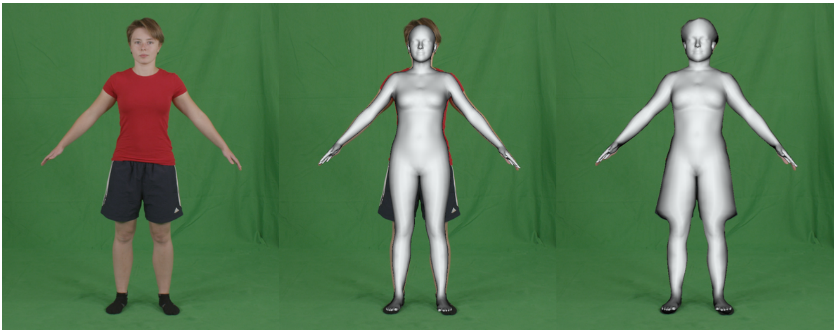

Figure B.5: Comparison to the monocular model-based method [26] (left to right) input frame, SMPLify, consensus shape. To make a fair comparison we extended [26] to multiple views as well. Compared to pure model-based methods, our approach captures also medium level geometry details from a single RGB camera.

Then, we compute a bi-directional vertex to surface distance between our reconstruction and the ground truth geometry. Per-vertex errors (in millimeters) on all sequences are provided in Tab. B.1. The heatmaps of per-vertex errors are shown in Fig. B.6. As can be seen, our method yields accurate reconstruction on all sequences including personalized details. To study the importance of the pose estimation component, we report the accuracy of our method using *ground truth poses* versus using estimated poses *full method*. Ground truth poses were obtained by registering SMPL to the 3D scans. The results of the ablation evaluation are also shown in Fig. B.6 and Tab. B.1. We can see that our complete pipeline achieved comparable accuracy with the one using ground truth poses which demonstrates robustness. Results show that there is still room for improvement in 3D pose reconstruction.

B.4.2 Qualitative results on RGB images

We also evaluate our method on real image sequences. The People-Snapshot dataset consists of 24 sequences of 11 subjects varying a lot in height and weight. The sequences are captured with a fixed camera, and we ask the subjects to rotate while holding an A-pose. To cover a variety of clothing, lighting conditions and background, the subjects were captured with varying sets of garments and with three different background scenes: in the studio with green screen, outdoor, and indoor with complex dynamic background. Some exam-

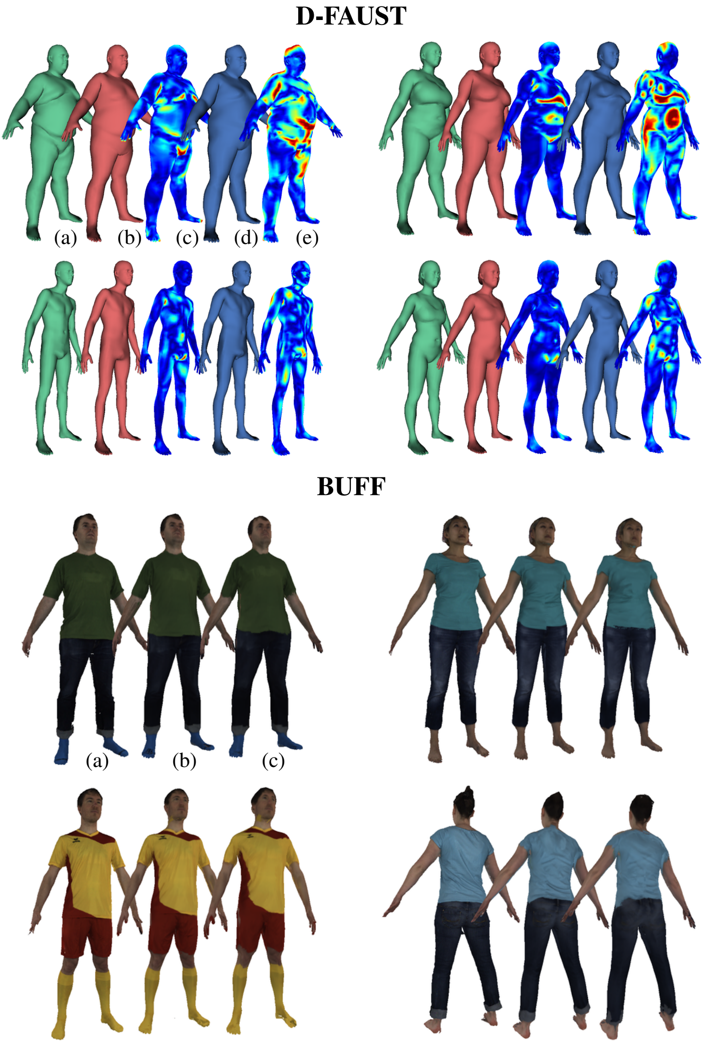

Figure B.6: Our results on image sequences from BUFF and D-FAUST datasets. D-FAUST: (a) ground truth 3D scan, (b) consensus shape with ground truth poses (consensus-p), (c) consensus-p heatmap, (d) consensus shape (consensus), (e) consensus heat-map (blue means 0mm, red means $\geq$ 2cm). Textured results on BUFF: (a) ground truth scan, (b) consensus-p (c) consensus.

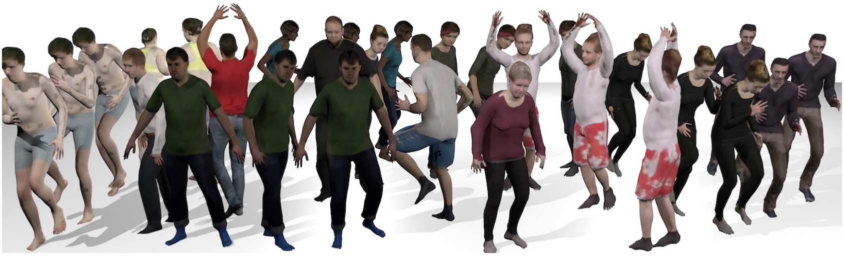

Figure B.7: Qualitative results: since the reconstructed templates share the topology with the SMPL body model we can use SMPL to change the *pose and shape* of our reconstructions. While SMPL does not model clothing deformations the deformed templates look plausible and maybe of sufficient quality for several applications.

ples of our reconstruction results are shown in Fig. B.7 and Fig. B.1. We show more example in the supplementary material and in the video. We can see that our method yields detailed reconstructions of similar quality as the results on rendered sequences, which demonstrates that our method generalizes well on the real world scenarios. The benefits of our method are further evidenced by overlaying the re-posed final reconstruction on to the input images. As shown in Fig. B.8, our reconstructions precisely overlay the body silhouettes in the input images.

B.4.3 Comparison with KinectCap

We compare our method to [25] on their collected dataset. Subjects were captured in both A-pose and T-poses in this dataset. Since T-poses (zero-pose in SMPL) are rather unnatural, they are not well captured in our general pose-prior. Hence, we adjust our pose prior to contain also T-poses. Note that their method relies on depth data, while ours only uses the RGB images. Notably, our method obtains comparable results qualitatively and quantitatively despite solving a much more ill-posed problem. This is further evidenced by the per-vertex errors in Tab. B.1.

B.4.4 Surface refinement using shading

As mentioned before, our method captures both body shape and medium level surface geometry. In contrast to pure model-based methods, we already add significant details (Fig. B.5). Using existing shape from shading methods the

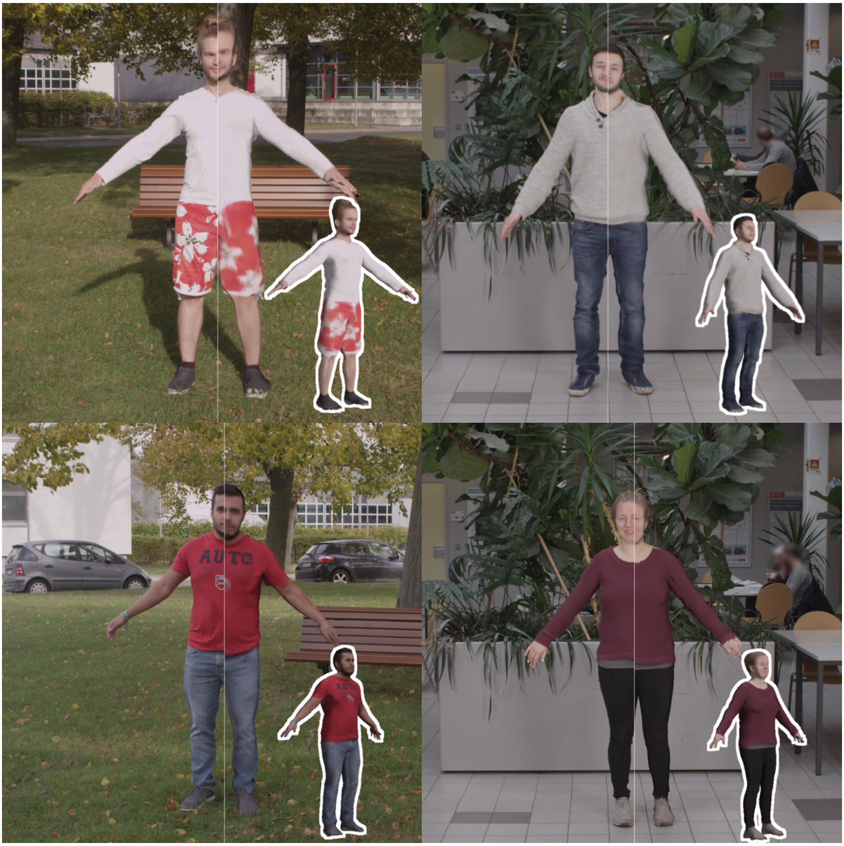

Figure B.8: Side-by-side comparison of our reconstructions (right) and the input images (left). As can be seen from the right side, our reconstructions precisely overlay on the input images. The reconstructed models rendered in a side view are shown at bottom right.

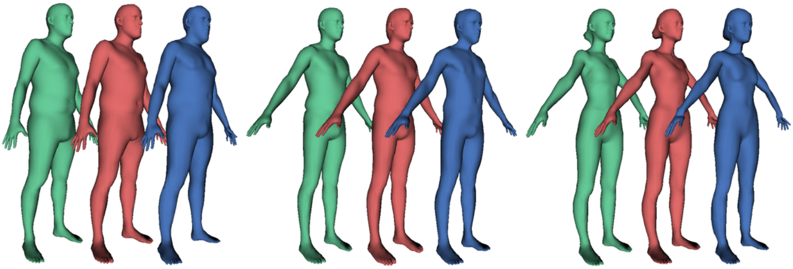

Figure B.9: Comparison to the RGB-D based method of [25] (red) and ground truth scans (green). Our approach (blue) achieves similar qualitative results despite using a monocular video sequence as opposed to a depth camera. Their approach is more accurate numerically 2.54 mm versus 3.97 mm but our results are comparable despite using a single RGB camera.

reconstruction can be further improved by adding the finer level details of the surface, e.g. folding and wrinkles. Fig. B.10 shows an example result of applying the shape from shading method of [256] to our reconstruction. This application further demonstrates the accuracy of our reconstruction, since such good result cannot be obtained without an accurate model-to-image alignment.

B.5 Discussion and Conclusion

We have proposed the first approach to reconstruct a personalized 3D human body model from a single video of a moving person. The reconstruction comprises personalized geometry of hair, body, and clothing, surface texture, and an underlying model that allows changes in pose and shape. Our approach combines a parametric human body model extended by surface displacements for refinement, and a novel method to morph and fuse the dynamic human silhouette cones in a common frame of reference. The fused cones merge the shape information contained in the video, allowing us to optimize a detailed model shape. Our algorithm not only captures the geometry and appearance of the surface, but also automatically rigs the body model with a kinematic skeleton enabling approximate pose-dependent surface deformation. Quantitative results demonstrate that our approach can reconstruct human body shape with an accuracy of 4.5mm and an ablation analysis shows robustness to noisy 3D pose estimates.

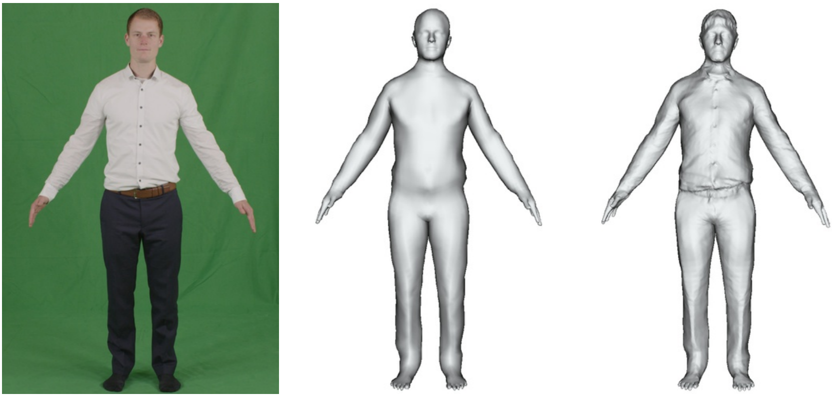

Figure B.10: Our reconstruction can be further improved by adding the finer level details of the surface using shape from shading.

The presented method finds its limits in appearances that do not share the same topology as the body: long open hair or skirts can not be modeled as an offset from the body. Furthermore, we can only capture surface details that are seen on the outline of at least one view. This means especially concave regions like armpits or inner thighs are sometimes not well handled. Strong fabric movement caused by fast skeletal motions will additionally result in decreased level of detail. In future work, we plan to incorporate illumination and material estimation alongside with temporally varying textures in our method to enable realistic rendering and video augmentation.

For the first time, our method can extract realistic avatars including hair and clothing from a moving person in a *monocular RGB video*. Since cameras are ubiquitous and low cost, people will be able to digitize themselves and use the 3D human models for VR applications, entertainment, biometrics or virtual try-on for online shopping. Furthermore, our method precisely aligns models with the images, which opens up many possibilities for image editing.

B.6 Appendix: Additional Results and Implementation Details

B.6.1 Implementation details

In this section, we present more implementation details of the presented method.

Optimization Parameters. The presented results are calculated using two empirically determined parameter sets: one for clothed subjects, one for subjects in minimal clothing. We found that the results are not very sensitive to optimization parameter weights and we select them so that the energy terms are balanced. The consensus objective function is defined as:

$$E_{\text{cons}} = E_{\text{data}} + w_{\text{lp}} E_{\text{lp}} + w_{\text{var}} E_{\text{var}} + w_{\text{sym}} E_{\text{sym}} \tag{B.13}$$

The method is initialized with $w_{\text{lp}} = 4.0$, $w_{\text{var}} = 0.6$ and $w_{\text{sym}} = 3.6$. For subjects in minimal clothing, we enforce a smoother surface with initializing $w_{\text{lp}} = 6.5$. We minimize E_{cons} with respect to model parameters and offsets. We update the point-to-line correspondences during optimization. An interesting direction to explore would be to extend [229] to continuously optimize line to surface correspondences, model parameters and offsets. In this work, we recompute correspondences during optimization. After each correspondence step, we re-initialize the three regularization terms E_{lp}, E_{var} and E_{sym}. To capture personal details, we gradually decrease the regularization weights.

Computation Time and Complexity. The results are calculated with Python code without highly parallel computation. No attempts for run-time optimization have been made. On an Intel Xeon E5-1630 v4 processor, the run-time for one frame of pose reconstruction is about 1 min including IO. Consensus shape estimation, meaning correspondence calculation and subsequent optimization on $F = 120$ frames, takes about 1:50 min.

Given, that the connectivity of the mesh is fixed and the maximum connectivity is bounded by constant k, the complexity of the regularization falls into $\mathcal{O}(N)$. As every new frame introduces more matches, the complexity of the optimization falls into $\mathcal{O}(FNP)$, with P being the number of pixels (upper bound for silhouette).

B.6.2 Scale ambiguity

Scale is an intrinsic ambiguity in monocular methods when the distance of the person to the camera is not known. Multiple views of the person in different poses help to mitigate the problem but we have observed that the ambiguity remains. The reason is that pose differences induce additional 3D ambiguities which cannot be uniquely decoupled from global size, even on multiple frames. Therefore, we perform an evaluation that is not sensitive to scale. Before calculating the per-vertex point to surface error, we adjust the one-dimensional scale parameter to match the ground truth. This step is necessary to evaluate the quality of the shape reconstructions as otherwise, almost all error would come from the scale miss-alignment.

B.6.3 Comparison with the depth camera based approach [25]

We compare our method against state-of-the-art RGB-D based approach [25] on their dataset which we refer to as KinectCap in the main paper. To make a fair comparison we also adjust the scale of their result to match the ground truth. In the original paper, they performed an evaluation that was based on scan to reconstructed mesh distance. Since the scan contains noise they had to filter out noise by not considering scan points that are further away than a given threshold. We tried to make the fairest comparison possible so we report in the main paper their result using this method, which was 2.54mm. Since we did not know what threshold to use to filter out noise in the scan and since different scan point sampling/density can produce very different results we followed the strategy explained in the main paper which was also followed in [274]. We first perform non-rigid registration regularized by the body model to obtain a ground truth registration (since registrations are regularized, they do not contain the noise in the scans). Then we compute a bi-directional surface to surface distance from the ground truth registration to the reconstructed shape. Following this strategy, their method achieves an accuracy of 3.2mm and ours 3.9mm. Our monocular approach is still not as accurate as approaches that use a depth camera [25] but produces comparable results despite using only a single RGB camera.

B.6.4 More results

We show all 9 reconstruction results on image sequences rendered from the DynamicFAUST dataset in Fig. B.11, and all 9 results from the BUFF scans

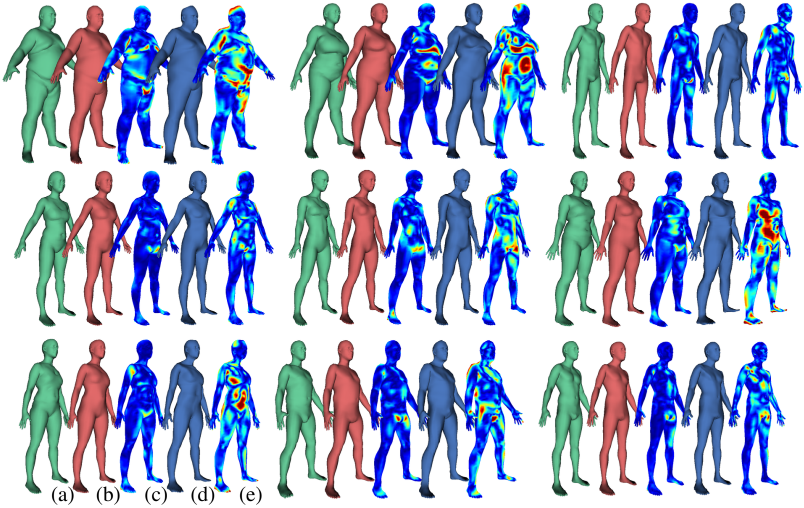

Figure B.11: Our results on image sequences from D-FAUST [27]. (a) ground truth 3D scan, (b) consensus shape with ground truth poses (consensus-p), (c) consensus-p heatmap, (d) consensus shape (consensus), (e) consensus heat-map (blue means 0mm, red means $\geq$ 2cm).

in Fig. B.12. It is worth noticing that the segmentation masks obtained from the scans in the BUFF dataset contain noise and missing data, which degrades the reconstruction quality of our method, especially for head, hands and feet. In addition, the pose reconstruction for the hip motion is less accurate than for people turning around. Note that the hip motion (in DynamicFAUST and BUFF) is probably not the most suitable motion pattern to reconstruct a static 3D person model but it allowed us to evaluate our approach numerically. Thus, the results using the rendered images of BUFF and DFAUST are slightly worse than results obtained with a real RGB camera. All the 24 reconstructed models in the People-Snapshot dataset are shown in Fig. B.13.

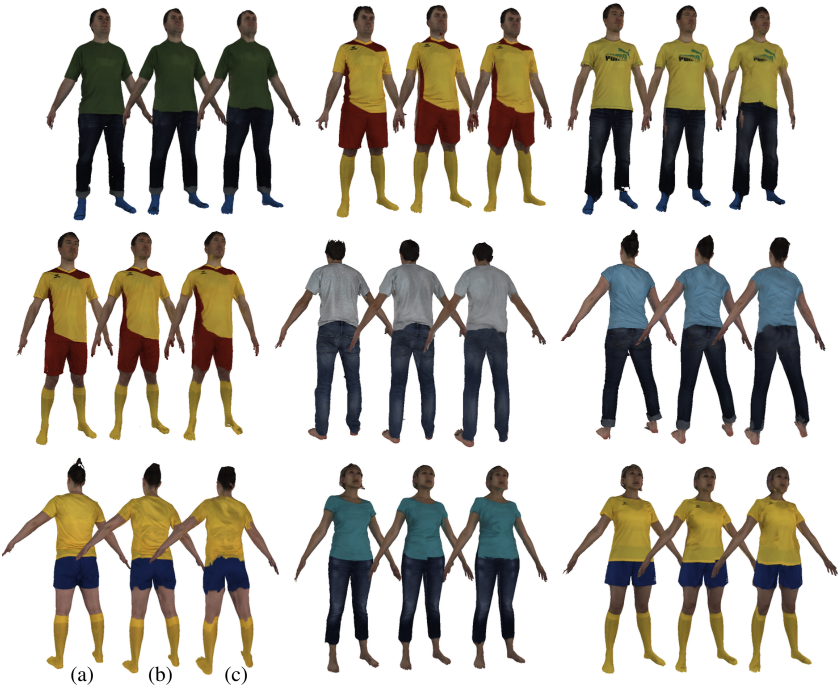

Figure B.12: Our results on image sequences from BUFF [274]. (a) ground truth scan, (b) consensus shape with ground truth poses and texture, (c) consensus shape with texture.

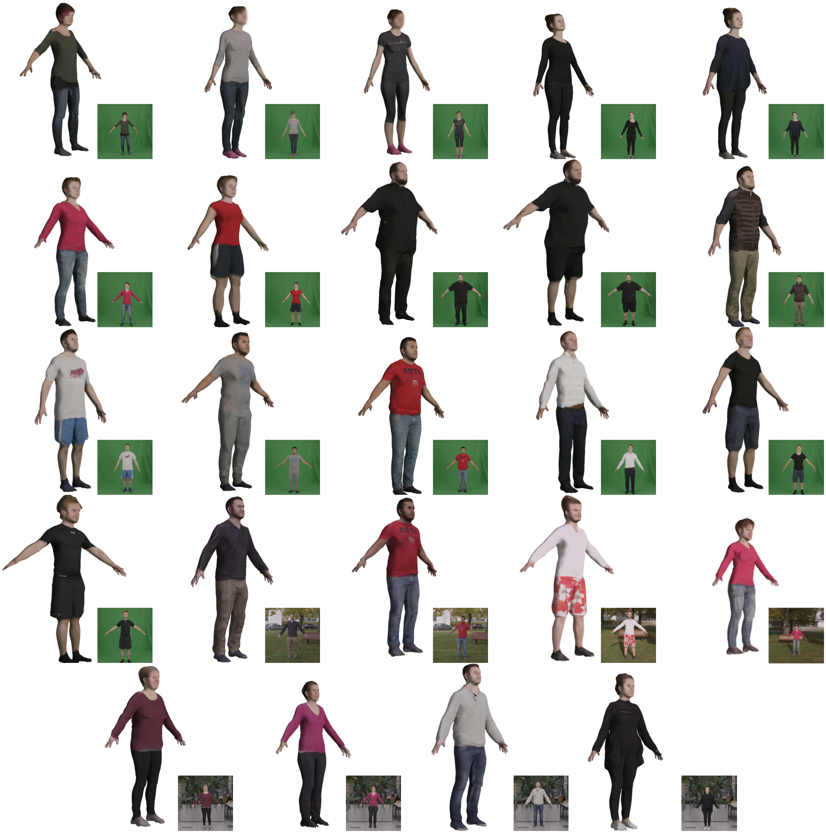

Figure B.13: Results on our People-Snapshot dataset. We blurred the faces for the subjects that did not give consent.

Errata

Compared to the original publication and in addition to editorial changes, the following corrections have been made:

- The consensus mesh in Fig. B.2 (c) has been corrected.
- The unit in Sec. B.6.3 has been changed to mm.

C | Detailed Human Avatars from Monocular Video

Thiemo Alldieck[1], Marcus Magnor[1], Weipeng Xu[2],
Christian Theobalt[2], and Gerard Pons-Moll[2]

[1] Computer Graphics Lab, TU Braunschweig

[2] Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

We present a novel method for high detail-preserving human avatar creation from monocular video. A parameterized body model is refined and optimized to maximally resemble subjects from a video showing them from all sides. Our avatars feature a natural face, hairstyle, clothes with garment wrinkles, and high-resolution texture. Our paper contributes facial landmark and shading-based human body shape refinement, a semantic texture prior, and a novel texture stitching strategy, resulting in the most sophisticated-looking human avatars obtained from a single video to date. Numerous results show the robustness and versatility of our method. A user study illustrates its superiority over the state-of-the-art in terms of identity preservation, level of detail, realism, and overall user preference.

Published in *International Conference on 3D Vision*. IEEE, 2018, pp. 98-109.

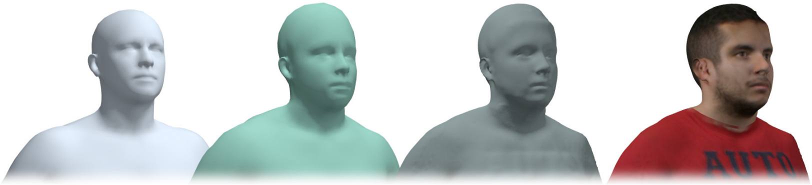

Figure C.1: Our method creates a detailed avatar from a monocular video of a person turning around. Based on the SMPL model, we first compute a medium-level avatar, then add subject-specific details and finally generate a seamless texture.

C.1 Introduction

The automatic generation of personalized 3D human models is needed for many applications, including virtual and augmented reality, entertainment, teleconferencing, virtual try-on, biometrics or surveillance. A personal 3D human model should comprise all the details that make us different from each other, such as hair, clothing, facial details and shape. Failure to faithfully recover all details results in users not feeling identified with their self-avatar.

To address this challenging problem, researchers have used very expensive recording equipment including 3D and 4D scanners [182, 27, 138] or multi-camera studios with controlled lighting [195, 127]. An alternative is to use passive stereo reconstruction [65, 161] with a camera moving around the person, but the person has to maintain a static pose which is not feasible in practice. Using depth data as input, the field has seen significant progress in reconstructing accurate 3D body models [25, 254, 275] or free-form geometry [287, 163, 170, 57] or both jointly [226]. Depth cameras are however much less ubiquitous than RGB cameras.

Monocular RGB methods are typically restricted to prediciting the parameters of a statistical body model [168, 112, 173, 26, 15, 85]. To the best of our knowledge, the only exception is a recent method [7] that can reconstruct shape, clothing and hair geometry from a monocular video sequence of a person rotating in front of the camera. The basic idea is to fuse the information from frame-wise silhouettes into a canonical pose, and optimize a free-form shape regularized by the SMPL body model [138]. While this is a significant step in 3D human reconstruction from monocular video, the reconstructions are overly smooth, lack facial details and the textures are blurry. This results in avatars that do not fully retain the identity of the real subjects.

