MODELLING 3D HUMANS: POSE, SHAPE, CLOTHING AND INTERACTIONS

by BHARAT LAL BHATNAGAR

A dissertation submitted towards the degree

Doctor of Engineering (Dr.-Ing.)

of the Faculty of Mathematics and Computer Science

of Saarland University

Saarbrücken, 2023

Date of Colloquium:	16 March, 2023
Dean of the Faculty:	Prof. Dr. Jürgen Steimle
Chair of the Committee:	Prof. Dr. Eddy Ilg
Reviewer, Advisor:	Prof. Dr. Gerard Pons-Moll
Reviewers:	Prof. Dr. Christian Theobalt
	Prof. Dr. Siyu Tang
	Prof. Dr. Jiajun Wu
Academic Assistant:	Dr. Jan Eric Lenssen

ABSTRACT

Digital humans are increasingly becoming a part of our lives with applications like animation, gaming, virtual try-on, Metaverse and much more. In recent years there has been a great push to make our models of digital humans *as real as possible*. In this thesis we present methodologies to model two key characteristics of real humans, their *appearance and actions*.

We propose *MGN*, the first approach to reconstruct 3D garments and body shape underneath, as separate meshes, from a few RGB images of a person. We extend the popular SMPL body model, which represents only undressed shapes, to also include garments (SMPL+G). SMPL+G can be dressed with garments which can be posed and shaped according to the SMPL model. This allows, for the first time, real world applications like texture transfer, garment transfer and virtual try-on in 3D, using just images. We also underline the key limitation of mesh based representation for digital humans, i.e. the ability to represent high frequency details.

Therefore, we explore the recent implicit function based representation as an alternative to the mesh based representation (including parametric models like SMPL) for digital humans. Typically, methods based on latter lack details while former lacks control. We propose *IPNet*, a neural network, that leverages implicit functions for detailed reconstruction and registers the reconstructed mesh with the parametric SMPL model to make it controllable. Thus leveraging the best of both worlds.

We further study the process of registering a parametric model, such as SMPL, to a 3D mesh. This decade old problem in computer vision and graphics, typically entails a two step process, i) establish correspondences between the model and the mesh, and ii) optimise the model to minimize the distance between the corresponding points. This two step process is not end-to-end differentiable. We propose *LoopReg*, that uses a novel implicit function based representation of the model and makes registration differentiable. Semi-supervised *LoopReg* outperforms contemporary supervised methods using $\sim$100x less supervised data.

Modelling the human appearance is necessary but not sufficient for building realistic digital humans. We must model not just how humans look but also how they interact with their surrounding objects. To this end, we present *BEHAVE* the first dataset of full body real interactions between humans and *movable objects*. We provide segmented multi-view RGBD frames along with registered SMPL and object fits as well as contact annotations in 3D. *BEHAVE* dataset contains $\sim$15k frames and its extension contains $\sim$400k frames with pseudo-ground truth annotations. Our *BEHAVE* method, uses this dataset to train a neural network to jointly track the human, the object and the contacts between them.

In this thesis we explore the aforementioned ideas and provide in-depth analysis of our key ideas and design choices. We also discuss the limitations of our ideas and propose future work to not just address these limitations but also extend the research further. All our code, *MGN* digital wardrobe and *BEHAVE* dataset are publicly available for further research.

ZUSAMMENFASSUNG

Der digitale Mensch wird immer mehr zu einem Teil unseres Lebens mit Anwendungen wie Animation, Spielen, virtuellem Ausprobieren, Metaverse und vielem mehr. In den letzten Jahren wurden große Anstrengungen unternommen, um unsere Modelle digitaler Menschen *so real wie möglich zu gestalten*. In dieser Arbeit stellen wir Methoden zur Modellierung von zwei Schlüsseleigenschaften echter Menschen vor: *ihr Aussehen und ihre Handlungen*.

Wir schlagen *MGN* vor, den ersten Ansatz zur Rekonstruktion von 3D-Kleidungsstücken und der darunter liegenden Körperform als separate Netze aus einigen wenigen RGB-Bildern einer Person. Wir erweitern das weit verbreitete SMPL-Körpermodell, das nur unbekleidete Formen darstellt, um auch Kleidungsstücke zu erfassen (SMPL+G). SMPL+G kann mit Kleidungsstücken bekleidet werden, die entsprechend dem SMPL-Modell posiert und geformt werden können. Dies ermöglicht zum ersten Mal reale Anwendungen wie Texturübertragung, Kleidungsübertragung und virtuelle Anprobe in 3D, wobei nur Bilder verwendet werden. Wir unterstreichen auch die entscheidende Einschränkung der netzbasierten Darstellung für digitale Menschen, nämlich die Fähigkeit, hochfrequente Details darzustellen.

Daher untersuchen wir die neue implizite funktionsbasierte Darstellung als Alternative zur netzbasierten Darstellung (einschließlich parametrischer Modelle wie SMPL) für digitale Menschen. Typischerweise mangelt es den Methoden, die auf letzteren basieren, an Details, während ersteren die Kontrolle fehlt. Wir schlagen *IPNet* vor, ein neuronales Netzwerk, das implizite Funktionen für eine detaillierte Rekonstruktion nutzt und das rekonstruierte Netz mit dem parametrischen SMPL-Modell registriert, um es kontrollierbar zu machen. Auf diese Weise wird das Beste aus beiden Welten genutzt.

Wir untersuchen den Prozess der Registrierung eines parametrischen Modells, wie z. B. SMPL, auf ein 3D-Netz. Dieses jahrzehntealte Problem im Bereich der Computer Vision und der Graphik erfordert in der Regel einen zweistufigen Prozess: i) Herstellung von Korrespondenzen zwischen dem Modell und dem Netz, und ii) Optimierung des Modells, um den Abstand zwischen den entsprechenden Punkten zu minimieren. Dieser zweistufige Prozess ist nicht durchgängig differenzierbar. Wir schlagen *LoopReg* vor, das eine neue, auf impliziten Funktionen basierende Darstellung des Modells verwendet und die Registrierung differenzierbar macht. Semi-überwachtes *LoopReg* übertrifft aktuelle überwachte Methoden mit $\sim$100x weniger überwachten Daten.

Die Modellierung des menschlichen Aussehens ist notwendig, aber nicht ausreichend, um realistische digitale Menschen zu schaffen. Wir müssen nicht nur modellieren, wie Menschen aussehen, sondern auch, wie sie mit ihren umgebenden Objekten interagieren. Zu diesem Zweck präsentieren wir mit *BEHAVE* den ersten Datensatz von realen Ganzkörper-Interaktionen zwischen Menschen und beweglichen Objekten. Wir stellen segmentierte Multiview-RGBD-Frames zusammen mit registrierten SMPL- und Objekt-Fits sowie Kontaktannotationen in 3D zur Verfügung. Der *BEHAVE*-Datensatz enthält $\sim$15k Frames und seine Erweiterung enthält $\sim$400k Frames mit Pseudo-Ground-Truth-Annotationen. Unsere *BEHAVE*-Methode verwendet diesen Datensatz, um ein neuronales Netz zu trainieren, das die Person, das Objekt und die Kontakte zwischen ihnen gemeinsam verfolgt.

In dieser Arbeit untersuchen wir die oben genannten Ideen und bieten eine eingehende Analyse unserer Schlüsselideen und Designentscheidungen. Wir erörtern auch die Grenzen unserer Ideen und schlagen künftige Arbeiten vor, um nicht nur diese Grenzen anzugehen, sondern auch die Forschung weiter auszubauen. Unser gesamter Code, die digitale Garderobe und der Datensatz sind für weitere Forschungen öffentlich zugänglich.

ACKNOWLEDGMENTS

I want to thank all the people who have helped me throughout this journey. This thesis would not have come to fruition without them.

First and foremost, I want to thank Prof. Gerard for giving me the opportunity to work with him at Max Planck Institute, Saarland Informatics Campus and University of Tübingen. I greatly admire his vision towards research, the persistent hard work and drive to achieve perfection. I hope some of it rubbed off on me during our work together.

I would also like to thank Bernt Schiele, for creating a highly conducive environment for research at our department at MPI, D2. I've always enjoyed his droll sense of humour and I've gained a lot of perspective from our open and frank conversations. I cannot thank my colleagues at D2 enough, for all the help and support I received.

I want to thank all my collaborators, they made this journey feel natural and effortless. Christian Theobalt has supported me since my very first project at MPI. I've also received a lot of support from his department, D4, in operating scanners, cameras and calibration.

I have thoroughly enjoyed our collaboration with Cristian Sminchisescu. Cristian has been incredibly supportive and always pushed me to have bigger vision for my research. I've learnt a lot from his insights into writing papers and presenting research.

I count myself fortunate to have worked at length with Xianghui Xie, Xiaohan Zhang, Keyang Zhou and Garvita Tiwari. It is inspiring to see them learn, grow and produce quality research. I also want to thank Thiemo Alldieck, Jan Eric Lenssen, Julian Chibane, Yongqin Xian, Shuo Wen, Zeynep Akata and Bernt Schiele for letting me hop on their projects. It has been a terrific learning experience to work along with these people on different problems. I've also greatly benefited from the highly supportive environment at our group, Real Virtual Humans. I had an absolute blast working with Ilya Petrov, Yuxuan Xue, Riccardo Marin, Aymen Mir, Vladimir Guzov, Verica Lazova, Chaitanya Patel and Zhou Liao. Our conversations were fun and discussions enriching. Our weekly meetings were not just a source of great ideas and critique but also helped me beat the Covid blues.

I was fortunate to do an internship at Meta Reality Labs during my PhD and I want to thank Tony Tung, Nikolaos Sarafianos and Yuanlu Xu for it. The team provided the perfect balance of support when I needed help and latitude when I wanted to try out crazy ideas.

A very big thanks to Connie Balzert and Jessica Endress, our administrative assistants at MPI and Tübingen, for making my life in Germany easier. And I cannot thank our IT service desk (IST) at MPI enough. I can very confidently say that I have never experienced a smoother computing infrastructure before.

I want to thank my friends, who have been a constant source of much needed fun. Big shout out to Klaara and Rakshith for always looking after me and keeping me well fed; Anusha and Tribhu for always having a place for me to vacation; Maria and Max for the amazing vegan food and our picnics; Anna for the best apple pies; Mo and Kata for hosting me over for wonderful dinners; Ayush, Adwait and Aditya for our planned but often impromptu chilling sessions. I also want to thank my friends who have been constantly checking-in on me and the status of this thesis. Thank you Abhijeet, Agni, Roopal, Aditya and Vedika.

Finally but not in the least, I want to thank Aamna and my family for always being there for me. Words fail me in expressing my gratitude for all that they have done for me.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Modeling 3D humans has garnered a lot of attention in the last few years due to several real world applications like virtual try-on, gaming, animation, motion capture and Metaverse. All these applications require building models of digital humans in the image of real ones.

Early works leveraging deep neural networks, on understanding humans started with estimating the 2D pose (Insafutdinov et al., 2016; Pishchulin et al., 2016) and subsequently 3D poses (Guzov et al., 2021; Ionescu et al., 2014; Mahmood et al., 2019; Marcard et al., 2018; Omran et al., 2018). The pose carries a lot of information about a person, for instance, it can help us with tracking (Andriluka et al., 2018; Reddy et al., 2021), understand human activities (Luvizon et al., 2018; Song et al., 2021) and also forecast actions (Bodla et al., 2021; Walker et al., 2017). But it still does not capture the full picture. For instance, it does not tell about the shape of the person and one cannot reason about fine grained interactions between a person and their environment with just pose.

This necessitates building models of human body shape. Early models used volumetric representations like mixture of gaussians to represent different body parts (Plänkers and Fua, 2003; Sminchisescu and Telea, 2002; Stoll et al., 2011). The volumetric models were later extended to obtain a mesh based representation (Hasler et al., 2009; Rhodin et al., 2016a) which has several advantages like explicit surface representation and ease of rendering.

These advances paved way for popular parametric body models like SCAPE (Anguelov et al., 2005; Pishchulin et al., 2017), SMPL (Loper et al., 2015; Pavlakos et al., 2019), GHUM (Xu et al., 2020) and their extensions. Building on the success of such models, more detailed models have been proposed for modelling soft tissue deformations (Pons-Moll et al., 2015), articulated hands (Romero et al., 2017) and faces (Li et al., 2017; Ranjan et al., 2018; Tewari et al., 2017). Despite their strengths and wide applicability, these methods model smooth *undressed* body shape without high frequency clothing details.

In order to capture clothing details, Alldieck et al. (2018b) proposed an extension to SMPL model (SMPL+D) by adding per-vertex displacements on the SMPL mesh (Alldieck et al., 2019a, 2018a, 2019b). We extensively explore this representation in this thesis and demonstrate that SMPL+D can be used to register scans and point clouds of dressed people (Bhatnagar et al., 2020a,b). This makes the 3D scans more amenable to editing shape and pose, which is essential for a lot of real world applications. Chapters 5 and 6 describe our work in more detail.

SMPL+D allows us to model deformations corresponding to the garments, but the deformations do not carry any semantic meaning corresponding to the garments, i.e. SMPL+D produces a single mesh without reasoning about the garments. This limits its use for tasks like garment modelling, texture transfer and virtual try-on. To address this Pons-Moll et al. (2017), proposed a method to register a garment template to a 4D sequence of a dressed person. Although quite powerful, this method requires high quality 4D data thus limiting practical application. Therefore we present a new extension to SMPL model, SMPL+Garments (SMPL+G) (Bhatnagar et al., 2019) that allows us to register garments from a large corpus of static 3D scans. Garments in SMPL+G can be used to dress SMPL model in arbitrary poses and shapes. We further learn Multi-Garment Network (MGN), that allows us to reconstruct the body shape as SMPL model and garments as separate meshes on top of it from just a few RGB images. We discus this in more detail in chapter 4.

Mesh based models like SMPL, SMPL+D, SMPL+G are really powerful as they provide control over the human surface via explicit parameters for pose, shape and non-rigid deformations but meshes have fixed resolution. This limits the resolution of details that these models can capture. Recent works on implicit function based reconstruction have shown great promise in this regard (Chibane et al., 2020a,b; He et al., 2021; Huang et al., 2020; Saito et al., 2019, 2020). Unfortunately, these works produce a static human mesh with no explicit control over the pose and shape which comes with the parametric modeling. In chapter 5, we present Implicit Part Network (IPNet) (Bhatnagar et al., 2020a) that uses implicit functions for detailed 3D reconstruction and it also registers the static implicit reconstruction with the SMPL model, making the reconstruction controllable. Another concurrent line of research, orthogonal to our work, is also looking into using implicit function to model not just the static shape but also articulation (Chen et al., 2021; Deng et al., 2020; Dong et al., 2022; Saito et al., 2021; Tiwari et al., 2021).

Modelling how humans look is necessary but not sufficient. Human beings are more than just what we look like. We can perceive our surroundings, plan actions and act upon our environment. This ability needs to be modeled in the next generation of digital humans we aim to build. Modelling human actions is also important in several other fields like robotics where we want to train agents to perform certain actions/tasks by looking at how humans do them (**robotics work**). And it is not just robots that can learn from the models of human interactions, but even human beings can learn things like how to operate tools, do different tasks etc. using these models. This is an especially challenging task as there is very limited data available to build models of human interaction. This is in big part because capturing human-object interactions is hard. There are sever occlusions, humans need to be free to move around which means that they can go outside the recording volume of scanners, and the scope of possible interactions with objects is huge. For instance there are many different ways of even sitting on a chair!

To this end, in chapter 7, we propose a portable, easy to setup and use multi-camera capture system to record human-object interactions. We use this setup to record natural and diverse human-object interactions and learn the first model that can automatically track the human, the object and the interactions between them.

Modelling humans is not only useful for real world applications like virtual try-on, gaming, learning in simulated environment etc. but it also provides a deeper insight into human nature and behaviour as well, eg. how we perceive ourselves and others, how we interact with our surroundings including other humans and how we envision the digital world to be, for instance Metaverse. In this thesis we propose several advancements in data driven modelling of 3D humans including pose, shape, clothing and interactions. We summarise our contributions chapter-wise below.

1.1 STRUCTURE AND CONTRIBUTIONS

This thesis is divided into eight chapters. Chapters 4, 5, 6 and 7 include the main technical contributions and their evaluations.

- Chapter 2 introduces relevant technical background, such as the parametric body model SMPL and introduction to implicit functions, used in the later chapters.
- Chapter 3 discusses the relevant prior work.

- Chapter 4 (published as Bhatnagar et al. (2019)) introduces Multi-garment Network (MGN), the first approach that can reconstruct the 3D body shape and garment meshes as separate layers from a few RGB images of a person. More importantly MGN is trained on *real data*. At the time of publication, there did not exist a dataset of registered 3D garments that could be used for training MGN. Therefore, we proposed the first approach to register 3D garments of a class, in arbitrary poses and topology, to a common garment template. With our registered 3D garments, we released a digital wardrobe fully compatible with the SMPL model and we show that we can dress SMPL in diverse poses and shapes with our garments. This allowed for the first time a straightforward approach to real world applications like texture transfer and virtual try-on.

- Chapter 5 (published as Bhatnagar et al. (2020a)) presents an approach for controllable 3D reconstruction of dressed humans using implicit functions. Concurrent works on implicit 3D reconstruction can only reconstruct static 3D meshes of dressed humans but a lot of real world applications like animation, gaming etc. require controllable 3D models of a person and not static meshes. This control is possible with parametric body models such as SMPL but these models lack high frequency details corresponding to clothing, face, hair etc. Our proposed Implicit Part Network (IPNet), leverages implicit functions to obtain detailed 3D reconstruction and also predicts correspondences to the SMPL body model. We use these correspondences to register the SMPL model to our implicit reconstruction therefore making it controllable.

- Chapter 6 (published as Bhatnagar et al. (2020b)) proposes an approach to fit a parametric model to a scan or point cloud of a person. We address the classic problem of model based fitting or 3D registration. 3D registration is a decade old problem in computer vision and graphics and is the corner stone of several real world tasks. Traditionally, registration is a two step process where we first (i) establish the correspondences between the input point cloud or scan and the parametric model, and then (ii) we optimise the model to minimize the distance between the corresponding points. This two step process is typically non-differentiable with respect to correspondences thus making the registration task not end-to-end differentiable. In this chapter, we use implicit function based representation of the SMPL model and propose a novel formulation that makes registration fully differentiable. Our semi-supervised approach can be trained with 1000 synthetic SMPL meshes as opposed to prior supervised approach, which requires 200k annotated 3D scans. More importantly, our approach works with dressed humans whereas prior work could only handle undressed shapes.

- Chapter 7 (published as Bhatnagar et al. (2022)) presents the first dataset and method to track humans, objects and their interactions in 3D. In the previous chapters we focused on modelling the appearance of 3D humans including pose, body shape and clothing. Although quite useful, these representations are not sufficient as they do not take the surroundings of the person into account. Human beings can perceive their surroundings and also act upon it. This ability needs to be captured in the models of digital humans that we build. Our BEHAVE dataset and method is a significant contribution in this direction.

- Chapter 8 discusses important insights from this thesis as well as opportunities for future work.

1.2 PUBLICATIONS

All the work presented in this thesis was also published in the following publications:

- Bharat Lal Bhatnagar et al. (2019). “Multi-Garment Net: Learning to Dress 3D People from Images.” In: *IEEE International Conference on Computer Vision (ICCV)*
- Bharat Lal Bhatnagar et al. (2020a). “Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction.” In: *European Conference on Computer Vision (ECCV)* (oral)
- Bharat Lal Bhatnagar et al. (2020b). “LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration.” In: *Advances in Neural Information Processing Systems (NeurIPS)* (oral)
- Bharat Lal Bhatnagar et al. (2022). “BEHAVE: Dataset and Method for Tracking Human Object Interactions.” In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*

In addition, contributions were made to the following publications which are, however, not part of this thesis:

- Thiemo Alldieck et al. (2019a). “Learning to Reconstruct People in Clothing from a Single RGB Camera.” In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Garvita Tiwari et al. (2020). “SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing.” In: *European Conference on Computer Vision (ECCV)*. Springer
- Keyang Zhou et al. (2020). “Unsupervised Shape and Pose Disentanglement for 3D Meshes.” In: *The European Conference on Computer Vision (ECCV)*
- Keyang Zhou et al. (2022a). “TOCH: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement.” In: *European Conference on Computer Vision (ECCV)*. Springer
- Xiaohan Zhang et al. (2022b). “COUCH: Towards Controllable Human-Chair Interactions.” In: *European Conference on Computer Vision (ECCV)*. Springer
- Xianghui Xie et al. (2022). “CHORE: Contact, Human and Object Reconstruction from a single RGB image.” In: *European Conference on Computer Vision (ECCV)*. Springer
- Yongqin Xian et al. (2022). “Any-Shot GIN: Generalizing Implicit Networks for Reconstructing Novel Classes.” In: *2022 International Conference on 3D Vision (3DV)*. IEEE
- Keyang Zhou et al. (2022b). “Adjoint Rigid Transform Network: Task-conditioned Alignment of 3D Shapes.” In: *2022 International Conference on 3D Vision (3DV)*. IEEE

2 BACKGROUND

This chapter introduces the relevant technical background for the thesis. The core technical contributions of the thesis are centered around proposing detailed and control-able representations for 3D humans, clothing and interactions with objects. To this end we use parametric body models, primarily SMPL (Loper et al., 2015) and deep learnt implicit functions. We will briefly cover these here.

2.1 IMPLICIT FUNCTIONS FOR REPRESENTING SURFACES.

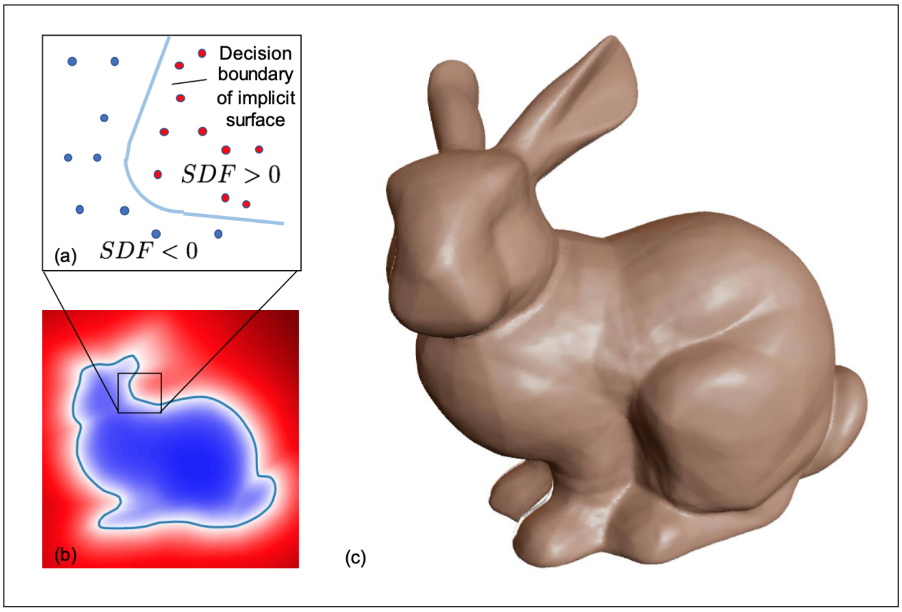


Figure 2.1: A triangulated surface (c) can be represented implicitly using a distance field (b). The points inside the surface will have an SDF<0 whereas SDF>0 outside (a). The boundary or the zero-level set of this implicit representation marks the surface. Source: Park et al. (2019).

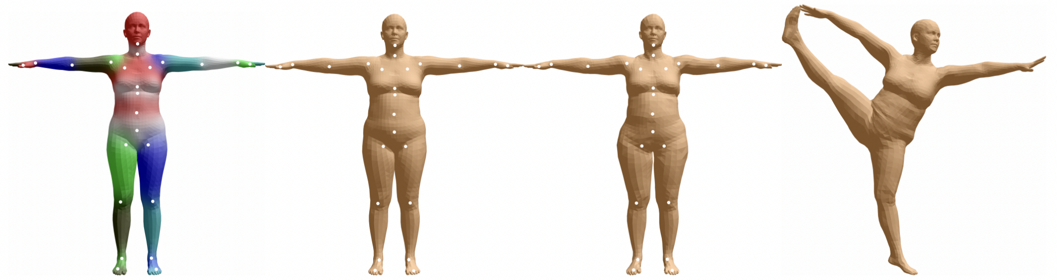
Traditionally, triangulated meshes and voxels have been the preferred representations for 3D surfaces because they are easy to use, interpret and render. Implicit representations on the other hand are quite compact but not as amenable. Let us understand implicit representations better with a simple example of 2D circle centered at $(a, b) \in \mathbb{R}^2$ with radius $r \in \mathbb{R}$. We can define a function $d(x, y) = (x - a)^2 + (y - b)^2 - r^2, d \in \mathbb{R}$ which gives us the distance of any arbitrary point $(x, y) \in \mathbb{R}^2$ from the surface of the circle. The circle can therefore be implicitly represented with function $d(\cdot, \cdot)$, with all points (x, y) such that $d(x, y) < 0$ lie inside the circle, $d(x, y) > 0$ lie outside the circle and points with $d(x, y) = 0$ (zero-level set) constitute the surface of the circle.

It is important to note that it is easy to come up with a manual implicit representation for a 2D circle but for arbitrary detailed 3D surfaces such a formulation might not be intuitive. This was one of the main reasons why implicit representations did not grow in popularity until recent times. Recent work proposed to use neural networks to learn implicit representations such as occupancy (Mescheder et al., 2019) and signed distance fields (Park et al., 2019) with neural networks as representations for 3D surfaces (see Fig. 2.1).

The typical framework involves a trainable neural network $f(\cdot)$, which conditioned on an "input", $\mathcal{S}$ (a learn-able latent code, a 2D image, 3D voxel grid/ scan/ point cloud, or even text), predicts "output", $o = f(\mathbf{p}|\mathcal{S}) \in \mathbb{R}$ (occupancy, signed/unsigned distance field etc.) for several "query points", $\mathbf{p} \in \mathbb{R}^d$. Where typically $d = 3$ for 3D space, although higher dimensional manifolds can also be learnt. The neural network and optional latent code are be trained with gradient based back-propagation.

Once trained, the network $f(\cdot)$ can be densely queried with points in $\mathbb{R}^d$ to obtain an occupancy/distance field. The surface is represented as the zero level set of this field and is typically extracted using marching cubes (Lorensen and Cline, 1987).

2.2 PARAMETRIC BODY MODEL: SMPL.



(a) $\bar{\mathbf{T}}, \mathcal{W}$ **(b)** $\bar{\mathbf{T}} + B_S(\vec{\beta}), J(\vec{\beta})$ **(c)** $T_P(\vec{\beta},\vec{\theta}) = \bar{\mathbf{T}}+B_S(\vec{\beta})+B_P(\vec{\theta})$ **(d)** $W(T_P(\vec{\beta},\vec{\theta}), J(\vec{\beta}),\vec{\theta},\mathcal{W})$

Figure 2.2: SMPL represents 3D humans as a function of pose and shape parameters. SMPL applies a series of linear displacements based on pose and shape to the base template and performs linear skinning based on the skeleton. Source: Loper et al. (2015).

The most commonly used representation for 3D humans are parametric meshes and SMPL(+X)/ SMPL+D body model (Loper et al., 2015; Pavlakos et al., 2019) is perhaps the most widely used. We briefly describe the working of SMPL model here and in Chapter 4.

SMPL+D, $M(\cdot)$, represents the human body as a parametric function of pose($\boldsymbol{\theta}$), shape($\boldsymbol{\beta}$), global translation($\boldsymbol{t}$) and optional per-vertex displacements ($\mathbf{D}$). The SMPL function applies a series of shape $B_s(\cdot)$ and pose $B_p(\cdot)$, dependent linear displacements to a base mesh $\mathbf{T}$ with $n = 6890$ vertices in a T-pose followed by standard skinning $W(\cdot)$. See Fig. 2.2.

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{2.1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D}. \tag{2.2}$$

$J(\cdot)$ denotes the pre-defined SMPL skeleton, and $\mathbf{W}$ represents the skinning blend weights.

3 RELATED WORK

In this thesis, we present our work on modelling 3D humans, including pose, body shape, clothing and even interactions with objects. These are very broad and active research areas and in this chapter we discuss various works addressing these directions. We first discuss various ways to represent 3D humans and clothing, covering classical parametric models, implicit function based representations and more recent hybrid representations (Sec. 3.1). We cover various advantages and the accompanying trade-offs, such as capacity to represent fine-grained details vs. controllability, with these representations.

We then discuss how can these models be used for real world tasks like reconstructing 3D humans from 2D images, videos, and 3D point clouds, scans (Sec. 3.2). We cover wide range of works dealing with modelling the pose, shape and even detailed clothing We put special emphasis on 3D registration of human point clouds and scans, with parametric human models as it is a core task in computer vision and graphics alike, and is the corner stone of several applications. Our work also makes significant contributions in this direction.

As also discussed in the introduction, we argue in this thesis that modelling the appearance of 3D humans is necessary but not sufficient. We should also look at the humans in the context of their surroundings. To that end, we discuss works on modelling human-object interactions in 3D. In the interest of brevity we primarily focus on recent relevant works but we do acknowledge the vast body of prior work that enabled these recent advances.

3.1 REPRESENTATIONS FOR 3D HUMANS

Attention that modelling 3D humans has received in the recent years is well warranted due to the many real world applications that it enables, ranging from gaming, virtual-try on, animation to the more ambitious full fledged Metaverse. The first step therefore, is to come up with suitable ways to represent the humans in 3D. Understandably, these representations vary depending on the downstream tasks and bring with them various trade-offs as to what they can and cannot do. In this section, we discuss these representations and their strengths and weaknesses.

3.1.1 *Parametric representations for 3D humans and clothing*

Parametric body models factorize deformations into shape and pose (Joo et al., 2018b; Loper et al., 2015; Xu et al., 2020), soft-tissue (Pons-Moll et al., 2015), and recently even clothing (Bhatnagar et al., 2019; Patel et al., 2020; Tiwari et al., 2020). This factorization constraints meshes to the space of humans. A major disadvantage of such parametric representations is that they are typically tied to a fixed topology thus limiting details. Fixed topology also makes it harder to generalise to diverse clothing, eg: it is difficult to model skirts and loose clothing with SMPL model.

3.1.2 *Implicit function based representations for 3D humans and clothing*

TSDFs (Curless and Levoy, 1996) can represent the human surface implicitly, which is common in depth-fusion approaches (Newcombe et al., 2015; Slavcheva et al., 2017). Such free-form representation has been combined with SMPL (Loper et al., 2015) to increase robustness and tracking (Yu et al., 2018). Alternatively, implicit functions can be parameterized with Gaussian primitives (Rhodin et al., 2016b; Stoll et al., n.d.). Since these approaches are not learning based, they can not reconstruct the occluded part of the surface in single view settings.

Voxels discretize the implicit occupancy function, which makes convolution operations possible. CNN based reconstructions using voxels (Gilbert et al., 2018; Varol et al., 2018; Zheng et al., 2019) or depth-maps (Gabeur et al., 2019; Leroy et al., 2018; Smith et al., 2019) typically produce more details than parametric models, but limbs are often missing. More importantly, unlike our method, the reconstruction quality is limited by the resolution of the voxel grid and increasing the resolution is hard as the memory footprint grows cubically.

Recent methods learn a continuous implicit function representing the object surface directly (Chen and Zhang, 2019; Mescheder et al., 2019; Park et al., 2019). However, these approaches have difficulties reconstructing articulated structures because they use a global shape code, and the networks tend to memorize typical object coordinates (Chibane et al., 2020a). The occupancy can be predicted based on local image features instead (Saito et al., 2019), which results in medium-scale wrinkles and details, but the approach has difficulties with out of image plane poses, and is designed for image-reconstruction and can not handle point clouds. Recently, IF-Nets (Chibane et al., 2020a) have been proposed for 3D reconstruction and completion from point clouds – a mutliscale grid of deep features is first computed from the point cloud, and a decoder network classifies the occupancy based on mutli-scale deep features extracted at continuous point locations. These recent approaches (Chibane et al., 2020a; Saito et al., 2019) make occupancy decisions based on local and global evidence, which results in more robust reconstruction of articulated and fine structures than decoding based on the X-Y-Z point coordinates and a global latent shape code (Chen and Zhang, 2019; Mescheder et al., 2019; Michalkiewicz et al., 2019; Park et al., 2019). However, they do not reconstruct shape under clothing and surfaces are not typically controllable. There are concurrent works (Deng et al., 2020; He et al., 2021; Huang et al., 2020) that can directly predict the skinning weights using implicit functions thus providing an alternative to our approach of registering the implicit reconstruction with a parametric model to enable re-animation.

IMPLICIT VS PARAMETRIC MODELLING Parametric models allow control over the surface and never miss body parts, but feed-forward prediction is hard, and reconstructions lack detail. Learning the implicit functions representing the surface directly is powerful because the output is continuous, details can be preserved better, and complex topologies can be represented. However, the output is not controllable, and can not guarantee that all body parts are reconstructed. Naive fitting of a body model to a reconstructed implicit surface often gets trapped into local minimal when the poses are difficult or clothing occludes the body (see Fig. 5.5). These observations motivate the design of our hybrid method, which retains the benefits of both representations: i) control, ii) detail, iii) alignment with the input point clouds.

3.2 DETAILED 3D HUMANS FROM 2D IMAGES

Perceiving humans from monocular RGB data (Bogo et al., 2016; Guler and Kokkinos, 2019; Habermann et al., 2020; Kanazawa et al., 2018; Kocabas et al., 2020; Kolotouros et al., 2019; Pavlakos et al., 2019, 2018; Saito et al., 2019; Zanfir et al., 2021) and under multiple views (Huang et al., 2017; Huang et al., 2018; Iskakov et al., 2019; Joo et al., 2018a; Rhodin et al., 2018) settings has been widely explored. Recent work tends to focus on reconstructing fine details like hand gestures and facial expressions (Choutas et al., 2020; Feng et al., 2021; Zanfir et al., 2020; Zhou et al., 2021b), self-contacts (Fieraru et al., 2021; Muller et al., 2021a), interactions between humans(Fieraru et al., 2020), and even clothing (Alldieck et al., 2019a; Bhatnagar et al., 2019).
Following the success of pixel-aligned implicit function learning (Saito et al., 2019, 2020), recent methods can capture human performance from sparse (Huang et al., 2018; Xu et al., 2021) or even a single RGB camera (Li et al., 2020a,b). However, capturing 3D humans from RGB data involves a fundamental ambiguity between depth and scale. Therefore, recent methods use RGBD (Pandey et al., 2019; Su et al., 2020; Tao et al., 2018; Wang et al., 2020; Yu et al., 2021) and even volumetric input for reliable human capture. We discuss these works next.

3.3 DETAILED 3D HUMANS FROM 3D SCANS AND POINTCLOUDS

3.3.1 *Static 3D Humans*

TSDFs (Curless and Levoy, 1996) can represent the human surface implicitly, which is common in depth-fusion approaches (Newcombe et al., 2015; Slavcheva et al., 2017). Such free-form representation has been combined with SMPL (Loper et al., 2015) to increase robustness and tracking (Yu et al., 2018). Alternatively, implicit functions can be parameterized with Gaussian primitives (Rhodin et al., 2016b; Stoll et al., n.d.). Since these approaches are not learning based, they can not reconstruct the occluded part of the surface in single view settings.
Voxels discretize the implicit occupancy function, which makes convolution operations possible. CNN based reconstructions using voxels (Gilbert et al., 2018; Varol et al., 2018; Zheng et al., 2019) or depth-maps (Gabeur et al., 2019; Leroy et al., 2018; Smith et al., 2019) typically produce more details than parametric models, but limbs are often missing. More importantly, unlike our work, the reconstruction quality is limited by the resolution of the voxel grid and increasing the resolution is hard as the memory footprint grows cubically.
Recent methods learn a continuous implicit function representing the object surface directly (Chen and Zhang, 2019; Mescheder et al., 2019; Park et al., 2019). However, these approaches have difficulties reconstructing articulated structures because they use a global shape code, and the networks tend to memorize typical object coordinates. Recent works use CNNs to perform localised predictions (Chibane et al., 2020a,b; Saito et al., 2019, 2020) to mitigate this issue.

3.3.2 *Controllable 3D Humans*

In the context of articulated humans, classical ICP based alignment to parametric models such as SMPL (Loper et al., 2015) has been widely used for registering human body shapes (Bogo et al., 2014, 2017; Dyke et al., 2019; Hirshberg et al., 2012; Pishchulin et al., 2017; Pons-Moll et al., 2015; Pons-Moll and Rosenhahn, 2011) and even detailed 3D garments (Bhatnagar et al., 2019; Pons-Moll et al., 2017). Incorporating additional knowledge such as precomputed 3D joints, facial landmarks (Alldieck et al., 2019a; Lazova et al., 2019) and part segmentation (Bhatnagar et al., 2020a) significantly improves the registration quality but these pre-processing steps are prone to error at various steps. Our work on the other hand, can reconstruct controllable 3D humans without relying on these additional inputs.

Orthogonal to our work, an interesting research direction to use implicit functions for also learn skinning along with 3D shape (Chen et al., 2021; Dong et al., 2022; Mihajlovic et al., 2021; Saito et al., 2021; Tiwari et al., 2021). This direction is quite promising as it allows us to learn control-able 3D models with garment deformations (Ma et al., 2021a, 2022, 2021b; Wang et al., 2021) in arbitrary topology.

3.4 INTERACTIONS BETWEEN HUMANS AND OBJECTS IN 3D

Reconstructing rigid objects and more importantly humans from images, videos and point clouds has been an active research area in the last decade. These works focus primarily on modelling the appearance of the human. Few works try to reason about the human in the context of its environment but are largely restricted to static scenes. Research on dynamic human-object interactions has either lacks real scenes or is restricted to just hand-object interactions. This is in large parts due to the unavailability of large scale datasets to understand and benchmark full-body human-object interactions.

In this section, we first briefly review work focused on object and human reconstruction, in isolation from their environmental context. Such methods focus on modelling appearance and do not consider interactions. Next, we cover methods focused on humans in static scenes and finally discuss closer-related work to ours, for modelling dynamic human-object interactions.

3.4.1 *Modelling 3D objects*

Most existing work on reconstructing 3D objects from RGB (Choy et al., 2016; Lei et al., 2020; Mescheder et al., 2019; Tzionas and Gall, 2015; Wu et al., 2017, 2018) and RGBD (Kundu et al., 2018; Muller et al., 2021b; Yang et al., 2017; Zhou et al., 2021a) data does so in isolation, without the human involvement or the interaction. While challenging, it is arguably more interesting to reconstruct objects in a dynamic setting under severe occlusions from the human.

3.4.2 *Humans in static scenes*

Modelling how humans act in a scene is both important and challenging. Tasks like placement of humans into static scenes (Hassan et al., 2021b; Li et al., 2019b; Zhang et al., 2020b), motion prediction (Cao et al., 2020; Hassan et al., 2021a), human pose reconstruction under scene constrains (Chen et al., 2019; Hassan et al., 2019; Weng and Yeung, 2020; Zanfir et al., 2018; Zhang et al., 2021; Zhao et al., 2022), and human-object interactions (Savva et al., 2016; Wu et al., 2022; Zhang et al., 2022a; Zhang et al., 2022b), have been investigated extensively in recent years. These methods are relevant but restricted to modelling humans interacting with *static* objects. We address a more challenging problem of jointly tracking human-object interactions in *dynamic* environments where objects are manipulated.

3.4.3 *Modelling dynamic human-object interactions in 3D*

Recently, there has been a strong push on modeling hand-object interactions based on 3D (Karunratanakul et al., 2021, 2020; Taheri et al., 2020; Zhou et al., 2022a), 2.5D (Brahmbhatt et al., 2019, 2020) and 2D (Corona et al., 2020; Ehsani et al., 2020; Grady et al., 2021; Hasson et al., 2019; Yang et al., 2021) data. Although powerful, these methods are currently restricted to modelling only *hand-object* interactions. In contrast, we are interested in *full body* capture. Methods for dynamic full body human object interaction approach the problem via 2D action recognition (Hu et al., 2017; Liu et al., 2019) or reconstruct 3D object trajectories during interactions (Dabral et al., 2021). Despite being impressive, such methods either lack full 3D reasoning (Hu et al., 2017; Liu et al., 2019) or are limited to specific objects (Dabral et al., 2021).

More recent work reconstructs and tracks human-object interactions from RGB (Sun et al., 2021) or RGBD streams (Su et al., 2021), but does not consider contact prediction, thus missing a component necessary for accurate interaction estimates.

Very relevant to our work, PHOSA (Zhang et al., 2020a) reconstructs humans and objects from a single image. PHOSA uses hand crafted heuristics, instance specific optimization for fitting, and pre-defined contact regions, which limits generalization to diverse human-object interactions. Using our BEHAVE dataset, our work (Xie et al., 2022) learns to reconstruct humans and objects from data and outperforms PHOSA. After our work BEHAVE (Bhatnagar et al., 2022), Huang et al. (2022) proposed a similar dataset to capture human-object interaction with articulated hands.

4 MULTI-GARMENT NET: LEARNING TO DRESS 3D PEOPLE FROM IMAGES

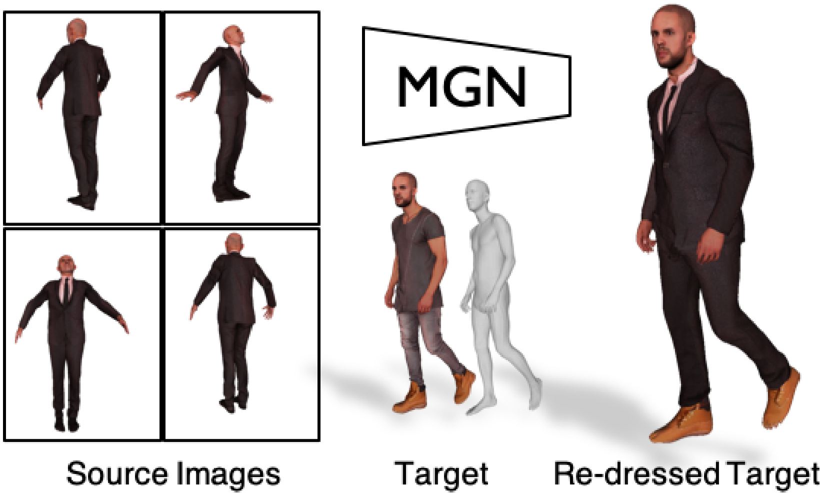


Figure 4.1: Garment re-targeting with Multi-Garment Network (MGN). Left to right: images from source subject, body from the target subject, target dressed with source garments. From one or more images, MGN can reconstruct the body shape and each of the garments separately. We can transfer the predicted garments to a novel body including geometry and texture.

In this chapter, we present Multi-Garment Network (MGN), a method to predict body shape and clothing, layered on top of the SMPL (Loper et al., 2015) model from a few frames (1-8) of a video. Several experiments demonstrate that this representation allows higher level of control when compared to single mesh or voxel representations of shape. Our model allows to predict garment geometry, relate it to the body shape, and transfer it to new body shapes and poses. To train MGN, we leverage a digital wardrobe containing 712 digital garments in correspondence, obtained with a novel method to register a set of clothing templates to a dataset of real 3D scans of people in different clothing and poses. Garments from the digital wardrobe, or predicted by MGN, can be used to dress any body shape in arbitrary poses. Using this digital wardrobe, we synthesize images of people in different poses and clothing. Several experiments demonstrate the potential uses of predicting separate meshes, which has never been explored to reconstruct people in images. We will make publicly available the digital wardrobe, the MGN model, and code to *dress* SMPL with the garments at: http://virtualhumans.mpi-inf.mpg.de/mgn.

4.1 INTRODUCTION

The 3D reconstruction and modelling of humans from images is a central problem in computer vision and graphics. Although a few recent methods (Alldieck et al., 2019a, 2018a,b; Habermann et al., 2019; Natsume et al., 2019; Saito et al., 2019) attempt reconstruction of people with clothing, they lack realism and control. This limitation is in great part due to the fact that they use a single surface (mesh or voxels) to represent both clothing and body. Hence they can not

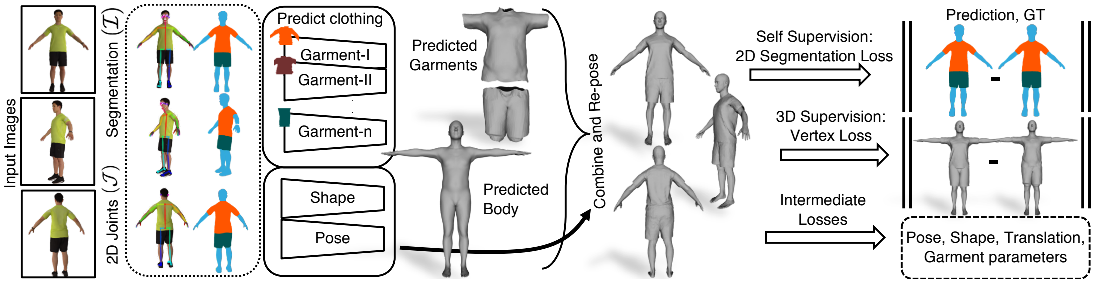

Input Images
Segmentation ($\mathcal{I}$)
2D Joints ($\mathcal{J}$)
Predict clothing
Garment-I
Garment-II
Garment-n
Shape
Pose
Predicted Garments
Predicted Body
Combine and Re-pose
Self Supervision: 2D Segmentation Loss
3D Supervision: Vertex Loss
Intermediate Losses
Prediction, GT
Pose, Shape, Translation, Garment parameters

Figure 4.2: Overview of our approach. Given a small number of RGB frames (currently 8), we pre-compute semantically segmented images ($\mathcal{I}$) and $2D$ Joints ($\mathcal{J}$). Our Multi-Garment Network (MGN), takes $\{\mathcal{I}, \mathcal{J}\}$ as input and infers separable garments and the underlying human shape in a canonical pose. We repose these predictions using our per-frame pose predictions. We train MGN with a combination of 2D and 3D supervision. The 2D supervision can be used for online refinement at test time.

capture the clothing separately from the subject in the image, let alone map it to a novel body shape.

In this paper, we introduce Multi-Garment Network (MGN), the first model capable of inferring human body and layered garments on top as separate meshes from images directly. As illustrated in Fig. 4.1 this new representation allows full control over body shape, texture and geometry of clothing and opens the door to a range of applications in VR/AR, entertainment, cinematography and virtual try-on.

Compared to previous work, MGN produces reconstructions of higher visual quality, and allows for more control: 1) we can infer the 3D clothing from one subject, and dress a second subject with it, (see Fig. 4.1, 4.8) and 2) we can trivially map the garment texture captured from images to any garment geometry of the same category (see Fig.4.7).

To achieve such level of control, we address two major challenges: learning per-garment models from 3D scans of people in clothing, and learning to reconstruct them from images.

We define a discrete set of garment templates (according to the categories long/short shirt, long/short pants and coat) and register, for every category, a single template to each of the scan instances, which we automatically segmented into clothing parts and skin. Since garment geometry varies significantly within one category (*e.g.* different shapes, sleeve lengths), we first minimize the distance between template and the scan boundaries, while trying to preserve the Laplacian of the template surface. This initialization step only requires solving a linear system, and nicely stretches and compresses the template globally, which we found crucial to make subsequent non-rigid registration work. Using this, we compile a *digital wardrobe* of real 3D garments worn by people, (see Fig. 4.3). From such registrations, we learn a vertex based PCA model per garment. Since garments are naturally associated with the underlying SMPL body model, we can transfer them to different body shapes, and re-pose them using SMPL. From the digital wardrobe, MGN is trained to predict, given one or more images of the person, the body pose and shape parameters, the PCA coefficients of each of the garments, and a displacement field on top of PCA that encodes clothing detail. At test time, we refine this bottom-up estimates with a new top-down objective that forces projected garments and skin to explain the input semantic segmentation. This allows more fine-grained image matching as compared to standard silhouette matching. Our contributions can be summarized as:

- A novel data driven method to infer, for the first time, separate body shape and clothing from just images (few RGB images of a person rotating in front of the camera).

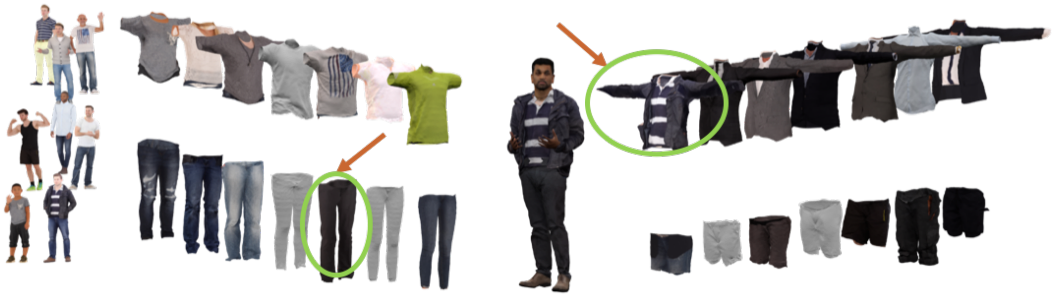


Figure 4.3: Digital 3D wardrobe. We use our proposed multi-mesh registration approach to register garments present in the scans (left) to fixed garment templates. This allows us to build a digital wardrobe and dress arbitrary subjects (center) by picking the garments (marked) from the wardrobe.

- A robust pipeline for 3D scan segmentation and registration of garments. To the best of our knowledge, there are no existing works capable of automatically registering a single garment template set to multiple scans of real people with clothing.
- A novel top-down objective function that forces the predicted garments and body to fit the input semantic segmentation images.
- We demonstrate several applications that were not previously possible such as dressing avatars with predicted $3D$ garments from images, and transfer of garment texture and geometry.
- We will make publicly available the MGN to predict $3D$ clothing from images, the digital wardrobe, as well as code to "dress" SMPL with it.

4.2 MULTI-GARMENT NETWORK: METHOD

In order to learn a model to predict body shape and garment geometry directly from images, we process a dataset of 356 scans of people in varied clothing, poses and shapes. Our data pre-processing (Sec. 4.2.1) consists of the following steps: SMPL registration to the scans, body aware scan segmentation and template registration. We obtain, for every scan, the underlying body shape, and the garments of the person registered to one of the 5 garment template categories: shirt, t-shirt, coat, short-pants, long-pants. The obtained digital wardrobe is illustrated in Fig. 4.3. The garment templates are defined as regions on the SMPL surface; the original shape follows a human body, but it deforms to fit each of the scan instances after registration. Since garment registrations are naturally associated to the body represented with SMPL, they can be easily reposed to arbitrary poses. With this data, we train our Multi-Garment Network to estimate the body shape and garments from one or more images of a person, see Sec. 4.2.2.

4.2.1 *Data Pre-Processing: Scan Segmentation and Registration*

Unlike ClothCap (Pons-Moll et al., 2017) which registers a template to a $4D$ scan sequence of a *single subject*, our task is to register single template across instances of varying styles, geometries, body shapes and poses. Since our registration follows the ideas of Pons-Moll et al. (2017), we describe the main differences here.