In this work, we extend [7] in several important ways to improve the quality of the 3D reconstructions and textures. Specifically, we incorporate information from facial landmark detectors, shape-from-shading, and we introduce a new algorithm to efficiently stitch partial textures coming from frames of the moving person. Since the person is moving, information (projection rays from face landmarks and normal fields from shading cues) can not be directly fused into a single reconstruction. Hence, we track the person's pose using SMPL [138]; then we apply an inverse pose transformation to frame-wise projection rays and normal fields to fuse all the evidence in a canonical T-pose; in that space, we optimize a high-resolution shape regularized by SMPL. Precisely, with respect to previous work, our approach differs in four important aspects that allow us better preserve subject identity and details in the reconstructions:

Facial landmarks: Since the face is a crucial part of the body, we incorporate 2D facial landmark detections into the 3D reconstruction objective. To gain robustness against misdetections, we fuse temporal detections by transforming the landmark projection rays into the joint T-pose space.

Illumination and shape-from-shading: Shading is a strong cue to recover fine details such as wrinkles. Most shape-from-shading approaches focus on adding detail to static objects. Here, we perform shape-from-shading at every frame, obtaining frame-wise partial 3D normal fields that are then fused in T-pose space for final reconstruction.

Efficient texture stitching: Seamless stitching of partial textures from different camera views is particulary hard for moving articulated objects. To prevent blurry textures, one typically assigns the RGB value of one the views to each texture pixel (texel), while preserving spatial smoothness. Such assignment problem can be formulated as a multi-labeling assignment, where number possible labels grows with the number of views. Consequently, the computational time and memory becomes intractable for a large number of labels – we define a novel *texture update energy function* which can be minimized efficiently with a graph cut for every new incoming view.

Semantic texture stitching: Aside from stitching artifacts, texture spilling is another common problem. For example texture that corresponds to the clothing often floods into the skin region. To minimize spilling we add an additional semantic term into the texture update energy. The term penalizes updating a texel with an RGB value that is unlikely under a part-based appearance distribution. This semantic appearance term significantly reduces spilling, and implicitly "connects" texels belonging to the same part.

The result is the most sophisticated method to obtain detailed 3D human shape reconstructions from single monocular video. Since metric based evaluations such as scan to mesh distances do not reflect the perceptual quality, we performed a user study to assess the improvement of our method. The results show that users prefer our avatars over state-of-the-art $89.64\ \%$ of the times and they think our reconstructions are more detailed 95.72% of the times.

C.2 Related Work

Modeling the human body is a long-standing problem in computer vision. Given a densely distributed multi-camera system, one can make use of multi-view stereo methods [120] for reconstructing the human body [65, 67, 107, 277]. More advanced systems allow reconstruction of body shape under clothing [274, 265, 260], joint shape, pose and clothing reconstruction [184], or capture body pose and facial expressions [109]. However, such setups are expensive and require complicated calibration.

Hence, monocular 3D reconstruction methods [162, 161] are appealing but require depth images from many view points around a static object and humans can not hold a static pose for a long time. Therefore, nonrigid deformation of the human body has to be taken into account. Many methods are based on depth sensors and require the subject to hold the same pose. For example, in [131, 46, 209, 273], the subject alternatively makes a certain pose and rotates in front of the sensor. Then, several depth snapshots taken from different view points are fused to generate a complete 3D model. Similarly, [237] proposes to use a turntable to rotate the subject to minimize pose variations. In contrast, the methods of [25, 254, 275] allow a user to move freely in front of the sensor. In recent years, real time nonrigid depth fusion has been achieved [163, 100, 218]. These methods usually maintain a growing template and consist of two alternating steps, i.e. a registration step, where the current template is aligned to the new frame, and a fusion step, where the observation in the new frame is merged to the template. However, these methods typically suffer from “phantom surfaces” artifacts during fast motion. In [226], this problem is alleviated by using SMPL to constraint tracking. Model based monocular methods [26, 55, 76, 15, 85, 183, 179] have recently been integrated with deep learning [168, 112, 173]. However, they are restricted to predicting the parameters of a statistical body model [138, 14, 84, 288, 182]. There are two exceptions, that recover clothing and shape from a single image [79, 39] but these methods require manual initialization of pose and clothing parameters. [7] is the first method capable of reconstructing full 3D shape and clothing geometry from a single RGB video. Users can freely rotate in front of the camera while roughly holding the A-pose. Unfortunately, this approach is restricted to recover only medium-level details. The fine-level details such as garment wrinkles, subtle geometry on the clothes and facial features, which are essential elements for preserving the identity information, are missing. Our goal is to recover the missing fine-level details of the geometry and improve the texture quality such that the appearance identity information can be faithfully recovered.

Another branch of work in human body reconstruction is more focused on capturing the dynamic motion of the character. Works either recover articulated skeletal motion [224, 146, 69, 212, 5, 95], or surfaces with deformed clothing, usually called performance capture. In performance capture many approaches reconstruct a 3D model for each individual frame [223, 127, 45] or fuse a window of frames [170, 57]. However, these methods cannot generate a temporal coherent representation of the model, which is an important characteristic for many applications. To solve this, methods register a common model to results of all frames [34], use volumetric representation for surface tracking [4, 94], or assume a pre-built static template. Again, most of those methods are based on multi-view images [51, 66, 178, 37, 192, 195]. There are attempts on reducing the number of cameras, such as the stereo method [259], single view depth based method [287] and the recent monocular RGB based method [264]. Note that the result of our method can be used as the initial template for above-mentioned template based performance capture methods.

Shape-from-shading is also highly related to our method. A comprehensive survey can be found in [276]. We only discuss the application of shape-from-shading in the context of human body modeling. Geometric details, e.g. folds in the non-textured region, are difficult to capture with silhouette or photometric information. In contrast, shape-from-shading captures such details [259, 257, 81]. There are also approaches for photometric stereo which recover the shape using controlled light stage setup [244].

Texture generation is an essential task for modeling a realistic virtual character, since a texture image can describe the material properties that cannot be modeled by the surface geometry. The key of a texture generation method is how to combine texture fragments created from different views. Many early works blend the texture fragments using weighted averaging across the entire surface [20, 52, 166, 176]. Others make use of mosaicing strategies, which yields sharper results [18, 126, 164, 196]. [125] is the first to formulate texture stitching as a graph cut problem. Such formulation has been commonly used in texture generation for multi-view 3D reconstruction. However, without accurately reconstructed 3D geometry and registered images, these methods usually suffer from blurring or ghosting artifacts. To this end, many methods focus on compensating registration errors [59, 21, 248, 64, 279]. In our scenario, the registration misalignment problem is even more severe, due to our challenging monocular nonrigid setting. Therefore, we propose to take advantage of semantic information to better constrain our problem.

C.3 Method

In this paper, our goal is to create a detailed avatar from an RGB video of a subject rotating in front of the camera. The focus lies hereby on fine-level details, that model a subject's identity and individual appearance. As shown in Fig. C.2, our method reconstructs a textured mesh model in a coarse-to-fine manner, which consists of three steps: First we estimate a rough body shape of the subject, similar to [7], where the medium-level geometry of the clothing and skin is reconstructed. Then we add fine-level geometric details, such as garment wrinkles and facial features, based on shape-from-shading. Finally, we compute a seamless texture to capture the texel-level appearance details. In the following, we first describe our body shape model, and then discuss the details of our three steps.

C.3.1 Subdivided SMPL body model

Our method is based on the SMPL body model [138]. However, the original SMPL model is too coarse to model fine-level details such as garment wrinkles and fine facial features. To this end, we adapt the model as follows.

The SMPL model is a parameterized human body model described by a function of pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ returning $N = 6890$ vertices and $F = 13776$

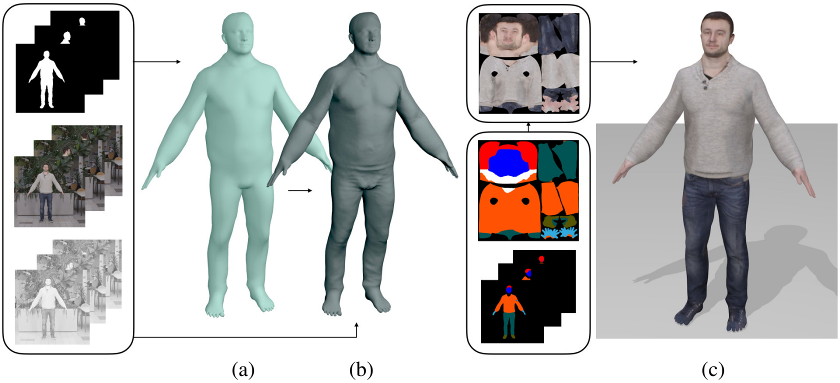

Figure C.2: Our method 3-step method: We first estimate a medium level body shape based on segmentations (a), then we add details using shape-from-shading (b). Finally we compute a texture using a semantic prior and a novel graph cut optimization strategy (c).

faces. As SMPL only models naked humans, we use the extended formulation from [7] allowing offsets $\mathbf{D}$ from the template $\mathbf{T}$:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{C.1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D} \tag{C.2}$$

where W is a linear blend-skinning function applied to a rest pose $T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$ based on the skeleton joints $J(\boldsymbol{\beta})$ and after pose $B_p(\boldsymbol{\theta})$ and shape dependent $B_s(\boldsymbol{\beta})$ deformations. The inverse function $M^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$ *unposes* the model and brings the vertices back into the canonical T-pose. As we aim for fine details and a subject's identity, we further extent the formulation. As shown in Fig. C.3, we subdivide every edge of the the SMPL model twice. Every new vertex is defined as:

$$\boldsymbol{v}_{N+e} = 0.5(\boldsymbol{v}_i + \boldsymbol{v}_j) + s_e \boldsymbol{n}_e, \quad (i, j) \in \mathcal{E}_e \tag{C.3}$$

where $\mathcal{E}$ defines the pairs of vertices forming an edge and $\boldsymbol{n}_e$ is the average normal between the normals of the vertex pair. $s \in \mathbf{s}$ defines the displacement in normal direction $\boldsymbol{n}_e$. $\boldsymbol{n}_e$ is calculated at initialization time in unposed space and can be posed according to W. The new finer model $M_f(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}, \mathbf{s})$ consists of $N = 110210$ vertices and $F = 220416$ faces. To recover the high-res smooth surface we calculate an initial set $\mathbf{s}_0 = \{s_0, \ldots, s_e\}$ by minimizing

$$\arg\min_{\mathbf{s}} \Big(\mathbf{L} M_f = \sum_{j \in \mathcal{N}(i)} w_{ij} (\boldsymbol{v}_i - \boldsymbol{v}_j) \Big) \tag{C.4}$$

where $\mathbf{L}$ is the Laplace matrix with cotangent weights w_{ij} and $\mathcal{N}(i)$ defines the neighbors around $\boldsymbol{v}_i$.

C.3.2 Medium-level body shape reconstruction

In recent work, a pipeline to recover a subject's body shape, hair and clothing in the same setup as ours has been presented [7]. They first select a number of key-frames ($K \approx 120$) evenly distributed over the sequence and segment them into foreground and background using a CNN [33]. Then they recover the 3D pose for each selected frame based on 2D landmarks [36]. At the core of their method they transform the silhouette cone of every key-frame back into the canonical T-pose of the SMPL model using the inverse formulation of SMPL. This allows efficient optimization of the body shape independent of pose. We follow their pipeline and optimize for the subjects body shape in unposed space. However, we notice that the face estimation of [7] is not accurate enough. This prevents us from further recovering fine-level facial features in the following steps, since precise face alignment is necessary for that. To this end, we propose a new objective for body shape estimation (dependency on parameters removed for clarity):

$$\arg\min_{\boldsymbol{\beta}, \mathbf{D}} E_{\text{silh}} + E_{\text{face}} + E_{\text{regm}} \tag{C.5}$$

The silhouette term E_{silh} measures the distance between boundary vertices and silhouette rays. See [7] for details and regularization E_{regm}. The face alignment term E_{face} penalizes the distance between the 2D facial landmark detections and the 2D projection of 3D facial landmarks. We use OpenPose [216] to detect 2D facial landmarks for every key-frame. In order to incorporate the detections into the method, we establish a static mapping between landmarks and points on the mesh. Every landmark $\boldsymbol{l}$ is mapped to the surface via barycentric interpolation of neighboring vertices. During optimization, we measure the point to line distance between the landmark $\boldsymbol{l}$ on the model and the corresponding camera ray $\mathbf{r}$ describing the 2D landmark detection in unposed space:

$$\delta(\boldsymbol{l}, \mathbf{r}) = \boldsymbol{l} \times \boldsymbol{r}_n - \boldsymbol{r}_m \tag{C.6}$$

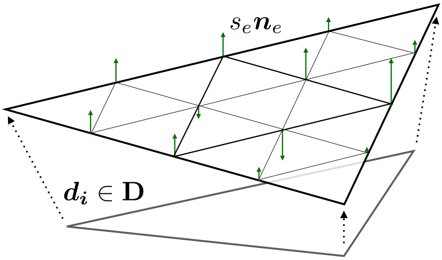

Figure C.3: One face of the new SMPL formulation. The displacement field vectors $\boldsymbol{d}_*$ and the normal displacements $s_* \boldsymbol{n}_*$ form the subdivided surface.

where $\mathbf{r} = (\boldsymbol{r}_m, \boldsymbol{r}_n)$ is given in Plucker coordinates. The face alignment term finally is:

$$E_{\text{face}} = \sum_{l,r \in \mathcal{L}} w_l \rho(\delta(\boldsymbol{l}_l, \mathbf{r}_r)) \tag{C.7}$$

where $\mathcal{L}$ defines the mapping between mesh points and landmarks, w is the confidence of the landmark given by the CNN and ρ is the Geman-McClure robust cost function. To speed up computation time, we use the coarse SMPL model formulation (Eq. C.1) for the medium-level shape estimation.

C.3.3 Modeling fine-level surface details

In Sec. C.3.2, we capture the medium-level details by globally integrating the silhouette information from all key-frames. Now our goal is to obtain fine-level surface details, which cannot be estimated from silhouette, based on shape-from-shading. Note that estimating shape-from-shading globally over all frames would lead to a smooth shape without details, due to fabric movement and misalignments. Thus, we first capture the details for a number of key-frames individually, and then incrementally merge the details into the model as new triangles become visible in a consecutive key-frame. We found that the number of key-frames can be lower than in the first step and choose $K = 60$. Now we describe how to capture the fine-level details for a single key-frame k based on shape-from-shading. To make this process robust, we estimate shading normals individually in a window around the key-frame and then jointly optimize for the surface.

Shape-from-shading: For each frame, we first decompose the image into reflectance $\mathbf{I}_r$ and shading $\mathbf{I}_s$ using the CNN based intrinsic decomposition method of [158]. The function $H_{\boldsymbol{c}}$ calculates the shading of a vertex with spherical harmonic components $\boldsymbol{c}$. We estimate spherical harmonic components $\boldsymbol{c}$ that minimize the difference between the simulated shading and the observed image shading $\mathbf{I}_s$ jointly for the given window of frames [256]:

$$\arg\min_{\boldsymbol{c}} \sum_{i \in \mathcal{V}} \left| H_{\boldsymbol{c}}(\boldsymbol{n}_i) - \mathbf{I}_s(\mathbf{P}\boldsymbol{v}_i) \right|, \tag{C.8}$$

where $\mathcal{V}$ denotes the subset of visible vertices, i.e. the angle between the normal and the viewing direction is $0 < \alpha \leq \alpha_{\text{max}}$. $\mathbf{P}$ is the projection matrix. Having the scene illumination and the shading for every pixel, we can now estimate auxiliary normals $\tilde{\mathbf{N}} = \{\tilde{\boldsymbol{n}}_0, \dots, \tilde{\boldsymbol{n}}_N\}$ for every vertex per frame:

$$\arg\min_{\tilde{\mathbf{N}}} E_{\text{grad}} + w_{\text{lapn}} E_{\text{lapn}}. \tag{C.9}$$

The Laplacian smoothness term $E_{\text{lapn}} = \mathbf{L}\tilde{\mathbf{N}}$ enforces the normals to be locally smooth. E_{grad} penalizes shading errors by calculating the difference between the gradient between a shaded vertex and its neighbors $\mathcal{N}$ and the image gradient at the projected vertex positions:

$$E_{\text{grad}} = \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}(i) \cap \mathcal{V}} ||\Delta_{H_{\boldsymbol{c}}}(\tilde{\boldsymbol{n}}_i, \tilde{\boldsymbol{n}}_j) - \Delta_{\mathbf{I}_s}(\mathbf{P}\boldsymbol{v}_i, \mathbf{P}\boldsymbol{v}_j)||^2 \tag{C.10}$$

with $\Delta_f(a, b) = f(a) - f(b)$.

Surface reconstruction: In order to merge information about all estimated normals within the window, we transform the normals back into the canonical T-pose using the inverse pose function of SMPL M^{-1}. Then we optimize for the surface which explains the merged normals. Further, we include the silhouette term and face term of Sec. C.3.2 to enforce the surface to be well aligned to the images. Specifically, we minimize:

$$\arg\min_{\mathbf{D},\mathbf{s}} \sum_{j \in \mathcal{C}} (\lambda_j E_{\text{silh},j} + \lambda_j w_{\text{face}} E_{\text{face},j}) + w_{\text{sfs}} E_{\text{sfs}} + E_{\text{regf}} \tag{C.11}$$

with weights w_* and $\lambda_j = 1$ for $j = k$ and $\lambda_j < 1$ otherwise. E_{silh} and E_{face} are evaluated over a number of control frames $\mathcal{C}$ and matches in E_{silh} are limited to

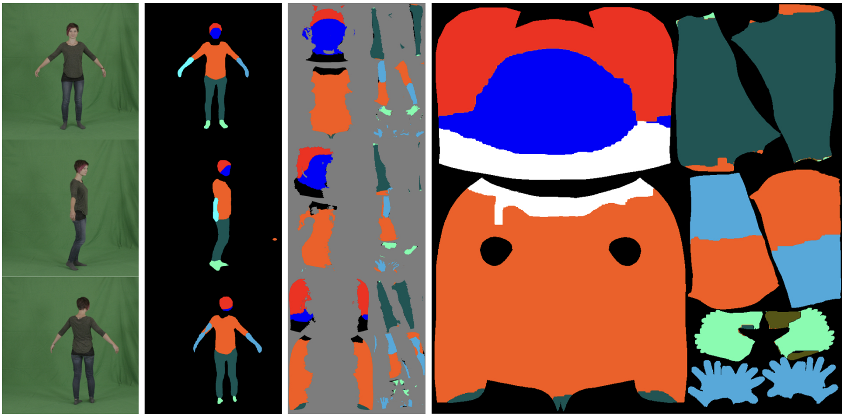

Figure C.4: We calculate a semantic segmentation for every key frame. The semantic labels are mapped into texture space and combined into a semantic texture prior.

vertices in the original SMPL model. The shape-from-shading term is defined as:

$$E_{\text{sfs}} = \sum_{f=k-m}^{k+m} \sum_{i \in \mathcal{V}} ||\boldsymbol{n}_i - \tilde{\boldsymbol{n}}_i^f||^2 \tag{C.12}$$

where k is the current key-frame and m specifies the window size, usually $m = 1$. $\tilde{\boldsymbol{n}}_i^f$ denotes the auxiliary normal of vertex i calculated from frame f. All normals are in T-pose space. E_{regf} regularizes the optimization as described in the following:

$$E_{\text{regf}} = w_{\text{match}} E_{\text{match}} + w_{\text{lap}} E_{\text{lap}} + w_{\text{struc}} E_{\text{struc}} + w_{\text{cons}} E_{\text{cons}} \tag{C.13}$$

E_{match} penalizes the discrepancy between two neighboring key-frames. Specifically, for a perfect estimation, the following assumption should hold: When warping a key-frame into a neighboring key-frame based on the warp-field described by the projected vertex displacement, the warped frame and the target frame should be similar. E_{match} describes this metric: First we calculate the described warp. Then we calculate warping errors based on optical flow [31]. Based on the sum of the initial warp-field and the calculated error, we establish a grid of correspondences between neighboring key-frames. Every correspon-

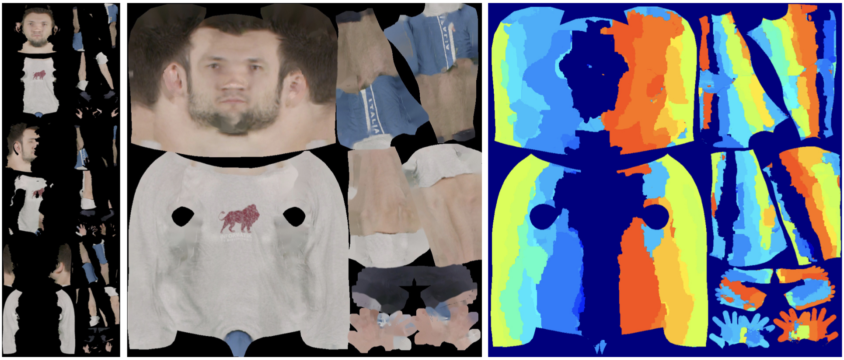

Figure C.5: Based on part textures from key frames (left), we stitch a complete texture using graph-cut based optimization. The associated key frames for each texel are shown as colors on the right.

dence c should be explained by a particular point of the mesh surface. We first find a candidate for every correspondence:

$$\underset{i \in \mathcal{V}}{\arg\min} \frac{\cos(\alpha_k^i)\delta(\boldsymbol{v}_i^k, \mathbf{r}_c^k) + \cos(\alpha_j^i)\delta(\boldsymbol{v}_i^j, \mathbf{r}_c^j)}{\cos(\alpha_k^i) + \cos(\alpha_j^i)} \tag{C.14}$$

where α_k^i is the viewing angle under which the vertex i has been seen in key-frame k and $\mathbf{r}_c^k$ is the projection ray of correspondence c in posed space of key-frame k. Then we minimize point to line distance in unposed space:

$$E_{\text{match}} = \sum_{i,c \in \mathcal{M}} \rho(\delta(\boldsymbol{v}_i, \mathbf{r}_c)) \tag{C.15}$$

where $\mathcal{M}$ is the set of matches established in Eq. C.14.

The remaining regularization terms of Eq.C.13 are as follows: E_{lap} is the Laplacian smoothness term with anisotropic weights [256]. E_{struc} aims to keep the structure of the mesh by pruning edge length variations. E_{cons} prunes large deviations from the consensus shape.

We optimize using a *dog-leg* trust region method using the chumpy autodifferentiation framework. We alternate minimizing and finding silhouette point to line correspondences. Regularization is reduced step-wise.

C.3.4 Texture generation

A high quality texture image is an essential component for a realistic virtual character, since it can describe the material properties that cannot be modeled by the surface geometry. In order to obtain a sharp and seamless texture, we solve the texture stitching on a per texel level (Fig. C.5), in contrast to that on a per face level as in other works [125]. In other words, our goal is to color each pixel in the texture image with a pixel value taken from one out of K key-frames. However, this makes the scale of our problem much larger, and therefore does not allow us to perform global optimization. To this end, we propose a novel texture merging method based on graph cut, which translates our problem to a series of binary labeling subproblems that can be efficiently solved. Furthermore, meshes and key-frames are not perfectly aligned. To reduce color spilling and artifacts caused by misalignments, we compute a semantic prior before stitching the final texture (Fig. C.4).

Partial texture generation: For every key-frame, we first project all visible surface points to the frame and write the color at the projected position into the corresponding texture coordinates. In order to factor out the illumination in the texture images, we *unshade* the input images by dividing them with the shading images as used in Sec. C.3.3. The partial texture calculation can easily be achieved using the OpenGL rasterization pipeline. Apart from the partial color texture image, we calculate two additional texture maps for the merging step, i.e. the viewing-angle map and the semantic map. For the viewing-angle map, we compute the viewing angle α_k^t under which the surface point t has been seen in key-frame k.

The semantic prior is generated by re-projecting the human semantic segmentation to the texture space. Specifically, we first calculate a semantic label for every pixel in the input frames using a CNN based human parsing method [133]. Each frame is segmented into 10 semantic classes such as *hair*, *face*, *left leg* and *upper clothes*. Then the semantic information of all frames is fused into the global semantic map by minimizing for labeling $\boldsymbol{x}$:

$$\arg\min_{\boldsymbol{x}} \sum_{t=0}^{T} \varphi_t(x_t) + \sum_{t,q\in\mathcal{N}} \psi(x_t, x_q) \tag{C.16}$$

$$\varphi_t(x_t) = 1 - \frac{\sum_{k=0}^{K} X_k(\cos^2 \alpha_k^t)}{K} \tag{C.17}$$

Here φ is the energy term describing the compatibility of a label x with the texel t, where X_k returns the given value if the texel was labeled with x in view k and 0 otherwise. ψ gives the label compatibility of neighboring texels t and q. We solve Eq. C.16 by multi-label graph-cut optimization with alpha-beta swaps [29]. While constructing the graph, we connect every texel not only with its neighbors in texture space but with all neighbors on the surface. In particular this means texels are connected across texture seams. To have a strong prior for the texture completion, we calculate Gaussian mixture models (GMM) of the colors in HSV space per label using the part-textures and corresponding labels.

Texture merging: Next, we calculate the complete texture by merging the partial textures. While keeping the same graph structure, the objective function is:

$$\arg\min_{\boldsymbol{u}} \sum_{t=0}^{T} \theta_t(u_t) + \sum_{t,q \in \mathcal{N}} \eta_{t,q}(u_t, u_q) \tag{C.18}$$

where the labeling $\boldsymbol{u}$ assigns every texel to a partial texture k. The first term seeks to find the best image for each texel:

$$\theta_t(k) = w_{\text{vis}} \sin^2 \alpha_k^t + w_{\text{gmm}} m(\mathbf{U}_k^t, x_t) + w_{\text{face}} d(\mathbf{U}_k^t) + w_{\text{silh}} E_{\text{silh},k} \tag{C.19}$$

with weights w_*. m returns the Mahalanobis distance between the color value for t in part-texture k given the semantic label x_t. d calculates the structural dissimilarity between the first and the given key-frame. d is only evaluated on texels belonging to the facial region and ensures consistent facial expression over the texture.

The smoothness-term η ensures similar colors for neighboring texels. For neighboring texels assigned to different key-frames $u_t \neq u_q$, while belonging to the same semantic region $x_t = x_q$, $\eta_{t,q}$ equals the gradient magnitude between the texel colors $||\mathbf{U}_{u_t}^t - \mathbf{U}_{u_q}^q||$.