BODY-AWARE SCAN SEGMENTATION We first automatically segment the scans into three regions: skin, upper-clothes and pants (we annotate the garments present for every scan). Since even SOTA image semantic segmentation (Gong et al., 2018) is inaccurate, naive lifting

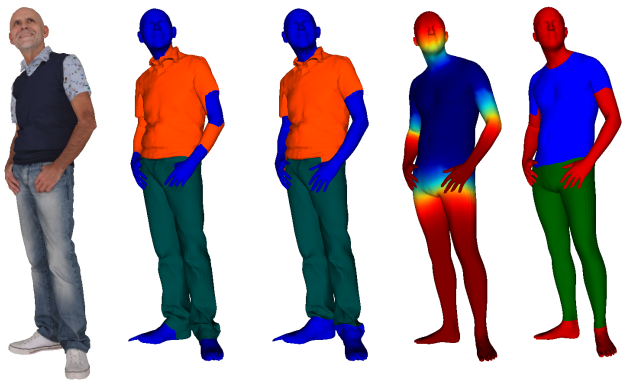


Figure 4.4: Left to right: Scan, segmentation with MRF and CNN unaries, MRF with CNN unaries + garment prior + appearance terms, the garment(t-shirt) prior based on geodesics and the template. Notice how the garment prior is crucial to obtain robust results.

to 3D is not sufficient. Hence, we incorporate *body specific garment priors* and segment scans by solving an MRF on the UV-map of the SMPL surface after non-rigid alignment.

A garment prior (for garment g) derives from a set of labels $\mathbf{l}_g^i \in \{0, 1\}$ indicating the vertices $\mathbf{v}_i \in \mathcal{S}$ of SMPL that are likely to overlap with the garment. The aim is to penalize labelling vertices as g outside this region, see Fig 4.4. Since garment geometry varies significantly within one category (e. g.t-shirts of different sleeve lengths), we define a cost increasing with the geodesic distance $\text{dist}_{\text{geo}}(\mathbf{v}) : \mathcal{S} \mapsto \mathbb{R}$ from the garment region boundary – efficiently computed based on heat flow (Crane et al., 2013). Conversely, we define a similar penalty for labeling vertices in the garment region with a label different than g. As data terms, we incorporate CNN based semantic segmentation (Gong et al., 2018), and appearance terms based Gaussian Mixture Models in *La* color space. The influence of each term is illustrated in Fig. 4.4, for more details we refer to the supp. mat.

After solving the MRF on the SMPL UV map, we can segment the scans into 3 parts by transferring the labels from the SMPL registration to the scan.

GARMENT TEMPLATE We build our garment template on top of SMPL+D, $M(\cdot)$, which represents the human body as a parametric function of pose($\boldsymbol{\theta}$), shape($\boldsymbol{\beta}$), global translation($\boldsymbol{t}$) and optional per-vertex displacements ($\mathbf{D}$):

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{4.1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D}. \tag{4.2}$$

The basic principle of SMPL is to apply a series of linear displacements to a base mesh $\mathbf{T}$ with n vertices in a T-pose, and then apply standard skinning $W(\cdot)$. Specifically, $B_p(\cdot)$ models pose-dependent deformations of a skeleton J, and $B_s(\cdot)$ models the shape dependent deformations. $\mathbf{W}$ represents the blend weights.

For each garment class g we define a template mesh, $\mathbf{G}^g$ in T-pose, which we subsequently register to explain the scan garments. We define $\mathbf{I}^g \in \mathbb{Z}^{m_g \times n}$ as an indicator matrix, with $\mathbf{I}^g_{i,j} = 1$ if garment g vertex $i \in \{1 \ldots m_g\}$ is associated with body shape vertex $j \in \{1 \ldots n\}$. In our experiments, we associate a single body shape vertex to each garment vertex. We compute displacements to the corresponding SMPL body shape $\boldsymbol{\beta}^g$ under the garment as

$$\mathbf{D}^g = \mathbf{G}^g - \mathbf{I}^g T(\boldsymbol{\beta}^g, \mathbf{0}\grave{}, \mathbf{0}_{\mathbf{D}}) \tag{4.3}$$

Consequently, we can obtain the garment shape (unposed), T^g for a new shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ as

$$T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g) = \mathbf{I}^g T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0}) + \mathbf{D}^g \tag{4.4}$$

To pose the vertices of a garment, each vertex uses the skinning function in Eq. 4.1 of the associated SMPL body vertex.

$$G(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g) = W(T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{4.5}$$

GARMENT REGISTRATION Given the segmented scans, we non-rigidly register the body and garment templates (upper-clothes, lower-clothes) to scans using the multi-part alignment proposed in Pons-Moll et al. (2017). The challenging part is that garment geometries vary significantly across instances, which makes the multi-part registration fail (see supplementary). Hence, we first initialize by deforming the vertices of each garment template with the shape and pose of SMPL registrations, obtaining deformed vertices $\mathbf{G}^g_{\text{init}}$. Note that since the vertices defining each garment template are fixed, the clothing boundaries of the initially deformed garment template will not match the scan boundaries. In order to globally deform the template to match the clothing boundaries in a single shot, we define an objective function based on Laplacian deformation (Sorkine, 2005).

Let $\mathbf{L}^g \in \mathbb{R}^{m_g \times m_g}$ be the graph Laplacian of the garment mesh, and $\boldsymbol{\Delta}_{\text{init}} \in \mathbb{R}^{m_g \times 3}$ the differential coordinates of the initially deformed garment template $\boldsymbol{\Delta}_{\text{init}} = \mathbf{L}\,\mathbf{G}^g_{\text{init}}$. For every vertex $\mathbf{s}_i \in \mathcal{S}_b$ in a scan boundary $\mathcal{S}_b$, we find its closest vertex in the corresponding template garment boundary, obtaining a matrix of scan points $\mathbf{q}_{1:C} = \{\mathbf{q}_1, \ldots, \mathbf{q}_C\}$ with corresponding template vertex indices $j_{1:C}$. Let $\mathbf{I}_{C \times m_g}$ be a selector matrix indicating the indices in the template corresponding to each $\mathbf{q}_i$. With this, we minimize the following least squares problem:

$$\begin{bmatrix} \mathbf{L}^g \\ w\mathbf{I}_{C \times m_g} \end{bmatrix} \mathbf{G}^g = \begin{bmatrix} \boldsymbol{\Delta}_{\text{init}} \\ w\mathbf{q}_{1:C} \end{bmatrix} \tag{4.6}$$

with respect to the template garment vertices $\mathbf{G}^g$, where the first block $\mathbf{L}^g\mathbf{G}^g = \boldsymbol{\Delta}_{\text{init}}$ forces the solution to keep the local surface structure, while the second block $w\mathbf{I}_{C \times m_g}\mathbf{G}^g = w\mathbf{q}_{1:C}$ makes the boundaries match. The nice property of the linear system solve is that the garment template globally stretches or compresses to match the scan garment boundaries, which would take many iterations of non-linear non-rigid registration (Pons-Moll et al., 2017) with the risk of converging to bad local minima. After this initialization, we non-linearly register each garment $\mathbf{G}^g$ to fit the scan surface. We build on top of the proposed multi-part registration in Pons-Moll et al. (2017) and propose additional loss terms on garment vertices, $v_k \in \mathbf{G}^g$, to facilitate better garment unposing, E_{unpose}, and minimize interpenetration, E_{interp}, with the underlying SMPL body surface, $\mathcal{S}$.

$$E_{\text{interp}} = \sum_g \sum_{\mathbf{v}_k \in \mathbf{G}^g} d(\mathbf{v}_k, \mathcal{S}) \tag{4.7}$$

$$d(\mathbf{x}, \mathcal{S}) = \begin{cases} 0, & \text{if } \mathbf{x} \text{ outside } \mathcal{S} \\ w * |\mathbf{x} - \mathbf{y}|_2, & \text{if } \mathbf{x} \text{ inside } \mathcal{S} \end{cases} \tag{4.8}$$

where w is a constant ($w = 25$ in our experiments), $\mathbf{v}_k$ is the k^{th} vertex of $\mathbf{G}^g$ and $\mathbf{y}$ is the point closest to $\mathbf{x}$ on $\mathcal{S}$.

Our garment formulation allows us to freely repose the garment vertices. We can use this to our advantage for applications such as animating clothed virtual avatars, garment re-targeting

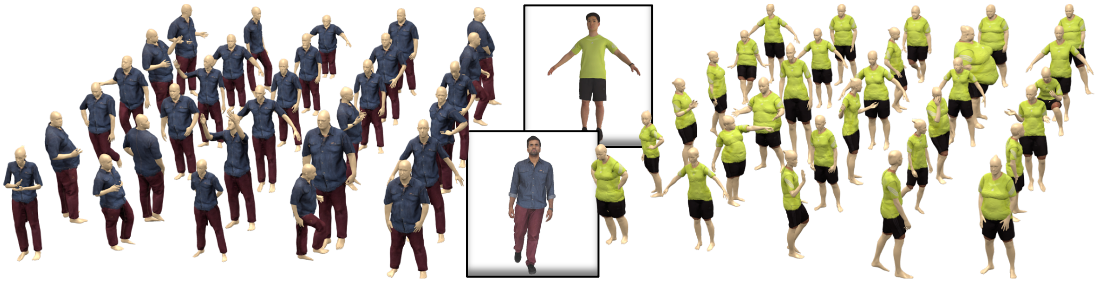


Figure 4.5: Dressing SMPL with just images. We use MGN to extract garments from the images of a source subject (middle) and use the inferred 3D garments to dress arbitrary human bodies in various poses from SMPL shape subjects. The two sets correspond to male (left) and female (right) body shapes respectively.

etc. However, posing is highly non-linear and can lead to undesired artefacts, specially when re-targeting garments across subjects with very different poses. Since we re-target the garments in unposed space, we reduce distortion by forcing distances from garment vertices to the body to be preserved after unposing:

$$E_{\text{unpose}} = \sum_{g} \sum_{\mathbf{v}_k \in \mathbf{G}^g} (d(\mathbf{v}_k, \mathcal{S}) - d(\mathbf{v}_k^0, \mathcal{S}^0))^2 \tag{4.9}$$

where $d(\mathbf{x}, \mathcal{S})$ is the L_2 distance between point $\mathbf{x}$ and surface $\mathcal{S}$. $\mathbf{v}_k^0$ and $\mathcal{S}^0$ denote garment vertex and body surface in unposed space, using Eq. 4.5 and 4.1 respectively.

DRESSING SMPL The SMPL model has proven very useful for modelling unclothed shapes. Our idea is to build a wardrobe of digital clothing compatible with SMPL to model clothed subjects. To this end we propose a simple extension that allows to *dress* SMPL. Given a garment $\mathbf{G}^g$, we use Eq. 4.3, 4.4, 4.5 to pose and skin the garment vertices. The dressed body including body shape (encoded as G_1) will be given by stacking the L individual garment vertices $[G_1(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}_1)^T, \ldots, G_L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}_L)^T]^T$. We define the function $C(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D})$ which returns the posed and shaped vertices for the skin, and each of the garments combined.
See Fig. 4.5 and supplementary for results on re-targeting garments using MGN across different SMPL bodies.

4.2.2 *From Images to Garments*

From registrations, we learn a shape space of garments, and generate a synthetic training dataset with pairs of images and body+3D garment pairs. From this data we train MGN:*Multi-Garment Net*, which maps images to 3D garments and body shape.

GARMENT SHAPE SPACE In order to factor out pose deformations from garment shape, we "unpose" the j^{th} garment registrations $\mathbf{G}_j^g \in \mathbb{R}^{m_g \times 3}$, similar to Pons-Moll et al. (2017) and Zhang et al. (2017a). Since the garments of each category are all in correspondence, we can easily compute PCA directly on the unposed vertices to obtain pose-invariant shape basis ($\mathbf{B}^g$). Using this, we encode a garment shape using 35 components $z^g \in \mathbb{R}^{35}$, plus a residual vector of offsets $\mathbf{D}_j^{\text{hf},g}$, mathematically: $\mathbf{G}_j^g = \mathbf{B}^g \mathbf{z}_j^g + \mathbf{D}_j^{\text{hf},g}$. From each scan, we also extract the body shape *under* clothing similarly as in Zhang et al. (2017a), which is essential to re-target a garment from one body to another.

MGN: MULTI-GARMENT NET The input to the model is a set of semantically segmented images, $\mathcal{I} = \{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_F - 1\}$, and corresponding $2D$ joint estimates, $\mathcal{J} = \{\mathbf{J}_0, \mathbf{J}_1, ..., \mathbf{J}_F - 1\}$, where F is the number of images used to make the prediction. Following Alldieck et al. (2019a) and Gong et al. (2018), we abstract away the appearance information in RGB images and extract semantic garment segmentation (Gong et al., 2018) to reduce the risk of over-fitting, albeit at the cost of disregarding useful shading signal. For simplicity, let now $\boldsymbol{\theta}$ denote both the joint angles $\boldsymbol{\theta}$ and translation $\boldsymbol{t}$.

The base network, f_w, maps the 2D poses $\mathcal{J}$, and image segmentations $\mathcal{I}$, to per frame latent code $(\boldsymbol{l}_{\mathcal{P}})$ corresponding to 3D poses

$$\boldsymbol{l}_{\mathcal{P}} = f_w^{\theta}(\mathcal{I}, \mathcal{J}), \tag{4.10}$$

and to a common latent code corresponding to body shape $(\boldsymbol{l}_{\beta})$ and garments $(\boldsymbol{l}_{\mathcal{G}})$ by averaging the per frame codes

$$\boldsymbol{l}_{\beta}, \boldsymbol{l}_{\mathcal{G}} = \frac{1}{F} \sum_{f=0}^{F-1} f_w^{\beta,\mathcal{G}}(\mathbf{I}_f, \mathbf{J}_f). \tag{4.11}$$

For each garment class, we train separate branches, $M_w^g(\cdot)$, to map the latent code $\boldsymbol{l}_{\mathcal{G}}$ to the un-posed garment $\mathbf{G}^g$, which itself is reconstructed from low-frequency PCA coefficients $\mathbf{z}^g$, plus $\mathbf{D}^{\text{hf},g}$ encoding high-frequency displacements

$$M_w^g(\boldsymbol{l}_{\mathcal{G}}, \mathbf{B}^g) = \mathbf{G}^g = \mathbf{B}^g \boldsymbol{z}^g + \mathbf{D}^{\text{hf},g}. \tag{4.12}$$

From the shape and pose latent codes $\boldsymbol{l}_{\beta}, \boldsymbol{l}_{\boldsymbol{\theta}}$, we predict body shape parameters $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$ respectively, using a fully connected layer. Using the predicted body shape $\boldsymbol{\beta}$ and geometry $M_w^g(\boldsymbol{l}_{\mathcal{G}}, \mathbf{B}^g)$ we compute displacements as in Eq. 4.3:

$$\mathbf{D}^g = M_w^g(\boldsymbol{l}_{\mathcal{G}}, \mathbf{B}^g) - \mathbf{I}^g T(\boldsymbol{\beta}, \mathbf{0}_{\backprime}, \mathbf{0}_{\mathbf{D}}). \tag{4.13}$$

Consequently, the final predicted $3D$ vertices posed for the f^{th} frame are obtained with $C(\boldsymbol{\beta}, \boldsymbol{\theta}_f, \mathbf{D})$, from which we render 2D segmentation masks

$$\mathbf{R}_f = R(C(\boldsymbol{\beta}, \boldsymbol{\theta}_f, \mathbf{D}), c), \tag{4.14}$$

where $R(\cdot)$ is a differentiable renderer (Henderson and Ferrari, 2018), $\mathbf{R}_f$ the rendered semantic segmentation image for frame f, and c denotes the camera parameters that are assumed fixed while the person moves. The rendering layer in Eq. (4.14) allows us to compare predictions against the input images. Since MGN predicts body and garments separately, we can predict a semantic segmentation image, leading to a more fine-grained $2D$ loss, which is not possible using a single mesh surface representation (Alldieck et al., 2019a). Note that Eq. 4.14 allows to train with self-supervision.

4.2.3 *Loss functions*

The proposed approach can be trained with $3D$ supervision on vertex coordinates, and with self supervision in the form of $2D$ segmented images. We use upper-hat for variables that are known and used for supervision during training. We use the following losses to train the network in an end to end fashion:

- $3D$ vertex loss in the canonical T-pose ($\boldsymbol{\theta} = \mathbf{0}_{\boldsymbol{\theta}}$):

$$\mathcal{L}^{3D}_{\mathbf{0}_{\boldsymbol{\theta}}} = ||C(\boldsymbol{\beta}, \mathbf{0}_{\boldsymbol{\theta}}, \mathbf{D}) - C(\hat{\boldsymbol{\beta}}, \mathbf{0}_{\boldsymbol{\theta}}, \hat{\mathbf{D}})||^2, \tag{4.15}$$

  where, $\mathbf{0}_{\boldsymbol{\theta}}$ represents zero-vector corresponding to zero pose.

- $3D$ vertex loss in posed space:

$$\mathcal{L}^{3D}_{\mathcal{P}} = \sum_{f=0}^{F-1} ||C(\boldsymbol{\beta}, \boldsymbol{\theta}_f, \mathbf{D}) - C(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}_f, \hat{\mathbf{D}})||^2 \tag{4.16}$$

- $2D$ segmentation loss: Unlike Alldieck et al. (2019a) we do not optimize silhouette overlap, instead we jointly optimize the projected per-garment segmentation against the input segmentation mask. This ensures that each garment explains its corresponding mask in the image:

$$\mathcal{L}^{2D}_{seg} = \sum_{f=0}^{F-1} ||\mathbf{R}_f - \mathbf{I}_f||^2, \tag{4.17}$$

- Intermediate losses: We further impose losses on intermediate pose, shape and garment parameter predictions: $\mathcal{L}_{\boldsymbol{\theta}} = \sum_{f=0}^{F-1} ||\hat{\boldsymbol{\theta}}_f - \boldsymbol{\theta}_f||^2, \mathcal{L}_{\boldsymbol{\beta}} = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2, \mathcal{L}_{\boldsymbol{z}} = \sum_{g=0}^{L-1} ||\hat{\boldsymbol{z}}^g - \boldsymbol{z}^g||^2$ where F, L are the number of images and garments respectively. $\hat{\boldsymbol{z}}$ are the ground truth PCA garment parameters. While such losses are a bit redundant, they stabilize learning.

4.2.4 *Implementation details*

BASE NETWORK (f^*_w): We use a CNN to map the input set $\{\mathcal{I}, \mathcal{J}\}$ to the body shape, pose and garment latent spaces. It consists of five, $2D$ convolutions followed by max-pooling layers. Translation invariance, unfortunately, renders CNNs unable to capture the location information of the features. In order to reproduce garment details in 3D, it is important to leverage 2D features as well as their location in the 2D image. To this end, we adopt a strategy similar to Liu et al. (2018), where we append the pixel coordinates to the output of every CNN layer. We split the last convolutional feature maps into three parts to individuate the body shape, pose and garment information. The three branches are flattened out and we append 2D joint estimates to the pose branch. Three fully connected layers and average pooling on garment and shape latent codes, generate $\boldsymbol{l}_{\boldsymbol{\beta}}, \boldsymbol{l}_{\boldsymbol{\theta}}$ and $\boldsymbol{l}_{\mathcal{G}}$ respectively. See supplementary for more details.

GARMENT NETWORK (M^g_w): We train separate garment networks for each of the garment classes. The garment network consists of two branches. The first predicts the overall mesh shape, and second one adds high frequency details. From the garment latent code ($\boldsymbol{l}_{\mathcal{G}}$), the first branch, consisting of two fully connected layers (sizes=1024, 128), regresses the PCA coefficients. Dot product of these coefficients with the PCA basis generates the base garment mesh. We use the second fully connected branch (size = m^g) to regress displacements on top of the mesh predicted in the first branch. We restrict these displacements to $\leq 1cm$ to ensure that overall shape is explained by the PCA mesh and not these displacements.

4.3 DATASET AND EXPERIMENTS

DATASET We use 356 3D scans of people with various body shapes, poses and in diverse clothing. We held out 70 scans for testing and use the rest for training. Similar to Alldieck et al.

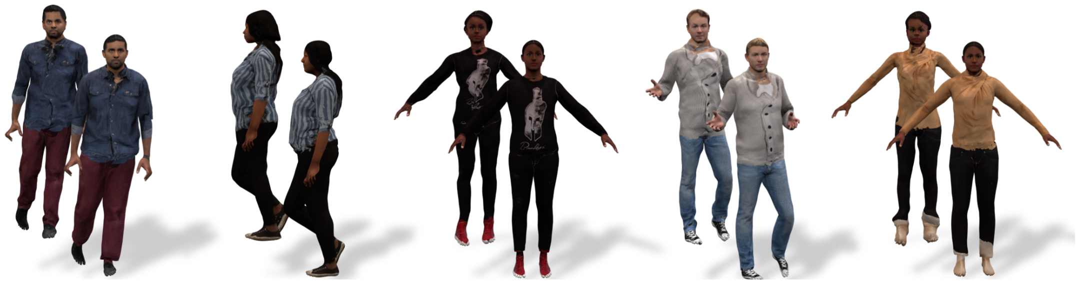


Figure 4.6: Qualitative comparison with Alldieck et al. (2019a). In each set we visualize 3D predictions from Alldieck et al. (2019a)(left) and our method (right) for five test subjects. Since our approach explicitly models garment geometry, it preserves more garment details, as is evident from minimal distortions across all the subjects. For more results see supplementary.

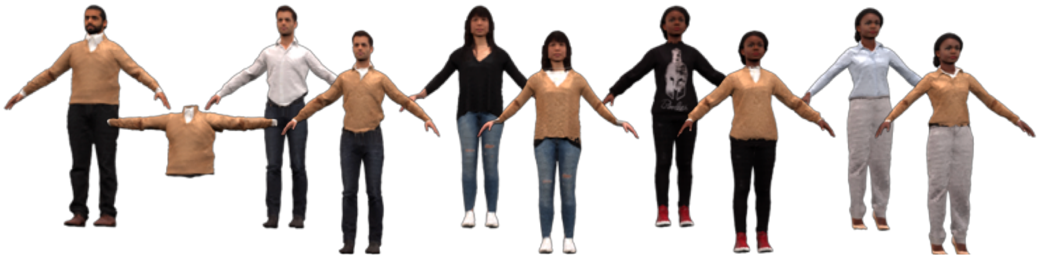


Figure 4.7: Texture transfer. We model each garment class as a mesh with fixed topology and surface parameterization. This enables us to transfer texture from any garment to any other registered instance of the same class. The first column shows the source garment mesh, while the subsequent images show original and transferred garment texture registrations.

(2019a, 2018b), we also restrict our setting to the scenario where the person is turning around in front of the camera. We register the scans using multi-mesh registration, SMPL+G. This enables further data augmentation since the registered scans can now be re-posed and re-shaped.
We adopt the data pre-processing steps from Alldieck et al. (2019a) including the rendering and segmentation. We also acknowledge the scale ambiguity primarily present between the object size and the distance to the camera. Hence we assume that the subjects in 3D have a fixed height and regress their distance from the camera. Same as Alldieck et al. (2019a), we also ignore the effect of camera intrinsics.

4.3.1 *Experiments*

In this section we discuss the merits of our approach both qualitatively and quantitatively. We also show real world applications in the form of texture transfer (Fig. 4.7), where we maintain the original geometry of the source garment but map novel texture. We also show garment re-targeting from images using MGN in Fig. 4.8.

QUALITATIVE COMPARISONS: We compare our method against Alldieck et al. (2019a) on our scan dataset. For fair comparison we re-train the models proposed by Alldieck et al. (2019a) on our dataset and compare against our approach (Dataset used by Alldieck et al. (2019a) is not publicly available). Figure 4.6 indicates the advantage of incorporating the garment model in structured prediction over simply modelling free form displacements. Explicit garment modelling allows us to predict sharper garment boundaries and minimize distortions (see Fig. 4.6). More examples are shown in the supplementary material.

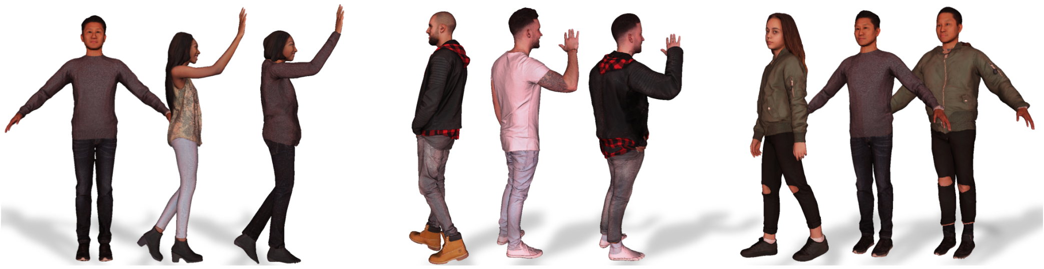


Figure 4.8: Garment re-targeting by MGN using 8 RGB images. In each of the three sets we show the source subject, target subject and re-targeted garments. Using MGN, we can re-target garments including both texture and geometry.