Since the number of combinations in η is very high, it is computationally not feasible to solve Eq. C.18 as a multi label graph-cut problem. Thus, we propose the following strategy for an approximate solution: We convert the multi-label problem to a binary labeling decision $b \in \{\textit{update}, \textit{keep}\}$. We initialize the texture with $\mathbf{M} = \mathbf{U}_0$. Then we randomly choose a key-frame k and test it against the current solution. The likelihood of selecting a key-frame is inversly pro-

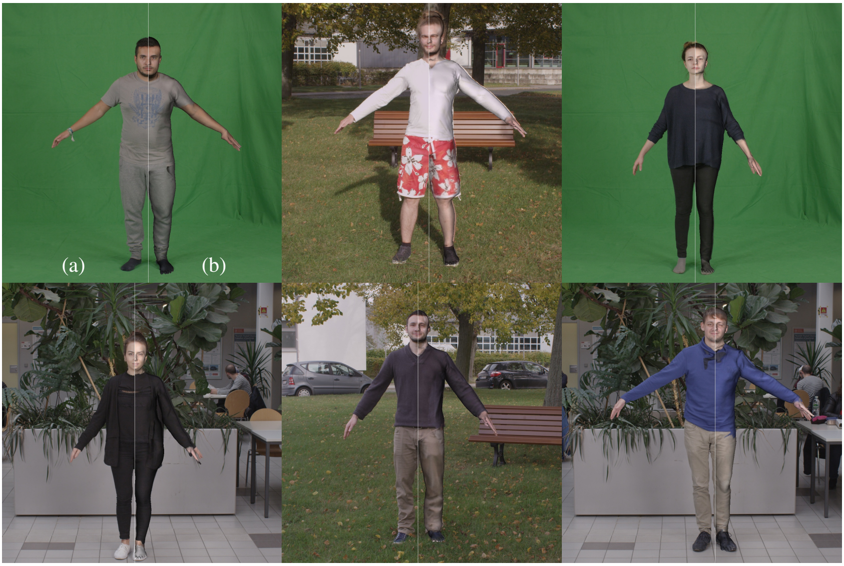

Figure C.6: Side-by-side comparisons of our reconstructions (b) and the input frame (a). As can be seen from (b), our method closely resembles the subject in the video (a).

portional to its remaining silhouette error $E_{\text{silh},k}$ in order to favor well-aligned key-frames. Further, η is approximated with:

$$\eta_{t,q} = \begin{cases} \max(||\mathbf{M}^t - \mathbf{U}_k^q||, ||\mathbf{M}^q - \mathbf{U}_k^t||), & \text{if } b_t \neq b_q \wedge x_t = x_q \\ 0, & \text{otherwise} \end{cases} \tag{C.20}$$

Convergence is usually reached between $2K$ to $3K$ iterations. Finally, we cross-blend between different labels to reduce visible seams. The run-time per iteration on 1000×1000 px with Python code using a standard graph cut library is $\sim$2 sec. No attempts for run-time optimization have been made.

C.4 Experiments

We evaluate our method on two publicly available datasets: The People-Snapshot dataset [7] and the dataset used in [25]. To validate the perceived quality of our results we performed a user study.

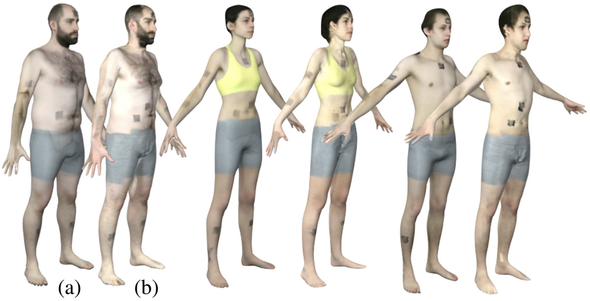

Figure C.7: Our results (b) in comparison against the RGB-D method [25] (a). Note that the texture prior has not been used (see Sec. C.4.1).

C.4.1 Qualitative results and comparisons

We compare our method to the recent method of [7] on their People-Snapshot dataset. The approach of [7] is the only other monocular 3D person reconstruction method. The People-Snapshot dataset consists of 24 sequences of different subjects rotating in front of the camera while roughly holding an A-pose. In Fig. C.6, we show some examples of our reconstruction results, which precisely overlay the subjects in the image. Note that the level of detail of the input images is captured by our reconstructed avatars. In Fig. C.11, we show side-by-side comparison to [7]. Our results (right) reconstruct the face better and preserve many more details, e.g. clothing wrinkles and t-shirt stamps.

Additionally, we compare against the state-of-the-art RGB-D method [25], also using their dataset of people in minimal clothing[1]. While their method relies on depth data, we only use the RGB video which makes the problem much harder. Despite this, as shown in Fig. C.7, our results are comparable in quality to theirs.

C.4.2 Face similarity measure

One goal of our method was to preserve the individual appearance of subjects in their avatars. Since the face is crucial for this, we leverage facial landmarks detections and shape-from-shading. As seen in Fig. C.11 our method adds a significant level of detail to the facial region in comparison to state-of-the-art.

---

[1] The deep learning based segmentation [73] only works for fully clothed people so we had to deactivate the semantic prior in this dataset.

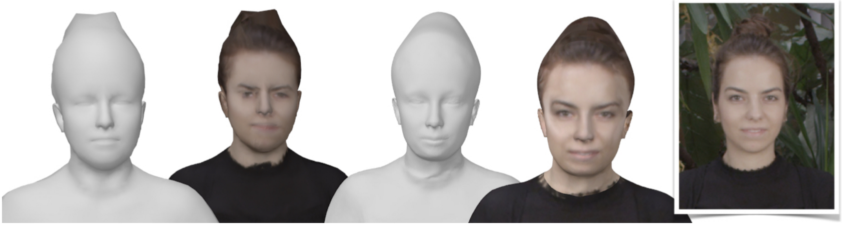

Figure C.8: In comparison to the method of [7] (left), the faces in our results (right) have finer details in the mesh and closely resemble the subject in the photograph.

In Fig. C.8 we show the same comparison also for untextured meshes. Our result closely resembles the subject in the photograph. To further demonstrate the effectiveness of our method for face similarity preservation, we perform the following experiment: FaceNet [206] is a deep network, that is trained to map from face images to an Euclidean space where distance corresponds to face similarity. We use FaceNet trained on the CASIA WebFace dataset [269] to measure the similarity between photos of the subjects in the People-snapshot dataset and their reconstructions. Two distinct subjects in the dataset have a mean similarity distance of 1.33 ± 0.13. Same subjects in different settings differ by 0.55 ± 0.18. Our reconstructions feature a mean distance of 0.99 ± 0.11 to their photo counterparts. Reconstructions of [7] perform significantly worse with a mean distance of 1.09 ± 0.15. While our reconstructions can be reliable identified using FaceNet, reconstructions of [7] have a similarity distance close to a distance of distinct people, making them less likely to be identified correctly.

C.4.3 Ablation analysis

In the following we qualitatively demonstrate the effectiveness of further design choices of our method.

Shape-from-shading: In order to render the avatars under different illuminations, detailed geometry should be present in the mesh. In Fig. C.9, we demonstrate the level of detail added to the meshes by shape-from-shading. While the mesh on the left only describes the low-frequency shape, our refined result on the right contains fine-grained details such as wrinkles and buttons.

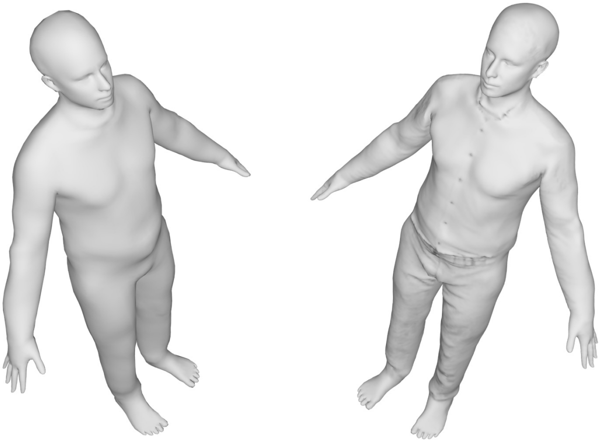

Figure C.9: Comparison of a result of our method before (left) and after (right) applying shape-from-shading based detail enhancing.

Influence of the texture prior: In Fig. C.10 we show the effectiveness of the semantic prior for texture stitching. While the texture on the left computed without the prior contains noticeable color spills on the arms and hands, the final texture on the right contains no color spills and less stitching artifacts along semantic boundaries.

C.4.4 User study

Finally, we conducted a user study in order to validate the visual fidelity of our results. Each participant was asked four questions about 6 randomly chosen results out of the 24 reconstructed subjects in People-Snapshot dataset. The avatars shown to each participant and the questions asked were randomized. In every question the participants had to decide between our method, and the method of [7]. The four question were:

- Which avatar preserves the identity of the person in the image better? (*identity*)
- Which avatar has more detail? (*detail*)
- Which avatar looks more real to you? (*realism*)
- Which avatar do you like better? (*preference*)

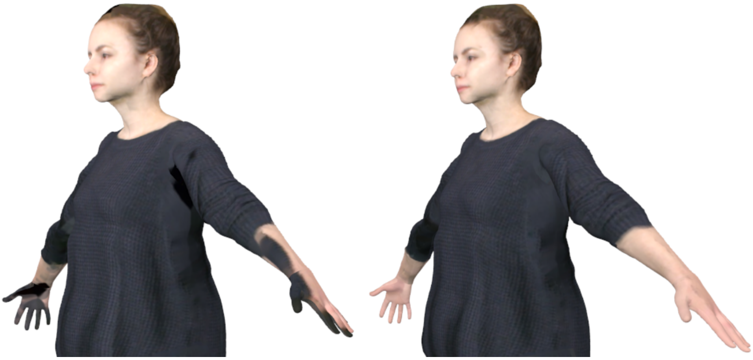

Figure C.10: The semantic prior for texture stitching successfully removes color spilling (left) in our final texture (right).

We presented the users renderings of the meshes in consistent pose and illumination. The users were allowed to zoom into the images. At questions *identity* and *realism* we showed the participants either textured or untextured meshes. For *identity* comparison we additionally showed a photo of the subject next to the renderings. When asking for *detail* we only showed untextured meshes, and when asking for *preference* we only showed textured results. Additionally, we asked for the level of experience with 3D data (*None*, *Beginner*, *Proficient*, *Expert*). 74 people participated in our online survey, covering the whole range of expertise.

The results of the study are summarized in Table C.1. The participants clearly preferred our results in all scenarios over current state-of-the-art. Admittedly, when asked about identity preservation in untextured meshes, users preferred our method, but this time only 65.70%. Further inspection of the results shows that users with high experience with 3D data think our method preserves the identity better with 90.48% versus 60.49% for novice users. We hypothesize that unexperienced users find it more difficult to recognize people from 3D meshes without textures. Most importantly, by a large margin, our results are perceived as more realistic (92.27%), preserve more details (95.72%) and where preferred 89.64% of the times.

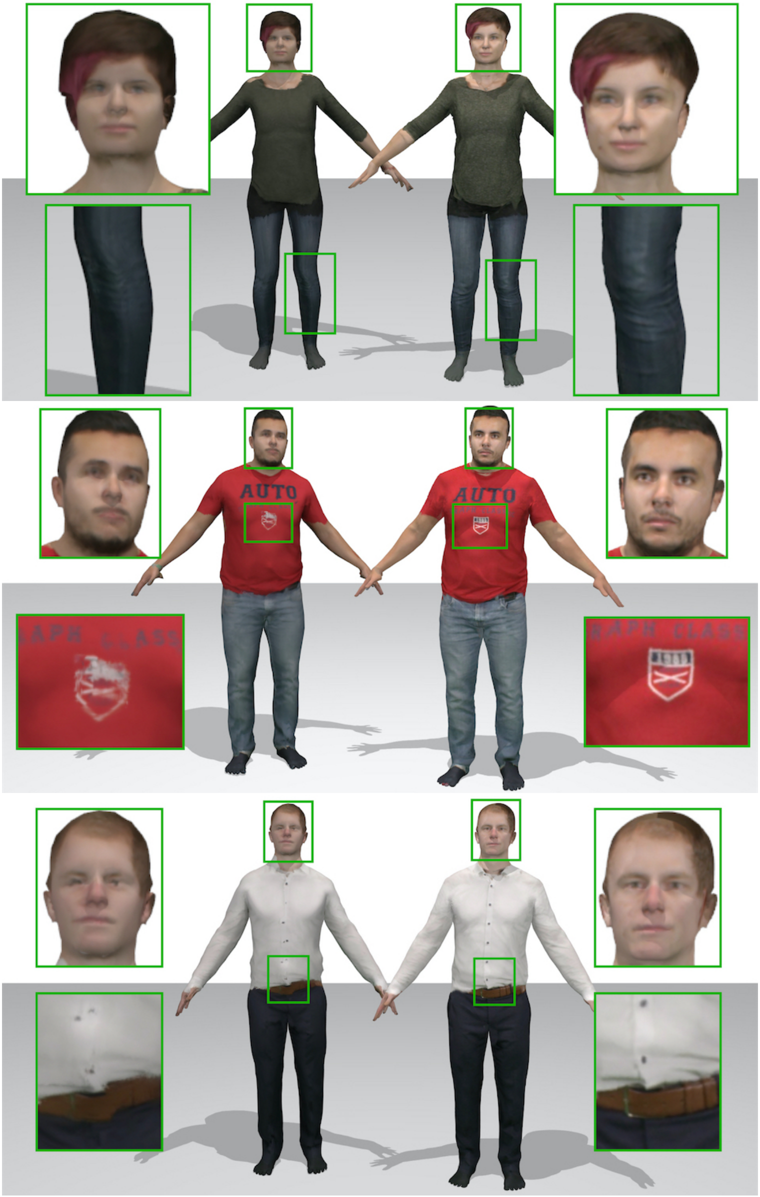

Figure C.11: In comparison to the method of [7] (left), our results (right) look much more natural and have finer details.

C.5 Discussion and Conclusion

We have proposed a novel method to create highly detailed personalized avatars from monocular video. We improve over the state-of-the-art in several important aspects: Our optimization scheme allows to integrate face landmark detections and shape-from-shading from multiple frames. Experiments demonstrate that this results in better face reconstruction and better identity preservation. This is also confirmed by our user study, which shows that people think our method preserves identity better 83.12% of the times, and capture more details 95.72% of the times.

We introduced a new texture stitching binary optimization, which allows us to efficiently merge the appearance of multiple frames into a single coherent texture. The optimization includes a semantic texture term that incorporates appearance models for each semantic segmentation part. Results demonstrate that the common artifact of color spilling from skin to clothing or viceversa gets reduced.

We have argued for a method to capture the subtle, but very important details to make avatars look *realistic*. Indeed *details matter*, the user study shows that users think our results are more realistic than the state of the art 92.7% of the times, and prefer our avatars 89.64% of the times.

Future work should address capture of subjects wearing clothing with topology different from the body, including skirts and coats. Furthermore, to obtain full texturing, subjects have to be seen from all sides – it may be possible to infer occluded appearance using sufficient training data. Another avenue to explore is reconstruction in an un-cooperative setting, e.g. from online videos of people.

Having cameras all around us, we can now serve the growing demand for personalized avatars in virtual and augmented reality applications e.g. in the fields of entertainment, communication or e-commerce.

	Identitiy	Details	Realism	Preference
Textured Avatars	83.12 %	-	92.27 %	89.64 %
Untextured Avatars	65.70 %	95.72 %	89.73 %	-

Table C.1: Results of the user study. Percentage of answers where users preferred our method over [7]. We asked for four different aspects. See Sec. C.4.4 for details.

D | Learning to Reconstruct People in Clothing from a Single RGB Camera

Thiemo Alldieck[1], Marcus Magnor[1], Bharat Lal Bhatnagar[2],
Christian Theobalt[2], and Gerard Pons-Moll[2]

[1] Computer Graphics Lab, TU Braunschweig

[2] Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

We present Octopus, a learning-based model to infer the personalized 3D shape of people from a few frames (1-8) of a monocular video in which the person is moving with a reconstruction accuracy of 4 to 5mm, while being orders of magnitude faster than previous methods. From semantic segmentation images, our Octopus model reconstructs a 3D shape, including the parameters of SMPL plus clothing and hair in 10 seconds or less. The model achieves fast and accurate predictions based on two key design choices. First, by predicting shape in a canonical T-pose space, the network learns to encode the images of the person into pose-invariant latent codes, where the information is fused. Second, based on the observation that feed-forward predictions are fast but do not always align with the input images, we predict using both, bottom-up and top-down streams (one per view) allowing information to flow in both directions. Learning relies only on synthetic 3D data. Once learned, Octopus can take a variable number of frames as input, and is able to reconstruct shapes even from a single image with an accuracy of 5mm. Results on 3 different datasets demonstrate the efficacy and accuracy of our approach.

Published in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 1175-1186.

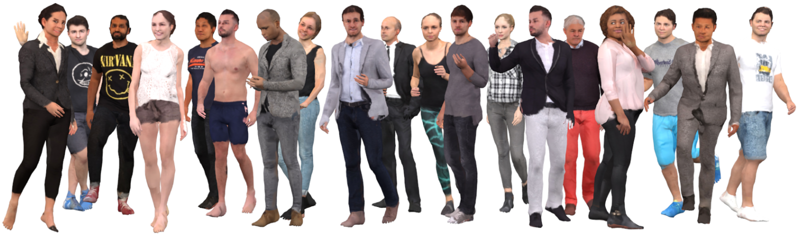

Figure D.1: We present a deep learning based approach to estimate personalized body shape, including hair and clothing, using a single RGB camera. The shapes shown above have been calculated using only 8 input images, and re-posed using SMPL.

D.1 Introduction

The automatic acquisition of detailed 3D human shape and appearance, including clothing and facial details is required for many applications such as VR/AR, gaming, virtual try-on, and cinematography.

A common way to acquire such models is with a scanner or a multi-view studio [1, 130]. The cost and size prevent the wide-spread use of such setups. Therefore, numerous works address capturing body shape and pose with more practical setups, e.g. from a low number of video cameras [192], or using one or more depth cameras, either specifically for the human body [25, 254, 275] or for general free-form surfaces [287, 163, 170, 100, 57, 226]. The most practical but also challenging setting is capturing from a single monocular RGB camera. Some methods attempt to infer the shape parameters of a body model from a single image [112, 168, 26, 55, 15, 85, 280, 106, 173], but reconstructed detail is constrained to the model shape space, and thus does not capture personalized shape detail and clothing geometry. Recent work [7, 6] estimates more detailed shape, including clothing, from a video sequence of a person rotating in front of a camera while holding a rough A-pose. While reconstructed models have high quality, the optimization approach takes around 2 minutes only for the shape component. More importantly, the main bottleneck is the pre-processing step, which requires *fitting* the SMPL model to each of the frame silhouettes using time-consuming non-linear optimization ($\approx$ 120 min for 120 frames). This is impractical for many applications that require fast acquisition such as telepresence and gaming.

In this work, we address these limitations and introduce *Octopus*, a convolutional neural network (CNN) based model that learns to predict 3D human shapes in a canonical pose given a few frames of a person rotating in front of a *single camera*. Octopus predicts using both, bottom-up and top-down streams (one per view) allowing information to flow in both directions. It can make bottom-up predictions in 50ms per view, which are effectively refined top-down using the same images in 10s. Inference, both bottom-up and top-down, is performed fully-automatically using the same model. *Octopus* is therefore easy to use and more practical than previous work [7]. Learning only relies on synthetic 3D data, and on semantic segmentation images and keypoints derived from synthesized video sequences. Consequently, *Octopus* can be trained without paired data – real images with ground truth 3D shape annotations – which is very difficult to obtain in practice.

Octopus predicts SMPL body model parameters, which represent the undressed shape and the pose, plus additional 3D vertex offsets that model clothing, hair, and details beyond the SMPL space. Specifically, a CNN encodes F frames of the person (in different poses) into F latent codes that are fused to obtain a single shape code. From the shape code, two separate network streams predict the SMPL shape parameters, and the 3D vertex offsets in the canonical T-pose space, giving us the "unpose" shape or T-shape. Predicting the T-shape forces the F latent codes to be pose-invariant, which is necessary to fuse the shape information contained in each frame. Octopus also predicts a pose for each frame, which allows to "pose" the T-shape and render a silhouette to evaluate the overlap against the input images in a top-down manner during both training and inference. Specifically, since bottom-up models do not have a feedback loop, the feed-forward 3D predictions are correct but do not perfectly align with the input images. Consequently, we refine the prediction top-down by optimizing the F poses, the T-shape, and the vertex offsets to maximize silhouette overlap and joint re-projection error.

Experiments on a newly collected dataset (LifeScans), the publicly available PeopleSnapshot dataset [7], and on the dataset used in [25] demonstrate that our model infers shapes with a reconstruction accuracy of 4mm in less than 10 seconds. In summary, *Octopus* is faster than purely optimization-based fitting approaches such as [7], it combines the advantages of bottom-up and top-down methods in a single model, and can reconstruct detailed shapes and clothing from a few video frames. Examples of reconstruction results are shown in Fig. D.1. To foster further research in this direction, we made Octopus available for research purposes[1].

---

[1] http://virtualhumans.mpi-inf.mpg.de/octopus/

D.2 Related Work

Methods for 3D human shape and pose reconstruction can be broadly classified as top-down or bottom-up. Top-down methods either fit a *free-form* surface or a statistical body model (*model-based*). Bottom-up methods directly infer a surface or body model parametrization from sensor data. We will review bottom-up and top-down methods for human reconstruction.

Top-down, free-form methods non-rigidly deform meshes [37, 51, 34] or volumetric shape representations [94, 4]. These methods are based on multi-view stereo reconstruction [120], and therefore require multiple RGB or depth cameras, which is a practical barrier for many applications. Using *depth cameras*, KinectFusion [104, 162] approaches reconstruct 3D scenes by incrementally fusing frame geometry, and appearance [279], in a canonical frame. Several methods build on KinectFusion for body scanning [209, 131, 273, 46]. The problem is that these methods require the person to stand still while the camera is turned around. DynamicFusion [163] generalized KinectFusion to non-rigid objects by combining non-rigid tracking and fusion. Although template-free approaches [162, 100, 218] are flexible, they can only handle very careful motions. Common ways to add robustness are pre-scanning the template [287], or using multiple kinects [57, 170] or multi-view [223, 127, 45]. These methods, however, do not register the temporal 3D reconstructions to the same template and focus on other applications such as streaming or telepresence [170]. Estimating shape by compensating for pose changes can be tracked back to Cheung *et al.* [43, 44], where they align visual hulls over time to improve shape estimation. To compensate for articulation, they merge shape information in a coarse voxel model. However, they need to track each body part separately and require multi-view input. All free-form works require multi-view input, depth cameras or cannot handle moving humans.

Top-down, model-based methods exploit a parametric body model consisting of pose and shape [14, 84, 138, 288, 182, 109] to regularize the fitting process. Some *Depth-based* methods [254, 86, 268, 275, 25] exploit the temporal information by optimizing a single shape and multiple poses (jointly or sequentially). This leads to expensive optimization problems. Using *mutli-view*, some works achieve fast performance [192, 194] at the cost of using a coarser body model based on Gaussians [224], or a pre-computed template [270]. Early RGB-based methods were restricted to estimating the parameters of a body model, and required multiple views [15] or manually clicked points [76, 280, 106, 198]. Shape and clothing have been recovered from RGB images [79, 39], depth [40], or scan data [184], but require manual intervention or clothing is limited to a pre-defined set of templates. In [266] a fuzzy vertex association from clothing to body surface is introduced, which allows complex clothing modeled as body offsets. Some works are in-between free-form and model-based methods. In [66, 243], authors pre-scan a template and insert a skeleton, and in [226] authors combine the SMPL model with a volumetric representation to track the clothed human body from a depth camera.

Bottom-up. Learning of features for multi-view photo-consistency [128], and auto-encoders combined with visual hulls [72, 238] have shown to improve free-form performance capture. These works, however, require more than one camera view. Very few works learn to predict personalized human shape from images–lack of training data and the lack of a feedback loop between feed-forward predictions and the images makes the problem hard. Variants of random forests and neural networks have been used [55, 53, 54, 240] to regress shape from silhouettes. The problem here is that predictions tend to look over-smooth, are confined to the model shape space, and do not comprise clothing. Garments are predicted [49] from a single image, but a single model for every new garment needs to be trained, which makes it hard to use in practice. Recent pure bottom-up approaches to human analysis [146, 147, 185, 283, 225, 236, 197] typically predict shape represented as a coarse stick figure or bone skeleton, and can not estimate body shape or clothing.

Hybrid methods. A recent trend of works combines bottom-up and top-down approaches–a combination that has been exploited already in earlier works [221]. The most straightforward way is by fitting a 3D body model [138] to 2D pose detections [26, 124]. These methods, however, can not capture clothing and details beyond the model space. Clothing, hair and shape [7, 6] can be inferred by fusing dynamic silhouettes (predicted bottom-up) of a video to a canonical space. Even with good 2D predictions, these methods are susceptible to local minima when not initialized properly, and are typically slow. Furthermore, the 2D prediction network and the model fitting is de-coupled. Starting with a feed-forward 3D prediction, semantic segmentation, keypoints and scene constraints are integrated top-down in order to predict the pose and shape of multiple people [272]. Other recent works integrate the SMPL model, or a voxel representation [241], as a layer within a network architecture [112, 173, 168, 239]. This has several advantages: (i) predictions are constrained by a shape space of humans, and (ii) bottom-up 3D predictions can be verified top-down using 2D keypoints and silhouettes during training. However, the shape estimates are confined to the model shape space and tend to be close to the average. The focus of these works is rather on robust pose estimation, while we focus on personalized shapes. We also integrate SMPL within our architecture but our work is different in several aspects. First, our architecture fuses the information of several images of the same person in different poses. Second, our model incorporates a fast top-down component during training *and* at test time. As a result, we can predict clothing, hair and personalized shapes using a single camera.

D.3 Method

The goal of this work is to create a 3D model of a subject from a few frames of a monocular RGB video, and in less than 10 seconds. The model should comprise body shape, hair, and clothing and should be animatable. We take inspiration from [7] and focus on the cooperative setting with videos of people rotating in front of a camera holding a rough A-pose – this motion is easy and fast to perform, and ensures that non-rigid motion of clothing and hair is not too large. In contrast to previous work [7], we aim for fast and fully automatic reconstruction. To this end, we train a novel convolutional neural network to infer a 3D mesh model of a subject from a small number of input frames. Additionally, we train the network to reconstruct the 3D pose of the subject in each frame. This allows us to refine the body shape by utilizing the decoder part of the network for instance-specific optimization (Fig. D.2).

In Sec. D.3.1 we describe the shape representation used in this work followed by its integration into the used predictor (Sec. D.3.2). In Sec. D.3.3 we explain the losses, that are used in the experiments. We conclude by describing the instance-specific top-down refinement of results (Sec. D.3.4).

D.3.1 Shape representation

Similar to previous work [274, 7], we represent shape using the SMPL statistical body model [138], which represents the undressed body, and a set of offsets modeling instance specific details including clothing and hair.