QUANTITATIVE COMPARISON: In this experiment we do a quantitative analysis of our approach against the state of the art $3D$ prediction method, Alldieck et al. (2019a). We compute a symmetric error between the predicted and GT garment surfaces similar to Alldieck et al. (2019a). We report per-garment error, E^g (supplementary), and overall error, i.e. mean of E^g over all the garments

$$E^g = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|\hat{\mathbf{S}}_i^g|}\sum_{\mathbf{v}_k\in\hat{\mathbf{S}}_i^g} d(\mathbf{v}_k,\mathcal{S}_i^g) + \frac{1}{|\mathbf{S}_i^g|}\sum_{\mathbf{v}_k\in\mathbf{S}_i^g} d(\mathbf{v}_k,\hat{\mathcal{S}}_i^g)\right), \tag{4.18}$$

where N is the number of meshes with garment g. $\mathbf{S}_i^g$ and $\mathcal{S}_i^g$ denote the set of vertices and the surface of the i^{th} predicted mesh respectively, belonging to garment g. Operator $(\hat{\cdot})$ denotes GT values. $d(\mathbf{v}_k, \mathcal{S})$ computes the L_2 distance between the vertex $\mathbf{v}_k$ and surface $\mathcal{S}$.

This criterion is slightly different than Alldieck et al. (2019a) because we do not evaluate error on the skin parts. We reconstruct the 3D garments with mean vertex-to-surface error of 5.78 mm with 8 frames as input. We re-train Alldieck et al. (2019a) on our dataset and the resulting error is 5.72mm.

We acknowledge the slightly better performance of Alldieck et al. (2019a) and attribute it to the fact that the single mesh based approaches do not bind vertices to semantic roles, i.e these approaches can pull vertices from any part of the mesh to explain $3D$ deformations where as our approach ensures that only semantically correct vertices explain the $3D$ shape.

It is also worth noting that MGN predicts garments as linear function (PCA coefficients) of latent code, whereas Alldieck et al. (2019a) deploys GraphCNN. PCA based formulation though easily tractable is inherently biased towards smooth results. Our work paves the way for further exploration into building garment models for modelling the variations in garment geometry over a fixed topology.

We report the results for using varying number of frames in the supplementary.

GT VS PREDICTED POSE: The $3D$ vertex predictions are a function of pose and shape. In this experiment we do an ablation study to isolate the effect of errors in pose estimation on vertex predictions. This experiment is important to better understand the strengths and weaknesses of the proposed approach in shape estimation by marginalizing over the errors due to pose fitting. We study two scenarios, first where we predict the 3D pose and second, where we have access to GT pose. We report mean vertex-to-surface error of 5.78mm with GT poses and 11.90mm with our predicted poses.

4.3.2 *Re-targeting*

Our multi-mesh representation essentially decouples the underlying body and the garments. This opens up an interesting possibility to take garments from source subject and virtually dress a novel subject. Since the source and the target subjects could be in different poses, we first unpose the source body and garments along with the target body. We drop the $(.)^0$ notation for the unposed space in the following section for clarity. Below we propose and compare two garment re-targeting approaches. After re-targeting the target body and re-targeted garments are re-posed to their original poses.

NAIVE RE-TARGETING: The simplest approach to re-target clothes from source to target is to extract the garment offsets, $\mathbf{D}^{s,g}$ from the source subject using Eq. 4.13 and dress a target subject using Eq. 4.5.

BODY AWARE RE-TARGETING: The naive approach is problematic because it relies on non-local pre-set vertex association between the garment and the body (I^g). This results in inaccurate association between the body blend shapes, $B_{p,s}$ and the garment vertices. This eventually leads to incorrect estimation of source offsets, $\mathbf{D}^{s,g}$ and in turn leads to higher inter-penetrations between the re-targeted garment and the body (see supplementary). In order to mitigate this issue, we compute the new k^{th} target garment vertex location, $\mathbf{v}^t_k$ as follows

$$\mathbf{v}^t_k = \mathbf{v}^s_k - \mathbf{S}^s_{\mathrm{I}_k} + \mathbf{S}^t_{\mathrm{I}_k} \tag{4.19}$$

$$\mathrm{I}_k = \underset{\mathrm{I}\in[0,\,|\mathbf{S}^s|-1]}{\operatorname{argmin}} \; ||\mathbf{v}^s_k - \mathbf{S}^s_{\mathrm{I}}||_2, \tag{4.20}$$

where $\mathbf{v}^s_k$ is the source garment vertex, $\mathbf{S}^s_{\mathrm{I}_k}$ is the vertex (indexed by I_k) among the source body vertices, $\mathbf{S}^s$, closest to $\mathbf{v}^s_k$ and $\mathbf{S}^t_{\mathrm{I}_k}$ is the corresponding vertex among the target body vertices.

MGN allows us to predict separable body shape and garments in 3D, allowing us to do garment re-targeting (as described above) using just images. To the best of our knowledge this is the first method to do so. See Fig. 4.8 for results on garment re-targeting by MGN. See supplementary for more results.

4.4 LIMITATIONS AND FUTURE WORKS

In this section (and Fig. 4.9) we discuss some of the research avenues that our approach opens up or shows unsatisfactory performance. We hope that this would simulate further research into the direction of modelling 3D garments, underlying body and their interactions.

- The proposed approach does not deal with pose dependent deformations.
- Skinning garments though convenient, often leads to artifacts while re-posing in case of extreme poses (Fig 4.9 a).
- Re-targeting relies heavily on segmentation. In case we wrongly segment part of skin as garment our approach incorrectly moves the skin along with the garment. (Fig 4.9 b)
- Current approach cannot impaint the skin texture underneath the garments. This creates artifacts when re-targeting short garments (eg: t-shirt) on a body which was previously wearing long garments (eg: coat) (Fig 4.9 c)
- In its current form MGN does not model hair.

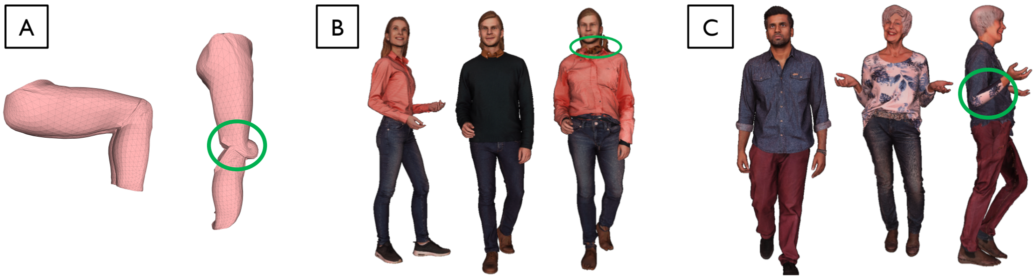


Figure 4.9: Ours is the first approach to infer separable 3D garments from images. Though very promising, the proposed approach has shortcomings. In this figure we present some interesting challenges for future work. From left to right: A) Unposing artifacts due to skinning, B) part of source hair got moved along with the garments due to incorrect segmentation at the boundary, C) Current approach cannot impaint texture under clothing.

4.5 CONCLUSIONS

We introduce MGN, the first model capable of jointly reconstructing from few images, body shape and garment geometry as layered meshes. Experiments demonstrate that this representation has several benefits: it is closer to how clothing layers on top of the body in the real world, which allows control such as re-dressing novel shapes with the reconstructed clothing. Additionally, we introduce for the first time, a dataset of registered *real* garments from real scans obtained with a robust registration pipeline. When compared to more classical single mesh representations, it allows more control and qualitatively the results are very similar. In summary, we think that MGN provides a first step in a promising research direction. The MGN model and the digital wardrobe are publicly available to stimulate research in this direction.

5

COMBINING IMPLICIT FUNCTION LEARNING AND PARAMETRIC MODELS FOR 3D HUMAN RECONSTRUCTION

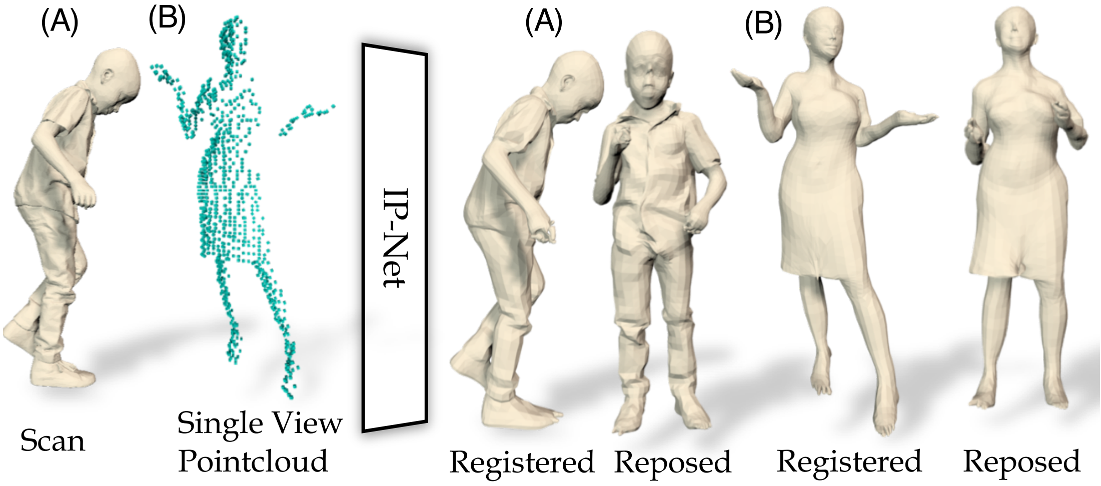


Figure 5.1: We combine implicit functions and parametric modeling for detailed and controllable reconstructions from sparse point clouds. IP-Net predictions can be registered with SMPL+D model for control. IP-Net can also register (A) 3D scans and (B) single view point clouds.

Implicit functions represented as deep learning approximations are powerful for reconstructing 3D surfaces. However, they can only produce static surfaces that are not controllable, which provides limited ability to modify the resulting model by editing its pose or shape parameters. Nevertheless, such features are essential in building flexible models for both computer graphics and computer vision. In this work, we present methodology that combines detail-rich implicit functions and parametric representations in order to reconstruct 3D models of people that remain controllable and accurate even in the presence of clothing. Given sparse 3D point clouds sampled on the surface of a dressed person, we use an Implicit Part Network (IP-Net) to jointly predict the *outer* 3D surface of the dressed person, the *inner* body surface, and the semantic correspondences to a parametric body model. We subsequently use correspondences to fit the body model to our inner surface and then non-rigidly deform it (under a parametric body + displacement model) to the outer surface in order to capture garment, face and hair detail. In quantitative and qualitative experiments with both full body data and hand scans we show that the proposed methodology generalizes, and is effective even given incomplete point clouds collected from single-view depth images. Our models and code are available at: http://virtualhumans.mpi-inf.mpg.de/ipnet.

5.1 INTRODUCTION

The sensing technology for capturing unstructured 3D point clouds is becoming ubiquitous and more accurate, thus opening avenues for extracting detailed models from point cloud data. This is important in many 3D applications such as shape analysis and retrieval, 3D content generation,

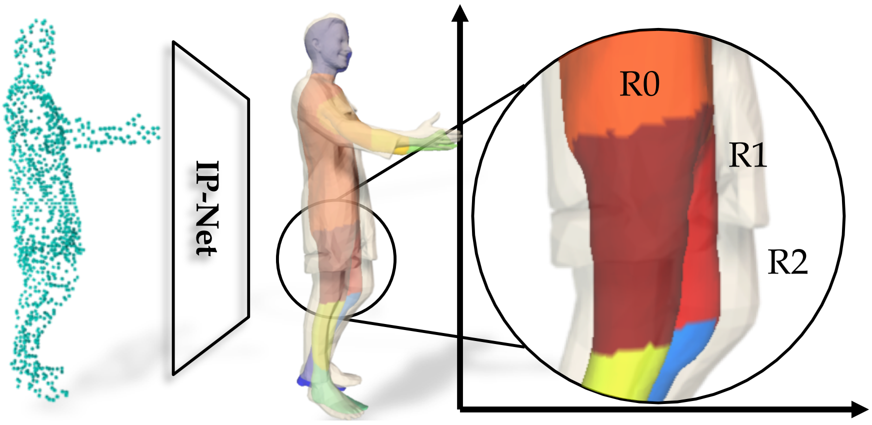


Figure 5.2: Unlike typical implicit reconstruction methods, IP-Net predicts a double layered surface, classifying the points as lying inside the body (R0), between the body and the clothing (R1) and outside the clothing (R2). IP-Net also predicts part correspondences to the SMPL model.

3D human reconstruction from depth data, as well as mesh registration, which is the workhorse of building statistical shape models (Joo et al., 2018b; Loper et al., 2015; Xu et al., 2020).

Therefore obtaining complete and accurate 3D representations of the surface geometry from such data, and processing those in order to obtain semantic information, such as object or human body parts, or clothing, is an increasingly important problem. The problem is extremely challenging as the body can be occluded by clothing, hence identifying body parts given a point cloud is often ambiguous, and reasoning-with (or filling-in) missing data often requires non-local analysis. In this paper, we focus on the reconstruction of human models from sparse or incomplete point clouds, as captured by body scanners or depth cameras. In particular, we focus on extracting detailed 3D representations, including models of the underlying body shape and clothing, in order to make it possible to seamlessly re-pose and re-shape (*control*) the resulting dressed human models. To avoid ambiguity, we refer to static implicit reconstructions as *reconstruction* and our controllable model fit as *registration*. Note that the *registration* involves both reconstruction (explaining the given point cloud geometry) and registration, as it is obtained by deforming a predefined model.

Learning-based methods are well suited to process sparse or incomplete point clouds, as they can leverage prior data to fill in the missing information in the input, but the choice of output representation limits either the resolution, when working with voxels or meshes, or the surface control, for implicit shape representations (Chen and Zhang, 2019; Chibane et al., 2020a; Mescheder et al., 2019; Park et al., 2019).

The main limitation of learning an implicit function is that the output is “just” a static surface with *no explicit model to control its pose and shape*. In contrast, parametric body models, such as SMPL (Loper et al., 2015) allow control, but the resulting meshes are overly-smooth and accurately regressing parameters directly from a point cloud is difficult (see Table 5.1). Furthermore, the surface of SMPL can not represent clothing, which makes registration difficult. Non-rigidly registering a less constrained parametric model to point clouds using non-linear optimization is possible, but only yields good results when provided with very good initialization close to the data (without local assignment ambiguity) and the point cloud is reasonably complete (see Table 5.1 and Fig. 5.5).

The main idea in this work is to take advantage of the best of both representations (implicit and parametric), and learn to predict body under clothing (including body part labels) in order to make subsequent optimization-based registration feasible. Specifically, we introduce a novel architecture which jointly learns 2 implicit functions for (i) the joint occupancy of the outer (body+clothing) and the inner (body) surfaces and (ii) body part labels. Following recent work Chibane et al., 2020a, we compute a 3-dimensional multi-scale tensor of deep features from the input point cloud, and make predictions at continuous query points. Unlike recent work that only predicts the occupancy of a single surface (Chen and Zhang, 2019; Chibane et al., 2020a; Mescheder et al., 2019; Park et al., 2019), we jointly learn a continuous implicit function for the inner/outer surface prediction and another classifier for body part label prediction. Our key insight is that since the inner surface (body) can be well approximated by a parametric body model (SMPL), and the predicted body parts constrain the space of possible correspondences, fitting SMPL to the predicted inner surface is very robust. Starting from SMPL fitted to the inner surface, we register it to the outer surface (under an additional displacement model, SMLP+D Alldieck et al., 2019a; Lazova et al., 2019), which in turn allows us to *re-pose and re-shape* the implicitly reconstructed outer surface.

Our experiments show that our implicit network can accurately predict body shape under clothing, the outer surface, and part labels, which makes subsequent parametric model fitting robust. Results on the Renderpeople dataset [1] demonstrate that our tandem of implicit function and parametric fitting yields detailed outer reconstructions, which are controllable, along with an estimation of body shape under clothing. We further achieve comparable performance on body shape under clothing on the BUFF dataset (Zhang et al., 2017b) without training on BUFF and without using temporal information. To show that our model can be useful in other domains, we train it on the MANO dataset (Romero et al., 2017) and show accurate registration using sparse and single view point clouds. Our key contributions can be summarized as follows:

- We propose a unified formulation which combines implicit functions and parametric modelling to obtain high quality controllable reconstructions from partial/ sparse/ dense point clouds of articulated dressed humans.
- Ours is the first approach to jointly reconstruct body shape under clothing along with full dressed reconstruction using a double surface implicit function, in addition to predicting part correspondences to a parametric model.
- Results on a dataset of articulated clothed humans and hands (MANO, Romero et al. (2017)) show the wide applicability of our approach.

5.2 IMPLICIT PART NETWORK: METHOD

We introduce IP-Net, a network to generate detailed 3D reconstruction from an unordered sparse point cloud. IP-Net can additionally infer body shape under clothing and the body parts of the SMPL model. Training IP-Net requires supervision on three fronts, i) an outer dressed surface occupancy–directly derived from 3D scans, ii) an inner body surface–we supervise with an optimization based body shape under clothing registration approach and iii) correspondences to the SMPL model–obtained by registering SMPL to scans using custom optimization.

5.2.1 *Training Data Preparation*

To generate training data, we require non-rigidly registering SMPL (Alldieck et al., 2019a; Lazova et al., 2019) to 3D scans and estimating body shape under clothing, which is extremely

[1] https://renderpeople.com

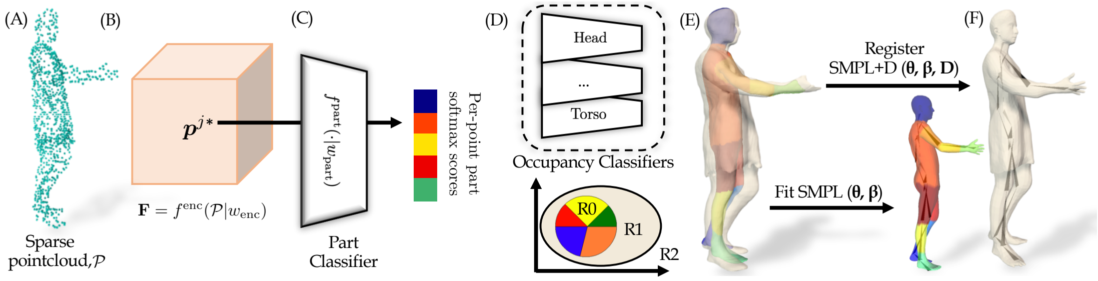

(A)
(B)
(C)
(D)
(E)
(F)
Sparse pointcloud, $\mathcal{P}$
p^{j*}
$\mathbf{F} = f^{\text{enc}}(\mathcal{P}|w_{\text{enc}})$
$f^{\text{part}}(\cdot|w_{\text{part}})$
Part Classifier
Per-point part softmax scores
Head
...
Torso
Occupancy Classifiers
R0
R1
R2
Register SMPL+D (θ, β, D)
Fit SMPL (θ, β)

Figure 5.3: The input to our method is (A) sparse point cloud $\mathcal{P}$. IP-Net encoder $f^{\text{enc}(\cdot)}$ generates an (B) implicit representation of $\mathcal{P}$. IP-Net predicts, for each query point $\boldsymbol{p}^j$, its (C) part label and double layered occupancy. IP-Net uses (D) occupancy classifiers to classify the points as lying inside the body (R0), between the body and the clothing (R1) and outside the body (R2), hence predicting (E) full 3D shape $\mathcal{S}_o$, body shape under clothing $\mathcal{S}_{in}$ and part labels. We register IP-Net predictions with (F) SMPL+D model to make implicit reconstruction controllable for the first time.

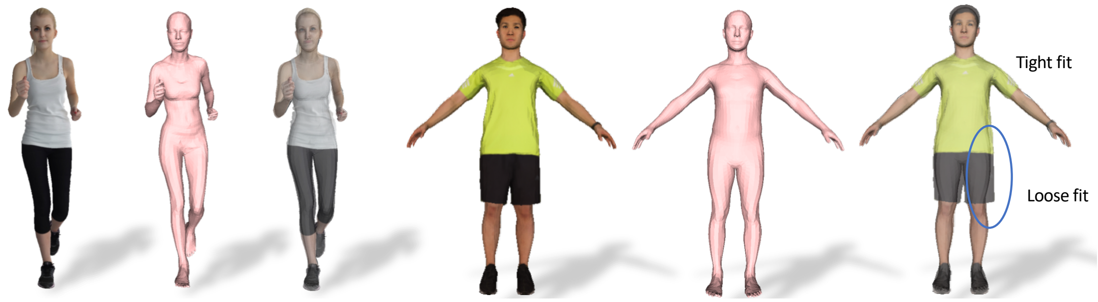


Figure 5.4: To train IP-Net we require estimating body shape under clothing from a dressed scan. We propose an optimization based approach to obtain body shape under clothing from dressed 3D scans. We show (L to R) input scan, estimated body shape, estimated body overlayed with scan.

challenging for the difficult poses in our dataset. Consequently, we first render the scans in multiple views, detect keypoints and joints, and integrate these as viewpoint landmark constraints to regularize registration similarly as in Alldieck et al. (2019a) and Lazova et al. (2019). To non-rigidly deform SMPL to scans, we leverage SMPL+D (Alldieck et al., 2019a; Lazova et al., 2019), which is an extension to SMPL that adds per-vertex free-form displacements on top of SMPL to model deformations due to garments and hair. Once SMPL+D has been registered to the scans, we transfer body part labels from the SMPL model to the scans.
Next we look at obtaining body shape under clothing which is required to train the body surface prediction in IP-Net.

BODY SHAPE UNDER CLOTHING REGISTRATION. For the body shape under clothing, we build on top of Zhang et al. (2017b) and propose a similar optimization based approach integrating viewpoint landmarks.

Similar to Zhang et al. (2017b), we model the inner body surface $\mathcal{B}$ using a modified SMPL, $M(\cdot)$, which uses pose($\boldsymbol{\theta}$), shape($\boldsymbol{\beta}$) and translation($\boldsymbol{t}$) parameters to model undressed humans in 3D.

$$B(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{t}) = W(T(\boldsymbol{\beta},\boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) + \boldsymbol{t} \tag{5.1}$$

$$T(\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}), \tag{5.2}$$

where $\mathbf{T}$ is a base template mesh with 6890 vertices in a canonical T-pose. $B_p(\cdot)$ represents the pose dependent deformations of a skeleton $J(\boldsymbol{\beta})$. $B_s(\cdot)$ represents the shape dependent deformations. The model is skinned, $W(\cdot)$, with blend weights, $\mathbf{W}$.

We further make the template $\mathbf{T}$ optimizable to model surface variations outside the PCA shape space of the SMPL model. We incorporate translation, $\boldsymbol{t}$ in pose parameters, $\boldsymbol{\theta}$ for brevity in further notation.

$$\mathcal{B} = B(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{5.3}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}). \tag{5.4}$$

We first segment garment and skin parts on the scans using the approach proposed by Bhatnagar et al. (2019) and initialize the pose and shape parameters using registration proposed in Alldieck et al. (2019a) and Lazova et al. (2019). We use a similar objective E_{skin} (Eq. 3 in Zhang et al. (2017b)) to register the visible skin parts on the scans. To register skin parts underneath the garments we make slight modifications to the E_{cloth} term in Eq. 4 from Zhang et al. (2017b) by replacing the Geman-McClure cost function by a hinge cost and also add a geodesic term to force smoothness near the garment boundaries. The objective can be formally written as follows:

$$E_{cloth} = \sum_{\boldsymbol{v}_i \in \mathcal{S}} g_i * (1 - l_i) * (H(d_1(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i), c) + d_2(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i)) \tag{5.5}$$

$$H(x, c) = \begin{cases} x & \text{if } x < c \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

$$d_1(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i) = \begin{cases} d(\boldsymbol{v}_i, \mathcal{B}) & \text{if } \boldsymbol{v}_i \text{ is outside } \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

$$d_2(\boldsymbol{v}_i, \hat{\boldsymbol{p}}_i) = \begin{cases} w * d(\boldsymbol{v}_i, \mathcal{B}) & \text{if } \boldsymbol{v}_i \text{ is inside } \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

where $H(\cdot)$ acts as a hinge for loose clothing, $d_2(\cdot)$ and $d_1(\cdot)$ are the scaled distance functions to ensure that 'body is brought close to the garment surface' and 'body should not intersect the garment surface' respectively. l_i and g_i are the skin identifier label and normalised geodesic cost respectively. We use $w = 20$ and $c = 0.01$.

We additionally enforce facial landmark matching to register better facial details. To get 3D facial landmarks for a scan we render it from multiple viewpoints and run OpenPose [2] to get 2D facial landmarks on images. We then solve graphcut to lift the multi-view landmarks to 3D (this is similar to what Bhatnagar et al. (2019) use for lifting 2D segmentation to scans). We use the following objective to match facial landmarks between the scan and the body.

$$E_{face} = |\mathbf{L} - \mathbf{L}^{\text{face}} \cdot \mathcal{B}|_2, \tag{5.9}$$

where $\mathbf{L}$, $\mathbf{L}^{\text{face}}$ are facial landmarks on scan and SMPL facial landmark regressor respectively.

In order to ensure that $\mathcal{B}$ is smooth and retains human body like appearance we add the following regularization term. For the skin vertices it is important to ensure that the surface near

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose

the garment boundary is tightly coupled to the underlying body where as vertices away from the boundary can deform to explain hair, hands etc.

$$\begin{aligned} E_{lap} = \sum_{\boldsymbol{v}_i \in \mathcal{B}} \Big\{ &(1 - l_i) * |L_i(\boldsymbol{v}_i^{init}) - L_i(\boldsymbol{v}_i)|_2 + \\ &l_i * (1 - g_i) * |L_i(\boldsymbol{v}_i^{init}) - L_i(\boldsymbol{v}_i)|_2 \Big\}. \end{aligned} \tag{5.10}$$

Here L_i is the laplacian operator at vertex $\boldsymbol{v}_i$. l_i and g_i are the skin label and normalised geodesic cost respectively.

OVERALL OBJECTIVE: We jointly optimise the SMPL parameters ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$) and the template $\mathbf{T}$, to minimise the objectives described above.