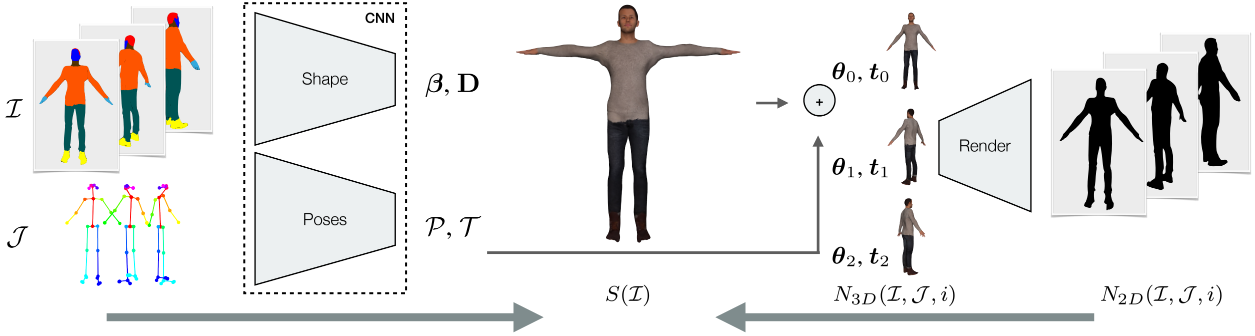
$\mathcal{I}$
$\mathcal{J}$
CNN
Shape
Poses
$\boldsymbol{\beta}, \mathbf{D}$
$\mathcal{P}, \mathcal{T}$
$S(\mathcal{I})$
+
$\boldsymbol{\theta}_0, \boldsymbol{t}_0$
$\boldsymbol{\theta}_1, \boldsymbol{t}_1$
$\boldsymbol{\theta}_2, \boldsymbol{t}_2$
$N_{3D}(\mathcal{I}, \mathcal{J}, i)$
Render
$N_{2D}(\mathcal{I}, \mathcal{J}, i)$

Figure D.2: Overview of our method: Our novel CNN predicts 3D human shapes from semantic images in an canonical pose together with per-image pose information calculated from 2D joint detections (left to center). The pose information can be used to refine the shape via 'render and compare' optimization using the same predictor (right to center).

SMPL is a function $M(\cdot)$ that maps pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ to a mesh of $V = 6890$ vertices. By adding offsets $\mathbf{D}$ to the template $\mathbf{T}$, we obtain a posed shape instance as follows:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{D.1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D}, \tag{D.2}$$

where linear blend-skinning $W(\cdot)$ with weights $\mathbf{W}$, together with pose-dependent deformations $B_p(\boldsymbol{\theta})$ allow to pose the T-shape ($\mathbf{T} + B_s(\boldsymbol{\beta})$) based on its skeleton joints $J(\cdot)$. SMPL plus offsets, denoted as SMPL+D, is fully differentiable with respect to pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$ and free-form deformations $\mathbf{D}$. This allows us to directly integrate SMPL as a fixed layer in our convolutional architecture.

D.3.2 Model and data representation

Given a set of images $\mathcal{I} = \{\mathbf{I}_0, \ldots, \mathbf{I}_{F-1}\}$ depicting a subject from different sides with corresponding 2D joints $\mathcal{J} = \{\mathbf{J}_0, \ldots, \mathbf{J}_{F-1}\}$, we learn a predictor f_w^* that infers the body shape $\boldsymbol{\beta}$, personal and scene specific body features $\mathbf{D}$, and 3D poses $\mathcal{P} = \{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{F-1}\}$ along with 3D positions $\mathcal{T} = \{\boldsymbol{t}_0, \ldots, \boldsymbol{t}_{F-1}\}$ for each image. $f_w^* : (\mathcal{I}, \mathcal{J}) \mapsto (\boldsymbol{\beta}, \mathbf{D}, \mathcal{P}, \mathcal{T})$ is a CNN parametrized by network parameters w.

Input modalities. Images of humans are highly diverse in appearance, requiring large datasets of annotated images in the context of deep learning. Therefore, to abstract away as much information as possible while still retaining shape and pose signal, we build on previous work [74, 36] to simplify each RGB image to a semantic segmentation and 2D keypoint detections. This allows us to train the network using only synthetic data and generalize to real data.

Model parametrization. By integrating the SMPL+D model (Sec. D.3.1) into our network formulation, we can utilize its mesh output in the training of f_w^*. Concretely, we supervise predicted SMPL+D parameters in three ways: Imposing a loss directly on the mesh vertices $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$, on the predicted joint locations $J(\boldsymbol{\beta})$ and their projections on the image, and densely on a rendering of the mesh using a differential renderer [87].

The T-shape ($\mathbf{T} + B_s(\boldsymbol{\beta}) + \mathbf{D}$) in Eq. D.2 is now predicted from the set of semantic images $\mathcal{I}$ with the function:

$$S(\mathcal{I}) = \mathbf{T} + B_s(f_w^{\boldsymbol{\beta}}(\mathcal{I})) + f_w^{\mathbf{D}}(\mathcal{I}), \tag{D.3}$$

where f_w^* are the regressors to be learned. Similarly, the posed mesh $N_{3D}(\mathcal{I}, \mathcal{J}, i)$ is predicted from the image $\mathbf{I}_i$ and 2D joints $\mathbf{J}_i$ with the function:

$$N_{3D}(\mathcal{I}, \mathcal{J}, i) = W(P(\mathcal{I}, \mathcal{J}, i), J(f_w^{\boldsymbol{\beta}}(\mathcal{I})), f_w^{\boldsymbol{\theta}_i}(\mathcal{I}, \mathcal{J}), \mathbf{W}) \tag{D.4}$$

$$P(\mathcal{I}, \mathcal{J}, i) = S(\mathcal{I}) + B_p(f_w^{\boldsymbol{\theta}_i}(\mathcal{I}, \mathcal{J})), \tag{D.5}$$

from which the 3D Joints are predicted with the linear regressor J_{B25}:

$$N_{J_{3D}}(\mathcal{I}, \mathcal{J}, i) = J_{\mathrm{B25}}(N_{3D}(\mathcal{I}, \mathcal{J}, i)) \tag{D.6}$$

J_{B25} has been trained to output 25 joint locations consistent with the BODY_25 keypoint ordering[1]. The estimated posed mesh N_{3D} can be rendered in uniform color with the image formation function $R(\cdot)$ paramerized by camera c:

$$N_{2D}(\mathcal{I}, \mathcal{J}, i) = R_c(N_{3D}(\mathcal{I}, \mathcal{J}, i)) \tag{D.7}$$

Similarly, we can project the the joints $N_{J_{3D}}$ to the image plane by perspective projection π:

$$N_{J_{2D}}(\mathcal{I}, \mathcal{J}, i) = \pi_c(N_{J3D}(\mathcal{I}, \mathcal{J}, i)) \tag{D.8}$$

All these operations are differentiable, which we can conveniently use to formulate suitable loss functions.

D.3.3 Loss functions

Our architecture permits two sources of supervision: (i) 3D supervision (in our experiments, from synthetic data derived by fitting SMPL+D to static scans), and (ii) 2D supervision from video frames alone. In this section, we discuss different loss functions used to train the predictors f_w^*.

Losses on body shape and pose. For a paired sample in the dataset $\{(\mathcal{I}, \mathcal{J}), (\boldsymbol{\beta}, \mathbf{D}, \mathcal{P}, \mathcal{T})\}$ we use the following losses between our estimated model N_{3D} and the ground truth model $M(\cdot)$ scan:

---

[1] https://github.com/cmu-perceptual-computing-lab/openpose

- Per-vertex loss in the canonical T-pose $\mathbf{0}_{\boldsymbol{\theta}}$. This loss provides a useful 3D supervision on shape independently of pose:

$$\mathcal{L}_S = ||S(\mathcal{I}) - M(\boldsymbol{\beta}, \mathbf{0}_{\boldsymbol{\theta}}, \mathbf{D})||^2 \tag{D.9}$$

- Per-vertex loss in posed space. This loss supervises both pose and shape on the Euclidean space:

$$\mathcal{L}_{N_{3D}} = \sum_{i=0}^{F-1} ||N_{3D}(\mathcal{I}, \mathcal{J}, i) - M(\boldsymbol{\beta}, \boldsymbol{\theta}_i, \mathbf{D})||^2 \tag{D.10}$$

- Silhouette overlap:

$$\mathcal{L}_{N_{2D}} = \sum_{i=0}^{F-1} ||R_c(N_{3D}(\mathcal{I}, \mathcal{J}, i)) - b(\mathbf{I}_i)||^2, \tag{D.11}$$

where $b(\mathbf{I}_i)$ is the binary segmentation mask and R_c is the image formation function defined in Eq. D.7. $\mathcal{L}_{N_{2D}}$ is a weakly supervised loss as it does not require 3D annotations and $b(\mathbf{I}_i)$ can be estimated directly from RGB images. In the experiments, we investigate whether such self-supervised loss can reduce the amount 3D supervision required (see D.4.4). Additionally, we show that N_{2D} can be used at test time to refine the bottom-up predictions and capture instance specific details in a top-down manner (see D.3.4).

- Per-vertex SMPL undressed body loss:

The aforementioned losses only penalize the final SMPL+D 3D shape. It is useful to include an "undressed-body" $(\hat{S})$ loss to force the shape parameters $\boldsymbol{\beta}$ to be close to the ground truth

$$\mathcal{L}_{\hat{S}} = ||\hat{S}(\mathcal{I}) - M(\boldsymbol{\beta}, \mathbf{0}_{\boldsymbol{\theta}}, \mathbf{0}_{\mathbf{D}})||^2 \tag{D.12}$$

$$\hat{S}(\mathcal{I}) = \mathbf{T} + B_s(f_w^{\boldsymbol{\beta}}(\mathcal{I})), \tag{D.13}$$

where $\mathbf{0}_{\mathbf{D}}$ are vectors of length 0. This also prevents that the offsets $\mathbf{D}$ explain the overall shape of the person.

Pose specific losses. In addition to the posed space $\mathcal{L}_{N_{3D}}$ and silhouette overlap $\mathcal{L}_{N_{2D}}$ losses, we train for the pose using a direct loss on the predicted parameters $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{t}}$

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{t}} = \sum_{i=0}^{F-1} \left(||\mathbf{R}(f_w^{\boldsymbol{\theta}_i}) - \mathbf{R}(\boldsymbol{\theta}_i)||^2 + ||f_w^{\boldsymbol{t}_i} - \boldsymbol{t}_i||^2\right), \tag{D.14}$$

where $\mathbf{R}$ are vectorized rotation matrices of the 24 joints. Similar to [168, 124, 173], we use differentiable SVD to force the predicted matrices to lie on the manifold of rotation matrices. This term makes the pose part of the network converge faster.

Losses on joints. We further regularize the pose training by imposing a loss on the joints in Euclidean space:

$$\mathcal{L}_{J_{3D}} = \sum_{i=0}^{F-1} ||N_{J_{3D}}(\mathcal{I}, \mathcal{J}, i) - J_{\text{B25}}(M(\boldsymbol{\beta}, \boldsymbol{\theta}_i, \mathbf{D}))||^2 \tag{D.15}$$

Similar to the 2D image projection loss on model $\mathcal{L}_{N_{2D}}$ (Eq. D.11), we also have a weakly supervised 2D joint projection loss $\mathcal{L}_{J_{2D}}$

$$\mathcal{L}_{J_{2D}} = \sum_{i=0}^{F-1} ||N_{J_{2D}}(\mathcal{I}, \mathcal{J}, i) - \pi_c(J_{\text{B25}}(M(\boldsymbol{\beta}, \boldsymbol{\theta}_i, \mathbf{D})))||^2. \tag{D.16}$$

D.3.4 Instance-specific top-down optimization

The bottom-up predictions of the neural model can be refined top-down at test time to capture instance specific details. It is important to note that this step requires no 3D annotation as the network fine-tunes using only 2D data. Specifically, at the test time, given a subject's images $\mathcal{I}$ and 2D joints $\mathcal{J}$ we optimize a small set of layers in f_w^* using image and joint projection losses $\mathcal{L}_{N_{2D}}, \mathcal{L}_{J_{2D}}$ (see D.4.1). By fixing most layers of the network and optimizing only latent layers, we find a compromise between the manifold of shapes learned by the network and new features, that have not been learned. We further regularize this step using Laplacian smoothness, face landmarks, and symmetry terms from [7, 6]. Table D.1 illustrates the performance of the pipeline before and after optimization (see D.4.2, D.4.3).

D.4 Experiments

The following section focuses on the evaluation of our method. In Sec. D.4.1 we introduce technical details of the used dataset and network architecture. The following sections describe experiments for quantitative and qualitative evaluation as well as ablation and parameter analysis.

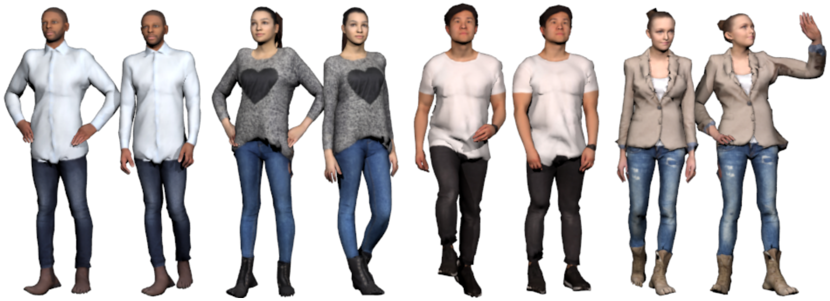

Figure D.3: Sample scans from the *LifeScans* dataset.

D.4.1 Experimental setup

Dataset. To alleviate the lack of paired data, we use 2043 static 3D scans of people in clothing. We purchased 163 scans from renderpeople.com and 54 from axyz-design.com. 1826 scans were kindly provided from Twindom[1]. Unfortunately, in the 2043 there is not enough variation in pose and shape to learn a model that generalizes. Hence, we generate synthetic 3D data by non-rigidly registering SMPL+D to each of the scans. This allows us to change the underlying body shape and pose of the scan using SMPL, see Fig. D.3. Like [7], we focus on a cooperative scenario where the person is turning around in front of the camera. Therefore, we animate the scans with turn-around poses and random shapes and render video sequences from them. We call the resulting dataset *LifeScans*, which consists of rendered images *paired* with 3D animated scans in various shapes and poses. Since the static scans are from real people, the generated images are close to photo-realistic, see Fig D.3. To prevent overfitting, we use semantic segmentation together with keypoints as intermediate image representation, which preserve shape and pose signatures while abstracting away appearance. This reduces the amount of appearance variation required for training. To be able to render synthetic semantic segmentation, we first render the LifeScans subjects from different viewpoints and segment the output with the method of [74]. Then we project the semantic labels back in the SMPL texture space and fuse different views using graph cut-based optimization. This final step enables full synthetic generation of paired training data.

---

[1] https://web.twindom.com/

Scale ambiguity. Scale is an inherent ambiguity in monocular imagery. Three factors determine the size of an object in an image: distance to the camera, camera intrinsics, and the size of the object. As it is not possible to decouple this ambiguity in a monocular set-up with moving objects, we fix two factors and regress one. In other works [168, 112, 173] authors have assumed fixed distance to the camera. We cannot make this assumption, as we leverage multiple images of the same subject, where the distances to the camera may vary. Consequently, we fix the size of the subject to average body height. Precisely, we make SMPL height independent, by multiplying the model by 1.66m divided by the y-axis distance of vertices describing ankles and eyes. Finally, we fix the focal length to sensor height.

Network architecture. In the following we describe details of the convolutional neural network f_w^*. An overview is given in Fig. D.4. The input to f_w^* is a set of 1080x1080px semantically segmented images $\mathcal{I}$ and corresponding 2D joint locations $\mathcal{J}$. f_w^* encodes each image $\mathbf{I}_i$ with a set of five, 3x3 convolutions with ReLU activations followed by 2x2 max-pooling operations into a pose invariant latent code l_i^{inv}. In our experiments we fixed the size of l_i^{inv} to 20. The *pose branch* maps both joint detections $\mathbf{J}_i$ and output of the last convolutional layer to a vector of size 200 and finally to the pose-dependent latent code l_i^{pose} of size 100 via fully connected layers. The *shape branch* aggregates pose invariant information across images and computes mean l^{inv}. Note that this formulation allows us to aggregate pose-dependent and invariant information across an arbitrary and varying number of views. The shape branch goes on to predict SMPL shape parameters $\boldsymbol{\beta}$ and free-form deformations $\mathbf{D}$ on the SMPL mesh. $\boldsymbol{\beta}$ is directly calculated from l^{inv} with a linear layer. In order to predict per-vertex offsets from the latent code l^{inv}, we use a four-step graph convolutional network with Chebyshev filters and mesh upsampling layers similar to [190]. Each convolution is followed by ReLU activation. We prefer a graph convolutional network over a fully connected decoder due to memory constraints and in order to get structured predictions.

Training scheme. The proposed method, including rendering, is fully differentiable and end-to-end trainable. Empirically we found it better to train the pose branch before training the shape branch. Thereafter, we optimize the network end-to-end. We use a similar training schedule for our pose branch as [173], where we first train the network using losses on the joints and pose parameters ($\mathcal{L}_{J_{3D}}$, $\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{t}}$) followed by training using losses on the vertices and pose

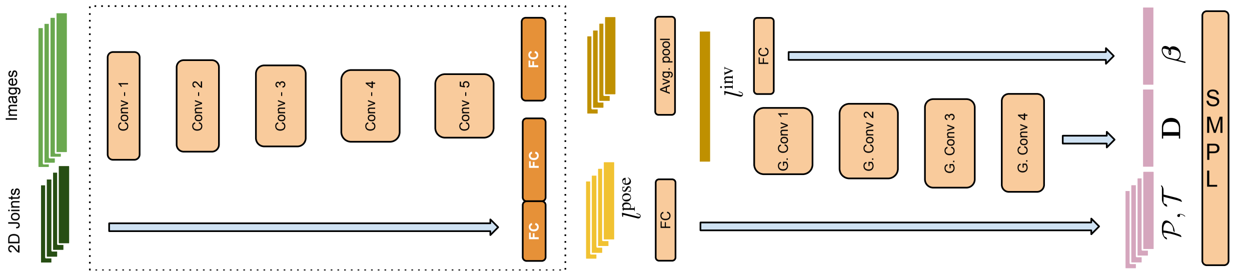

Figure D.4: Network architecture: Our bottom-up inference network first encodes the input (semantically segmented images $\mathcal{I}$ and 2D joints $\mathcal{J}_{2D}$) into a decoupled, pose dependent l^{pose} and pose invariant l^{inv} latent space. The pose branch subsequently infers per-frame pose and translation parameters $\mathcal{P}$ and $\mathcal{T}$ from l^{pose}. The shape branch infers the body shape $\boldsymbol{\beta}$ and free-form deformations $\mathbf{D}$ in T-pose from l^{inv}. We use a graph convolution based decoder, to learn per-vertex offsets $\mathbf{D}$. The entire model is end-to-end trainable. The orange FC layers and the final graph convolution layer can be fine-tuned at test time to better model instance-specific details (see Sec. D.3.4).

	Before optimization	After optimization
Full Pipeline	4.47 ±4.45	4.00 ±3.94
GT Poses	4.47 ±4.41	3.17 ±3.41

Table D.1: Mean vertex error (mm) of 55 test samples computed on $F = 8$ input images. The *full method* with inferred poses produces comparable results to using *GT poses*. Both variants gain accuracy from subsequent optimization.

parameters ($\mathcal{L}_{N_{3D}}, \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{t}}$). We also experiment with various training schemes, and show that weakly supervised training can significantly reduce the dependence on 3D annotated data (see Sec. D.4.4). For that experiment, we train the model with alternating full ($\mathcal{L}_S, \mathcal{L}_{\hat{S}}, \mathcal{L}_{N_{3D}}, \mathcal{L}_{J_{3D}}$) and weak supervision ($\mathcal{L}_{N_{2D}}, \mathcal{L}_{J_{2D}}$). During instance-specific optimization we keep most layers fixed and only optimize latent pose l^{pose}, latent shape l^{inv} and the last graph convolutional layer, that outputs free-form displacements $\mathbf{D}$.

D.4.2 Numerical evaluation

We quantitatively evaluate our method on a separated test set of the LifeScans dataset containing 55 subjects. We use $F = 8$ semantic segmentation images and 2D poses as input and optimize the results for a maximum budget of 10 seconds. All results have been computed without intensive hyper-parameter tuning. To quantify shape reconstruction accuracy, we adjust the pose of the estimation to match the ground truth, following [274, 25]. This disentangles errors in pose from errors in shape and allows to quantify shape accuracy. Finally, we compute the bi-directional vertex to surface distance between scans and reconstructions. We report mean errors in millimeters (mm) across the test set in Tab. D.1. We differentiate between *full method* and *ground truth (GT) poses*. Full method refers to our method as described in Sec. D.4.1. The latter is a variant of our method that uses ground truth poses, which allows to study the effect of pose errors. In Fig. D.5 we display subjects in the test set for both variants along with per-vertex error heatmaps. Visually the results look almost indistinguishable, which is corroborated by the fact that the numerical error increases only by $\approx$ 1mm between GT and predicted pose models. This demonstrates the robustness of our approach. We show more examples with the corresponding texture for qualitative assessment in Fig. D.1. The textures have been computed using graph cut-based optimization using semantic labels as described in [6].

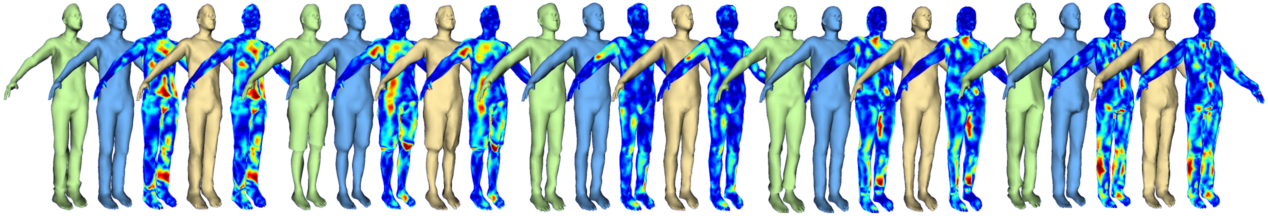

Figure D.5: Results from LifeScans in comparison to ground truth shapes (green). We show results computed with ground truth poses (blue) and results of the full method (yellow) with corresponding error heatmaps with respect to ground truth shapes (red means $\geq$ 2cm).

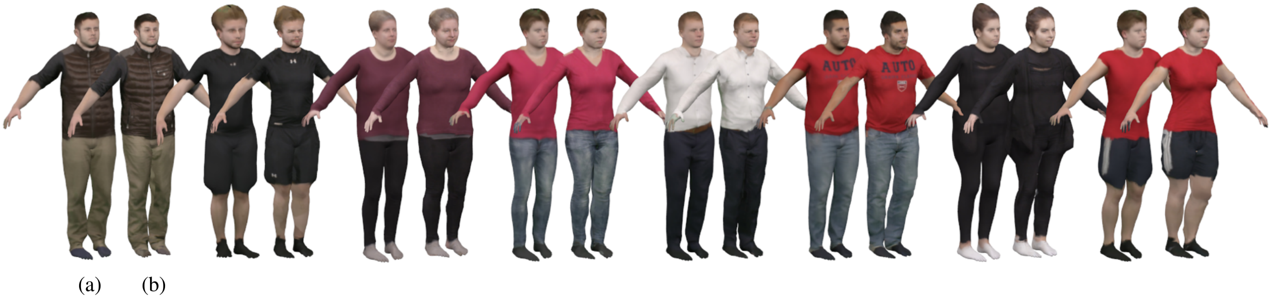

Figure D.6: Comparison to the state-of-the-art optimization based method [7]. Their method (a) uses 120 frames, while ours (b) only uses 8 images and is several magnitudes faster.

D.4.3 Analysis of key parameters

Our method comes with two key hyper-parameters, namely number of input images F, and number of optimization steps. In the following section, we study these parameters and how they affect the performance of our approach. We also justify our design choices.

Fig. D.7 illustrates the performance of our method with growing number of optimization steps. While the performance gain saturates at around $70-80$ steps, we use 25 steps in following experiments as a compromise between accuracy and speed. For the case of $F = 8$ input images optimization for 25 steps takes $\approx$ 10s on a single Volta V100 GPU. We believe 10s is a practical waiting time and a good compromise for many applications. Therefore we fix the time budget to 10s for the following experiments.

Including more input views at test time can potentially improve the performance of the method. However, in practice, this means more data preprocessing and longer inference times. Fig. D.8 illustrates the performance with different number of input images. Perhaps surprisingly, the performance saturates already at around 5 images before optimization. After optimization, the error saturates at around 8 images. While more images potentially means better supervision, we cannot see improved results for optimization on many images. This can be explained with the fixed time budget in this experiment, where more images mean fewer optimization steps. While we could potentially use fewer images, we found $F = 8$ views as a practical number of input views. This has the following reason: A calculated avatar should not only be numerically accurate but also visually appealing. Results based on more number of views show more fine details and most importantly allow accurate texture calculation.

D.4.4 Type of supervision

Since videos are easier to obtain than 3D annotations, we evaluate to which extent they can substitute full 3D supervision to train our network. To this end, we split the LifeScans dataset. One part is used for full supervision, the other part is used for weak supervision in form of image masks and 2D keypoints. All forms of supervision can be synthetically generated from the LifeScans dataset. We train f_w^* with 10%, 20%, 50%, and 100% full supervision and compare the performance on the test set in Tab. D.2. In order to factor out the effect of problematic poses during the training, we used ground truth poses in this experiment. The results suggest that f_w^* can be trained with only minimal amount of

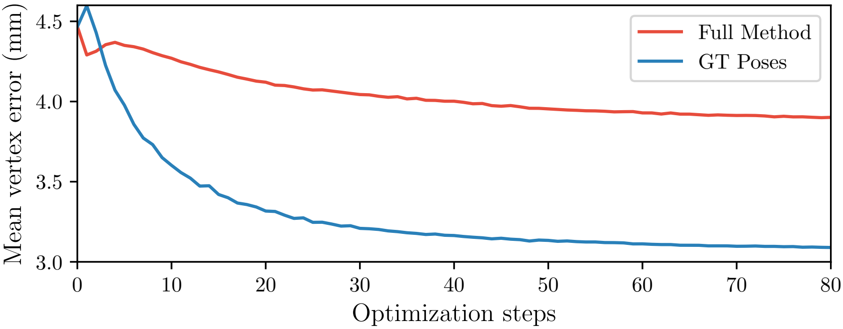

Figure D.7: Error decrease of the test set with increased number of optimization steps computed on $F = 8$ input images.

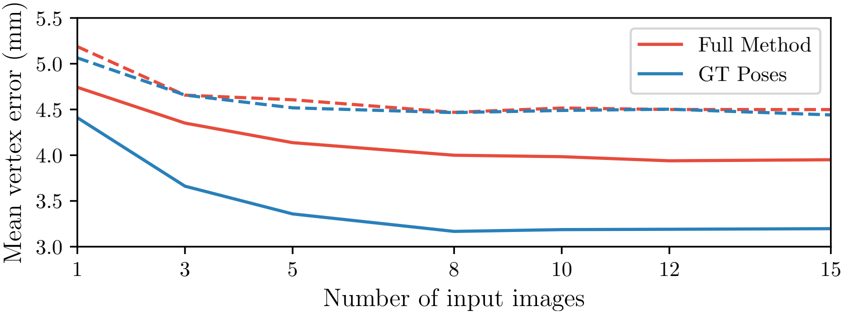

Figure D.8: Error development on the test set with increased number of input views F before (dashed) and after optimization (solid). Optimization has been limited by a time budget of 10s allowing very few gradient steps for large numbers of views, which explains why the error plateaus for more than 8 views.

	Before optimization	After optimization
100%	4.47 ±4.41	3.17 ±3.41
50%	4.57 ±4.52	3.19 ±3.43
20%	4.74 ±4.65	3.29 ±3.53
10%	4.73 ±4.56	3.46 ±3.62

Table D.2: Mean vertex error (mm) of 55 test samples with different amount of *full supervision* during training of the shape branch. f_w^* can be trained with only 10% full supervision with minimal accuracy lose.

full supervision, given strong pose predictions. The performance of the network decreases only slightly for less than 100% full supervision. Most interestingly, the results are almost identical for 10%, 20%, and 50% full supervision. This experiment suggests that we could potentially improve performance by supervising our model with additionally recorded videos. We leave this for future work.

D.4.5 Qualitative results and comparisons

We qualitatively compare our method against the most relevant work [7] on their *PeopleSnapshot* dataset. While their method leverages 120 frames, we still use $F = 8$ frames for our reconstructions. For a fairer comparison, we optimize for $\approx$ 20s in this experiment. This is still several magnitudes faster than the 122min needed by [7]. Their method needs 2 minutes for shape optimization plus 1 minute per frame for the pose. In Fig. D.6 we show side-by-side comparison to [7]. Our results are visually still on par while requiring a fraction of the data.

We also compare our method against [25], a RGB-D based optimization method. Their dataset displays subjects in minimal clothing rotating in front of the camera in T-pose. Unfortunately, the semantic segmentation network is not able to successfully segment subjects in minimal clothing. Therefore we sightly change the set-up for this experiment. We segment their dataset using the semi-automatically approach [33] and re-train our predictor to be able to process binary segmentation masks. Additionally, we augment the LifeScans dataset with T-poses. We show side-by-side comparisons in Fig. D.9. Again our results are visually similar, despite the use of less and only monocular data.

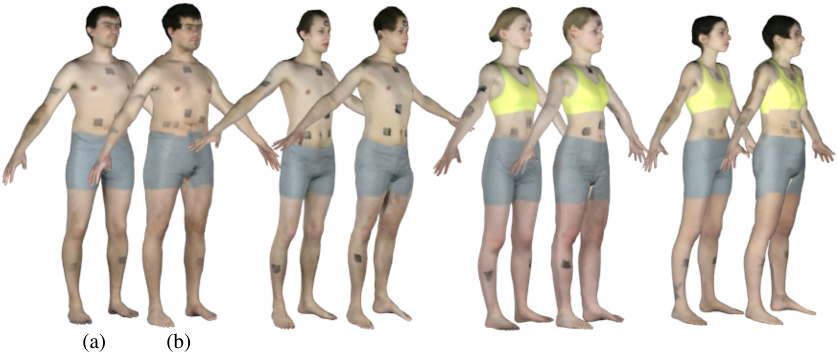

Figure D.9: Comparison to the RGB-D method [25] (a). Our method (b) is visually on par, despite using only 8 RGB images as input.

D.5 Discussion and Conclusion

We have proposed a novel method for automatic 3D body shape estimation from only $1-8$ frames of a monocular video of a person moving. Our *Octopus* model predicts mesh-based pose invariant shape and per-image 3D pose from a flexible number of views. Experiments demonstrate that the feed-forward predictions are already quite accurate (4.5mm), but often lack detail and do not perfectly overlap with the input images. This motivates refining the estimates with top-down optimization against the input images. Refining brings the error down to 4mm and aligns the model with the input image silhouettes, which allows texture mapping. In summary, we improve over the state-of-the-art in the following aspects: Our method allows, for the first time, to estimate full body reconstructions of people in clothing in a fully automatic manner. We significantly reduce the number of needed images at test time, and compute the final result several magnitudes faster than state-of-the-art (from hours to seconds). Extensive experiments on the LifeScans dataset demonstrate the performance and influence of key parameters of the predictor. While our model is independent on the number of input images and can be refined for different numbers of optimization steps, we have shown that using 8 views and refining for 10 seconds are good compromises between accuracy and practicability. Qualitative results on two real-world datasets demonstrate generalization to real data, despite training from synthetic data alone.

Future work should enable the proposed method for scenarios where the subject is not cooperating, for example from Youtube videos, or legacy movie material. Furthermore, clothing with geometry far from the body, such as skirts and coats or hairstyles like ponytails will require a different formulation.

By enabling fully automatic 3D body shape reconstruction from a few images in only a few seconds, we prepare the ground for wide-spread acquisition of personalized 3D avatars. People are now able to quickly digitize themselves using only a webcam and can use their model for various VR and AR applications.

D.6 Appendix: Implementation Details

In the following, we present implementation details of the presented method. In Sec. D.6.1, we explain of the parametrization of the body model. In Sec. D.6.2, we discuss the top-down optimization in detail.

D.6.1 Body model parametrization

The SMPL body model $M(\cdot)$ is a function of pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ (see Sec. 3.1 in main paper). The pose $\boldsymbol{\theta}$ is parametrized by 23 ball-joints and global rotation in axis-angle representation modeled in a kinematic chain. Additionally, we move SMPL in space by 3D translation $\boldsymbol{t}$. The shape $\boldsymbol{\beta}$ is parameterized by the first 10 coefficients of a PCA shape space learned from body scans. This parametrization comes in three variants: A PCA space describing only female body shapes, a space for male shapes, and a joint shape space describing body shapes across genders. In order to keep the fully-automatic property of our method, we estimate body shapes based on the joint shape space.

D.6.2 Instance-specific top-down optimization

During top-down optimization (see Sec. 3.4 in main paper), we refine the predictor f^*_w for a specific instance using weak supervision. Besides losses on input images $\mathcal{L}_{N_{2D}}$ and 2D joints $\mathcal{L}_{J_{2D}}$, we further regularize the optimization with losses on the vertices, that preserve mesh sanity, inspired by [7]. Hereby, body regions that are typically unclothed are more regularized by per-vertex weights $\boldsymbol{\sigma}, \boldsymbol{\tau}$. We enforce smooth meshes with Laplacian mesh regularization [222]. Similar smoothness as the undressed body shape is achieved by:

$$\mathcal{L}_{\text{lp}} = \sum_{j=0}^{V-1} \boldsymbol{\sigma}_j ||L(\boldsymbol{v}_j) - L(\hat{\boldsymbol{v}}_j)||^2, \tag{D.17}$$

where $\boldsymbol{v}_*$ are vertices of $S(\mathcal{I})$, $\hat{\boldsymbol{v}}_*$ are vertices of the undressed body $\hat{S}(\mathcal{I})$, and L is the Laplace operator calculated with cotangent weights. Furthermore, we take advantage of the fact, that bodies are mostly axially symmetrical with respect to the Y-axis:

$$\mathcal{L}_{\text{sym}} = \sum_{i=0}^{F-1} \sum_{(j,k)\in\mathcal{S}} \boldsymbol{\tau}_{j,k} ||[-1, 1, 1]^T \cdot \boldsymbol{v}_{i,j} - \boldsymbol{v}_{i,k}||^2, \tag{D.18}$$

where $\boldsymbol{v}_{i,*}$ are vertices of $N_{3D}(\mathcal{I}, \mathcal{J}, i)$ and $\mathcal{S}$ contains all pairs of Y-symmetric vertices.

To further improve the visual fidelity of the results, we include 2D face keypoint matching from [6]. To this end, we compute face keypoints $\mathcal{K} = \{\boldsymbol{K}_0, \ldots, \boldsymbol{K}_{F-1}\}$ per image and include a 2D reprojection loss:

$$\mathcal{L}_{\text{face}} = \sum_{i=0}^{F-1} ||\pi_c(K_{70}(N_{3D}(\mathcal{I}, \mathcal{J}, i))) - \boldsymbol{K}_i||^2, \tag{D.19}$$

where K_{70} has been trained to output 70 face keypoints. As demonstrated in [6], this loss enforces the produced results to look more like their human counterpart. The loss used for fine-tuning f_w^* finally is:

$$\mathcal{L}_{\text{opt}} = w_{N_{2D}}\mathcal{L}_{N_{2D}} + w_{J_{2D}}\mathcal{L}_{J_{2D}} + \mathcal{L}_{\text{reg}} \tag{D.20}$$

$$\mathcal{L}_{\text{reg}} = w_{\text{lp}}\mathcal{L}_{\text{lp}} + w_{\text{sym}}\mathcal{L}_{\text{sym}} + w_{\text{face}}\mathcal{L}_{\text{face}}, \tag{D.21}$$

with weights w_* that balance the influence of different terms.

Errata

Compared to the original publication and in addition to editorial changes, the following corrections have been made:

- The call of Equation D.5 has been corrected in Equation D.4.

E Tex2Shape: Detailed Full Human Body Geometry From a Single Image

Thiemo Alldieck[1], Gerard Pons-Moll[2], Christian Theobalt[2],
and Marcus Magnor[1]

[1] Computer Graphics Lab, TU Braunschweig

[2] Max Planck Institute for Informatics, Saarland Informatics Campus

Abstract

We present a simple yet effective method to infer detailed full human body shape from only a single photograph. Our model can infer full-body shape including face, hair, and clothing including wrinkles at interactive frame-rates. Results feature details even on parts that are occluded in the input image. Our main idea is to turn shape regression into an aligned image-to-image translation problem. The input to our method is a partial texture map of the visible region obtained from off-the-shelf methods. From a partial texture, we estimate detailed normal and vector displacement maps, which can be applied to a low-resolution smooth body model to add detail and clothing. Despite being trained purely with synthetic data, our model generalizes well to real-world photographs. Numerous results demonstrate the versatility and robustness of our method.

Published in *IEEE International Conference on Computer Vision*, IEEE, 2019, pp. 2293-2303.

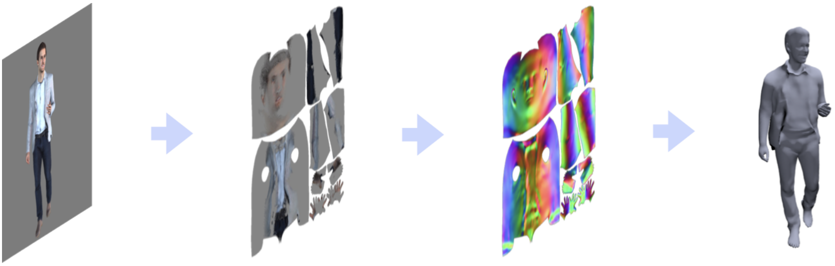

Figure E.1: We present an image-to-image translation model for detailed full human body geometry reconstruction from a single image.

E.1 Introduction

In this paper, we address the problem of automatic *detailed* full-body human shape reconstruction from a single image. Human shape reconstruction has many applications in virtual and augmented reality, scene analysis, and virtual try-on. For most applications, acquisition should be quick and easy, and visual fidelity is important. Reconstructed geometry is most useful if it shows hair, face, and clothing folds and wrinkles at sufficient detail – what we refer to as detailed shape. Detail adds realism, allows people to feel identified with their self-avatar and their interlocutors, and often carries crucial information.

While a large number of papers focus on recovering pose, and rough body shape from a single image [168, 112, 173, 26], much fewer papers focus on recovering detailed shapes. Some recent methods recover pose and non-rigid deformation from monocular video [264], even in real-time [80]. However, they require a pre-captured static template of each subject. Other recent works [7, 8] recover static body shape, and clothing as displacements on top of the SMPL body model [138] (model-based), or use a voxel representation [240, 156]. Voxel-based methods [240, 156] often produce errors at the limbs of the body and require fitting a model post-hoc [240]. Model-based methods are more robust, but results tend to lack fine detail. We hypothesize there are three reasons for this. Firstly, they rely mostly on silhouettes for either fitting [7], or CNN-based regression plus fitting [8], ignoring the rich illumination and shading information contained in RGB values. Secondly, the regression from image pixels directly to 3D mesh displacements is hard because inputs and outputs are *not aligned*. Furthermore, prediction of high-resolution meshes requires mesh-based neural networks, which are very promising but are harder to train than standard 2D CNNs. Finally, they rely on 3D pose estimation, which is hard to obtain accurately.

Based on these observations, our idea is to turn the shape regression into an *aligned* image-to-image translation problem (see Fig. E.1). To that end, we map input and output pairs to the pose-independent UV-mapping of the SMPL model. The UV-mapping unfolds the body surface onto a 2D image such that every pixel corresponds to a 3D point on the body surface. Similar to [160], we map the visible image pixels to the UV space using DensePose [12] obtaining a partial texture map image, which we use as input. Instead of regressing details directly on the mesh, we propose to regress shape as UV-space displacement and normal maps. Every pixel stores a normal and a displacement vector from a smooth shape (in the space of SMPL) to the detailed shape. We call our model to Tex2Shape.

We train Tex2Shape with a dataset of 2043 3D scans of people in varying clothing, poses, and shapes. To map all scan shapes to the UV-space, we non-rigidly register SMPL to each scan, optimizing for model shape parameters and free-form displacements, and store the latter in a displacement map. Registration is also useful for augmentation; using SMPL, we render multiple images of varying pose and camera view. We further augment the renderings with realistic illumination, which is a strong cue in this problem. Assuming a Lambertian reflectance model, we know that color forms from the dot product of light direction and the surface normal times albedo. Shape-from-shading [276] allows to invert the process and estimate the surface from shading, which was used before to refine geometry of stereo-based [259] or multi-view-based human performance capture results [258, 130]. After synthesizing image pairs, we train a Pix2Pix network [103] to map from partial texture maps to complete normal and displacement maps and a second small network for estimating SMPL body shape parameters.

Several experiments demonstrate that our proposed data pre-processing undoubtedly pays-off. Trained only from synthetic images, our model can robustly produce, in one shot, *full 3D shapes* of people with varied clothing, shape, and hair. In contrast to models that produce normals or shading only for the visible image part, Tex2Shape hallucinates the shape also for the *occluded* part – effectively performing translation and completion together. In summary, our contributions are:

- We turn a hard full-body shape reconstruction problem into an easier 3D pose-independent image-to-image translation one. To the best of our knowledge, this is the first method to infer detailed body shape as image-to-image translation.

- From a single image, our model can regress full 3D clothing, hair and facial details in 50 milliseconds.
- Experiments demonstrate that, while very simple, Tex2Shape is very effective and is capable of regressing full 3D clothing, hair and facial details in a static reference pose in one shot.
- Tex2Shape is available for research purposes[1].

E.2 Related Work

Human shape reconstruction is a wide field of research, often jointly approached with pose reconstruction. In the following, we review methods for human pose and shape reconstruction from monocular image and video. Full body methods are often inspired by methods for face geometry estimation. Hence, we include face reconstruction in our review. When it comes to detailed reconstruction, clothing plays an important role. Therefore, we conclude with a brief overview of garment reconstruction and modeling.

Pose and shape reconstruction. Methods for monocular pose and shape reconstruction often utilize parametric body models to limit the search space [14, 84, 138, 182, 109], or use a pre-scanned static template to capture pose and non-rigid surface deformation [264, 80]. To recover pose and shape, the 3D body model is fitted against 2D poses. In early works 2D poses have been entirely or partially manually clicked [76, 280, 106, 198], later the process was automated [26, 124] with 2D landmark detections from deep neural networks [177, 101, 36]. In recent work, the SMPL [138] model has been integrated into network architectures [112, 173, 168, 239]. This further automates and robustifies the process. All these works focus mostly on robust pose detection. Shape estimation is often limited to surface correlations with bone lengths. Most importantly, the shape is limited to the model space. In contrast, we focus only on shape and estimate geometry details beyond the model space.

Clothing and hair can be obtained by optimization-based methods [7, 6]. From a video of a subject turning around in A-pose, silhouettes are fused in

---

[1] http://virtualhumans.mpi-inf.mpg.de/tex2shape/

canonical pose. In the same setting, the authors in [8] present a hybrid learning and optimization-based method, that makes the process completely automatic, fast, and dependent only on a handful of images. However, all these methods can only process A-poses and depend on robust pose detection. The method in [255] loosens this restriction and creates humanoid shapes from a single image via 2D warping of SMPL parameters, but only partially handles self-occlusion. Another recent line of research estimates pose and shape in form of a voxel representation [241, 105, 156], which allows for more complex clothing but limits the level of detail. In [278] the authors alleviate this limitation by augmenting the visible parts with a predicted normal map. In contrast, we present 3D pose-independent shape estimation in a reference pose with high-resolution details also on non-visible parts.

Several previous methods exploited shading cues in high-frequency texture to estimate high-frequency detail. For instance, they estimated lighting and reflectance to compute shape-from-shading-refined geometry of a human template from stereo [259] or multi-view imagery [258, 130].

Face reconstruction. Several recent monocular face reconstruction and performance capture methods use shading-based refinement for geometry improvement, e.g., in analysis-by-synthesis fitting [207] or refinement, or in a trained neural network [208, 99]. Also related to our approach are recent works integrating a differentiable face renderer in a neural network to estimate instance correctives of geometry and albedo relative to a base model [230], or learn an identity geometry and albedo basis from scratch from video [232].

Garment reconstruction and modeling. Body shape under clothing has been estimated without [274] and jointly with a separate clothing layer [184] from 3D scans and from RGB-D [227]. [266] introduces a technique, which allows complex clothing to be modeled as offsets from the naked body. The work in [251] describes a model that encodes shape, garment sketch, and garment model, in a single shared latent code, which enables interactive garment design. High frequency wrinkles are predicted as a function of pose either in UV space using a CNN [123, 108] or directly in 3D using a data-driven optimization method [186]. All these methods [123, 266, 108] target realistic *animation* of clothing and can only predict garments in isolation [123, 108]. Learning based normals and depth recovery [19] or meshes [49] has been demonstrated but again only for single garments. In contrast, our approach is the first to reconstruct the detailed shape of a *full-body* from a *single image* by learning an image-to-image mapping.

E.3 Method

The goal of this work is to create an animatable 3D model of a subject from a single photograph. The model should reflect the subject's body shape and contain details such as hair and clothing with garment wrinkles. Details should be present also on body parts that have not been visible in the input image, e.g. on the back of the person. In contrast to previous work [156, 255, 8] we aim for fully automatic reconstruction which does not require accurate 3D pose. To this end, we train a Pix2Pix-style [103] convolutional neural network to infer normals and vector displacement (*UV shape-images*) on top of the SMPL body model [138]. To align the input image with the output UV-shape images, we extract a partial UV texture map of the visible area using off-the-shelf methods [12, 112]. An overview is given in Fig. E.2. A second small CNN infers SMPL shape parameters from the image (see Sec. E.5.1). In Sec. E.3.1 we describe the parametric body model used in this work, and in Sec. E.3.2 we explain our parameterization of appearance, normals, and displacements.

E.3.1 Parametric body model

SMPL is a parameterized body model learned from scans of subjects in minimal clothing. It is defined as a function of pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ returning a mesh of $N = 6890$ vertices and $F = 13776$ faces. Shape $\boldsymbol{\beta}$ corresponds to the first 10 principal components of the training data subjects. Since scale is an inherent ambiguity in monocular images, we made $\boldsymbol{\beta}$ independent of body height in this work. Our method estimates $\boldsymbol{\beta}$ with a standardized height and is independent of pose $\boldsymbol{\theta}$. Details that go beyond the SMPL shape space are added via UV displacement and normal maps (UV shape-images), as described in Sec. E.3.2. During the dataset generation (see Sec. E.4), we use SMPL to synthesize images of humans posing in front of the camera.

E.3.2 UV parameterization

The SMPL model describes body shapes with a mesh containing 6890 vertices. Unfortunately, this resolution is not high enough to explain fine details, such

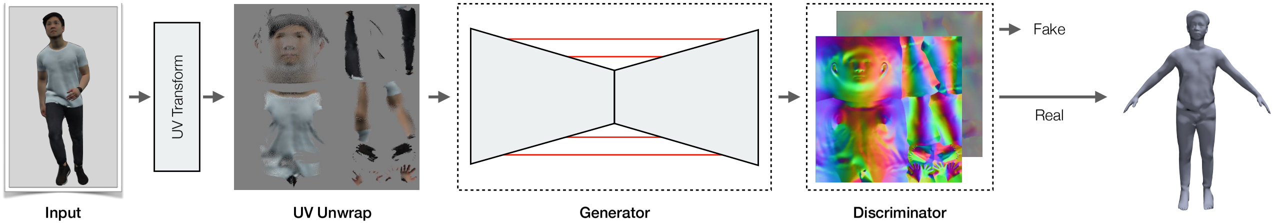

Figure E.2: Overview of the key component of our method: A single photograph of a subject is transformed into a partial UV texture map. This map is then processed with a U-Net with skip connections that preserve high-frequent details. A PatchGAN discriminator enforces realism. The generated normals and displacements can be applied to the SMPL model using standard rendering pipelines.

as garment wrinkles. Another problem is that meshes do not live on a regular 2D grid like images, and consequently require taylored solutions [30] that are not yet as effective as standard CNNs on the image domain. To leverage the power of standard CNNs, we propose to use a well-established parameterization of mesh surfaces: UV mapping [24]. A UV map unwraps the surface onto an image, allowing to represent functions defined on the surface as images. Hereby, U and V denote the 2 axes of the image. The mapping is defined once per mesh topology and assigns every pixel in the map to a point on the surface via barycentric interpolation of neighboring vertices. By using a UV map, a mesh can be augmented with geometric details of a resolution proportional to the UV map resolution.

We augment SMPL using two UV maps, namely normal map and vector displacement map. A normal map contains new surface normals, that can add or enhance visual details through shading. A vector displacement map contains 3D vectors that displace the underlying surface. Displacements and normals are defined on the canonical T-pose of SMPL. The input to our neural network is a partial texture map of the visible pixels on the input photograph (see Sec. E.5.3).

E.4 Dataset Generation

To learn our model we synthesize a varied dataset from real 3D scans of people. Specifically, we synthesize images of humans in various poses under realistic illumination paired with normal maps, displacement maps, and SMPL shape parameters $\boldsymbol{\beta}$. The large majority of scans (1826) was kindly provided from Twindom (https://web.twindom.com/). We additionally purchased 163 scans from renderpeople.com and 54 from axyz-design.com. These scans do not share the same mesh layout, and therefore we can not directly compute coherent normal and displacement maps. To this end, we non-rigidly register the SMPL model against each of the scans. This ensures that all vertices share the same contextual information across the dataset. Furthermore, we can change the pose of the scans using SMPL. Unfortunately, non-rigid registration of clothed people is a very challenging problem itself (see Sec. E.4.1), and often results in unnatural shapes. Hence, we manually selected 2043 high quality registrations. Unfortunately, our current dataset is slightly biased towards men because registration currently fails more often for women, due to long hair, skirts and dresses. Of the 2043 scans, we reserve 20 scans for validation and 55 scans for testing.

In the following, we explain our non-rigid registration procedure in more detail and describe the synthetization of the paired dataset for training of the models.

E.4.1 Scan registration

As discussed in Sec. E.3.1, $N = 6890$ vertices are not enough to explain fine details. To this end, we sub-divide each face in SMPL into four, resulting in a new mesh consisting of $N = 27554$ vertices and $F = 55104$ faces. This high-resolution mesh can better explain fine geometric details in the scans. While joint optimization is generally desirable, registration is much more robust when done in stages: we first compute 3D pose, then body shape and finally non-rigid details. We start the registration by reconstructing the pose of the scan subject. Therefore, we find 3D landmarks by rendering the scan from multiple cameras and minimizing the 2D re-projection error to 2D joint OpenPose detections [36]. Then we optimize the SMPL pose parameters $\boldsymbol{\theta}$ to explain the estimated 3D joint locations. Next, we optimize for shape parameters $\boldsymbol{\beta}$ to minimize scan to SMPL surface distance. Here, we make sure SMPL vertices stay inside the scan by paying a higher cost for vertices outside the scan since SMPL can only reliable explain the naked body shape. Finally, we recover fine-grained details by optimizing the location of SMPL vertices. The resulting registrations explain high-frequency details of the scans with the subdivided SMPL mesh layout and can be re-posed.

E.4.2 Spherical harmonic lighting

For a paired dataset, we first need to synthesize images of humans. For realistic illumination, we use spherical harmonic lighting. Spherical harmonics (SH) are orthogonal basis functions defined over the surface of the sphere. For rendering SH are used to describe the directions from where light is shining into the scene [189]. We follow the standard procedure and describe the illumination with the first 9 SH components per color. To produce a large variety of realistic illumination conditions, we convert images of the *Laval Indoor HDR dataset* [68] into diffuse SH coefficients, similar to [111]. For further augmentation, we rotate the coefficients randomly around the Y-axis.

E.4.3 UV map synthetization

To complete our dataset, we calculate UV maps that explain details of the 3D registrations. In UV mapping every face of the mesh has a 2D counterpart in the UV image. Hence, UV mapping is essentially defined through a 2D mesh. Given a 3D mesh and a set of per-vertex information, a UV map can be synthesized through standard rendering. Information between vertices is filled through barycentric interpolation. This means, given the high-resolution registrations, we can simply render detailed UV displacement and normal maps. The displacement maps encode the free-form offsets, that are not part of SMPL. The normal maps contain surface normals in canonical T-pose. These maps are used to augment the standard-resolution naked SMPL, which eliminates the need for higher mesh-resolution or per-vertex offsets. We use the standard-resolution SMPL augmented with the UV maps in all our experiments.

E.5 Model and Training

In the following, we explain the used network architectures, losses, and training schemes in more detail. Further, we explain how a partial texture can be obtained from DensePose [12] results.

E.5.1 Network architectures

Our method consists of two CNNs – one for normal and displacement maps and one for SMPL shape parameters $\boldsymbol{\beta}$. The main component of our method is the Tex2Shape-network as depicted in Fig. E.2. The network is a conditional Generative Adversarial Network (Pix2Pix) [103] consisting of a U-Net generator and a PatchGAN discriminator. The U-Net features each seven convolution-ReLU-batchnorm down- and up-sampling layers with skip connections. The discriminator consists of four of such down-sampling layers. We condition on 512×512 partial textures, based on two observations: First, when mapping pixels from an HD 1024×1024 image to UV, the resolution is high enough to contain most pixels from the foreground, and not too high to prevent large unoccupied regions. Second, using the mesh resolution of the training set, larger UV maps would only contain more interpolated data. See supplemental material for an ablation experiment using smaller UV maps.

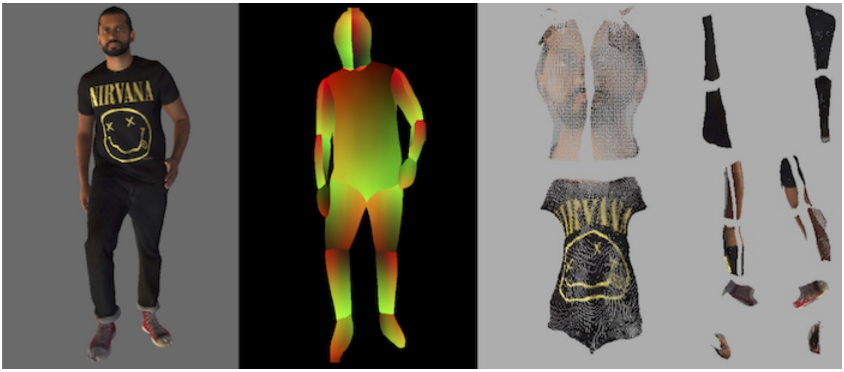

Figure E.3: To create the input to our method, we first process the input image (left) with DensePose. The DensePose result (middle) contains UV coordinates, that can be used to map the input image into a partial texture (right).