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}) = w_{skin}E_{skin} + w_{face}E_{face} + w_{cloth}E_{cloth} + w_{lap}E_{lap}, \tag{5.11}$$

where w are the weights associated with the corresponding objectives. We found scheduling of weights important for a smooth registration process.

$$\begin{aligned} w_{\{skin/cloth/face\}} &= c_{\{skin/cloth/face\}} * k, \\ w_{lap} &= c_{lap}/k \end{aligned} \tag{5.12}$$

In our experiments we keep c_{skin}, c_{face}, c_{garm} and c_{lap} as 5, 1, 5 and 100 respectively. k denotes optimization iteration. Qualitative results obtained from our body shape under clothing registration are shown in Fig. 5.4

This process to generate training data is fairly robust, but required a lot of engineering to make it work. It also requires rendering multiple views of the scan, and does not work for sparse point clouds or scans without texture.

One of the key contributions of this work is to replace this tedious process with IP-Net, which quickly predicts a double layer implicit surface for body and outer surface, and body part labels to make subsequent registration using SMPL+D easy.

We describe our network IP-Net, that infers detailed geometry and SMPL body parts from sparse point clouds next.

5.2.2 *IP-Net: Overview*

IP-Net $f(\cdot|w)$ takes in as input a sparse point cloud, $\mathcal{P}$ ($\sim$5k points), from articulated humans in diverse shapes, poses and clothing. IP-Net learns an implicit function to jointly infer outer surface, $\mathcal{S}_o$ (corresponding to full dressed 3D shape) and the inner surface $\mathcal{S}_{in}$ (corresponding to underlying body shape), of the person. Since we intend to register SMPL model to our implicit predictions, IP-Net additionally predicts, for each query point $\boldsymbol{p}^j \in \mathbb{R}^3$, the SMPL body part label $I^j \in \{0, \ldots, N-1\}$ (N=14) . We define I^j as a label denoting the associated body part on the SMPL mesh.

IP-NET: FEATURE ENCODING. Recently, IF-Net, Chibane et al. (2020a) achieved SOTA 3D mesh reconstruction from sparse point clouds. Their success can be attributed to two key insights: using a multi-scale, grid of deep features to represent shape, and predicting occupancy using features extracted at continuous point locations, instead of using the point coordinates. We build our IP-Net encoder $f^{\text{enc}}(\cdot|w_{\text{enc}})$ in the spirit of IF-Net encoder. We denote our multi-scale grid-aligned feature representation as $\mathbf{F} = f^{\text{enc}}(\mathcal{P}|w_{\text{enc}})$ and the features at point $\boldsymbol{p}^j = (x, y, z)$ as $\mathbf{F}^j = \mathbf{F}(x, y, z)$.

IP-NET: PART CLASSIFICATION. Next, we train a multi-class classifier $f^{\text{part}}(\cdot|w_{\text{part}})$ that predicts, for each point $\boldsymbol{p}^j$, its part label (correspondence to nearest SMPL part) conditioned on its feature encoding. More specifically, $f^{\text{part}}(\cdot|w_{\text{part}})$ predicts a per part score vector $\boldsymbol{D}^j \in [0,1]^N$ at every point $\boldsymbol{p}^j$

$$\boldsymbol{D}^j = f^{\text{part}}(\mathbf{F}^j|w_{\text{part}}). \tag{5.13}$$

Then, we classify a point with the part label of maximum score

$$I^j = \underset{I\in\{0,\ldots,N-1\}}{\arg\max} \ (\boldsymbol{D}^j_I). \tag{5.14}$$

IP-NET: OCCUPANCY PREDICTION. Previous implicit formulations (Chibane et al., 2020a; Mescheder et al., 2019; Park et al., 2019; Saito et al., 2019) train a deep neural network to classify points as being inside or outside a *single* surface. In addition, they minimize a classification/ regression loss over sampled points, which biases the network to perform better for parts with large surface area (more points) over smaller regions like hands (less points).

The key distinction between IP-Net and previous implicit approaches is that it classifies points as belonging to 3 different regions: 0-inside the body, 1-between body and clothing and 2-outside. This allows us to recover two surfaces (inner $\mathcal{S}_{in}$ and outer $\mathcal{S}_o$), see Fig. 7.1 and 5.3. Furthermore, we use an ensemble of occupancy classifiers $\{f^I(\cdot|w_I)\}_{I=0}^{N-1}$, where each $f^I(\cdot|w_I) : \mathbf{F}^j \mapsto \mathbf{o}^j \in [0,1]^3$ is trained to classify a point $\boldsymbol{p}^j$ with features $\mathbf{F}^j$ into the three regions $o^j \in \{0,1,2\}$, $o_j = \arg\max_i \mathbf{o}^j_i$. The idea here is to train the ensemble such that $f^I(\cdot|w_I)$ performs best for part I, and predict the final occupancy o^j as a sum weighted by the part classification scores $\mathbf{D}^j_I \in \mathbb{R}$ at point $\boldsymbol{p}^j$

$$o^j = \arg\max_i \mathbf{o}^j_i, \qquad \mathbf{o}^j = \sum_{I=0}^{N-1} \boldsymbol{D}^j_I \cdot f^I(\mathbf{F}^j|w_I), \tag{5.15}$$

thereby reducing the bias towards larger body parts. After dividing the space in 3 regions the double-layer surface is extracted from the two decision boundaries.

IP-NET: LOSSES IP-Net is trained using categorical cross entropy loss for both part-prediction (f^{part}) and occupancy prediction ($\{f^I\}_{I=0}^{N-1}$).

IP-NET: SURFACE GENERATION We use marching cubes (Lorensen and Cline, 1987) on our predicted occupancies to generate a triangulated mesh surface.

5.2.3 *Registering SMPL to IP-Net Predictions*

Implicit based approaches can generate details at arbitrary resolutions but reconstructions are static and not controllable. This makes these approaches unsuitable for re-shaping and re-posing. We propose the first approach to combine implicit reconstruction with parametric modelling which lifts the details from the implicit reconstruction onto the SMPL+D model (Alldieck et al., 2019a; Lazova et al., 2019) to obtain an editable surface. We describe our registration using IP-Net predictions next. We use SMPL to denote the parametric model constrained to undressed shapes, and SMPL+D (SMPL plus displacements) to represent details like clothing and hair.

FIT SMPL TO IMPLICIT BODY: We first optimize the SMPL shape, pose and translation parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t})$ to fit our inner surface prediction $\mathcal{S}_{in}$.

$$E_{\text{data}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t}) = \frac{1}{|\mathcal{S}_{in}|} \sum_{\boldsymbol{v}_i \in \mathcal{S}_{in}} d(\boldsymbol{v}_i, \mathcal{M}) + w \cdot \frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{v}_j \in \mathcal{M}} d(\boldsymbol{v}_j, \mathcal{S}_{in}), \tag{5.16}$$

where $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ denote vertices on $\mathcal{S}_{in}$ and SMPL surface $\mathcal{M}$ respectively. $d(\boldsymbol{p}, \mathcal{S})$ computes the distance of point $\boldsymbol{p}$ to surface $\mathcal{S}$. In our experiments we set $w = 0.1$

Additionally, we use the part labels predicted by IP-Net to ensure that correct parts on the SMPL mesh explain the corresponding regions on the inner surface $\mathcal{S}_{in}$. This term is critical to ensure correct registration (see Table 5.2 and Fig. 5.6)

$$E_{\text{part}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t}) = \frac{1}{|\mathcal{S}_{in}|} \sum_{I=0}^{N-1} \sum_{\boldsymbol{v}_i \in \mathcal{S}_{in}} d(\boldsymbol{v}_i, \mathcal{M}^I) \delta(I^i = I), \tag{5.17}$$

where $\mathcal{M}^I$ denotes the surface of the SMPL mesh corresponding to part I and I^i denotes the predicted part label of vertex $\boldsymbol{v}_i$. The final objective can be written as follows

$$E(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t}) = w_{\text{data}} E_{\text{data}} + w_{\text{part}} E_{\text{part}} + w_{\text{lap}} E_{\text{lap}}, \tag{5.18}$$

where E_{lap} denotes a Laplacian regularizer. In our experiments we set the balancing weights $w_{\text{data/part/lap}}$ to 100, 10 and 1 respectively based on experimentation.

REGISTER SMPL+D TO FULL IMPLICIT RECONSTRUCTION: Once we obtain the SMPL body parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t})$ from the above optimization, we jointly optimize the per-vertex displacements $\mathbf{D}$ to fit the outer implicit reconstruction $\mathcal{S}_o$.

$$E_{\text{data}}(\mathbf{D}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{t}) = \frac{1}{|\mathcal{S}_o|} \sum_{\boldsymbol{v}_i \in \mathcal{S}_o} d(\boldsymbol{v}_i, \mathcal{M}) + w \cdot \frac{1}{|\mathcal{M}|} \sum_{\boldsymbol{v}_j \in \mathcal{M}} d(\boldsymbol{v}_j, \mathcal{S}_o) \tag{5.19}$$

5.2.4 *Implementation details*

The input to IP-Net is a 3D voxel grid obtained by voxelizing the sparse input point cloud into a 128x128x128 grid. IP-Net encoder $f^{\text{enc}}(\cdot|w_{\text{enc}})$, consists of 3x$\{$Conv3D, Conv3D+stride$\}$ layers. IP-Net part predictor $f^{\text{part}}(\cdot|w_{\text{part}})$ and IP-Net part-conditioned classifiers $\{f^p(\cdot|w_p)\}_{p=0}^{N-1}$, each consist of 2x$\{FC\}$ layers. All except the final layer of IP-Net have Relu activation. We use categorical cross-entropy losses with Adam optimizer for training.

5.3 DATASET AND EXPERIMENTS

5.3.1 *Dataset*

We train IP-Net on a dataset of 700 scans from Twindon [3] and Treedys [4] and test on held out 50 scans from RenderPeople [5]. We normalize our scans to a bounding box of size 1.6m. To train IP-Net we need paired data of sparse point clouds (input) and the corresponding outer

[3] https://web.twindom.com
[4] https://www.treedys.com
[5] https://renderpeople.com

surface, inner surface and correspondence to SMPL model (output). We generate the sparse point clouds by randomly sampling 5k points on our scans, which we voxelize into a grid of size 128x128x128 for our input. We use the normalized scans directly as our ground truth dressed meshes and use our method for body shape registration under scan to get the corresponding body mesh $\mathcal{B}$. For SMPL part correspondences, we manually define 14 parts (left/right forearm, left/right mid-arm, left/right upper-arm, left/right upper leg, left/right mid leg, left/right foot, torso and head) on SMPL mesh and use the fact that our body mesh $\mathcal{B}$, is a template with SMPL-topology registered to the scan; this automatically annotates $\mathcal{B}$ with the part labels. The part label of each query point in $\mathbb{R}^3$, is the label of the nearest vertex on the corresponding body mesh $\mathcal{B}$. Note that part annotations do not require manual effort.

We evaluate the implicit outer surface reconstructions against the GT scans. We use the optimization based approach described in Sec. 5.2.1 to obtain ground truth registrations.

Register SMPL+D	**Outer reg.**	**Inner reg.**
(a) Sparse point cloud	14.85	NP*
(b) IF-Net	13.88	NP*
(c) Regress SMPL+D parameters	32.45	NP*
(d) **IP-Net (Ours)**	**3.67**	**3.32**

Table 5.1: IP-Net predictions, i.e. the outer/ inner surface and correspondences to SMPL are key to high quality SMPL+D registration. We compare the quality (vertex-to-vertex error in cm) of registering to (a) point cloud, (b) implicit reconstruction by Chibane et al. (2020a), (c) regressing SMPL+D parameters and (d) IP-Net predictions. NP* means 'not possible'.

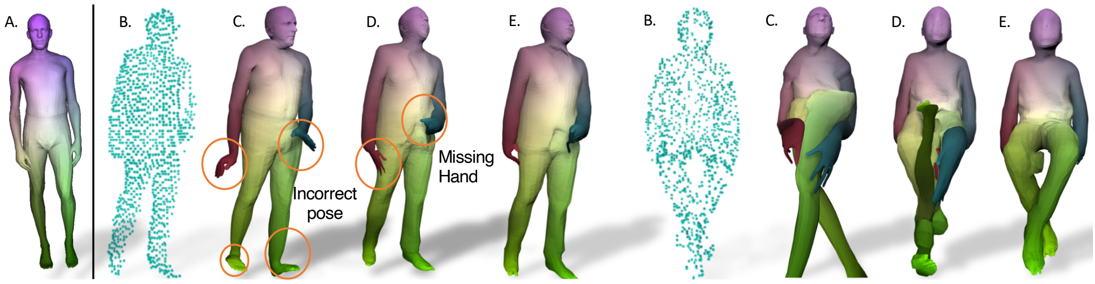


Figure 5.5: We compare quality of SMPL+D registration for various alternatives to IP-Net. We show A) colour coded reference SMPL, B) the input point cloud, C) registration directly to sparse PC, D) registration to Chibane et al. (2020a) prediction and E) registration to IP-Net predictions. It is important to note that poses such as sitting (second set) are difficult to register without explicit correspondences to the SMPL model.

5.3.2 *Outer surface reconstruction.*

For the task of outer surface reconstruction, we demonstrate that IP-Net performs better or on par with state of the art implicit reconstruction methods, Mescheder et al. (2019) and IF-Net, Chibane et al. (2020a). We report the average bi-directional vertex-to-surface error of 9.86mm, 4.86mm and 4.95mm for Mescheder et al. (2019), Chibane et al. (2020a) and IP-Net respectively. Unlike Chibane et al. (2020a) and Mescheder et al. (2019) which predict only the outer surface, we infer body shape under clothing and body part labels with the same model.

	Outer	**Inner**
(a) outer only	11.84	11.62
(b) outer+inner	11.54	11.14
(c) **outer+inner+parts**	**3.67**	**3.32**

Table 5.2: We compare three possibilities of registering the SMPL model to the implicit reconstruction produced by IP-Net. (a) registering SMPL+D to outer implicit reconstruction, (b) registering SMPL+D using the body prediction and (c) registering SMPL+D using body and part predictions. We report vertex-to-vertex error (cm) between the GT and predicted registered meshes.

Register, single view PC	**Outer reg.**	**Inner reg.**
Sin. view PC	15.90	NP*
Sin. view PC + IP-Net correspondences (Ours)	14.43	NP*
IP-Net (Ours)	**5.11**	**4.67**

Table 5.3: Depth sensors can provide single depth view point clouds. We report registration accuracy (vertex-to-vertex distance in cm) on such data and show that registration using IP-Net predictions is significantly better than alternatives. NP* implies 'not possible'.

5.3.3 *Comparison to Baselines*

The main contribution of our method is to make implicit reconstructions controllable. We do so by registering SMPL+D model (Alldieck et al., 2018b; Lazova et al., 2019) to IP-Net outputs: outer surface, inner surface and part correspondences.

This raises the the following questions, "Why not a) register SMPL+D directly to the input sparse point cloud?, b) register SMPL+D to the surface generated by an existing reconstruction approach, Chibane et al. (2020a)? c) directly regress SMPL+D parameters from the point cloud? and d) How much better is it to register using IP-Net predictions?".

Table 5.1 and Fig. 5.5 show that option d) (our method) is significantly better than the other baselines (a,b and c). To regress SMPL+D parameters (Option c), we implement a feed forward network that uses a similar encoder as IP-Net, but instead of predicting occupancy and part labels, produces SMPL+D parameters. We notice that the error for this method is dominated by misaligned pose and overall scale of the prediction. If we optimise the global orientation and scale of the predictions, this error is reduced from 32.45cm to 7.25cm which is still very high as compared to IP-Net based registration (3.67cm) which requires no such adjustments. This experiment provides two key insights, i) it is significantly better to make local predictions using implicit functions and later register a parametric model, than to directly regress the parameters of the model and ii) directly registering a parametric model to an existing reconstruction method Chibane et al. (2020a) yields larger errors than registering to IP-Net outputs (13.88cm vs 3.67cm).

Register **with IP-Net** correspondences	**Outer reg.**	**Inner reg.**
Sparse point cloud	13.93	NP*
Scan	3.99	NP*
IP-Net (Ours)	**3.67**	**3.32**

Table 5.4: An interesting use for IP-Net is to fit the SMPL+D model to sparse point clouds or scans using its part labels. This is useful for scan registration as we can retain the details of the high resolution scan and make it controllable. We report vertex-to-vertex error in cm. See Fig. 5.9 for qualitative results. NP* implies 'not possible'.

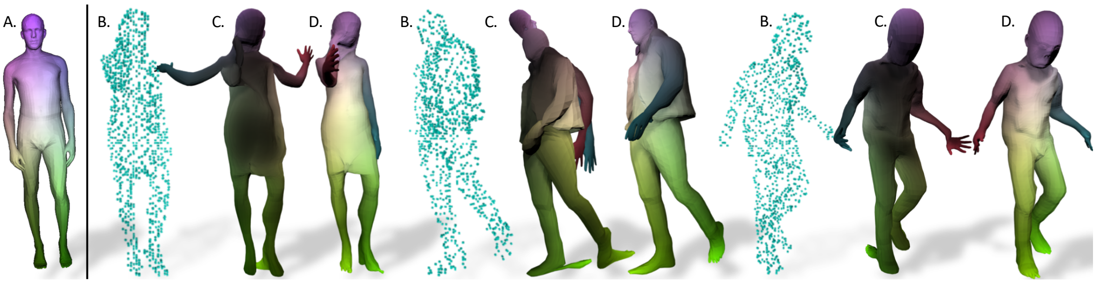


Figure 5.6: We highlight the importance of IP-Net predicted correspondences for accurate registration. We show A) color coded SMPL vertices to appreciate registration quality and three sets of comparative results. In each set, we visualize B) the input point cloud, C) registration without using IP-Net correspondences and D) registration with IP-Net correspondences. It can be seen that without correspondences we find problems like 180° flips (dark colors indicate back surface), vertices from torso being used to explain arms etc. These results are quantitatively corroborated in Table 5.2.

5.3.4 *Body Shape under Clothing*

We quantitatively evaluate our body shape predictions on BUFF dataset (Zhang et al., 2017b). Given a sparse point cloud generated from BUFF scans, IP-Net predicts the inner and outer surfaces along with the correspondences. We use our registration approach, as described in Sec. 5.2.3 to fit SMPL to our inner surface prediction and evaluate the error as per the protocol described in Zhang et al. (2017b). It is important to note that the comparison is unfair to our approach on several counts:

1. Our network uses sparse point clouds whereas Zhang et al. (2017b) use 4D scans for their optimization based approach.

2. Our network was not trained on BUFF (noisier scans, missing soles in feet).

3. The numbers reported by Zhang et al. (2017b) are obtained by jointly optimizing the body shape over entire 4D sequence, whereas our network makes a per-frame prediction without using temporal information.

We also compare our method to Yang et al. (2016). We report the following errors (mm): (Zhang et al. (2017b) male: 2.65, female: 2.48), (Yang et al. (2016) male: 17.85, female: 18.19) and (Ours male: 3.80, female: 6.17). Note that we did not have gender annotations for training IP-Net and hence generated our training data by registering all the scans to the 'male' SMPL model. This leads to significantly higher errors in estimating the body shape under clothing for 'female' subjects (we think this could be fixed by fitting gender specific models during training

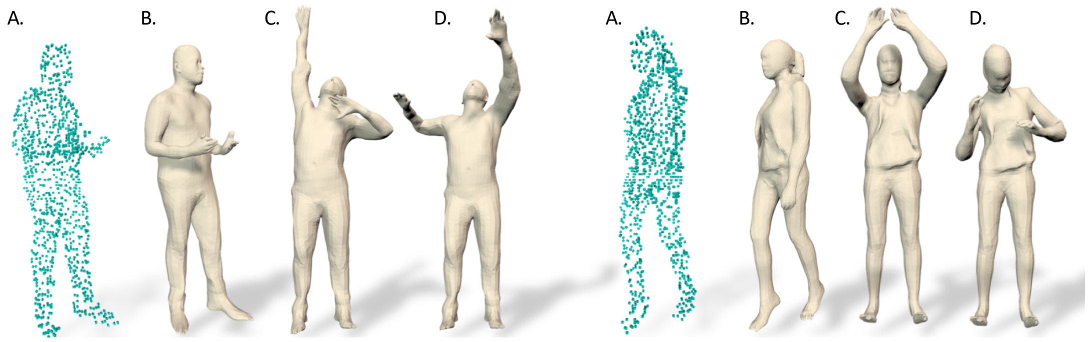


Figure 5.7: Implicit predictions by IP-Net can be registered with SMPL+D model and hence reposed. We show, A) input point cloud, B) corresponding SMPL+D registration and C,D) two instances of new poses.

data generation). We show that our approach can accurately infer body shape under clothing using just a sparse point cloud and is on par with approaches which use much more information.

5.3.5 *Why is correspondence prediction important?*

In this experiment, we demonstrate that inner surface reconstruction and part correspondences predicted by IP-Net are key for accurate registration. We discuss three obvious approaches for this registration:

(a) Register SMPL+D directly to the implicit outer surface predicted by IP-Net. This approach is simple and can be used with any other existing implicit reconstruction approaches.

(b) Register SMPL to the inner surface predicted by IP-Net and then non-rigidly register to the outer surface (without leveraging the correspondences).

(c) (Ours) First fit the SMPL model to the inner surface using correspondences and then non-rigidly register SMPL+D model to the implicit outer surface.

We report our results for the aforementioned approaches in Table 5.2 and Fig. 5.6. It can clearly be seen (Fig. 5.6, first set) that the arms of the SMPL model have not snapped to the correct pose. This is to be expected when arms are close to the body and no joint or correspondence information is present. In the second set, we see that vertices from torso are being used to explain the arms while SMPL arms are left hanging out. Third set is the classic case of 180° flipped fitting (dark color indicates back surface). This experiment highlights the importance of inner body surface and part correspondence prediction.

5.3.6 *Why not independent networks for inner & outer surfaces?*

IP-Net jointly predicts the inner and the outer surface for a human with clothing. Alternatively, one could train two separate implicit reconstruction networks using an existing SOTA approach. This has a clear disadvantage that one surface cannot reason about another, leading to severe inter-penetrations between the two. We report the average surface area of intersecting mesh faces which is 2000.71mm^2 for the two independent network approach, whereas with IP-Net the number is 0.65mm^2, which is four orders of magnitude smaller. See Fig. 5.8 for qualitative results. Our experiment demonstrates that having a joint model for inner and outer surfaces is better.

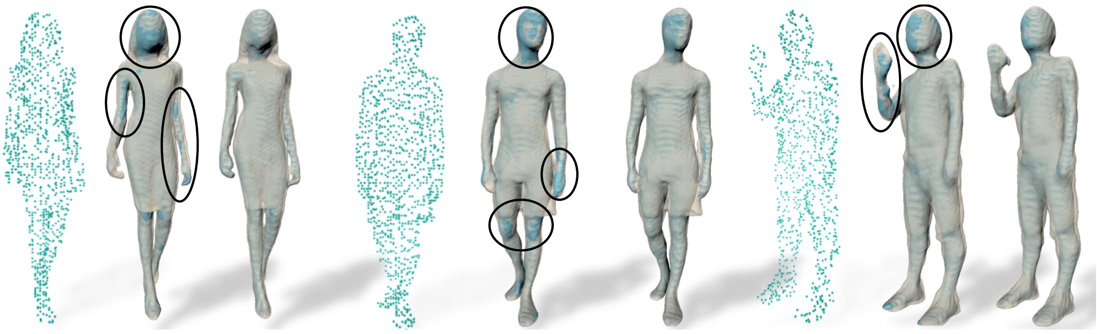


Figure 5.8: Advantage of IP-Net being a joint model for inner body surface and outer surface. In each set (L to R) we show input point cloud, inner (blue) and outer (off-white) surface reconstruction by two independent networks, IP-Net reconstructions. reconstructions from IP-Net have visibly fewer interpenetrations.

5.3.7 *Using IP-Net Correspondences to Register Scans*

A very powerful use case for IP-Net is scan registration. Current state-of-the art registration approaches (Alldieck et al., 2019a; Lazova et al., 2019) for registering SMPL+D to 3D scans are tedious and cumbersome (as described in Sec. 5.2.1). We provide a simple alternative using IP-Net. We sample points on our scan and generate the voxel grid used by IP-Net as input. We then run our pre-trained network and estimate the inner surface corresponding to the body shape under clothing. We additionally predict correspondences to the SMPL model for *each vertex on the scan*. We then use our registration (Sec. 5.2.3) to fit SMPL to the inner surface and then non-rigidly register SMPL+D to the scan surface, hence replacing the requirement for accurate 3D joints with IP-Net part correspondences. We show the scan registration and reposing results in Fig. 5.9 and Table 5.4. This is a useful experiment that shows that feed-forward IP-Net predictions can be used to replace tedious bottlenecks in scan registration.

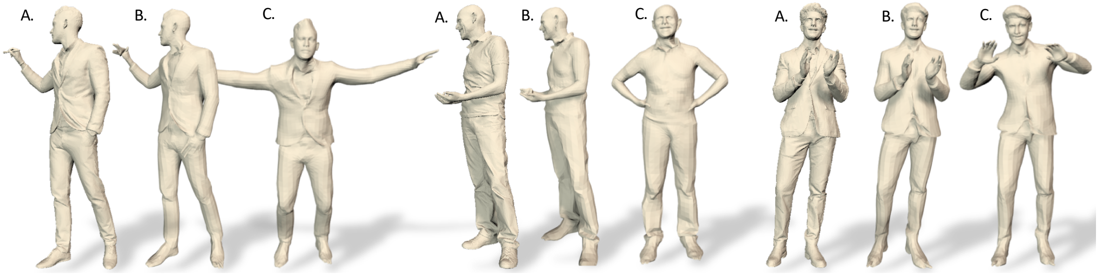


Figure 5.9: IP-Net can be used for scan registration. As can be seen from Table 5.1, registering SMPL+D directly to scan is difficult. We propose to predict the inner body surface and part correspondences for every point on the scan using IP-Net and subsequently register SMPL+D to it. This allows us to retain outer geometric details from the scan while also being able to animate it. We show A) input scan, B) SMPL+D registration using IP-Net, C) scan in a novel pose.

5.3.8 *Registration From Point Clouds Obtained from a Single View*

We show that IP-Net can be trained to process sparse point clouds from a single view (such as from Kinect). We show qualitative and quantitative results in Fig. 5.10 and Table 5.3, which demonstrate that IP-Net predictions are crucial for successful fitting in this difficult setting. This experiment highlights the general applicability of IP-Net to a variety of input modalities ranging from dense point clouds such as scans to sparse point clouds to single view point clouds.

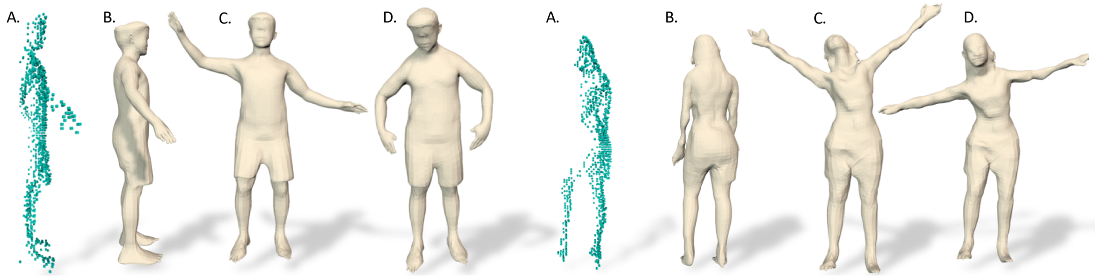


Figure 5.10: Single depth view point clouds (A) are becoming increasingly accessible with devices like Kinect. We show our registration using IP-Net (B) and reposing results (C,D) with two novel poses using such data.

5.3.9 *Hand Registration*

We show the wide applicability of IP-Net by using it for hand registration. We train IP-Net on the MANO hand dataset (Romero et al., 2017) and show hand registrations to full and single view point cloud in Fig. 5.11. We report an avg. vertex-to-vertex error of 4.80mm and 4.87mm in registration for full and single view point cloud respectively. This experiment shows that the idea of predicting implicit correspondences to a parametric model can be generically applied to different domains.

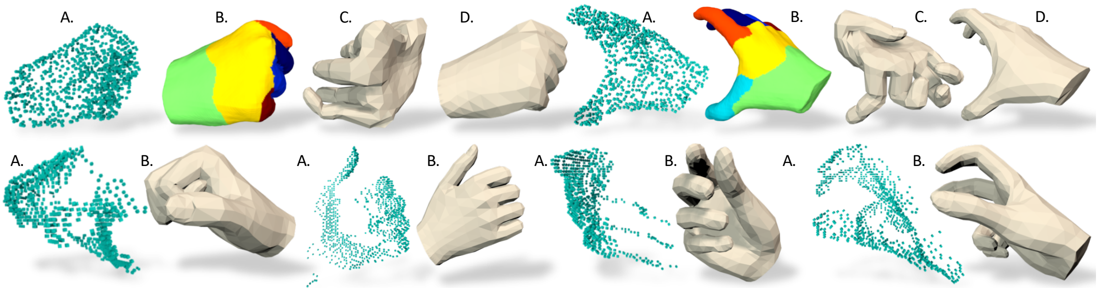


Figure 5.11: We extend our idea of predicting implicit correspondences to parametric models to 3D hands. Here, we show results on MANO hand dataset Romero et al., 2017. In the first row we show A) input PC, B) surface and part labels predicted by IP-Net, C) registration without part correspondences, and D) our registration. Registration without part labels is ill-posed and often leads to wrong parts explaining the surface. In the second row we show A) input single-view PC and B) corresponding registrations using IP-Net.

5.4 LIMITATIONS AND FUTURE WORKS

During our experiments we found IP-Net does not perform well with poses that were very different than the training set. In Fig. 5.12 first set, we have a person bending forward, and a similar pose was not present in our training set. Another limitation of our approach is that we rely on SMPL model for registration. This means that we are limited by SMPL topology and our registration suffers in the presence of non-clothing objects that cannot be handled by SMPL. We also feel that the reconstructed details can be further improved especially around the face.

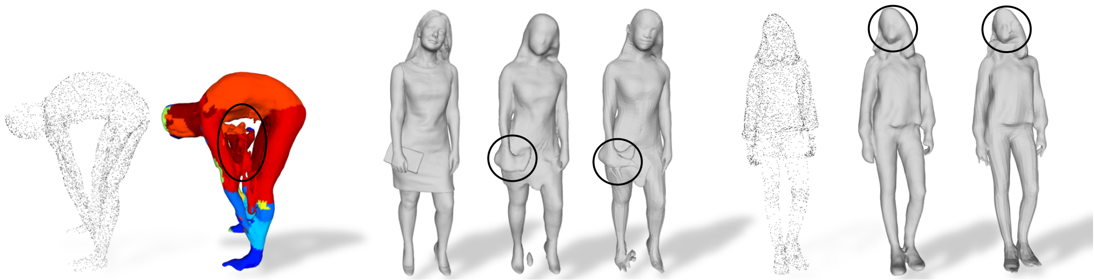


Figure 5.12: We present some of the failure cases of our proposed approach. In the first set, we show the input point cloud and the generated surface reconstruction by IP-Net. Unseen poses are difficult for IP-Net. In the second set we show the GT scan with the person holding an object, the IP-Net reconstruction and the resultant registration with artefacts around the hand. Our approach cannot deal with non-clothing objects. In the third set we show the input point cloud, the IP-Net generated surface and the registration. Notice that facial details are missing.

5.5 CONCLUSIONS

Learning implicit functions to model humans has been shown to be powerful but the resulting representations are not amenable to control or reposing which are essential for both animation and inference in computer vision. We have presented methodology to combine expressive implicit function representations and parametric body modelling in order to produce 3D reconstructions of humans that remain controllable even in the presence of clothing.

Given a sparse point cloud representing a human body scan, we use implicit representation obtained using deep learning in order to jointly predict the *outer* 3D surface of the dressed person and the *inner* body surface as well as the semantic body parts of the parametric model. We use the part labels to fit the parametric model to our inner surface and then non-rigidly deform it (under a body prior + displacement model) to the outer surface in order to capture garment, face and hair details. Our experiments demonstrate that 1) predicting a double layer surface is useful for subsequent model fitting resulting in reconstruction improvements of 3mm and 2) leveraging semantic body parts is *crucial* for subsequent fitting and results in improvements of 8.17cm. The benefits of our method are paramount for difficult poses or when input is incomplete such as single view sparse point clouds, where the double layer implicit reconstruction and part classification is essential for successful registration. Our method generalizes well to other domains such as 3D hands (as evaluated on the MANO dataset) and even works well when presented with incomplete point clouds from a single depth view, as shown in extensive quantitative and qualitative experiments.

6 LOOPREG: SELF-SUPERVISED LEARNING OF IMPLICIT SURFACE CORRESPONDENCES, POSE AND SHAPE FOR 3D HUMAN MESH REGISTRATION

In this chapter, we address the problem of fitting 3D human models to 3D scans of dressed humans. Classical methods optimize both the data-to-model correspondences and the human model parameters (pose and shape), but are reliable only when initialized close to the solution. Some methods initialize the optimization based on fully supervised correspondence predictors, which is not differentiable end-to-end, and can only process a single scan at a time. Our main contribution is *LoopReg*, an end-to-end learning framework to register a corpus of scans to a common 3D human model. The key idea is to create a self-supervised loop. A backward map, parameterized by a Neural Network, predicts the correspondence from every scan point to the surface of the human model. A forward map, parameterized by a human model, transforms the corresponding points back to the scan based on the model parameters (pose and shape), thus closing the loop. Formulating this closed loop is not straightforward because it is not trivial to force the output of the neural network to be on the surface of the human model – outside this surface the human model is not even defined. To this end, we propose two key innovations. First, we define the canonical surface implicitly as the zero level set of a distance field in $\mathbb{R}^3$, which in contrast to more common UV parameterizations $\Omega \subset \mathbb{R}^2$, does not require cutting the surface, does not have discontinuities, and does not induce distortion. Second, we diffuse the human model to the 3D domain $\mathbb{R}^3$. This allows to map the NN predictions forward, even when they slightly deviate from the zero level set. Results demonstrate that we can train *LoopReg* mainly self-supervised – following a supervised warm-start, the model becomes increasingly more accurate as additional unlabelled raw scans are processed. Our code and pre-trained models can be downloaded for research at: http://virtualhumans.mpi-inf.mpg.de/loopreg.

6.1 INTRODUCTION

We propose a novel approach for model-based registration, i.e. fitting parametric model to 3D scans of articulated humans. Registration of scans is necessary to complete, edit and control geometry, and is often a precondition for building statistical 3D models from data (Anguelov et al., 2005; Loper et al., 2015; Pons-Moll et al., 2015; Xu et al., 2020).

Classical model-based approaches optimize an objective function over scan-to-model correspondences and the parameters of a statistical human model, typically pose, shape and non-rigid displacement. When properly initialized, such approaches are effective and generalize well. However, when the variation in pose, shape and clothing is high, they are vulnerable to local minima.

To avoid convergence to local minima, researchers proposed to use predictors to either initialize the latent parameters of a human model (Groueix et al., 2018), or the correspondences between data points and the model (Pons-Moll et al., 2013; Taylor et al., 2012). Learning to predict global latent parameters of a human model directly from a point-cloud is difficult and such initializations to standard registration are not yet reliable. Instead, learning to predict correspondences to a 3D human model is more effective (Alp Güler et al., 2018; Bhatnagar et al., 2020a).

Several important limitations are apparent with current approaches. First, supervising an initial regression model of correspondence requires labeled scans (Or et al., 2017; Pons-Moll et al., 2013; Taylor et al., 2012; Wei et al., 2016), which are hard to obtain. Second, although some approaches use predicted correspondences to initialize a subsequent, classical optimization-based registration, this process involves non-differentiable steps.

What is lacking is a joint end-to-end differentiable objective over correspondences and human model parameters, which allows to train the correspondence predictor, self-supervised, given a corpus of unlabeled scans. This is our motivation in introducing *LoopReg*.

Given a point-cloud, a backward map, parameterized by a neural network, transforms every scan point to a corresponding point on the canonical surface (the human model in a canonical pose and shape). A forward map, parameterized by the SMPL human model (Loper et al., 2015), transforms canonical points under articulation, shape and non-rigid deformation, to fit the original point-cloud (see Fig. 6.1). *LoopReg* creates a differentiable loop which supports the self-supervised learning of correspondences, along with pose, shape and non-rigid deformation, following a short supervised warm-start.

The design of *LoopReg* requires several technical innovations. First, we need a continuous representation for canonical points, to be on the human surface manifold. We define the surface implicitly as the zero level set of a distance field in $\mathbb{R}^3$ instead of the more common approach of using a 2D UV parameterization $\Omega \subset \mathbb{R}^2$, which typically relies on manual interaction, and inevitably has distortion and boundary discontinuities (Hormann et al., 2007). We follow a Lagrangian formulation; during learning, NN predictions which deviate from the implicit surface are penalized softly. Furthermore, we interpret the 3D human model as a function on the surface manifold. We diffuse the function onto the 3D domain via a distance transform (Fig. 6.3), which allows to map the NN predictions forward, when they slightly deviate from the surface during learning.

In summary, our key contributions are:

- *LoopReg* is the first end-to-end learning process jointly defined over a parametric human model *and* the data (scan / point cloud) to model correspondences.
- We propose an alternative to classical UV parameterization for correspondences. We define the canonical human surface implicitly as the zero levelset of a distance field, and diffuse the SMPL function to the 3D domain. The formulation is continuous and differentiable.
- *LoopReg* supports self-supervision. We experimentally show that registration accuracy can be improved as more unlabeled data is added.
- While most approaches for human scan registration are based on instance specific optimization, our approach registers a dataset of scans to a common template, thereby leveraging aggregated statistics.

6.2 LOOPREG: METHOD

Current state of the art human model based registration such as Alldieck et al. (2019a) and Lazova et al. (2019) require pre-computed 3D joints and keypoint/landmark detection. These approaches render the scans from multiple views, use OpenPose [1] or similar models for image based 2D joint detection, and lift the 2D joint detections to 3D. This is prone to error at multiple

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose

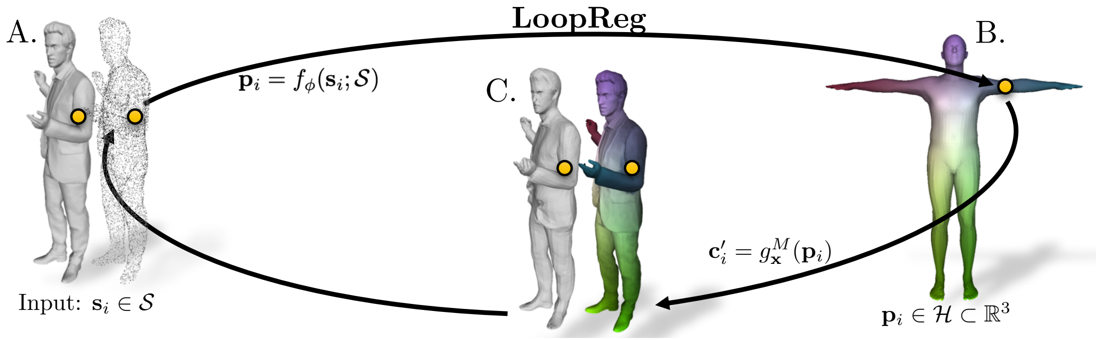

LoopReg
A.
$\mathbf{p}_i = f_\phi(\mathbf{s}_i; \mathcal{S})$
C.
B.
$\mathbf{c}'_i = g^M_{\mathbf{x}}(\mathbf{p}_i)$
Input: $\mathbf{s}_i \in \mathcal{S}$
$\mathbf{p}_i \in \mathcal{H} \subset \mathbb{R}^3$

Figure 6.1: Input to our method is a scan or point cloud (A) $\mathcal{S}$. For each input point $\mathbf{s}_i$, our network CorrNet $f_\phi(\cdot)$ predicts a correspondence $\mathbf{p}_i$ to a canonical model in $\mathcal{H} \subset \mathbb{R}^3$ (B). We use these correspondences to jointly optimize the parametric model (C) and CorrNet under self-supervised training.

levels. Per-view joint detection may be inconsistent across views. Furthermore, when scans are point-clouds instead of meshes, they can not be rendered. Fig. 6.4 and 6.5 show that accurate scan registration is not possible for complex poses without this information.

In *LoopReg*, we replace the pre-computed sparse joint information with continuous correspondences to a parametric human model. Our network *CorrNet* contains a backward map, that transforms scan points to corresponding points on the surface of a human model with canonical pose and shape. In a forward map, these corresponding points are deformed using the human model to fit the original scan, thereby creating a self-supervised loop. We start our description by reviewing the basic formulation of traditional model based fitting approaches, and follow with our self-supervised registration method.

6.2.1 *Classical Model-Based Fitting*

The classical way of fitting a 3D (human) model to a scan $\mathcal{S}$ is via minimization of an objective function. Let $M(\mathbf{v}_i, \mathbf{x}) : \mathcal{I} \times \mathcal{X}' \mapsto \mathbb{R}^3$, denote the human model which maps a 3D vertex $\mathbf{v}_i \in \mathcal{I}$, on the canonical human surface $\mathcal{M}_T \subset \mathbb{R}^3$ to a transformed 3D point after deforming according to model parameters $\mathbf{x} \in \mathcal{X}'$. For the SMPL+D model, which we use here, $\mathbf{x} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D}\}$ corresponds to pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and non-rigid deformation $\mathbf{D}$. The standard registration approach is to find a set of corresponding canonical model points $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_N\}$ (the *correspondences*) for the scan points $\{\mathbf{s}_1, ..., \mathbf{s}_N\}, \mathbf{s}_i \in \mathcal{S}$ and minimize a loss of the form:

$$L(\mathcal{C}, \mathbf{x}) = \sum_{\mathbf{s}_i \in \mathcal{S}} \mathrm{dist}(\mathbf{s}_i, M'(\mathbf{c}_i, \mathbf{x})), \tag{6.1}$$

where $\mathrm{dist}(\cdot, \cdot)$ is a distance metric in $\mathbb{R}^3$. Note that Eq. 6.1 uses continuous surface points $\mathbf{c}_i \in \mathcal{M}_T$, and $M'(\cdot)$ interpolates the model function $M(\cdot)$ defined for discrete model vertices $\mathbf{v}_i \in \mathcal{I}$ with barycentric interpolation.

Eq. 6.1 is minimized with non-linear ICP, which is a two step non-differentiable process. First, for every scan point a corresponding point on the human model is computed. Next, the model parameters are updated to minimize the distance between scan points and corresponding model points using gradient or Gauss-Newton optimizers. This alternating process is non-differentiable, which rules out end-to-end training. Our work is inspired by Taylor et al. (2016) which continuously optimizes the corresponding points and the model parameters. The trick is to parameterize the canonical surface with piece-wise mappings from a 2D space Ω to 3D $\mathbb{R}^3$ – per triangle mappings. This requires keeping track of the TriangleIDs and point location

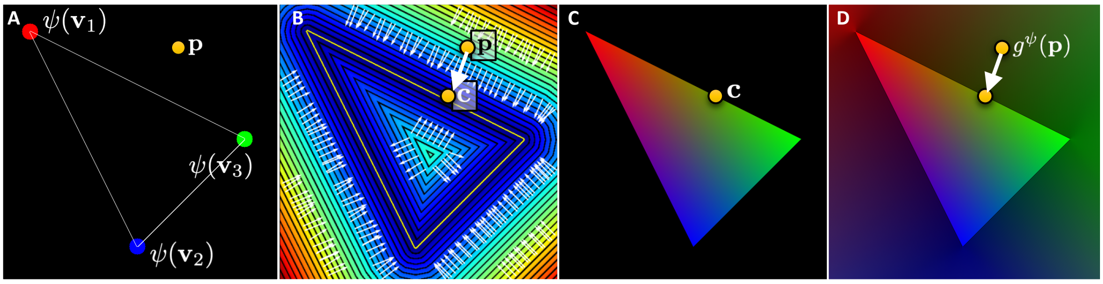


Figure 6.2: Illustration of the diffusion process. We diffuse the function $\psi(\cdot)$ (denoted as per-vertex colours), defined only on the vertices $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$, to any point $\mathbf{p} \in \mathbb{R}^3$. Within the surface (in this example just a triangle), the function $\psi(\cdot)$ is diffused using barycentric interpolation (sub-figure C). For a point $\mathbf{p} \in \mathbb{R}^3$ beyond the surface, the function $\psi(\cdot)$ is diffused by evaluating the barycentric interpolation at the closest point $\mathbf{c}$ to $\mathbf{p}$, implemented by pre-computing a distance transform (sub-figure B). The result is a diffused function $g^{\psi}(\mathbf{p})$ defined, not only on vertices, but over all $\mathbb{R}^3$. Fig. 6.3 explains the same process for a complete human mesh.

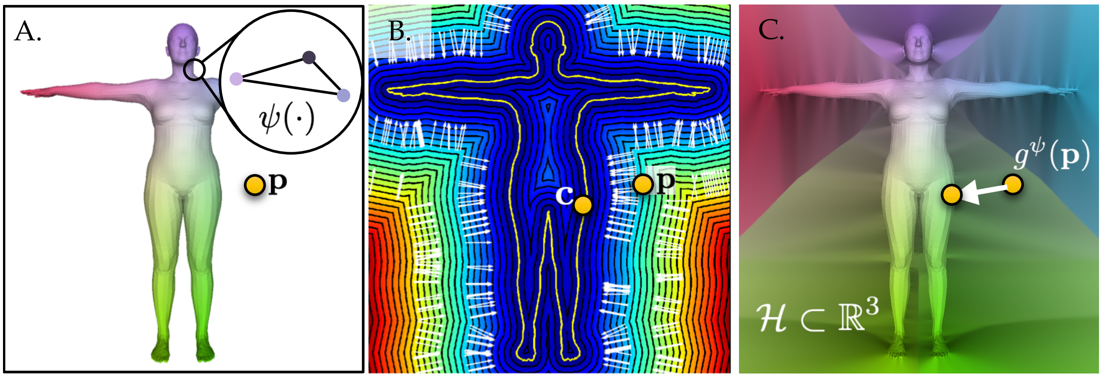


Figure 6.3: Illustration of the diffusion process on a human mesh. In sub-figure A, an arbitrary function $\psi(\cdot)$, defined over mesh vertices (illustrated as per vertex colors), is diffused to $\mathcal{H} \subset \mathbb{R}^3$ via a distance transform (sub-figure B). This results in a new function $g^{\psi}(\cdot)$ (sub-figure C).

within the triangle when correspondences shift across triangles. Apart from being difficult to implement, this is not a suitable representation for learning correspondences as TriangleIDs do not live in a continuous space with metric. Furthermore, the optimization in Taylor et al. (2016) is instance specific.

6.2.2 *Proposed Formulation*

Instead of instance specific optimization, we want to automatize the model fitting process by leveraging a corpus of 3D human scans. Our key idea is to create a differentiable registration loop motivated by classical model-based fitting Eq. 6.1. We learn a continuous and differentiable mapping $f_{\phi}(\mathbf{s}; \mathcal{S}) : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathcal{M}_T \subset \mathbb{R}^3$, with network parameters ϕ, from the scan points $\mathbf{s} \in \mathcal{S}$ to the canonical surface (of the human model in canonical pose and shape) $\mathcal{M}_T$. Let $\{\mathcal{S}_j\}_{j=1}^{N_u}$ be a set of unlabeled scans, and $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^{N_u}$ be the set of unknown instance specific latent parameters per scan. The following self-supervised loss creates a loop between the scans and the model

$$L(\phi, \mathcal{X}) = \sum_{j=1}^{N_u} \sum_{\mathbf{s}_i \in \mathcal{S}_j} \text{dist}(\mathbf{s}_i, M'(f_{\phi}(\mathbf{s}_i; \mathcal{S}_j), \mathbf{x}_j)), \tag{6.2}$$

where, in contrast to the instance specific Eq. 6.1, optimization in Eq. 6.2 is over a training set, and correspondences are predicted by the network, $\mathbf{c} = f_\phi(\mathbf{s}; \mathcal{S}_j)$. It is important to note that those correspondences network predicted, $\mathbf{c}$, need not be on the canonical surface. Outside the canonical surface, $M'(\cdot)$ is not even defined. The question then is how to minimize Eq. 6.2 end-to-end.

IMPLICIT SURFACE REPRESENTATION. To predict correspondences on $\mathcal{M}_T$, we need a continuous representation of its surface. One of our key ideas is to define the surface implicitly, as the zero levelset of a signed distance field. Let $d(\mathbf{p}) : \mathbb{R}^3 \mapsto \mathbb{R}$ be the distance field, defined as $d(\mathbf{p}) = \text{sign}(\mathbf{p}) \cdot \min_{\mathbf{c} \in \mathcal{M}_T} \|\mathbf{c} - \mathbf{p}\|$ taking a positive sign on the outside and negative otherwise. The surface is defined implicitly as $\mathcal{M}_T = \{\mathbf{p} \in \mathbb{R}^3 \mid d(\mathbf{p}) = 0\}$. But how can we satisfy the constraint $d(\mathbf{p}) = 0$ during learning? And how to handle network predictions that overshoot the surface?

DIFFUSING THE HUMAN MODEL FUNCTION TO $\mathbb{R}^3$. Imposing the hard constraint $d(\mathbf{p}) = 0$ during learning is not feasible. Hence, our second key idea is to diffuse the human body model to the full 3D domain (see Fig. 6.3, and Fig. 6.2 to better understand the diffusion process with a simple example). Without loss of generality, we use SMPL (Loper et al., 2015) as our human model throughout the paper, but the ideas apply generally, to other 3D statistical surface models (Blanz and Vetter, 1999; Li et al., 2017; Pishchulin et al., 2017; Romero et al., 2017; Xu et al., 2020).

The SMPL model applies a series of linear mappings to each vertex $\mathbf{v}_i \in \mathcal{I}$ of a template, followed by skinning. The per-vertex linear mappings are the pose blendshapes $b_P : \mathcal{I} \mapsto \mathbb{R}^{3\times|\boldsymbol{\theta}|}$, shape blendshapes $b_S : \mathcal{I} \mapsto \mathbb{R}^{3\times|\boldsymbol{\beta}|}$, applied to a canonical template, followed by linear blend skinning with parameters $w_i : \mathcal{I} \mapsto \mathbb{R}^K$. The i-th vertex $\mathbf{v}_i$ is transformed according to

$$\mathbf{v}'_i = \sum_{k=1}^{K} w(\mathbf{v}_i)_k G_k(\boldsymbol{\theta}, \boldsymbol{\beta}) \cdot (\mathbf{v}_i + b_P(\mathbf{v}_i) \cdot \boldsymbol{\theta} + b_S(\mathbf{v}_i) \cdot \boldsymbol{\beta}) \tag{6.3}$$

where $G_k(\boldsymbol{\theta}, \boldsymbol{\beta}) \in SE(3)$ is the 4×4 transformation matrix of part K, see Loper et al., 2015. Note that SMPL is a function defined on the vertices $\mathbf{v}_i$ of the template, while we need a continuous mapping in $\mathbb{R}^3$.