The β-network takes 1024×1024 DensePose detections as input. These are then again down-sampled with seven convolution-ReLU-batchnorm layers and finally mapped to 10 β-parameters by a fully-connected layer.

E.5.2 Losses and training scheme

The goal of our method is to create results with high *perceived quality*. We believe structure is more important than accuracy and therefore experiment with the following loss: The structural similarity index (SSIM) was introduced to predict the perceived quality of images. The multi-scale SSIM (MS-SSIM) [252] evaluates the image on different image scales. We maximize the structural similarity of ground truth and predicted normal and displacement maps by minimizing the dissimilarity (MS-DSSIM): $(1 - \text{MS-SSIM})/2$. We further train with the well-established L1-loss and the GAN-loss coming from the discriminator. Finally, the β-network is trained with an L2 parameter loss. We train both CNNs with the Adam optimizer [119] and decay the learning-rate once the losses plateau.

E.5.3 Input partial texture map

The partial texture forming the input to our method is created by transforming pixels from the input image to UV space based on DensePose detections, see

Fig. E.3. DensePose predicts UV coordinates of 24 body parts of the SMPL body model (Fig. E.3 middle). For easier mapping, we pre-compute a look-up table to convert from 24 DensePose UV maps to the single joint SMPL UV parameterization. Each pixel in the DensePose detection now maps to a coordinate in the SMPL UV map. Using this mapping, we compute a partial texture from the input image (Fig. E.3 right).

E.6 Experiments

In the following, we qualitatively and quantitatively evaluate our proposed method. Results on four different datasets and comparisons to state-of-the-art demonstrate the versatility and robustness of our method as well as the quality of results (Sec E.6.1). Further, we study the effect of different supervision losses (Sec. E.6.2), evaluate different methods for UV mapping (Sec. E.6.3), and measure the robustness for different visibility levels (Sec. E.6.4). Finally, in Sec. E.6.5 we demonstrate a potential application of our proposed method, namely garment transfer between subjects. More experiments and ablation studies can be found in the supplemental material. Due to scale ambiguity in monocular images, all results are up to scale. Also, our method does not compute pose. For better inspection, we depict results in ground truth or A-pose. Further, we color-code the results by the used method for UV-mapping (see Sec. E.6.3). Results using DensePose mapping are *green*, *blue* marks ground truth mapping, *red* indicates HMR-based [112] texture reprojection, and ground truth shapes are *grey*.

All results have been calculated at interactive frame-rates. Precisely, our method takes on average 50 ms for displacement map, normal map, and $\boldsymbol{\beta}$-estimation on an NVIDIA Tesla V100. UV mapping using DensePose can be performed in real-time.

E.6.1 Qualitative results and comparisons

We qualitatively compare our work against four relevant methods for monocular human shape reconstruction on the *PeopleSnapshot* dataset [7]. BodyNet [241] is a voxel-based method to estimate human pose and shape from only one image. SiCloPe [156] is voxel-based, too, but recovers certain details by relying on synthesized silhouettes of the subject. HMR [112] is a method to estimate pose and shape from single image using the SMPL body model. In [7] the authors present the first video-based monocular shape reconstruction method, that

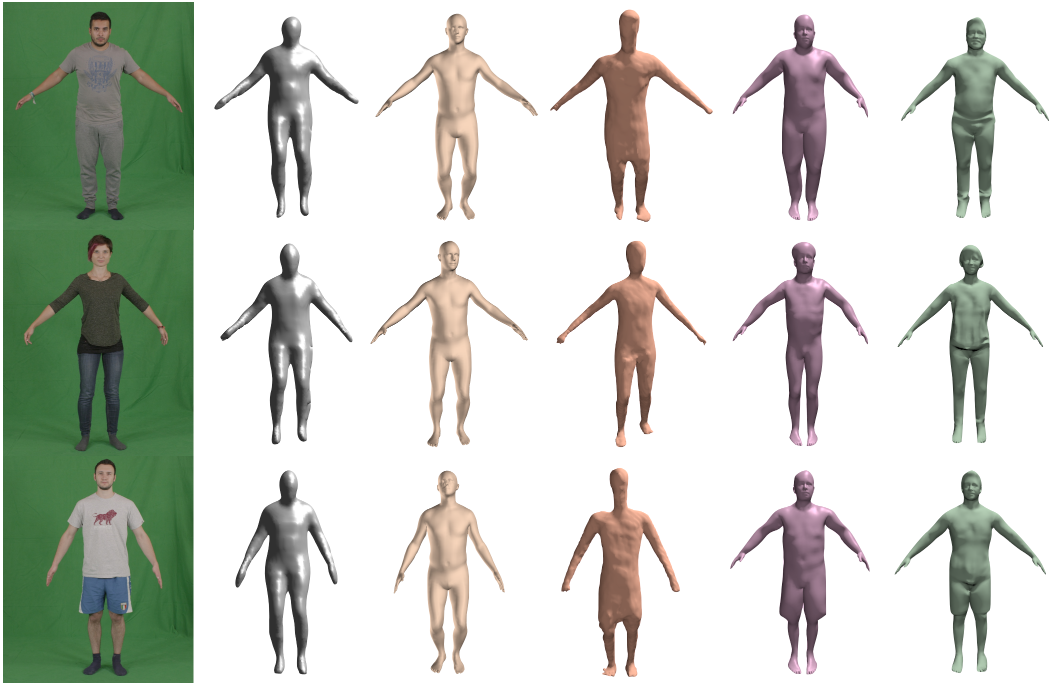

Figure E.4: Our method in comparison to other methods for human shape reconstruction. From left to right: Input image, BodyNet [241], HMR [112], SiCloPe [156], Video Shapes [7], and ours. Our method preserves the highest level of detail.

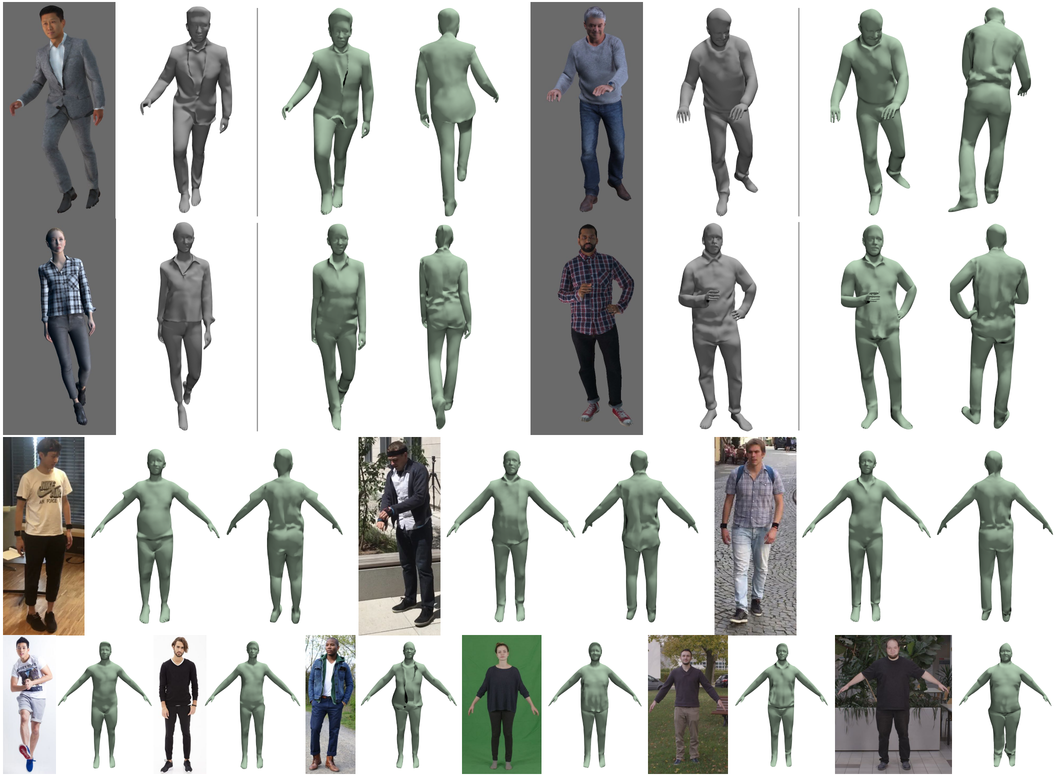

Figure E.5: Our 3D reconstruction results (green) on four different datasets. We compare against ground truth (grey) on our synthetic dataset (rows 1 and 2). Qualitative results on 3DPW (3rd row), DeepFashion (4th row left) and PeopleSnapshot (4th row right) demonstrate, that our model generalizes well to real-world footage. Details on the back of the models are hallucinated by our model.

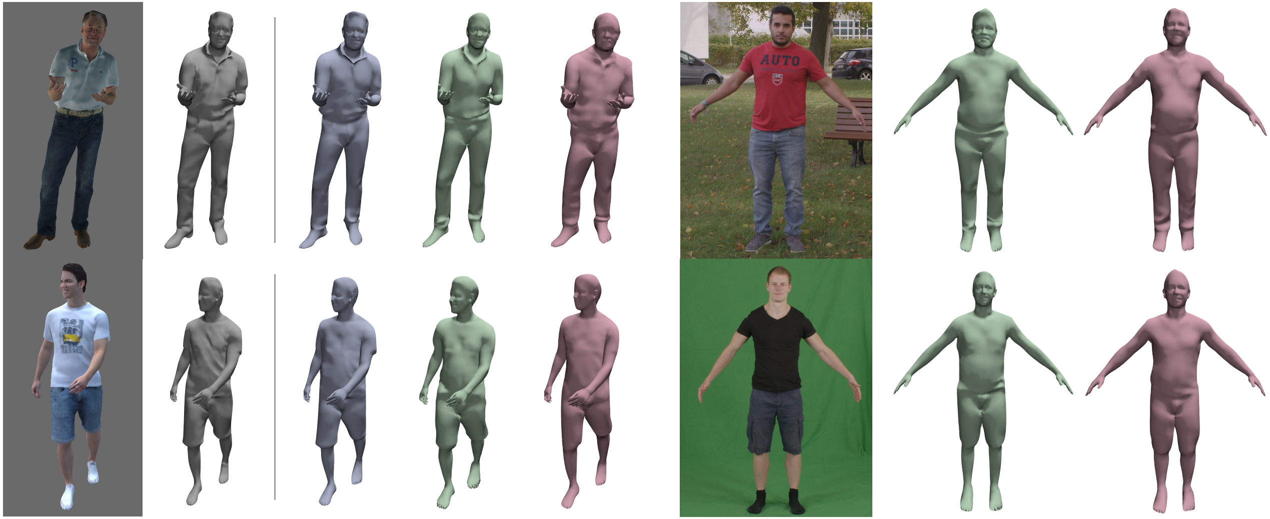

Figure E.6: Results using different UV mapping methods compared against input and ground truth (grey): ground truth UV mapping (blue), DensePose (green), HMR (red).

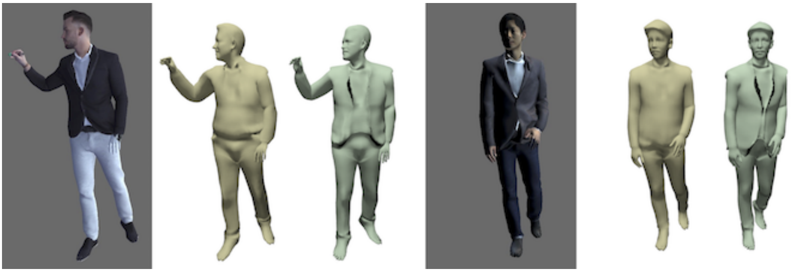

Figure E.7: After training with MS-DSSIM loss enabled (green) complex clothing is reconstructed more reliably, than after training with L1 loss only (yellow).

goes beyond the parameters of SMPL. They use 120 images of the same subject roughly posed in A-poses and fuse the silhouettes into a canonical representation. However, the method is optimization-based and requires to fit the pose in each frame first, which makes the process very slow. In Fig. E.4, we show a side-by-side comparison with our results. Our method clearly features the highest level of detail, even compared to [7] using 120 frames, while our method only takes a single image as input and runs at interactive frame-rates.

In Fig. E.17 we show more results of our method. We compare against ground truth on our own dataset and show qualitative results on *3DPW* [247], *DeepFashion* [134, 135], and *PeopleSnapshot* [7] datasets. Our method successfully generalizes to various real-world conditions. Please note how realistic garment wrinkles are hallucinated on the unseen back of the models. In general, we can see our method is able to infer realistic 3D models featuring hair, facial details, and various clothing including garment wrinkles from single image inputs.

E.6.2 Type of supervision

In Sec. E.5.2, we have introduced the MS-DSSIM loss. The intuition behind using this loss is that for visual fidelity structure is more important than accuracy. To evaluate this design decision, we train a variant of our Tex2Shape network with L1 and GAN losses only. Since it is not straight forward to quantify better structure, we closely inspect our results on a visual basis. We find, that the variant trained with MS-DSSIM loss is able to reconstruct complex clothing more reliably. Examples are shown in Fig. E.7. Note that the results computed with MS-DSSIM loss successfully reconstruct the jackets.

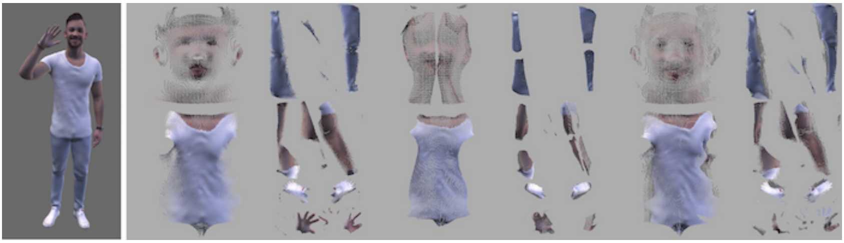

Figure E.8: Partial textures computed with different methods. From left to right: Input, ground truth UV mapping, DensePose, HMR.

E.6.3 Impact of UV mapping

Our method requires to first map an input image to a partial UV texture. We propose to use DensePose [12], which makes our method independent of the 3D pose of the subject. In the following, we evaluate the impact of the choice of UV mapping on our method. To this end, we train three variants of our network. Firstly, we train with ground truth UV mappings calculated from the scans. We render the scan's UV coordinates in image space, that are then used for UV mapping, similar to the mapping using DensePose (see Sec. E.5.3). In the following, we refer to this variant as *GT-UV*. Secondly, we train a variant that can be used with off-the-shelf 3D pose estimators. To this end, we render UV coordinates of the naked SMPL model without free-form offsets. This way only pixels that are covered by the naked SMPL shape are mapped, what simulates UV mapping as created from results of 3D pose detectors (*3D pose variant*). Finally, we compare with our standard training procedure using DensePose. A comparison of partial textures created with the three variants is given in Fig. E.8. Note how we lose large parts of the texture by using DensePose mapping.

To evaluate the 3D pose variant, we choose HMR [112] as 3D pose detector. Unfortunately, the results of HMR do not always align with the input image what produces large errors in the UV space. To this end, we refine the results by minimizing the 2D reprojection error of SMPL joints to OpenPose [36] detections. We choose dogleg optimization and optimize for 20 steps.

In Fig. E.6 we show a side-by-side comparison of the three variants. While GT-UV and DensePose variants are almost identical, the 3D pose variant lacks some detail and introduces noise in the facial region. This is caused by the fact, that perfect alignment is still not achieved even after pose-refinement. The GT-UV and DensePose variants differ the most in hairstyle and at the boundary

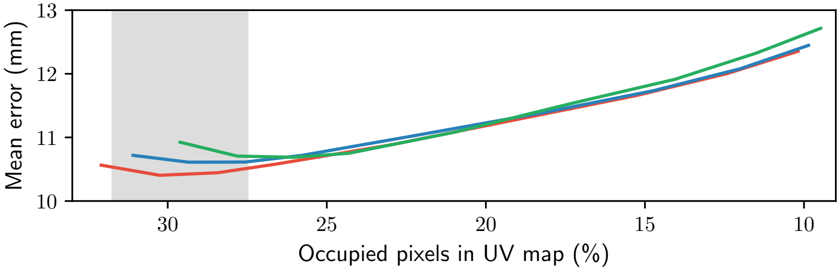

Figure E.9: Average displacement error for three different poses (red: A-pose, blue: walking, green: posing sideways with hands touching) and different distances to the camera. The shaded region marks the margin of trained UV map occupancy.

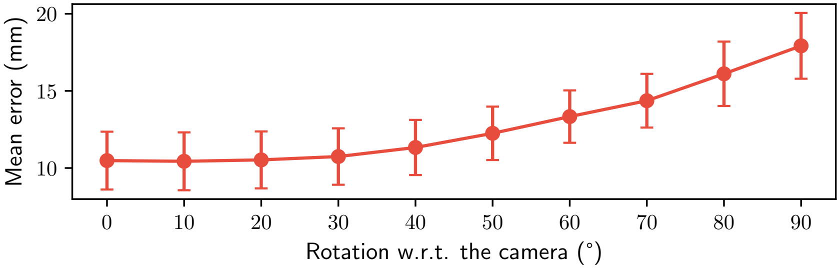

Figure E.10: Average displacement error for A-posed subjects and different rotations around Y-axis with respect to the camera. Our model has been trained on rotations $\pm 20^\circ$.

of the shorts, what is not surprising since hair and clothing are only partially mapped by DensePose. However, both variants closely resemble ground truth results. The DensePose and 3D pose mapping variants can directly be used on real-world footage, while only being trained with synthetic data.

E.6.4 Impact of visibility

In the following, we numerically evaluate the robustness of our method to different visibility settings caused by different poses and distances to the camera. The following results have been computed using GT UV mapping to factor out noise introduced by DensePose. Which pixels can be mapped to the UV partial texture is determined by the subject's pose and distance to the camera. Parts of the body might be not visible (e.g. the subject's back) or occluded by other body parts. If the subject is far away from the camera, it only covers only a small area of the image and thus only a small number of pixels can be mapped.

In Fig. E.9 we measure how this influences the accuracy of our results. Over a test-set with 55 subjects, we synthesize images of three different poses with various distances to the camera. The three poses are A-pose, walking towards the camera, and posing sideways with hands touching. We report the mean per-pixel error of 3D displacements maps (including unseen areas) against the percentage of occupied pixels in the partial texture. For all three poses, the error increases linearly, even for untrained texture occupations. Not surprisingly, the minimum of all three poses lies in the margin of trained occupations. Admittedly, for higher occupations, the error slightly goes up what is caused by the fact, that the network was not trained for scenarios where the subject fully covers the input image.

In Fig. E.10, we study the robustness of our method against unseen poses. We trained the network with images of humans roughly facing the camera. Therefore, we randomly sampled poses in our dataset and Y-axis rotations between $\pm 20°$. In this experiment, we rotate an A-pose around the Y-axis and report the mean per-pixel 3D displacement error. From $0°$ to $30°$, the error stays almost identical, after $30°$ it increases linearly. Again this behavior can be explained by the network not being trained for such angles.

Both experiments demonstrate the robustness of our method against scenarios not covered by our training set.

E.6.5 Garment transfer

In our final experiment, we want to demonstrate a potential application of our method, namely garment transfer or virtual try-on. We take several results of our method and use them to synthesize a subject in different clothing. To achieve this, we keep the SMPL shape parameters $\boldsymbol{\beta}$. Then we alter normal and displacement maps according to a different result. Hereby, we keep details in the facial region, to preserve the subject's identity and hair-style. Since we edit in UV space, this operation can simply be done using standard image editing techniques. In Fig. E.11 we show a subject in three different synthesized clothing styles.

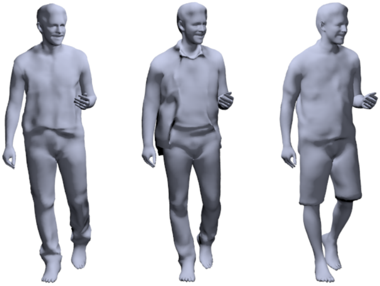

Figure E.11: Since all reconstructions share the same mesh layout, we can extract clothing styles and transfer them to other subjects.

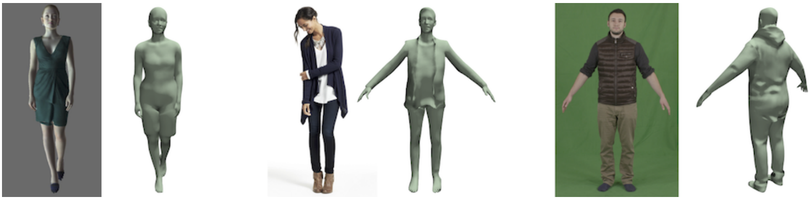

Figure E.12: Failure cases of our method: The predictor confuses a dress with short pants, a female subject with a male, and hallucinates a hood from a collar.

E.7 Conclusion

We have proposed a simple yet effective method to infer full-body shape of humans from a single input image. For the first time, we present single image shape reconstruction with fine details also on occluded parts. The key idea of this work is to turn a hard full-body shape reconstruction problem into an easier 3D pose-independent image-to-image translation one. Our model Tex2Shape takes partial texture maps created from DensePose as input and estimates details in the UV-space in form of normal and displacement maps. The estimated UV maps allow augmenting the SMPL body model with high-frequent details without the need for high mesh resolution. Our experiments demonstrate that Tex2Shape generalizes robustly to real-world footage, while being trained on synthetic data only.

Our method finds its limitations in hair and clothing that is not covered by the training set. This is especially the case for long hair and dresses since they cannot be modeled as vector displacement fields. Typical failure cases are depicted in Fig. E.12. These failures can be explained with garment-type or gender confusion, caused by missing training samples. In future work, we would like to further open up the problem of human shape estimation and explore shape representations that allow all types of clothing and even accessories.

We have shown, that by transferring a hard problem into a simple formulation, complex models can be outperformed. Our method lays the foundation for wide-spread 3D reconstruction of people for various applications and even from legacy material.

E.8 Appendix: Additional Results and Experiments

We show here additional experiments to understand the influence of illumination on our model and its robustness to varying camera intrinsics. We evaluate the β-regression network and perform an ablation of the UV map resolution. Finally, we present more qualitative results.

E.8.1 Influence of illumination

As already emphasized in the main paper, shading is potentially a strong cue for our model. In the following, we evaluate the illumination augmentation during training and the robustness of our model to varying illumination.

In order to evaluate the effect of the illumination augmentation during training, we re-trained our model with constant ambient illumination. This means we render the scans using the textures only. While being scanned, the subjects have been exposed to uniform lighting. However, shading is still present in wrinkles and smaller structures. This means, we cannot factor out shading effects completely. Nevertheless, in Fig. E.16 we can see more consistent details for our final method, especially for the faces.

Our model should produce the same or at least a very similar result when applied on two different photos of the same person in the same clothing but under varying illumination. To validate illumination invariance of our model, we took 9 photos of two subjects while rotating the light-source around the subject. In Fig. E.13 we show the different photos and a heat-map illustrating areas with high standard deviation. We see a consistent picture with varying details only in areas of likely fabric movement.

E.8.2 Influence of camera intrinsics

Camera intrinsics are mostly unknown at test time, especially for in-the-wild photos. The focal length is an important camera parameter, which can affect the results of our method. We have trained our model with a fixed focal length. To study the robustness of our method against varying focal length, we render our test set in A-poses with different focal length and distance to the camera. We keep the ratio between distance and focal length fixed, creating a *Vertigo Effect*. In Fig. E.14, we report the mean vertex-to-vertex error of the naked SMPL model under varying focal length. Although the lowest error is obtained

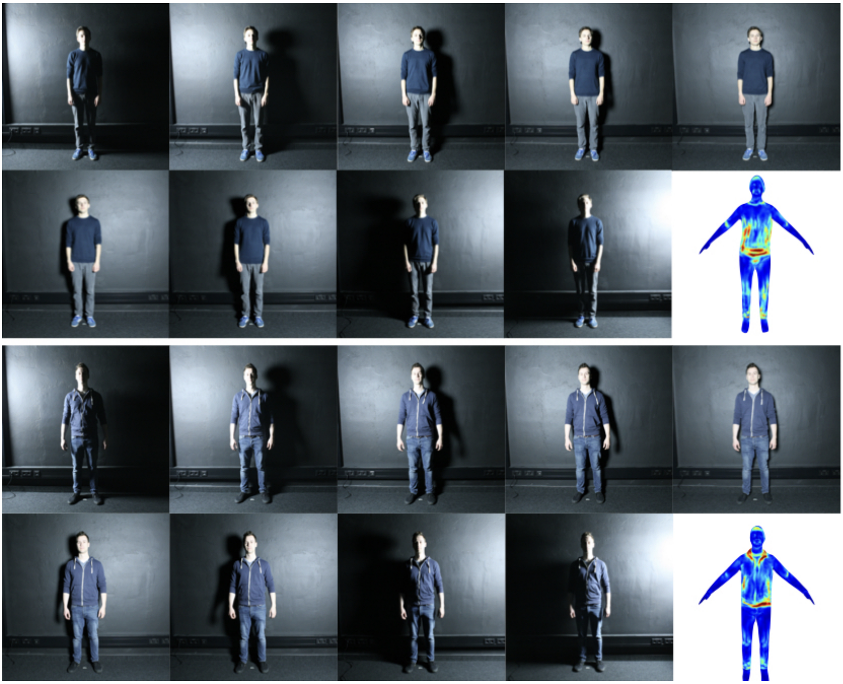

Figure E.13: Displacement reconstruction consistency under varying illumination. The heatmap illustrates the vector norm of per surface point standard deviation (dark-red means $\geq$ 4cm).

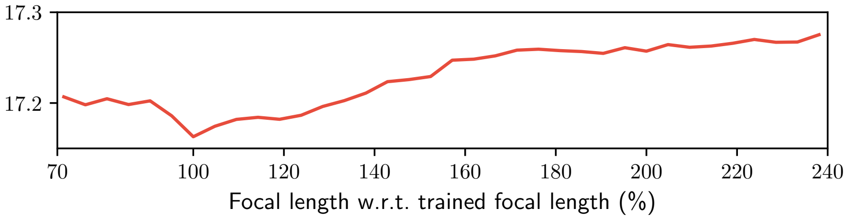

Figure E.14: Mean SMPL vertex-to-vertex error in mm (without added displacements) over the test-set for varying focal length.

for the focal length assumed during training, different focal lengths increase the error only slightly, which demonstrates the robustness of our model.

E.8.3 Numerical comparison with HMR

In order to evaluate the β-regression network, we compare our naked results without added displacements against HMR [112]. Since we do not estimate pose it has to be factored out before comparison. To this end, we follow the established procedure in [25] and adjust pose and scale of the results of both methods to match the ground truth scans. On our test-set, our method using DensePose mapping achieves a mean bi-directional vertex to surface error of 10.57 ± 10.68mm compared to the clothed scans. HMR achieves 16.28 ± 17.05mm. Our method can better estimate the body shapes. This is likely linked to the fact, that our method directly uses dense image-space detections, while HMR correlates surface with bone-lengths. With added displacements, our method achieves 5.19 ± 6.36mm. All results are up to scale.