Let $\psi : \mathcal{I} \mapsto \mathcal{Y}$ be a function defined on discrete vertices $\mathbf{v}_i \in \mathcal{I}$, with co-domain $\mathcal{Y}$. The idea is to derive a function $g^\psi(\mathbf{p}) : \mathbb{R}^3 \mapsto \mathcal{Y}$ which diffuses ψ to $\mathbb{R}^3$. ψ can trivially be diffused to the surface by barycentric interpolation and to $\mathbb{R}^3$ using the closest surface point $\mathbf{c} = \arg\min_{\mathbf{c} \in \mathcal{M}_T} \|\mathbf{c} - \mathbf{p}\|$,

$$g^\psi(\mathbf{p}) = \alpha_1 \psi(\mathbf{v}_l) + \alpha_2 \psi(\mathbf{v}_k) + \alpha_3 \psi(\mathbf{v}_m), \tag{6.4}$$

where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}^3$ are barycentric coordinates of $\mathbf{c} \in \mathcal{M}_T$, and $\mathbf{v}_l, \mathbf{v}_k, \mathbf{v}_m \in \mathcal{I}$ are corresponding canonical vertices, see Fig. 6.2-A,C,D.

During learning, we need to evaluate $g^\psi(\mathbf{p})$ and compute its spatial gradient $\nabla_{\mathbf{p}} g^\psi(\mathbf{p})$ efficiently. Hence, we pre-compute $g^\psi(\mathbf{p})$ in a 3D grid around the surface $\mathcal{M}_T$ (we use a unit cube with 64x64x64 resolution), and use tri-linear interpolation to obtain a continuous differentiable mapping in regions near the surface $\mathcal{H} \subset \mathbb{R}^3$.

LOOPREG. The aforementioned representation allows the formulation of correspondence prediction in a self-supervised loop, as a composition of the backward map f_ϕ and the forward map g^ψ. The backward map $f_\phi(\mathbf{s}; \mathcal{S}) : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathcal{H} \subset \mathbb{R}^3$ (implemented as a deep neural network) transforms every scan point $\mathbf{s}$ to the corresponding canonical point $\mathbf{p}$ in the unshaped and unposed space $\mathcal{H} \subset \mathbb{R}^3$. For clarity, we simply use $f_\phi(\mathbf{s})$ for the network prediction. The forward map $g^\psi(\mathbf{p})$ is the diffused SMPL function by setting $\psi = M$ (body model). Specifically, we diffuse the pose and shape blend-shapes, and skinning weights to the 3D region $\mathcal{H}$, obtaining functions g^{b_P}, g^{b_S}, g^w. We also obtain the function g^I which maps every point $\mathbf{p} \in \mathcal{H}$ to its closest surface point $\mathbf{c} \in \mathcal{M}_T$. The diffused SMPL function for $\mathbf{x} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D}\} \in \mathcal{X}'$ is obtained as

$$g^M : \mathcal{H} \times \mathcal{X}' \mapsto \mathbb{R}^3, \quad g^M(\mathbf{p}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D}) = \sum_{k=1}^{K} g_k^w(\mathbf{p}) G_k(\boldsymbol{\theta}, \boldsymbol{\beta})(g^I(\mathbf{p}) + g^{b_P}(\mathbf{p})\boldsymbol{\theta} + g^{b_S}(\mathbf{p})\boldsymbol{\beta}), \tag{6.5}$$

which is continuous and differentiable. This enables re-formulating Eq. 6.2 as a differentiable loss

$$L_{\text{self}}(\phi, \mathcal{X}) = \sum_{j=1}^{N_u} \sum_{\mathbf{s}_i \in \mathcal{S}_j} \text{dist}(\mathbf{s}_i, g^M_{\mathbf{x}_j}(f_\phi(\mathbf{s}_i))) + \lambda \cdot d(f_\phi(\mathbf{s}_i)), \tag{6.6}$$

where $f_\phi(\mathbf{s}) = \mathbf{p}$ is the corresponding point predicted by the network, and we used the notation $g^M_{\mathbf{x}_j}(\mathbf{p}) = g^M(\mathbf{p}_j, \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \mathbf{D}_j)$ for clarity, and $d(\cdot)$ is the distance transform. Network predictions which deviate from the surface are penalized with the term $\lambda \cdot d(f_\phi(\mathbf{s}_i))$ following a Lagrangian formulation of constraints. Note that this term is important in forcing predicted correspondences to be close to the template surface, as gradient updates far away from the surface may not be well behaved. Notice that $\nabla d(\mathbf{p}_j)$ points towards the closest surface point (see Fig. 6.3 B.), and along this direction $g^M_{\mathbf{x}_j}(\mathbf{p}_j)$ is constant, and hence $\nabla d(\mathbf{p}_j) \perp \nabla g^M_{\mathbf{x}_j}(\mathbf{p}_j)$; the terms are complementary and do not compete against each other. Unfortunately, directly minimizing Eq. 6.6 over network f_ϕ and instance specific parameters $\mathcal{X}$ is not feasible. The initial correspondences predicted by the network are random, which leads to unstable model fitting and a non-convergent process.

SEMI-SUPERVISED LEARNING. We propose the following semi-supervised learning strategy, where we warm-start the process using a small labeled dataset $\{\mathcal{S}_j, \mathcal{M}(\mathbf{x}_j)\}_{j=1}^{N_s}$, with $\mathcal{M}(\mathbf{x}_j)$ denoting the registered SMPL surface to the scan, and subsequently train with a larger un-labeled corpus of scans $\{\mathcal{S}_j\}_{j=1}^{N_u}$. Learning entails minimizing the following three losses over correspondence-network parameters ϕ and instance specific model parameters $\mathcal{X} = \{\mathbf{x}_j\}_{j=1}^{N_s}$:

$$L = L_{\text{unsup}} + L_{\text{sup}} + L_{\text{reg}}. \tag{6.7}$$

The unsupervised loss consists of a data-to-model term $L_{\text{d}\mapsto\text{m}} = \text{dist}(s, \mathcal{M}(\mathbf{x}))$ and the self-supervised loss L_{self}, in Eq. 6.6

$$L_{\text{unsup}}(\phi, \mathcal{X}) = \sum_{j=1}^{N_u} \sum_{\mathbf{s}_i \in \mathcal{S}_j} \text{dist}(s_i, \mathcal{M}(\mathbf{x}_j)) + \text{dist}(\mathbf{s}_i, g^M_{\mathbf{x}_j}(f_\phi(\mathbf{s}_i))) + \lambda \cdot d(f_\phi(\mathbf{s}_i)), \tag{6.8}$$

where $\mathcal{M}(\mathbf{x}_j)$ is the SMPL mesh deformed by the *unknown* parameters $\mathbf{x}_j$, and $\text{dist}(s, \mathcal{M}(\mathbf{x}))$ is a differentiable point-to-surface distance. The term $L_{\text{d}\mapsto\text{m}}$ pulls the deformed model $\mathcal{M}(\mathbf{x})$ to the data, which in turn makes learning the correspondence predictor f_ϕ more stable.

The supervised loss, L_{sup}, minimises the L_2 distance between network-predicted correspondences and ground truth ones $\hat{\mathbf{c}}_{ji}$ obtained from $\mathcal{M}(\mathbf{x}_j)$

$$L_{\text{sup}}(\phi) = \sum_{j=1}^{N_s} \sum_{\mathbf{s}_i \in \mathcal{S}_j} \| f_\phi(\mathbf{s}_i, \mathcal{S}_j) - \hat{\mathbf{c}}_{ji} \|_2. \tag{6.9}$$

The regularisation term L_{reg} consists of priors on the SMPL shape and pose (L_θ) parameters.

$$L_{\text{reg}}(\phi, \mathcal{X}) = \sum_{j=1}^{N_u} L_\theta(\boldsymbol{\theta}_j) + \|\boldsymbol{\beta}_j\|_2 \tag{6.10}$$

Our experiments show that with good initialization, our approach can be trained self-supervised, using only L_{unsup} and L_{reg}.

CORRNET: PREDICTING SCAN TO MODEL CORRESPONDENCES The aforementioned formulation is continuous and differentiable with respect to instance specific parameters $\mathbf{x}_j = \{\boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \mathbf{D}_j\}$ as well as global network parameters ϕ. In this section, we describe our correspondence prediction network, CorrNet. CorrNet $f_\phi(\cdot)$, is designed with a PointNet++ (Qi et al., 2017) backbone and regresses for each input scan point $\mathbf{s}_i$, its correspondence $\mathbf{p}_i \in \mathcal{H}$. We observe that the correspondence mapping is discontinuous when the scan has self-occlusion and contact. For example, when the hand touches the hip, nearby scan points need to be mapped to distant $\mathbf{p} = (x, y, z)$ coordinates, which is difficult. Inspired by Riza Alp Güler (2018), we first predict a body part label (we use $N = 14$ pre-defined parts on the SMPL mesh) for each scan point and subsequently regress the continuous x, y, z-correspondence only within that part. Similar to ensemble learning approaches we use a weighted sum of part-specific classifiers to regress correspondences. This way, our formulation is differentiable with respect to part classification, which would not be possible if we directly used arg max for hard part assignment:

$$f_\phi(\mathbf{s}) = \mathbf{p} = \sum_{k=1}^{N_{\text{parts}}} f_{\phi,k}^{\text{class}}(\mathbf{s}, \mathcal{S}) f_{\phi,k}^{\text{reg}}(\mathbf{s}, \mathcal{S}), \tag{6.11}$$

where $f_\phi^{\text{class}} : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathbb{R}^{N_{\text{parts}}}$ is the (soft) part-classification branch of the CorrNet and $f_{\phi,k}^{\text{reg}} : \mathbb{R}^3 \times \mathcal{S} \mapsto \mathcal{H} \subset \mathbb{R}^3$ is the part specific regressor.

6.3 DATASET AND EXPERIMENTS

In this section, we evaluate and show that our approach outperforms existing scan registration approaches. Moreover, our approach seamlessly generalizes across undressed (comparatively easier) and fully clothed (significantly more challenging) scans in complex poses. We show that our approach can be trained with self-supervision (with supervised warm-start) and performance improves noticeably as more and more raw scans are made available to our method.

6.3.1 *Dataset*

We use 3D scans of humans from RenderPeople [2], AXYZ [3] and Twindom [4]. To obtain reference registrations for evaluation, we fit the SMPL model to scans using Alldieck et al. (2019a) and Lazova et al. (2019) with pre-computed 3D joints lifted from 2D detections, facial landmarks and manually select the good fits. The input to our method are point clouds, which we extract from SMPL fits for undressed humans, and from the raw scans for clothed humans. We divide the SMPL fits in a supervised (1000 scans), unsupervised (1631 scans) and testing set (290 scans). We perform additional experiments on Faust (Bogo et al., 2014) which contains around 100 scans of undressed people and corresponding GT SMPL registrations.

[2] https://renderpeople.com

[3] https://secure.axyz-design.com

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

Registration Errors	**vertex-to-vertex** (cm)		**surface-to-surface** (mm)	
Dataset	Our	FAUST	Our	FAUST
(a) Optimization based	16.5	13.5	12.6	3.8
(b) 3D-CODED single init.	22.3	21.9	8.7	10.6
(c) 3D-CODED	2.0	3.1	2.4	2.6
(d) Ours	**1.4**	**2.2**	**1.0**	**2.4**

Table 6.1: We compare registration performance of our approach against instance specific optimization (a) Alldieck et al. (2019a), Lazova et al. (2019), and Zhang et al. (2017b) (without pre-computed joints and manual selection) and learning based Groueix et al. (2018) approach. Note that 3D-CODED requires multiple initializations (c) to find best global orientation, without which the approach easily gets stuck in a local minima (b). Since Groueix et al. (2018) freely deforms a template, for a fair comparison, we use SMPL+D model for our and Alldieck et al. (2019a), Lazova et al. (2019), and Zhang et al. (2017b) approaches, even though data contains undressed scans. We report results on our dataset and FAUST (Bogo et al., 2014)

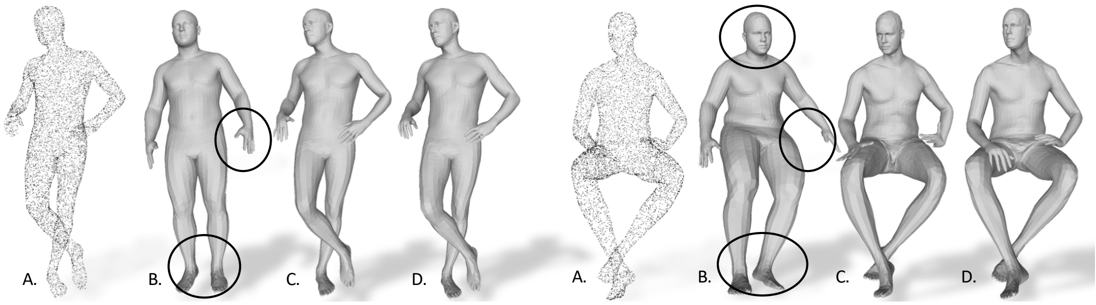


Figure 6.4: Comparison with existing scan registration approaches, Alldieck et al. (2019a) and Lazova et al. (2019). We show A) input point cloud, B) registration using Alldieck et al., 2019a; Lazova et al., 2019 without pre-computed joints/ landmarks. C) Our registration. D) GT registration using Alldieck et al. (2019a) and Lazova et al. (2019) + precomputed 3D joints + facial landmarks + manual selection. It can be seen that B) makes significant errors as compared to our approach C).

6.3.2 *Comparison with existing instance specific optimization based approaches*

REGISTERING UNDRESSED SCANS. One of the key strengths of our approach is the ability to register 3D scans without additional information such as pre-computed joint/ landmarks and manual intervention. In Fig. 6.4 we show that existing state of the art registration approaches (Alldieck et al., 2019a; Lazova et al., 2019; Zhang et al., 2017b) cannot perform accurate registration without these pre-processing steps. More qualitative results from our method can be seen in Fig. 6.6

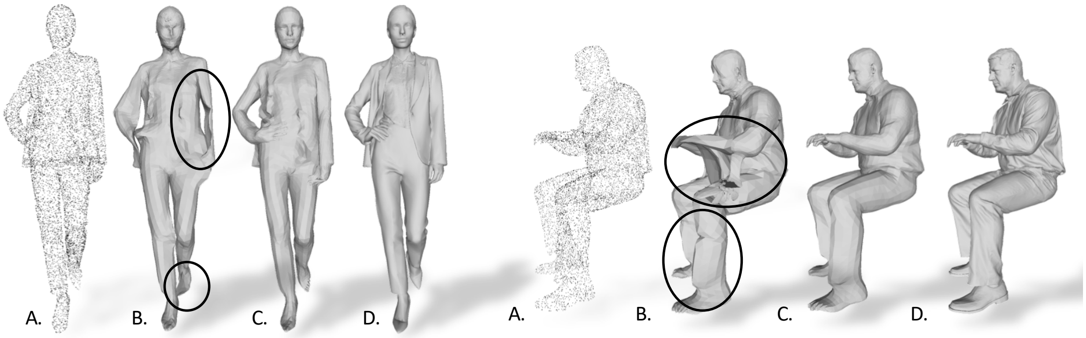


Figure 6.5: Comparison with existing scan registration approaches, Alldieck et al. (2019a) and Lazova et al. (2019). We show A) input point cloud, B) registration using Alldieck et al. (2019a) and Lazova et al. (2019) without pre-computed joints/ landmarks. C) Our registration and D) GT scan. It can be seen that B) makes significant errors as compared to our approach C).

Unsupervised %	**0%**	**10%**	**25%**	**50%**	**75%**	**100%**
(a) v2v (cm)	9.3	8.4	6.3	4.1	2.7	1.5
(b) s2s (mm)	6.8	6.6	6.2	5.5	5.1	4.2

Table 6.2: Performance of the proposed approach increases as we add more unsupervised data for training. Here 100% corresponds to 2631 scans. Out of the 2631 scans 1000 were also used for supervised warm-start. We report vertex-to-vertex (v2v) and bi-directional surface-to-surface (s2s) errors and clearly show that adding more unsupervised data improves registration performance, specially for the more demanding v2v metric.

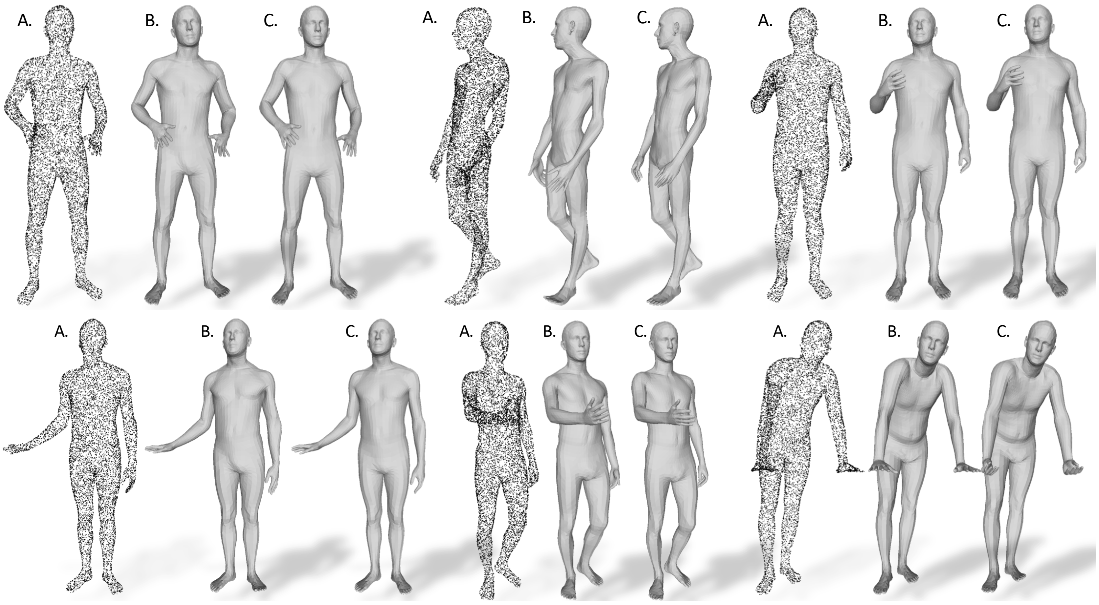


Figure 6.6: Qualitative results of our approach on undressed scans. We show A) input point cloud, B) SMPL registration from our approach and C) GT.

REGISTERING DRESSED SCANS. In Fig. 6.5 we show results with dressed scans and report an avg. surface-to-surface error after fitting the SMPL+D model of 2.2mm (Ours) vs 2.9mm (Alldieck et al., 2019a; Lazova et al., 2019; Zhang et al., 2017b). It can be clearly

Supervised %	**0%**	**10%**	**25%**	**50%**	**75%**	**100%**
(a) v2v (cm)	16.0	12.4	11.5	8.3	5.4	1.5
(b) s2s (mm)	13	7.8	8.5	7.7	6.4	4.2

Table 6.3: We study the effect of reducing the amount of available supervised data. Here 100% corresponds to 1000 scans used for supervised warm-start. We use additional 1631 scans for unsupervised training. We report vertex-to-vertex (v2v) and bi-directional surface-to-surface (s2s) errors.

Method	**Inter-class AE** (cm)	**Intra-class AE** (cm)
FMNet	4.83	2.44
FARM	4.12	2.81
LBS-AE	4.08	2.16
3D-CODED	2.87	1.98
Ours	**2.66**	**1.34**

Table 6.4: Comparison with existing correspondence prediction approaches. Our registration method clearly outperforms the existing supervised (Groueix et al., 2018; Or et al., 2017) and unsupervised (Li et al., 2019a; Marin et al., 2020) approaches.

seen that without pre-computed joints, prior approaches perform quite poorly, especially for complex poses. We quantitatively corroborate the same (for undressed scans) in Table 6.1. More qualitative results from our method can be seen in Fig. 6.7

6.3.3 *Comparison with existing learning based approach*

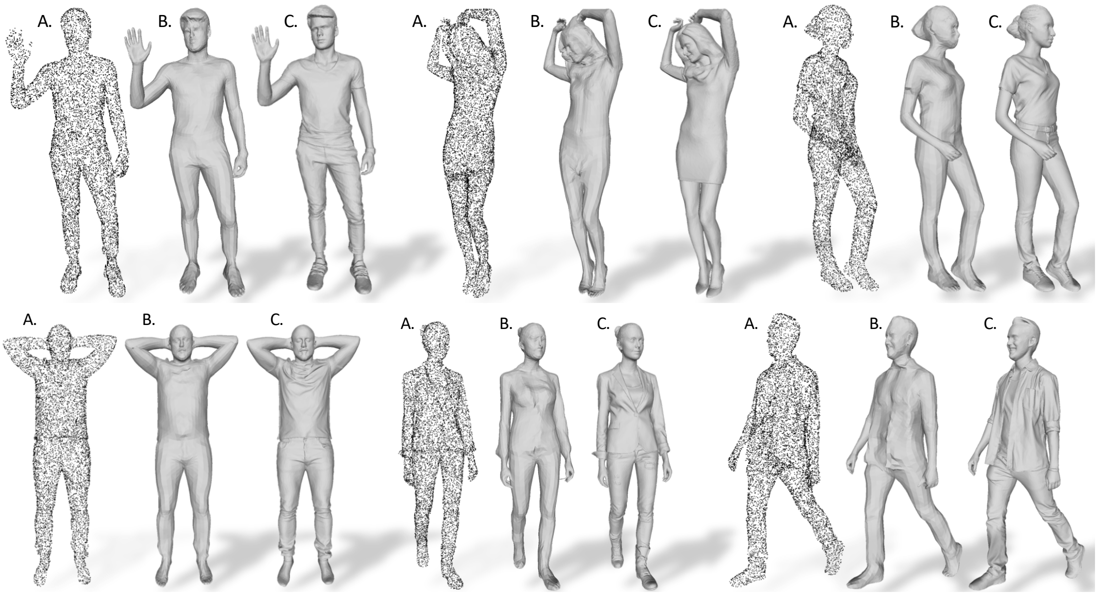


Figure 6.7: Qualitative results of our approach on dressed scans. We show A) input point cloud, B) SMPL+D registration from our approach and C) GT scan.

Conceptually, we found the work by Groueix et al. (2018) very related to our work, even though they require supervised training. For a fair comparison, we retrain their networks on our dataset and compare the registration error against our approach. Quantitative results in Table 6.1 clearly demonstrate the better performance of our approach. We also found that Groueix et al. (2018) is susceptible to bad initialization and hence requires multiple (global rotation) intializations for ideal performance. We report these numbers also in Table 6.1. Originally, the method in Groueix et al. (2018) was trained on SURREAL (Varol et al., 2017) with augmentation yielding a significantly larger dataset than ours. It is possible that the method of Groueix et al. (2018) requires a lot of data to perform well. Importantly, the supervised approach Groueix et al. (2018) does not deal with dressed humans whereas our approach works well for both dressed and undressed scans.

6.3.4 *Correspondence prediction*

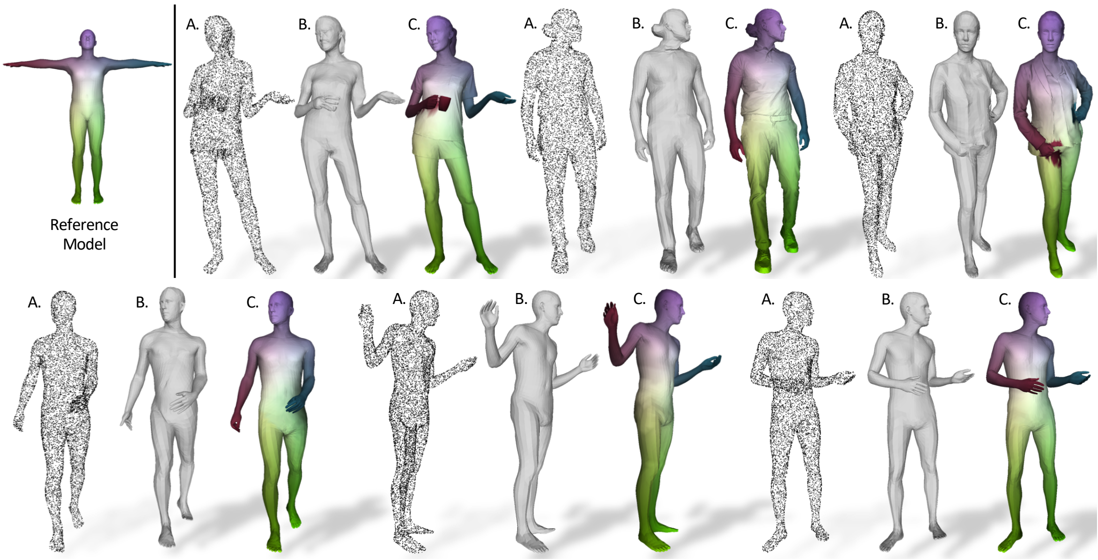


Figure 6.8: Our method predicts continuous correspondences between the input scan and the human model in canonical pose and shape. We visualize these correspondences here. Top row shows the reference human model and our predicted correspondences for dressed scans. In the bottom row we show the same for undressed humans. A) Input point cloud, B) our SMPL+D/SMPL registration and C) our predicted correspondences.