E.8.4 UV resolution ablation

To evaluate our choice of the UV resolution (512×512px), we train a variant of the network with 256×256px maps. The results look surprisingly good. A close inspection of the results reveals missing details and smoothed edges. An example is shown in Fig. E.15. However, this experiment demonstrates that Tex2Shape can be trained with lower resolution without largely decreased quality.

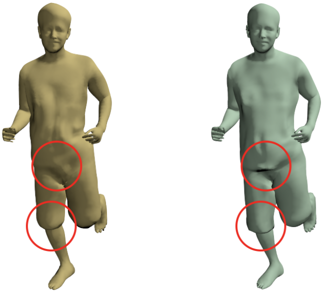

Figure E.15: Comparison of two variants of our network: Using 256×256px resolution (left) decreased the quality only sightly when compared to the original resolution of 512×512px (right).

E.8.5 Additional qualitative results

In Fig. E.17, we show more in-the-wild results of our method on *MonoPerfCap* [264] and *PeopleSnapshot* [7] datasets.

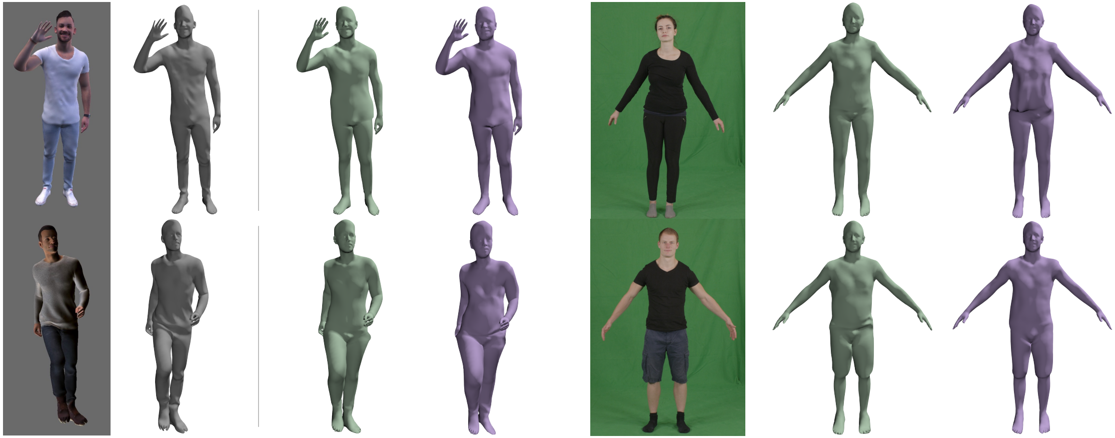

Figure E.16: Our method (green) compared to our method trained without illumination augmentation (purple) and ground truth (grey). Looking closely, we notice worse performance specially on the face region, and artifacts for the method without illumination augmentation. Notice for example the example on the bottom left, the face, legs shape, and chest region is more accurately reconstructed when using augmentation (green).

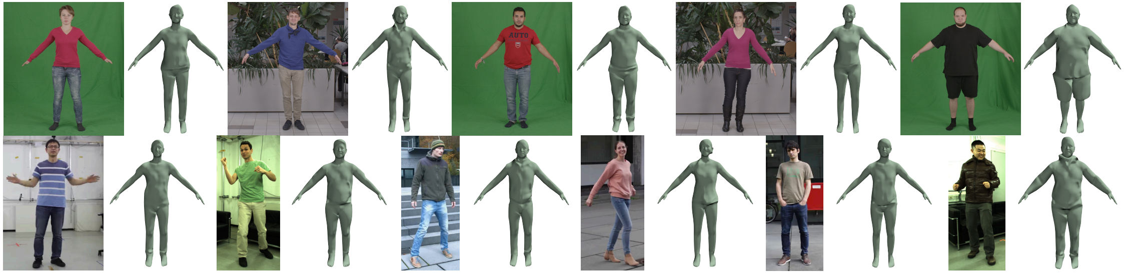

Figure E.17: 3D reconstruction results on two in-the-wild datasets: PeopleSnapshot [7] (1st row) and MonoPerfCap [264] (2nd row).

5 Conclusion and Future Perspectives

The publications presented in this dissertation have studied various aspects of 3D reconstruction of human pose and shape from monocular images. This chapter summarizes the methods, key insights, and contributions of the presented works. Finally, we give an outlook towards future research directions and possible applications.

5.1 Conclusions

This dissertation bundles five publications in the field of 3D reconstruction of human pose and shape from monocular images. While each publication tackles different aspects of the problem, all follow a similar concept: Instead of aiming at reconstructing the 3D shape of the observed human from scratch, we build upon a statistical body model. This way, we make use of a rich prior for the reconstruction process. Compared to competing work, our reconstructions often look more realistic, cf. Figure E.4. Additionally, our strategy ensures that the reconstructed avatars can directly be used by other applications, as our results already come rigged. We will now discuss the individual aspects of the presented works in the global context of 3D virtual human reconstruction from monocular images.

In the first work (Paper A) we presented a method to recover time-consistent 3D human motion from video by utilizing optical flow and silhouette cues. We have introduced a novel differential flow renderer that allows direct interpretation of optical flow via analysis-by-synthesis. Our results demonstrate that optical flow effectively regularizes the under-constrained problem of 3D human motion estimation by partially resolving ambiguities of pure silhouette- or landmark-based approaches. The advent of CNN-based 2D human joint landmark detectors, however, resulted in a paradigm shift in human pose estimation from images. Researchers now mainly focused on single image pose retrieval. Just recently, researchers began to again take time [113] and even optical flow [56] into consideration for regularizing 3D human motion estimation, what demonstrates the relevance of our ideas.

In the following work (Paper B), we shifted our focus to human shape estimation. We presented the first work to estimate 3D human shapes from monocular video of a moving person that goes beyond the parameter space of a parametric model. Our method estimates animatable 3D human avatars including hair, clothing, and surface texture. The method has been well received by the community and was also covered in the media[1],[2]. The key contribution of this work is the transformation of silhouette ray cones into a common frame of reference. We *unpose* the silhouette rays and thereby remove pose from the optimization problem. This significantly reduces memory consumption, speeds up the optimization process, and allows to combine information from many frames. Results on three different datasets demonstrate a mean surface reconstruction accuracy of 4.5mm.

The next project (Paper C) extended the preceding work. Instead of only unposing silhouette rays, we treat the unposing procedure as a general framework. In that sense, we include normals calculated from shape-from-shading and facial landmarks into the shape optimization. Further, we propose a novel texture stitching strategy that builds on a rich semantic prior with per-part appearance models. In order to validate the complex reconstruction method, we performed a user study. The results of the study clearly show that the effort paid off and details matter: 89.64% of the users preferred the enhanced avatars over those of our previous work, and 92.27% perceived them as more realistic.

After improving the visual fidelity of the reconstructions, we focused on the usability of our methods (Paper D). To that end, we trained a CNN that is capable of reconstructing the 3D shape of humans from semantic segmentation of only a few frames, down to a single frame. Knowing that details matter, we propose to refine the results using ‘render and compare’ supervised optimization at test time. Extensive experiments suggest that using eight frames as input and refining the results for 10 seconds results in high-quality reconstructions. Our method is flexible in both the number of input frames and refinement steps, thus the quality can be increased by using more input or longer refinement.

The above publications all introduced novel ideas and significantly advanced state-of-the-art. Nonetheless, we made three important observations:

---

[1] http://www.sciencemag.org/news/2018/04/watch-artificial-intelligence-create-3d-model-person-just-few-seconds-video

[2] https://www.facebook.com/ScienceChannel/videos/10155948465388387/

First, while building on silhouettes and semantic segmentation helps to make the problem tractable and applicable for real-world data, it also abstracts away valuable information. Second, regressing 3D vertex locations from 2D images is an *unaligned problem*. 2D images and 3D meshes can only be compared by projecting and rasterizing the meshes through rendering. Thus, 3D pose plays a significant role in supervision. Incorrect poses lead to decreased quality of the results, as we show in multiple ablation studies. These observations have been the inspiration for our last paper (Paper E). By transforming the input image into an incomplete texture in UV space, we are able to turn 3D shape reconstruction into a pose-independent 2D image-to-image translation problem. From the incomplete texture, we predict normal and displacement maps using a Pix2Pix network [103]. This set-up allowed, for the first time, to predict fine details, such as garment wrinkles, also on occluded parts. The method requires only a single image as input and predicts highly detailed results in only 50ms.

While we have been among the first to present detailed 3D human shape reconstruction methods from monocular images, the field has recently received more and more attention. Concurrent to our work, researchers have presented exciting new applications and reconstruction methods for 3D human shape reconstruction from monocular input [255, 156, 105, 202, 211, 286, 278]. Some of these works use alternative 3D shape representations, namely voxels and implicit functions. While voxel memory consumption grows cubically with respect to the resolution, and implicit functions are not straight-forward to parameterize, both resolve one limitation of our methods: By relying on the SMPL body model, our methods cannot model shapes with a different topology than the human body. Among others, skirts, dresses, long hair, braids, open jackets, ties, scarves, etc. cannot be well represented using a body-shaped mesh. In the following chapter, we discuss possible avenues for resolving this limitation along with additional ideas for extensions to our current methods. Finally, we give an outlook to potential research directions that go beyond our current methodology.

5.2 Future Work and Applications

The works in this dissertation followed the methodology of model-based reconstruction. As described earlier, this means the methods build on a rich prior – a parametric statistical body model. The model is tracked and heavily personalized in order to create virtual avatars of people as seen in the input images. Despite the great advances we have contributed to the field, both the modeling and the reconstruction aspect leave much room for future work, as we elaborate in the following.

When building an avatar from sparse and low-cost sensors, rich parametric models help to reduce the search space and make the problem tractable. State-of-the-art body models, however, only model very few aspects of the human body, mostly the distribution of body shapes and limb proportions among the training set. More detailed models of hands [210, 116, 201], faces [23, 13, 28, 132, 231] and even ears [48] exist. However, only recently have researchers developed joint models that model multiple aspects of the human body [109, 174]. While these models allow for joint reconstruction of human movements, gestures, and facial expressions, they are still far from resembling the true human. Probably the most obvious missing aspect is clothing. While some initial works on modeling clothing and its dynamics exist [77, 184, 123], a complete model of different clothing items or even single garments is yet to be presented. A garment model would not only model the shape variations of the particular garment type, but also wrinkle patterns based on pose changes or motion speed. Further, properties of the fabric, like stiffness and stretching, may be of interest. Similarly, reconstruction of hair can be performed using multi-view capturing set-ups, [141, 155] and first work on single-view reconstruction exist [284]. However, rich dynamic models would robustify the reconstruction process and at the same time would allow one to realistically animate the predicted avatar. Finally, humans wear glasses, shoes, and accessories. Again, first works handle those items explicitly [144] or implicitly [202]. However, no explicit parametrization of the reconstruction is provided. Generally speaking, rich parametric models of human shape and pose, hair, clothing, shoes, and accessories help to better model 3D virtual avatars from sparse and low-cost sensors by providing a rich prior and at the same time allow one to edit and animate the reconstructed avatars realistically and at high frame-rates.

The human body and its statistics have been successfully modeled using 3D meshes, see Section 2.3. However, using meshes as 3D representation in reconstruction problems introduces some limitations in the variety of shapes that be can modeled, as the mesh topology is often assumed to be given. This problem arises when we want to reconstruct objects that may exist in different genera or topologies, e.g. open and clothed shirts. To this end, the best representation for 3D data in a reconstruction problem is still subject of research. Besides meshes, three other forms of representation are commonly used: voxels, point clouds, and implicit surfaces. Simply put, voxels are the 3D extension to a pixel. A voxel represents a value in a regular Euclidian grid in 3D space. The Euclidian property of voxels make them well-suited for CNNs, however the memory consumption grows cubically with the sample density. While octrees and specialized network architectures [193] can alleviate the memory requirements to some extent, the possible level of detail is generally limited. Point clouds model the 3D shape by a sparse set of points in 3D. While this representation is memory efficient, it defines no object surface. Implicit surfaces [83] are another from of representing 3D data. Very recently, they have been rediscovered almost at the same time by multiple research groups for representing 3D data in neural networks [171, 148, 42, 150, 263]. Here the 3D shape is defined as a distance-based function between the object's surface and every point in 3D. Compared to meshes with a fixed surface topology, all three alternative 3D representations have less restrictions in the type of objects they can represent. However, meshes still have several advantages: Firstly, voxels, point clouds, and implicit surfaces have to be converted to meshes before they can be rendered in standard Computer Graphics pipelines. More importantly, the parametrization of meshes is well studied. Many works and tools for rigging, animating, and editing meshes exist, which is not the case for alternative representations. Finally, meshes are a very compact form of storing 3D data. Combining the advantages of all forms of 3D data representation into a new or extended form of a well-known representation is one important avenue to explore for future work.

Another aspect of 3D reconstruction is the reconstruction of the object's appearance. In Computer Graphics, the appearance of an object is modeled by its albedo color or texture, surface reflectance properties, and a light transport function. This means for relighting that the object's reflectance properties and albedo color need to be known. However, decomposition of appearance into reflectance properties and albedo color is not straight-forward. Additionally, only effects that are modeled in the light transport function can be synthesized. To this end, new forms of representing appearance in form of feature textures, point clouds, or volumes have been proposed recently [136, 217, 137, 3]. These representations use neural networks to synthesize images or textures with view-depended effects based on a learned appearance encoding. However, these learned appearances cannot be parameterized with changed illumination conditions and, most importantly, are object- or subject-specific. In future work, it would be interesting to see if these appearance encodings can generalize to new subjects and changing conditions. Finally, future work should enable to predict high detailed appearance on parts that have not been observed, e.g. on the backside of a person.

On the capturing side, our proposed methods are still quite restrictive. Here we intentionally leave much room for future work. Some example directions are simultaneous capturing of multiple people, in-the-wild capturing without any restrictions on performed actions or viewing angles, or capturing in challenging illumination conditions. Another interesting topic would be 3D reconstruction from analog legacy material, even black-and-white film. In an orthogonal direction, future work should not exclude disabled subjects and thus should be able to reconstruct people also with unusual body shapes or missing body parts. Our ultimate goal should be to make our methods work equally well for all diversity among humanity.

Building a realistic-looking avatar and being able to drive it via low-cost sensors is only one interesting aspect of virtual humans. Certain applications might need to work without a real human driving his or her virtual self, for example, virtual assistance. One could think of an Alexa, Siri, or Cortana embodied by an avatar of a real human. Even in a scenario where a real human drives the virtual avatar, the sensor might not be able to capture all subtle micro-expressions and social cues that we humans use to communicate. Yet, we expect the virtual avatar to perform these when talking to us for example in a collaborative multi-user application. In both cases, we need virtual humans that are more than just a 3D template of the real human. Our goal is full immersion in the virtual environment[1]. To achieve this goal, the virtual avatar needs to act, move, talk, and maybe even eventually think like his or her real template. To enable this, we need to understand more about real humans. We need to understand which social cues are important and how to extract and model motion patterns that make us unique.

Having built a virtual avatar of a person does not mean all work is done. To be able to synthesize convincing performances, the model needs to adapt to the real human. On a coarse scale this means the model needs to wear the same clothing, have the same hairstyle, wear the same make-up and so on. On a finer scale this means the model should reflect the current state of the real human. Is he or she tired or refreshed? Does he or she look healthy or sick? Is he or she happy or depressed? All this may reflect in our current appearance. Finally, the virtual human needs to age with the real human.

As realistic virtual humans have the potential to change the way we live and communicate, there are other aspects beyond 3D reconstruction and modeling that need to be investigated by scientists of other professions. These aspects are, but not limited to, security, ethics, social sciences, or display and sensor technologies.

---

[1] cf. German Science Foundation (DFG) project MA2555/15-1 “Digital Immersive Reality”

In this chapter, we have given an outlook on possible research directions for 3D virtual humans particularly and 3D reconstruction in general. We have discussed which aspects might become important once we can convincingly and indistinguishably digitize ourselves and which directions are potentially important to explore to achieve this. In this dissertation, we have presented fundamental and important steps to reconstructing virtual avatars of real humans from monocular video. Our methods allow, for the first time, to build detailed and animatable virtual humans using only a low-cost video camera. We have presented effective methods that enable easy-to-use self-digitization and that pave the path for exciting new applications, as, for example, new forms of communication, camera-based body monitoring, or virtual try-on for online shopping.

Bibliography

[1] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel, "Automatic generation of personalized human avatars from multi-view video," in *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, 2005, pp. 257–260.

[2] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 1446–1455.

[3] K.-A. Aliev, D. Ulyanov, and V. Lempitsky, "Neural point-based graphics," *arXiv preprint arXiv:1906.08240*, 2019.

[4] B. Allain, J.-S. Franco, and E. Boyer, "An Efficient Volumetric Framework for Shape Tracking," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 268–276.

[5] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor, "Optical flow-based 3D human motion estimation from monocular video," in *German Conference on Pattern Recognition*. Springer, 2017, pp. 347–360.

[6] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed human avatars from monocular video," in *International Conference on 3D Vision*. IEEE, 2018, pp. 98–109.

[7] ——, "Video based reconstruction of 3D people models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8387–8397.

[8] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 1175–1186.

[9] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *IEEE/CVF*

International Conference on Computer Vision. IEEE, 2019, pp. 2293–2303.

[10] B. Allen, B. Curless, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 587–594, 2003.

[11] B. Allen, B. Curless, Z. Popović, and A. Hertzmann, "Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis," in *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2006, pp. 147–156.

[12] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 7297–7306.

[13] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3D face recognition with a morphable model," in *8th IEEE International Conference on Automatic Face & Gesture Recognition*, 2008, pp. 1–6.

[14] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: shape completion and animation of people," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 408–416, 2005.

[15] A. O. Bălan and M. J. Black, "The naked truth: Estimating body shape under clothing," in *European Conference on Computer Vision*, 2008, pp. 15–29.

[16] A. O. Bălan, M. J. Black, H. Haussecker, and L. Sigal, "Shining a light on human pose: On shadows, shading and the estimation of pose and shape," in *IEEE International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[17] A. O. Bălan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker, "Detailed human shape and pose from images," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[18] A. Baumberg, "Blending images for texturing 3D models." in *British Machine Vision Conference*, vol. 3, 2002, p. 5.

[19] J. Bednarik, P. Fua, and M. Salzmann, "Learning to reconstruct textureless deformable surfaces from a single view," in *International Conference on 3D Vision*. IEEE, 2018, pp. 606–615.

[20] F. Bernardini, I. M. Martin, and H. Rushmeier, "High-quality texture reconstruction from multiple scans," *IEEE Transactions on Visualization and Computer Graphics*, no. 4, pp. 318–332, 2001.

[21] S. Bi, N. K. Kalantari, and R. Ramamoorthi, "Patch-based optimization for image-based texture mapping," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[23] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[24] J. F. Blinn and M. E. Newell, "Texture and reflection in computer generated images," *Communications of the ACM*, vol. 19, no. 10, pp. 542–547, 1976.

[25] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular RGB-D sequences," in *IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 2300–2308.

[26] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *European Conference on Computer Vision*. Springer, 2016.

[27] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black, "Dynamic FAUST: Registering human bodies in motion," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[28] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 5543–5552.

[29] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 11, pp. 1222–1239, 2001.

[30] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, 2017.

[31] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *European Conference on Computer Vision*, 2004, pp. 25–36.

[32] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel, "High accuracy optical flow serves 3-d pose tracking: exploiting contour and flow based constraints," in *European Conference on Computer Vision*, 2006, pp. 98–111.

[33] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[34] C. Cagniart, E. Boyer, and S. Ilic, "Probabilistic deformable surface tracking from multiple videos," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6314. Springer, 2010, pp. 326–339.

[35] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[36] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[37] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 569–577, 2003.

[38] J. E. Chadwick, D. R. Haumann, and R. E. Parent, "Layered construction for deformable animated characters," in *ACM Siggraph Computer Graphics*, vol. 23, no. 3, 1989, pp. 243–252.

[39] X. Chen, Y. Guo, B. Zhou, and Q. Zhao, "Deformable model for estimating clothed and naked human shapes from a single image," *The Visual Computer*, vol. 29, no. 11, pp. 1187–1196, 2013.

[40] X. Chen, B. Zhou, F. Lu, L. Wang, L. Bi, and P. Tan, "Garment modeling with a depth camera," *ACM Transactions on Graphics*, vol. 34, no. 6, p. 203, 2015.

[41] Y. Chen, T.-K. Kim, and R. Cipolla, "Inferring 3D shapes and deformations from single views," in *European Conference on Computer Vision*, 2010, pp. 300–313.

[42] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[43] G. K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2003, pp. I–I.

[44] ——, "Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2003, pp. II–375.

[45] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality streamable free-viewpoint video," *ACM Transactions on Graphics*, vol. 34, no. 4, p. 69, 2015.

[46] Y. Cui, W. Chang, T. Nöll, and D. Stricker, "Kinectavatar: fully automatic body capture using a single kinect," in *Asian Conference on Computer Vision*, 2012, pp. 133–147.

[47] H. Dai, N. Pears, W. A. Smith, and C. Duncan, "A 3d morphable model of craniofacial shape and texture variation," in *IEEE International Conference on Computer Vision*, 2017, pp. 3085–3093.

[48] H. Dai, N. Pears, and W. Smith, "A data-augmented 3D morphable model of the ear," in *13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 404–408.

[49] R. Daněřek, E. Dibra, C. Öztireli, R. Ziegler, and M. Gross, "Deepgarment: 3D garment shape estimation from a single image," *Computer Graphics Forum*, vol. 36, no. 2, pp. 269–280, 2017.

[50] E. De Aguiar, C. Theobalt, M. Magnor, H.-P. Seidel *et al.*, "Reconstructing human shape and motion from multi-view video," in *2nd European Conference on Visual Media Production (CVMP)*, 2005, pp. 42–49.

[51] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, “Performance capture from sparse multi-view video,” *ACM Transactions on Graphics*, p. 98, 2008.

[52] P. E. Debevec, C. J. Taylor, and J. Malik, “Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach,” in *Annual Conference on Computer Graphics and Interactive Techniques*, 1996, pp. 11–20.

[53] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross, “Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks,” in *International Conference on 3D Vision*. IEEE, 2016, pp. 108–117.

[54] E. Dibra, C. Öztireli, R. Ziegler, and M. Gross, “Shape from selfies: Human body shape estimation using cca regression forests,” in *European Conference on Computer Vision*, 2016, pp. 88–104.

[55] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross, “Human shape from silhouettes using generative HKS descriptors and cross-modal neural networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[56] C. Doersch and A. Zisserman, “Sim2real transfer learning for 3d human pose estimation: motion to the rescue,” in *Advances in Neural Information Processing Systems*, 2019, pp. 12 929–12 941.

[57] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor *et al.*, “Fusion4d: Real-time performance capture of challenging scenes,” *ACM Transactions on Graphics*, vol. 35, no. 4, p. 114, 2016.

[58] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” in *IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 726–733.

[59] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, “Floating textures,” *Computer Graphics Forum*, vol. 27, no. 2, pp. 409–418, 2008.

[60] A. Elhayek, E. de Aguiar, A. Jain, J. Thompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt, “Marconi—convnet-based marker-less motion capture in outdoor and indoor scenes,” *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, no. 3, pp. 501–514, 2017.

[61] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, “Vision-based hand pose estimation: A review,” *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 52–73, 2007.

[62] R. Fablet and M. J. Black, “Automatic detection and tracking of human motion with a view-based representation,” in *European Conference on Computer Vision*, 2002, pp. 476–491.

[63] K. Fragkiadaki, H. Hu, and J. Shi, “Pose from flow and flow from pose,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 2059–2066.

[64] Y. Fu, Q. Yan, L. Yang, J. Liao, and C. Xiao, “Texture mapping for 3D reconstruction with rgb-d sensor,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[65] S. Fuhrmann, F. Langguth, and M. Goesele, “Mve-a multi-view reconstruction environment.” in *Eurographics Workshops on Graphics and Cultural Heritage*, 2014, pp. 11–18.

[66] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, “Motion capture using joint skeleton tracking and surface estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1746–1753.

[67] S. Galliani, K. Lasinger, and K. Schindler, “Massively parallel multiview stereopsis by surface normal diffusion,” in *IEEE International Conference on Computer Vision*. IEEE, 2015, pp. 873–881.

[68] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde, “Learning to predict indoor illumination from a single image,” *ACM Transactions on Graphics*, vol. 9, no. 4, 2017.

[69] D. M. Gavrila and L. S. Davis, “3-d model-based tracking of humans in action: a multi-view approach,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1996, pp. 73–80.

[70] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter, “Morphable face models-an open framework,” in *13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 75–82.

[71] J. J. Gibson, *The perception of the visual world.* Houghton Mifflin, 1950.

[72] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton, “Volumetric performance capture from minimal camera viewpoints,” in *European Conference on Computer Vision*, 2018.

[73] K. Gong, X. Liang, X. Shen, and L. Lin, “Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[74] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, “Instance-level human parsing via part grouping network,” in *European Conference on Computer Vision*, 2018.

[75] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[76] P. Guan, A. Weiss, A. O. Bălan, and M. J. Black, “Estimating human shape and pose from a single image,” in *IEEE International Conference on Computer Vision*. IEEE, 2009, pp. 1381–1388.

[77] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, “Drape: Dressing any person.” *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 35–1, 2012.

[78] R. A. Guler and I. Kokkinos, “Holopose: Holistic 3D human reconstruction in-the-wild,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 10 884–10 894.

[79] Y. Guo, X. Chen, B. Zhou, and Q. Zhao, “Clothed and naked human shapes estimation from a single image,” *Computational Visual Media*, pp. 43–50, 2012.

[80] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, “Livecap: Real-time human performance capture from monocular video,” *ACM Transactions on Graphics*, vol. 38, no. 2, pp. 14:1–14:17, 2019.

[81] B. Haefner, Y. Quéau, T. Möllenhoff, and D. Cremers, “Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 164–174.

[82] E. P. Hanavan Jr, "A mathematical model of the human body," Air Force Aerospace Medical Research Lab Wright-Patterson AFB OH, Tech. Rep., 1964.

[83] J. C. Hart, "Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces," *The Visual Computer*, vol. 12, no. 10, pp. 527–545, 1996.

[84] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel, "A statistical model of human pose and body shape," *Computer Graphics Forum*, vol. 28, no. 2, pp. 337–346, 2009.

[85] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H.-P. Seidel, "Multilinear pose and body shape estismation of dressed subjects from image sets," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1823–1830.

[86] T. Helten, A. Baak, G. Bharaj, M. Muller, H.-P. Seidel, and C. Theobalt, "Personalization and evaluation of a real-time depth-based full body tracker," in *International Conference on 3D Vision*. IEEE, 2013, pp. 279–286.

[87] P. Henderson and V. Ferrari, "Learning to generate and reconstruct 3D meshes with only 2D supervision," in *British Machine Vision Conference*, 2018.

[88] N. Hesse, S. Pujades, M. J. Black, M. Arens, U. Hofmann, and S. Schroeder, "Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences," *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.

[89] A. Hilton, D. J. Beresford, T. Gentils, R. S. Smith, and W. Sun, "Virtual people: Capturing human models to populate virtual worlds," *Proceedings Computer Animation*, vol. 99, p. 174, 1999.

[90] D. A. Hirshberg, M. Loper, E. Rachlin, and M. J. Black, "Coregistration: Simultaneous alignment and modeling of articulated 3D shape," in *European Conference on Computer Vision*, 2012, pp. 242–255.

[91] D. Hogg, "Model-based vision: a program to see a walking person," *Image and Vision computing*, vol. 1, no. 1, pp. 5–20, 1983.

[92] B. K. Horn, “Shape from shading: A method for obtaining the shape of a smooth opaque object from one view,” Ph.D. dissertation, Massachusetts Inst. of Technology, 1970.

[93] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[94] C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer, “Volumetric 3d tracking by detection,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 3862–3870.

[95] Y. Huang, F. Bogo, C. Classner, A. Kanazawa, P. V. Gehler, I. Akhter, and M. J. Black, “Towards accurate markerless human shape and pose estimation over time,” in *International Conference on 3D Vision*. IEEE, 2017.

[96] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, “Deep inertial poser learning to reconstruct human pose from sparseinertial measurements in real time,” *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 185:1–185:15, 2018.

[97] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li, “Deep volumetric video from very sparse multi-view performance capture,” in *European Conference on Computer Vision*, 2018, pp. 336–354.

[98] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler, “A multiresolution 3d morphable face model and fitting framework,” in *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[99] L. Huynh, W. Chen, S. Saito, J. Xing, K. Nagano, A. Jones, P. Debevec, and H. Li, “Mesoscopic facial geometry inference using deep neural networks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8407–8416.

[100] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, “Volumedeform: Real-time volumetric non-rigid reconstruction,” in *European Conference on Computer Vision*, 2016.

[101] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schieke, “Deepercut: A deeper, stronger, and faster multi-person pose estimation model,” in *European Conference on Computer Vision*, 2016.

[102] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, “Arttrack: Articulated multi-person tracking in the wild,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[103] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1125–1134.

[104] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, “Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera,” in *ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[105] A. S. Jackson, C. Manafas, and G. Tzimiropoulos, “3d human body reconstruction from a single image via volumetric regression,” in *European Conference on Computer Vision*, 2018, pp. 64–77.

[106] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt, “MovieReshape: Tracking and reshaping of humans in videos,” *ACM Transactions on Graphics*, vol. 29, no. 6, p. 148, 2010.

[107] M. Jancosek and T. Pajdla, “Multi-view reconstruction preserving weakly-supported surfaces,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3121–3128.

[108] N. Jin, Y. Zhu, Z. Geng, and R. Fedkiw, “A pixel-based framework for data-driven clothing,” *arXiv preprint arXiv:1812.01677*, 2018.

[109] H. Joo, T. Simon, and Y. Sheikh, “Total capture: A 3D deformation model for tracking faces, hands, and bodies,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 8320–8329.

[110] I. A. Kakadiaris and D. Metaxas, “3d human body model acquisition from multiple views,” in *IEEE International Conference on Computer Vision*. IEEE, 1995.

[111] Y. Kanamori and Y. Endo, “Relighting humans: occlusion-aware inverse rendering for fullbody human images,” *ACM Transactions on Graphics*, vol. 37, no. 270, pp. 1–270, 2018.

[112] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, 2018.

[113] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, “Learning 3d human dynamics from video,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, 2019, pp. 5614–5623.

[114] L. Kavan and J. Žára, “Spherical blend skinning: a real-time deformation of articulated models,” in *Proceedings of the 2005 symposium on Interactive 3D graphics and games*, 2005, pp. 9–16.

[115] L. Kavan, S. Collins, J. Žára, and C. O’Sullivan, “Geometric skinning with approximate dual quaternion blending,” *ACM Transactions on Graphics*, vol. 27, no. 4, p. 105, 2008.

[116] S. Khamis, J. Taylor, J. Shotton, C. Keskin, S. Izadi, and A. Fitzgibbon, “Learning an efficient model of hand shape variation from depth images,” in *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2015, pp. 2540–2548.

[117] S. M. Khan and M. Shah, “Reconstructing non-stationary articulated objects in monocular video using silhouette information,” in *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2008, pp. 1–8.

[118] M. Kim, G. Pons-Moll, S. Pujades, S. Bang, J. Kim, M. J. Black, and S.-H. Lee, “Data-driven physics for human soft tissue animation,” *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[119] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, vol. 5, 2015.

[120] R. Koch, M. Pollefeys, and L. Van Gool, “Multi viewpoint stereo from uncalibrated video sequences,” in *European Conference on Computer Vision*, 1998, pp. 55–71.

[121] N. Kolotouros, G. Pavlakos, and K. Daniilidis, “Convolutional mesh regression for single-image human shape reconstruction,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* IEEE, 2019.

[122] V. Kraevoy, A. Sheffer, and M. van de Panne, “Modeling from contour drawings,” in *Eurographics Symposium on Sketch-Based interfaces and Modeling*, 2009, pp. 37–44.

[123] Z. Lahner, D. Cremers, and T. Tung, “Deepwrinkles: Accurate and realistic clothing modeling,” in *European Conference on Computer Vision*, 2018, pp. 667–684.

[124] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, “Unite the people: Closing the loop between 3d and 2d human representations,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[125] V. Lempitsky and D. Ivanov, “Seamless mosaicing of image-based texture maps,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–6.

[126] H. P. Lensch, W. Heidrich, and H.-P. Seidel, “A silhouette-based algorithm for texture registration and stitching,” *Graphical Models*, vol. 63, no. 4, pp. 245–262, 2001.

[127] V. Leroy, J.-S. Franco, and E. Boyer, “Multi-View Dynamic Shape Refinement Using Local Temporal Integration,” in *IEEE International Conference on Computer Vision*. IEEE, 2017.

[128] ——, “Shape reconstruction using volume sweeping and learned photoconsistency,” in *European Conference on Computer Vision*, 2018, pp. 796–811.

[129] J. P. Lewis, M. Cordner, and N. Fong, “Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation,” in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 165–172.

[130] G. Li, C. Wu, C. Stoll, Y. Liu, K. Varanasi, Q. Dai, and C. Theobalt, “Capturing relightable human performances under general uncontrolled illumination,” *Computer Graphics Forum*, vol. 32, pp. 1–8, 2013.

[131] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, “3d self-portraits,” *ACM Transactions on Graphics*, vol. 32, no. 6, p. 187, 2013.

[132] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, “Learning a model of facial shape and expression from 4D scans,” *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 194:1–194:17, 2017.

[133] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 12, pp. 2402–2414, 2015.

[134] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[135] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in *European Conference on Computer Vision*, 2016.

[136] S. Lombardi, J. Saragih, T. Simon, and Y. Sheikh, "Deep appearance models for face rendering," *ACM Transactions on Graphics*, vol. 37, no. 4, p. 68, 2018.

[137] S. Lombardi, T. Simon, J. Saragih, G. Schwartz, A. Lehrmann, and Y. Sheikh, "Neural volumes: Learning dynamic renderable volumes from images," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, p. 65, 2019.

[138] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[139] M. M. Loper and M. J. Black, "OpenDR: An approximate differentiable renderer," in *European Conference on Computer Vision*. Springer, 2014, pp. 154–169.

[140] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'81. Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[141] L. Luo, H. Li, S. Paris, T. Weise, M. Pauly, and S. Rusinkiewicz, "Multiview hair capture using orientation fields," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012.

[142] N. Magnenat-Thalmann and D. Thalmann, "The direction of synthetic actors in the film rendez-vous à montréal," *IEEE Computer Graphics and applications*, vol. 7, no. 12, pp. 9–19, 1987.

[143] M. A. Magnor, O. Grau, O. Sorkine-Hornung, and C. Theobalt, Eds., *Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality*. CRC Press, 2015.

[144] F. Maninchedda, M. R. Oswald, and M. Pollefeys, "Fast 3d reconstruction of faces with glasses," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6599–6608.

[145] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-based visual hulls," in *Annual Conference on Computer Graphics and Interactive Techniques*, 2000, pp. 369–374.

[146] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 44, 2017.

[147] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3D pose estimation from monocular rgb," in *International Conference on 3D Vision*. IEEE, 2018.

[148] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[149] D. Metaxas and D. Terzopoulos, "Shape and nonrigid motion estimation through physics-based synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 580–591, 1993.

[150] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktashmotlagh, and A. Eriksson, "Deep level sets: Implicit surface representations for 3D shape inference," *arXiv preprint arXiv:1901.06802*, 2019.

[151] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[152] S. C. Mölbert, A. Thaler, B. J. Mohler, S. Streuber, J. Romero, M. J. Black, S. Zipfel, H.-O. Karnath, and K. E. Giel, "Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: Attitudinal components rather than visual body size estimation are distorted," *Psychological Medicine*, vol. 48, no. 4, pp. 642–653, 2018.

[153] M. Mori, K. F. MacDorman, and N. Kageki, “The uncanny valley [from the field],” *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[154] M. Mustafa, S. Guthe, J.-P. Tauscher, M. Goesele, and M. Magnor, “How human am I? EEG-based evaluation of animated virtual characters,” in *Proceedings of the ACM Human Factors in Computing Systems*, May 2017, pp. 5098–5108.

[155] G. Nam, C. Wu, M. H. Kim, and Y. Sheikh, “Strand-accurate multi-view hair capture,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[156] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, “Siclope: Silhouette-based clothed people,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[157] L. P. Nedel and D. Thalmann, “Modeling and deformation of the human body using an anatomically-based approach,” in *Proceedings Computer Animation*, 1998, pp. 34–40.

[158] T. Nestmeyer and P. V. Gehler, “Reflectance adaptive filtering improves intrinsic image estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1771–1780.

[159] T. Neumann, K. Varanasi, N. Hasler, M. Wacker, M. Magnor, and C. Theobalt, “Capture and statistical modeling of arm-muscle deformations,” *Computer Graphics Forum*, vol. 32, no. 2, pp. 285–294, 2013.

[160] N. Neverova, R. Alp Guler, and I. Kokkinos, “Dense pose transfer,” in *European Conference on Computer Vision*, 2018.

[161] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, “Dtam: Dense tracking and mapping in real-time,” in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 2320–2327.

[162] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, “Kinectfusion: Real-time dense surface mapping and tracking,” in *IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2011, pp. 127–136.

[163] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 343–352.

[164] W. Niem and J. Wingbermuhle, "Automatic reconstruction of 3D objects using a mobile monoscopic camera," in *Proceedings International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, 1997, pp. 173–180.

[165] H. Ning, L. Wang, W. Hu, and T. Tan, "Model-based tracking of human walking in monocular image sequences," in *IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, vol. 1, 2002, pp. 537–540.

[166] E. Ofek, E. Shilat, A. Rappoport, and M. Werman, "Multiresolution textures from image sequences," *IEEE Computer Graphics and Applications*, vol. 17, no. 2, pp. 18–29, 1997.

[167] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox, "Deep learning for human part discovery in images," in *IEEE International Conference on Robotics and Automation*. IEEE, 2016.

[168] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *International Conference on 3D Vision*. IEEE, 2018.

[169] J. O'rourke and N. I. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 522–536, 1980.

[170] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou *et al.*, "Holoportation: Virtual 3D teleportation in real-time," in *Symposium on User Interface Software and Technology*, 2016, pp. 741–754.

[171] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[172] W. Paul, B. Janet, and D. Jackson Don, "Pragmatics of human communication. a study of interactional patterns, pathologies, and paradoxes," *New York and London: WW Norton & Co*, 1967.

[173] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, “Learning to estimate 3D human pose and shape from a single color image,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[174] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, “Expressive body capture: 3D hands, face, and body from a single image,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[175] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3D face model for pose and illumination invariant face recognition,” in *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

[176] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, “Synthesizing realistic facial expressions from photographs,” in *ACM SIGGRAPH 2006 Courses*, 2006, p. 19.

[177] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, “Deepcut: Joint subset partition and labeling for multi person pose estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[178] R. Plankers and P. Fua, “Articulated soft objects for video-based body modeling,” in *IEEE International Conference on Computer Vision*, no. CVLAB-CONF-2001-005. IEEE, 2001, pp. 394–401.

[179] G. Pons-Moll and B. Rosenhahn, *Model-Based Pose Estimation*. Springer, 2011, ch. 9, pp. 139–170.

[180] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H.-P. Seidel, and B. Rosenhahn, “Multisensor-fusion for 3D full-body human motion capture,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

[181] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn, “Posebits for monocular human pose estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 2345–2352.

[182] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, “Dyna: a model of dynamic human shape in motion,” *ACM Transactions on Graphics*, vol. 34, p. 120, 2015.

[183] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon, “Metric regression forests for correspondence estimation,” *International Journal of Computer Vision*, pp. 1–13, 2015.

[184] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, “ClothCap: Seamless 4D clothing capture and retargeting,” *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[185] A.-I. Popa, M. Zanfir, and C. Sminchisescu, “Deep multitask architecture for integrated 2D and 3D human sensing,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[186] T. Popa, Q. Zhou, D. Bradley, V. Kraevoy, H. Fu, A. Sheffer, and W. Heidrich, “Wrinkling captured garments using space-time data-driven deformation,” *Computer Graphics Forum*, vol. 28, no. 2, pp. 427–435, 2009.

[187] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, “3dpeople: Modeling the geometry of dressed humans,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2242–2251.

[188] V. Ramakrishna, T. Kanade, and Y. A. Sheikh, “Reconstructing 3D human pose from 2D image landmarks,” in *European Conference on Computer Vision*, 2012.

[189] R. Ramamoorthi and P. Hanrahan, “An efficient representation for irradiance environment maps,” in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 497–500.

[190] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, “Generating 3D faces using convolutional mesh autoencoders,” in *European Conference on Computer Vision*, 2018, pp. 725–741.

[191] A. Rehan, A. Zaheer, I. Akhter, A. Saeed, B. Mahmood, M. Usmani, and S. Khan, “Nrsfm using local rigidity,” in *Winter Conference on Applications of Computer Vision*. IEEE, 2014, pp. 69–74.

[192] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt, “General automatic human shape and motion capture using volumetric contour cues,” in *European Conference on Computer Vision*, 2016, pp. 509–526.

[193] G. Riegler, A. Osman Ulusoy, and A. Geiger, “Octnet: Learning deep 3D representations at high resolutions,” in *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2017.

[194] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt, “Model-based outdoor performance capture,” in *International Conference on 3D Vision.* IEEE, 2016.

[195] N. Robertini, D. Casas, E. De Aguiar, and C. Theobalt, “Multi-view performance capture of surface details,” *International Journal of Computer Vision*, pp. 1–18, 2017.

[196] C. Rocchini, P. Cignoni, C. Montani, and R. Scopigno, “Multiple textures stitching and blending on 3D objects,” in *Rendering Techniques’ 99.* Springer, 1999, pp. 119–130.

[197] G. Rogez, P. Weinzaepfel, and C. Schmid, “Lcr-net: Localization-classification-regression for human pose,” in *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2017.

[198] L. Rogge, F. Klose, M. Stengel, M. Eisemann, and M. Magnor, “Garment Replacement in Monocular Video Sequences,” *ACM Transactions on Graphics*, vol. 34, no. 1, pp. 6:1–6:10, 2014.

[199] K. Rohr, “Towards model-based recognition of human movements in image sequences,” *CVGIP: Image understanding*, vol. 59, no. 1, pp. 94–115, 1994.

[200] J. Romero, M. Loper, and M. J. Black, “Flowcap: 2D human pose from optical flow,” in *German Conference on Pattern Recognition.* Springer, 2015, pp. 412–423.

[201] J. Romero, D. Tzionas, and M. J. Black, “Embodied hands: Modeling and capturing hands and bodies together,” *ACM Transactions on Graphics*, vol. 36, no. 6, p. 245, 2017.

[202] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in *IEEE International Conference on Computer Vision.* IEEE, 2019.

[203] B. Sapp, D. Weiss, and B. Taskar, “Parsing human motion with stretchable models,” in *IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2011, pp. 1281–1288.

[204] H. Sattar, G. Pons-Moll, and M. Fritz, “Fashion is taking shape: Understanding clothing preference based on body shape from online sources,” in *IEEE Winter Conference on Applications of Computer Vision (WACV 2019)*, 2019.

[205] F. Scheepers, R. E. Parent, W. E. Carlson, and S. F. May, “Anatomy-based modeling of the human musculature,” in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 1997, pp. 163–172.

[206] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 815–823.

[207] M. Sela, E. Richardson, and R. Kimmel, “Unrestricted facial geometry reconstruction using image-to-image translation,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1576–1585.

[208] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, “Sfsnet: Learning shape, reflectance and illuminance of facesin the wild’,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 6296–6305.

[209] A. Shapiro, A. Feng, R. Wang, H. Li, M. Bolas, G. Medioni, and E. Suma, “Rapid avatar capture and simulation using commodity depth sensors,” *Computer Animation and Virtual Worlds*, vol. 25, no. 3-4, pp. 201–211, 2014.

[210] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei *et al.*, “Accurate, robust, and flexible real-time hand tracking,” in *Proceedings of the ACM Human Factors in Computing Systems*, 2015, pp. 3633–3642.

[211] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov *et al.*, “Textured neural avatars,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 2387–2397.

[212] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, “Tracking loose-limbed people,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2004, pp. I–421.

[213] L. Sigal, A. Balan, and M. J. Black, “Combined discriminative and generative articulated pose and non-rigid shape estimation,” in *Advances in Neural Information Processing Systems*, 2007, pp. 1337–1344.

[214] L. Sigal, A. O. Balan, and M. J. Black, “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,” *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[215] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, “Single image 3D human pose estimation from noisy observations.” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2673–2680.

[216] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[217] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer, “Deepvoxels: Learning persistent 3D feature embeddings,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[218] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, “Killingfusion: Non-rigid 3D reconstruction without correspondences,” in *IEEE Conference on Computer Vision and Pattern Recognition*, no. 4. IEEE, 2017, p. 7.

[219] C. Sminchisescu and A. Telea, “Human pose estimation from silhouettes. a consistent approach using distance level sets,” in *10th International Conference on Computer Graphics,Visualization and Computer Vision (WSCG ’02)*, 2002.

[220] C. Sminchisescu and B. Triggs, “Kinematic jump processes for monocular 3D human tracking,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2003, pp. I–I.

[221] C. Sminchisescu, A. Kanaujia, and D. Metaxas, “Learning joint top-down and bottom-up processes for 3D visual inference,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, pp. 1743–1752.

[222] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel, “Laplacian surface editing,” in *Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 2004, pp. 175–184.

[223] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE Computer Graphics and Applications*, vol. 27, no. 3, 2007.

[224] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt, "Fast articulated motion tracking using a sums of gaussians body model," in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 951–958.

[225] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *IEEE International Conference on Computer Vision*. IEEE, 2017.

[226] Y. Tao, Z. Zheng, K. Guo, J. Zhao, D. Quionhai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performance with inner body shape from a depth sensor," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[227] Y. Tao, Z. Zheng, Y. Zhong, J. Zhao, D. Quionhai, G. Pons-Moll, and Y. Liu, "Simulcap : Single-view human performance capture with cloth simulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[228] C. J. Taylor, "Reconstruction of articulated objects from point correspondences in a single uncalibrated image," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2000, pp. 677–684.

[229] J. Taylor, L. Bordeaux, T. Cashman, B. Corish, C. Keskin, T. Sharp, E. Soto, D. Sweeney, J. Valentin, B. Luff *et al.*, "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences," *ACM Transactions on Graphics*, vol. 35, no. 4, p. 143, 2016.

[230] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt, "Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

[231] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt, "Fml: face model learning from videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 10 812–10 822.

[232] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Perez, M. Zollhöfer, and C. Theobalt, "Fml: Face model learning from videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[233] D. Thalmann, J. Shen, and E. Chauvineau, "Fast realistic human body deformations for animation and VR applications," in *Proceedings of CG International'96*, 1996, pp. 166–174.

[234] C. Theobalt, J. Carranza, and M. A. Magnor, "Enhancing silhouette-based human motion capture with 3D motion fields," in *Proceedings of the 11th Pacific Conference onComputer Graphics and Applications*, 2003, pp. 185–193.

[235] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Facevr: Real-time gaze-aware facial reenactment in virtual reality," *ACM Transactions on Graphics*, 2018.

[236] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[237] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3d full human bodies using kinects," *IEEE Transactions on Visualization and Computer Graphics*, no. 4, pp. 643–650, 2012.

[238] M. Trumble, A. Gilbert, A. Hilton, and J. Collomosse, "Deep autoencoder for combined human pose estimation and body model upscaling," in *European Conference on Computer Vision*, 2018.

[239] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," in *Advances in Neural Information Processing Systems*, 2017, pp. 5236–5246.

[240] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[241] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "Bodynet: Volumetric inference of 3D human body shapes," in *European Conference on Computer Vision*, 2018.

[242] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, “Three-dimensional scene flow,” in *IEEE International Conference on Computer Vision*. IEEE, 1999, pp. 722–729.

[243] D. Vlasic, I. Baran, W. Matusik, and J. Popović, “Articulated mesh animation from multi-view silhouettes,” *ACM Transactions on Graphics*, vol. 27, no. 3, p. 97, 2008.

[244] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik, “Dynamic shape capture using multi-view photometric stereo,” *ACM Transactions on Graphics*, vol. 28, no. 5, p. 174, 2009.

[245] T. von Marcard, G. Pons-Moll, and B. Rosenhahn, “Human pose estimation from video and imus,” *Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2016.

[246] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, “Sparse inertial poser: Automatic 3D human pose estimation from sparse imus,” *Computer Graphics Forum*, pp. 349–360, 2017.

[247] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, “Recovering accurate 3D human pose in the wild using imus and a moving camera,” in *European Conference on Computer Vision*, 2018.

[248] M. Waechter, N. Moehrle, and M. Goesele, “Let there be color! large-scale texturing of 3D reconstructions,” in *European Conference on Computer Vision*, 2014, pp. 836–850.

[249] B. Wandt, H. Ackermann, and B. Rosenhahn, “3D human motion capture from monocular image sequences,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015.

[250] ——, “3D reconstruction of human motion from monocular image sequences,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[251] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra, “Learning a shared shape space for multimodal garment design,” *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 1:1–1:14, 2018.

[252] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Asilomar Conference on Signals, Systems & Computers*, vol. 2, 2003, pp. 1398–1402.

[253] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[254] A. Weiss, D. Hirshberg, and M. J. Black, "Home 3D body scans from noisy image and range data," in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 1951–1958.

[255] C.-Y. Weng, B. Curless, and I. Kemelmacher-Shlizerman, "Photo wake-up: 3d character animation from a single photo," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019.

[256] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt, "High-quality shape from multi-view stereo and shading under general illumination," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 969–976.

[257] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 1108–1115.

[258] C. Wu, K. Varanasi, and C. Theobalt, "Full body performance capture under uncontrolled and varying illumination: A shading-based approach," in *European Conference on Computer Vision*. Springer-Verlag, 2012, pp. 757–770.

[259] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt, "On-set performance capture of multiple actors with a stereo camera," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 161:1–161:11, 2013.

[260] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang, "Estimation of human body shape and posture under clothing," *Computer Vision and Image Understanding*, vol. 127, pp. 31–42, 2014.

[261] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 10 965–10 974.

[262] L. Xu, J. Jia, and Y. Matsushita, "Motion detail preserving optical flow estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 9, pp. 1744–1757, 2012.

[263] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann, “DISN: Deep implicit surface network for high-quality single-view 3D reconstruction,” in *Advances in Neural Information Processing Systems*, 2019, pp. 490–500.

[264] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, “Monoperfcap: Human performance capture from monocular video,” *ACM Transactions on Graphics*, 2018.

[265] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer, “Estimation of Human Body Shape in Motion with Wide Clothing,” in *European Conference on Computer Vision*, 2016.

[266] ——, “Analyzing clothing layer deformation statistics of 3D human motions,” in *European Conference on Computer Vision*, 2018, pp. 237–253.

[267] P. Yao, Z. Fang, F. Wu, Y. Feng, and J. Li, “Densebody: Directly regressing dense 3d human pose and shape from a single color image,” *arXiv preprint arXiv:1903.10153*, 2019.

[268] M. Ye and R. Yang, “Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 2345–2352.

[269] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning face representation from scratch,” *arXiv preprint arXiv:1411.7923*, 2014.

[270] R. Yu, C. Russell, N. D. F. Campbell, and L. Agapito, “Direct, dense, and deformable: Template-based non-rigid 3D reconstruction from RGB video,” in *IEEE International Conference on Computer Vision*. IEEE, 2015.

[271] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu, “Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera,” in *IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 910–919.

[272] A. Zanfir, E. Marinoiu, and C. Sminchisescu, “Monocular 3D pose and shape estimation of multiple people in natural scenes–the importance of multiple scene constraints,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 2148–2157.

[273] M. Zeng, J. Zheng, X. Cheng, and X. Liu, “Templateless quasi-rigid shape modeling with implicit loop-closure,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 145–152.

[274] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, “Detailed, accurate, human shape estimation from clothed 3D scan sequences,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017.

[275] Q. Zhang, B. Fu, M. Ye, and R. Yang, “Quality dynamic human body modeling using a single low-cost depth camera,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 676–683.

[276] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, “Shape-from-shading: a survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 8, pp. 690–706, 1999.

[277] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, “Patchmatch based joint view selection and depthmap estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1510–1517.

[278] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, “Deephuman: 3d human reconstruction from a single image,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7739–7749.

[279] Q.-Y. Zhou and V. Koltun, “Color map optimization for 3D reconstruction with consumer depth cameras,” *ACM Transactions on Graphics*, vol. 33, no. 4, p. 155, 2014.

[280] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han, “Parametric reshaping of human bodies in images,” *ACM Transactions on Graphics*, vol. 29, no. 4, p. 126, 2010.

[281] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, “3d shape estimation from 2D landmarks: A convex relaxation approach.” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 4447–4455.

[282] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, “Sparseness meets deepness: 3D human pose estimation from monocular video,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[283] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, “Towards 3D human pose estimation in the wild: A weakly-supervised approach,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 398–407.

[284] Y. Zhou, L. Hu, J. Xing, W. Chen, H.-W. Kung, X. Tong, and H. Li, “Hairnet: Single-view hair reconstruction using convolutional neural networks,” in *European Conference on Computer Vision*, 2018, pp. 235–251.

[285] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao, “Video-based outdoor human reconstruction,” *IEEE Transactions on Circuits and Systems for Video Technology*, no. 4, pp. 760–770, 2017.

[286] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, “Detailed human shape estimation from a single image by hierarchical mesh deformation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 4491–4500.

[287] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt *et al.*, “Real-time non-rigid reconstruction using an rgb-d camera,” *ACM Transactions on Graphics*, vol. 33, no. 4, p. 156, 2014.

[288] S. Zuffi and M. J. Black, “The stitched puppet: A graphical model of 3D human shape and pose,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3537–3546.

[289] S. Zuffi, J. Romero, C. Schmid, and M. J. Black, “Estimating human pose with flowing puppets,” in *IEEE International Conference on Computer Vision*. IEEE, 2013, pp. 3312–3319.

[290] S. Zuffi, A. Kanazawa, D. Jacobs, and M. J. Black, “3D menagerie: Modeling the 3D shape and pose of animals,” in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 5524–5532.