Establishing correspondences across 3D shapes is a challenging problem in computer graphics and vision community. Though our work does not directly predict correspondences between two shapes we can still register the two shapes with a common template. This allows us to establish correspondences between the shapes. We compare the performance of our approach on the correspondence prediction task on FAUST (Bogo et al., 2014). The FAUST test set contains 200 scans of undressed people in challenging poses and the scans themselves are noisy. The evaluation metric is based on the geodesic distance between the predicted correspondence and the GT correspondence. This metric heavily penalises the errors made due to self contacts on the body. In practice it can be seen that the overall distribution of errors is dominated by the contact errors making the evaluation less than ideal. Nonetheless we report the results as per the protocol in Table 6.4. For competing approaches we take the numbers from the corresponding papers. It can be clearly seen that our model trained primarily with self-supervision performs better than the competing approaches. Also note that none of these other approaches generalises to dressed humans where as in Fig. 6.8 we show that our method can predict correspondences for both dressed and undressed scans.

6.3.5 *Importance of our semi-supervised training*

ADDING MORE UNSUPERVISED DATA IMPROVES REGISTRATION. An advantage of our approach over existing approaches is the self-supervised training. We use a small amount of supervised data to warm start our method and subsequently the performance can be improved by throwing in raw scans (see Table 6.2). It can be clearly seen that performance improves significantly as more and more unlabeled scans are provided to our method.

IMPORTANCE OF SUPERVISED WARM-START. Good initialization is important for our network before it can adequately train using self-supervised data. In Table 6.3 we demonstrate the importance of good initialization using a supervised warm start. Note that we use only 1000 scans for supervised warm start where as methods such as 3D-CODED (Groueix et al., 2018) require an order of magnitude more supervised data for optimal performance.

6.4 LIMITATIONS AND FUTURE WORKS

Our formulation allows us to jointly differentiate through the correspondences and the instance specific human model parameters. This allows us to create a self-supervised loop for registration. But in practice we find that in order for this loop to not get stuck in a local minima, it is important that correspondences are initialized well. So even though our formulation does not require labeled data, in practice we find that a supervised warm-start with a small amount of data is important for subsequent self-supervised training.

As shown in our results, our approach performs significantly better than other competing approaches both qualitatively and quantitatively. We still find that our registration is not as high quality as Alldieck et al. (2019a) and Lazova et al. (2019) when they have access to precomputed 3D joints, facial landmarks and manual intervention (Note that our approach does not require this information). This is not necessarily a limitation as these additional cues can be integrated with our approach as well. We leave this as a potential future work.

6.5 CONCLUSIONS

We propose *LoopReg*, a novel approach to semi-supervised scan registration. Unlike previous work, our formulation is end-to-end differentiable with respect to both the model parameters and a learned correspondence function. While most of the current state of the art registration is based on instance specific optimization, our method can leverage information across a corpus of unlabeled scans. Experiments show that our formulation outperforms existing optimization and learning-based, approaches. Moreover, unlike prior work, we do not rely on additional information such as precomputed 3D joints or landmarks for each input, although these could be integrated in our formulation, as additional objectives, to improve results. Our second key contribution is representing parametric model as zero levelset of a distance field which allows us to diffuse the model function from the model surface to entire $\mathbb{R}^3$. Our formulation based on this representation can be useful for a wide range of methods as it removes the pre-requisite of computing a 2D surface parameterization. In contrast, we make predictions in the unconstrained $\mathbb{R}^3$ and subsequently map them to the model surface while still preserving differentiability. This makes our formulation easy to use, and potentially relevant for future work in learning based model fitting and correspondence prediction.

7

BEHAVE: DATASET AND METHOD FOR TRACKING HUMAN OBJECT INTERACTIONS

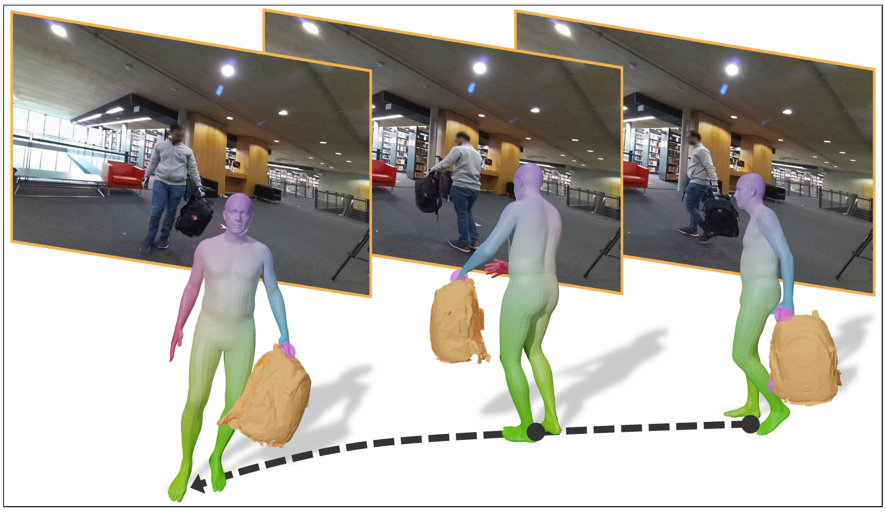


Figure 7.1: Given a multi-view RGBD sequence, our method tracks the human, the object and their contacts in 3D.

Modelling interactions between humans and objects in natural environments is central to many applications including gaming, virtual and mixed reality, as well as human behavior analysis and human-robot collaboration. This challenging operation scenario requires generalization to vast number of objects, scenes, and human actions. Unfortunately, there exist no such dataset. Moreover, this data needs to be acquired in diverse natural environments, which rules out 4D scanners and marker based capture systems. We present BEHAVE dataset, the first full body human-object interaction dataset with multi-view RGBD frames and corresponding 3D SMPL and object fits along with the annotated contacts between them. We record ∼15k frames at 5 locations with 8 subjects performing a wide range of interactions with 20 common objects. We use this data to learn a model that can jointly track humans and objects in natural environments with an easy-to-use portable multi-camera setup. Our key insight is to predict correspondences from the human and the object to a statistical body model to obtain human-object contacts during interactions. Our approach can record and track not just the humans and objects but also their interactions, modeled as surface contacts, in 3D. By combining optimization and learning techniques, we provide a solution to acquire annotated 3D human-object interaction data. Our code and data can be found at: http://virtualhumans.mpi-inf.mpg.de/behave.

7.1 INTRODUCTION

The last decade has seen rapid progress in modelling the appearance of humans ranging from body pose, shape (Loper et al., 2015; Pavlakos et al., 2019; Pons-Moll et al., 2017, 2015; Xu et al., 2020), faces (Tewari et al., 2017) and even detailed clothing (Alldieck et al., 2017, 2018a;

Bhatnagar et al., 2019; Patel et al., 2020; Saito et al., 2020). With various practical use cases like virtual try-on, personalised avatar creation, and several applications in augmented and mixed reality, or human-robot collaboration, the focus on humans is justified. In order to achieve the larger goal of building holistic human models, making the virtual humans look real is necessary but not sufficient. Human beings are much more than just their appearance with remarkable ability to perceive and act upon their surrounding environment. Our models of virtual humans should be able to do the same. Beyond modelling appearance, few methods have focused on capturing and synthesizing *human interactions* (human-object/scene interaction). There exists work to capture humans in a static 3D scene (Hassan et al., 2019), even without using external cameras (Guzov et al., 2021), and work to synthesize static poses (Hassan et al., 2021b; Li et al., 2019b), or full body movement (Hassan et al., 2021a; Ling et al., 2020; Starke et al., 2019) in a 3D scene.

These methods show growing interest in modelling human behavior, highlighting a need to capture real human interactions. Existing methods (Hassan et al., 2021a; Starke et al., 2019) however are learned from high quality curated data captured using optical marker based motion capture systems or wearable sensors. Unfortunately, such commercial systems are expensive, drastically limit the interactions that can be captured, and often fail when tracking humans and objects under occlusion. In addition, the recording volume is spatially confined and difficult to re-locate, thus limiting the activities, scenes, and objects that can be captured. Wearable sensors (Guzov et al., 2021) are not restricted in volume, but close range interaction can not be accurately captured. Altogether, the lack of diverse 3D interaction data, and the lack of accurate and flexible capture methods both constitute barriers in modelling human behavior.

With the goal of simplifying the data capture process and hence allowing faster progress in the field, we propose BEHAVE, a method to capture diverse 3D human interactions in natural environments, using a setup comprising of portable, cheap, and easy to use RGBD cameras. Tracking human interactions from sparse consumer grade cameras is however extremely challenging. Depth data is inherently noisy and incomplete. Moreover, the person and object occlude each other frequently during interactions. Furthermore, capturing interactions requires estimating human-object contacts accurately, which is difficult because contacts represent small regions in the image, close to the observable (resolution) limit. This requires innovation that goes significantly beyond the current state of the art trackers. We propose to track the human using a parametric human model (such as SMPL) and track objects using template meshes. Naively fitting the human model and an object 3D template to the point-cloud completely fails due to the aforementioned challenges. Our key idea is to train a neural model which jointly completes the human and object shape, represented with implicit surfaces, while predicting a correspondence field to the human, as well as an object orientation field. These rich outputs allow us to formulate a powerful human-object fitting objective which is robust to missing data, noise and occlusion.

To train and evaluate BEHAVE, we capture the *largest* dataset of human-object interactions in natural environments. The BEHAVE dataset contains 20 3D objects, 8 subjects (5 male, 3 female), 5 different locations and totals around 15.2k frames of recording. We provide ground truth SMPL and 3D object meshes as well as contacts.
Our contributions can be summarized as follows:

- We propose the first approach that can accurately 3D track humans, objects and contacts in natural environments using multi-view RGBD images.

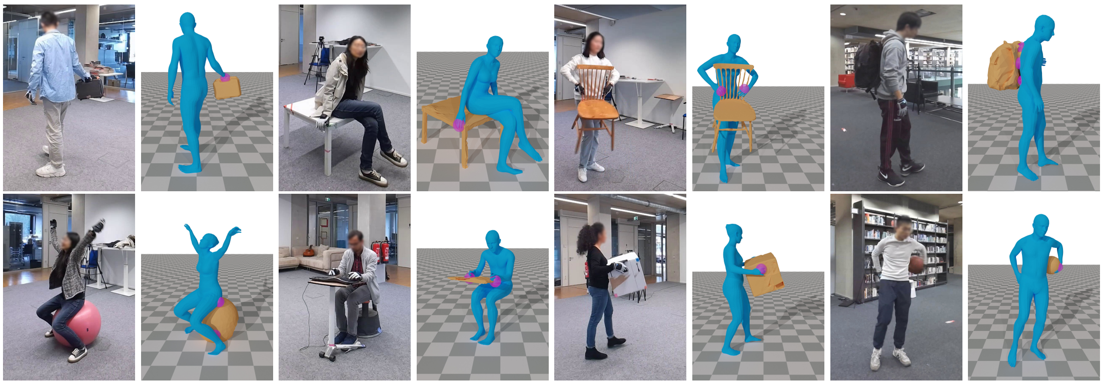


Figure 7.2: We present BEHAVE dataset, the *largest* dataset of human-object interactions in natural environments. BEHAVE contains multi-view RGBD sequences and corresponding 3D object and SMPL fits along with 3D contacts.

Dataset	**RGBD**	**Hum.**	**Ob.Cont.**	**Qual.**	**Scal.**
NTU (Liu et al., 2019)	✓	Jts.	X	NA	***
PiGr (Savva et al., 2016)	✓	Jts.	X	NA	**
GRAB (Taheri et al., 2020)	X	✓	✓	***	*
PROX (Hassan et al., 2019)	✓	✓	Stat.	*	**
Ours	✓	✓	✓	**	***

Table 7.1: We compare the proposed BEHAVE dataset with existing ones containing human-object interactions. Our criteria are based on availability of RGB input, 3D human, 3D contact with the object, quality (more stars, better), and scalability to capture at diverse locations (more stars, better). NTU-RGBD (Liu et al., 2019) and PiGraphs (Savva et al., 2016) do not provide full 3D human and object contacts and are hence unsuitable for modelling dynamic 3D interactions. GRAB (Taheri et al., 2020) uses a marker based capture system and hence contains the highest quality data but this also makes it difficult to scale. PROX (Hassan et al., 2019) is easier to scale as it uses a single Kinect based capture setup (although, scene needs to be pre-scanned) but this reduces the overall quality. More importantly it does not contain dynamic interactions. Ours is the first dataset that captures dynamic human-object interactions in diverse environments.

- We collect the *largest* dataset of multi-view RGBD sequences and corresponding human models, object and contact annotations. See Sec. 7.2 for details regarding its usefulness to the community.
- Since there exists no publicly available code and datasets to accurately track human-object interactions in natural environments, we will release our code and data for further research in this direction.

7.2 BEHAVE DATASET

We present BEHAVE dataset, the *largest* dataset of human-object interactions in natural environments, with 3D human, object and contact annotation, to date. See Table 7.1 for comparison with other datasets. Our dataset contains multi-view RGBD frames, with accurate pseudo-ground truth SMPL (Loper et al., 2015), object fits, human and object segmentation masks, and contact annotations.

7.2.1 *Recording multi-view RGBD data*

We setup and calibrate 4 Kinects at 4 corners of our square recording volume where all interactions are performed by 8 subjects (5 male, 3 female). Interactions are captured at 5 disparate indoor locations with 20 commonly used, yet diverse objects: 5 different boxes, 2 chairs, 2 tables, crate, backpack, trashcan, monitor, keyboard, suitcase, basketball, exercise ball, yoga mat, stool and a toolbox. We include common interactions such as lifting, carrying, sitting, pushing and pulling with hands and feet, as well as free interactions. In total, our dataset contains 10.7k frames for training and 4.5k frames for testing respectively.

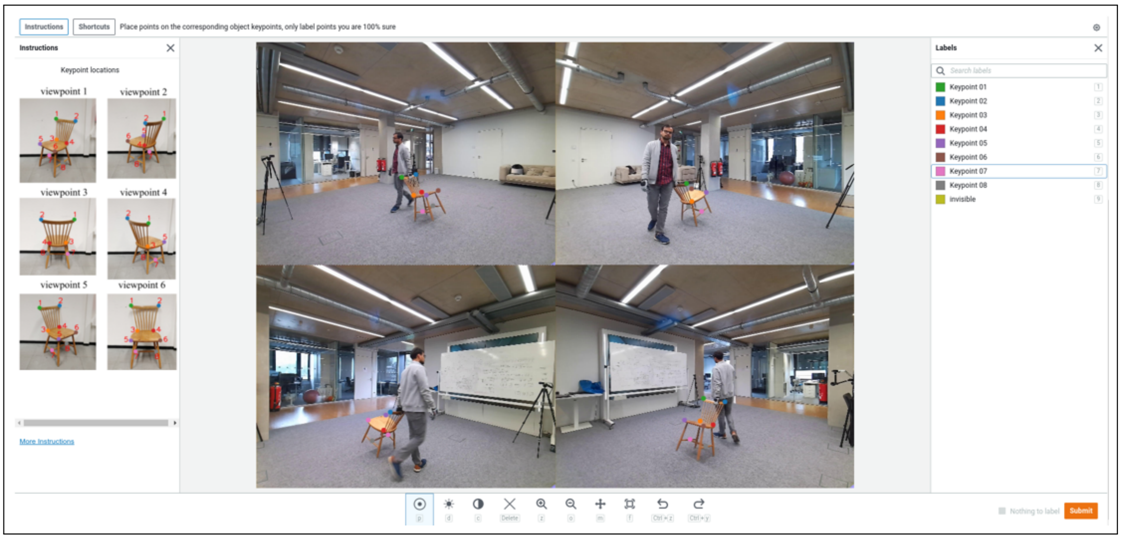
MULTI-CAMERA CALIBRATION. We use checkerboard to calibrate the relative poses between different Kinects in a pairwise manner. Specifically, we capture 20 pairs of RGB-D images from two Kinects and then register each color image with corresponding depth image such that they have the same resolution. We then use OpenCV to extract the checkerboard corners in the color images and obtain their 3D camera coordinates utilizing the registered depth map. Finally, we perform a Procrustes registration on these ordered 3D checkerboard corners to obtain the relative transformation between two Kinects. We obtain 3 pairs of relative transformation for 4 Kinects and combine them to compute the transformation under a common world coordinate.

MULTI-CAMERA SYNCHRONISATION. Both color and depth videos are captured at 30fps. Kinect cameras are synchronized through audio cables and the exact capture time of each image is saved for later processing. We extract synchronized frames and run depth-color registrations so that each depth image has the same resolution as color image. Frames are extracted at 10fps for better SMPL registration but our manual annotation is done at 1fps to maximize the diversity of annotated frames.

7.2.2 *Human segmentation and SMPL fitting*

HUMAN SEGMENTATION. We segment the human in our images by running DetectronV2 (Wu et al., 2019) followed by manual correction with Sofiiuk et al. (2020) on the segmentation masks. More specifically, we ask AMT workers to place 3-6 points on the erroneous segmentation regions. We use these clicks and run the interactive segmentation method Sofiiuk et al. (2020) to compute the corrected mask. We finally manually go over all corrected masks to filter out noisy segmentation. These corrected segmentation masks are then used to segment multi-view depth maps and lift human point clouds from 2D to 3D.

FITTING SMPL TO MULTI-VIEW KINECT POINT CLOUDS. We use FrankMocap (Rong et al., 2021) to initialize SMPL pose from the multi-view images and then use instance specific optimization (Alldieck et al., 2019a) to fit the SMPL model to the segmented human point cloud. For more accurate fitting, we additionally obtain the SMPL shape parameters of each subject from 3D scans using Bhatnagar et al. (2020a). We report a chamfer error of 1.80cm between the segmented Kinect point cloud and our SMPL fits.



Place points on the corresponding object keypoints, only label points you are 100% sure

Figure 7.3: Sample frame of our object keypoint annotation. Our predefined object keypoints are shown in the left, the frame to be annotated is shown in the middle. Annotators are asked to place keypoint labels (right panel) at their corresponding locations in the image.

7.2.3 *Object segmentation and fitting*

To obtain object segmentation, we pre-scan objects using a 3D scanner [1] [2]. We then use multi-view object keypoints, marked manually by AMT [3] annotators in images, to optimize the 6D pose of the pre-scanned object mesh to the given frame. We obtain the chamfer error of 2.42cm between the segmented Kinect point cloud and object fit.

ANNOTATING OBJECT KEYPOINTS. For each object, we pre-define 4-8 keypoints depending on the complexity of the object geometry. We show multi-view images of the captured human-object interaction frame together with example photos of our predefined object keypoints to AMT annotators and ask them to annotate only the visible keypoints in images, see Fig. 7.3. To ensure the annotation quality, we also manually go over all frames to filter out bad annotations. For all our AMT annotations, we first run a test batch to select good annotators and release the full batch only to the selected annotators.

OBJECT REGISTRATION AND SEGMENTATION MASK. Given template object meshes and annotated 2D keypoints from multi-views, we register the mesh to the images by optimizing the re-projection loss, similar to the method used in Pix3D registration Sun et al., 2018. To favor convergence, we down-sample the original scans to around 2000 faces and initialize the translation parameter as the center of human point clouds.
The segmentation masks are then obtained by projecting fitted object meshes to the images.

7.2.4 *Contact annotation*

Based on the pseudo-GT SMPL and object fits as described above, we automatically detect contacts if a point on the human surface (registered SMPL) is closer than 2cm to the object

[1] https://www.treedys.com
[2] https://www.agisoft.com
[3] https://www.mturk.com

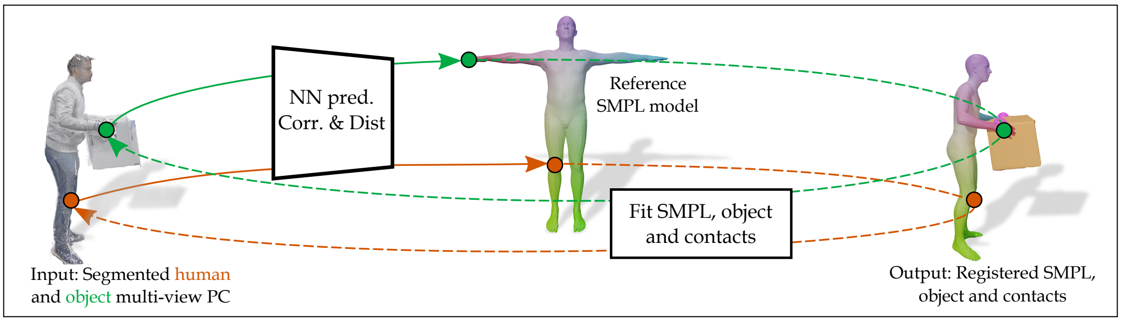


Figure 7.4: Given a sequence of multi-view images, we track the human and the object using SMPL and a template object mesh. We lift the segmented multi-view RGBD frames to 3D and obtain a human and object point cloud. As shown here, our network predicts correspondences between the human point cloud and the body model, which allows us to fit SMPL. We also predict correspondences from the object to the body model, thus allowing us to model contacts. Our network predictions (see Sec. 7.3) allow us to register SMPL and the object mesh to a video, making an accurate joint tracking of the human and object possible.

surface. For every object point, we store a binary contact label (whether there is a contact or not) and correspondence to the human (contact location on the surface).

7.2.5 *How will this dataset be useful to the community?*

We devote significant effort in recording the *largest*, so far, dataset of natural, full body, day-to-day human interactions with common objects in different natural environments. We propose following tasks/ challenges with BEHAVE dataset:

- **Tracking human-object interactions.** Track humans and objects using multi-view RGBD data. This can further be extended to track with just multi-view RGB, no-depth, and eventually just a single camera.
- **Reconstruction from a single image.** Joint 3D reconstruction of 3D humans and objects from a single RGB image. Currently, there is no dataset that can be used for benchmarking let alone to learn such a model.
- **Pose and shape estimation.** Benchmarking pose and shape estimation methods in challenging natural environments where the person is heavily occluded by the interacting object.
- **Motion synthesis.** BEHAVE dataset covers a diverse range of natural human-object interactions. This can be used to synthesise full-body human-object interaction motion.

Apart from these tasks, the research community is free to explore other applications of the BEHAVE dataset.

7.3 BEHAVE: METHOD

We present BEHAVE, a method to jointly track humans, objects and their interactions (represented as surface contacts) based on multi-view RGBD input. We formulate our method as an extended per-frame registration problem: we register the human (using SMPL) and the object (using its pre-scanned object mesh), and predict contacts as correspondences between SMPL and object meshes. See Sec. 7.4 for the overview of our method.

Our formulation must obey three properties, (i) the SMPL model $M(\cdot)$ should fit the human in the multi-view input, (ii) the object mesh $\mathbf{W}^o$ should fit the input object and, (iii) SMPL model and object should satisfy contacts.

To facilitate joint reasoning of the human, object and contacts directly in 3D, we lift the human $\mathcal{S}^h$, and object $\mathcal{S}^o$ point clouds to 3D using multi-view depth and semantic segmentation. Our joint formulation fits SMPL $M(\cdot)$ and the object $\mathbf{W}^o$ to multi-view RGB-D data at each time step, using explicit contacts. This takes the following form

$$E(\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{R}^o,\mathbf{t}^o) = E^{\text{SMPL}}(\mathcal{S}^h, M(\boldsymbol{\theta},\boldsymbol{\beta})) + E^{\text{obj}}(\mathcal{S}^o,\mathbf{W}^o) + E^{\text{contact}}(\mathbb{1}^c\mathbf{W}^o, M(\boldsymbol{\theta},\boldsymbol{\beta})). \quad (7.1)$$

The SMPL model is parameterized by pose $\boldsymbol{\theta}$, and shape $\boldsymbol{\beta}$. For notation brevity, we include the global SMPL translation into the pose parameters. We assume the template object $\mathbf{W}$ be rigid and only estimate the rotation $\mathbf{R}^o$, and translation $\mathbf{t}^o$, to fit the object mesh $\mathbf{W}^o = \mathbf{R}^o\mathbf{W} + \mathbf{t}^o$, to the object point cloud.

The indicator matrix, $\mathbb{1}^c$, selects the vertices on the object mesh $\mathbf{W}^o$, that are in contact with the SMPL model. This ensures that contact locations on the object and the human mesh adequately align in 3D.

The term $E^{\text{SMPL}}(\mathcal{S}^h, M(\boldsymbol{\theta},\boldsymbol{\beta}))$ is designed to accurately fit SMPL to the human point cloud $\mathcal{S}^h$. The term $E^{\text{obj}}(\mathcal{S}^o,\mathbf{W}^o)$ is designed to fit the object mesh to the object point cloud and $E^{\text{contact}}(\mathbb{1}^c\mathbf{W}^o, M(\boldsymbol{\theta},\boldsymbol{\beta}))$ ensures that contacts between the human and object match (align). We explain each term in detail next.

7.3.1 *Fitting human model to the human point cloud*

Fitting SMPL to the human point cloud $\mathcal{S}^h$ requires, (i) that distance between the SMPL model and the human point cloud should be minimized and (ii) the correct SMPL parts fit the corresponding body parts of the point cloud. The latter is important to avoid degenerate cases such as 180° flipped fitting, where the left hand is erroneously matched to the right side of the body or vice-versa (Bhatnagar et al., 2020a). With these considerations, we design our SMPL fitting objective as:

$$E^{\text{SMPL}} = \boldsymbol{d}(\mathcal{S}^h, M(\boldsymbol{\theta},\boldsymbol{\beta})) + E^{\text{corr}} + E^{\text{reg}}, \quad (7.2)$$

where $\boldsymbol{d}(\mathcal{S}^h, M(\boldsymbol{\theta},\boldsymbol{\beta}))$ minimizes the point-to-mesh distance between the input human point cloud $\mathcal{S}^h$ and the SMPL model. To avoid sub-optimal local minima during fitting (Bhatnagar et al., 2020a,b), we train a neural network that predicts dense correspondences from the input to the SMPL model. This ensures that correct SMPL parts explain corresponding input regions, using the term E^{corr}.

Specifically, we train an encoder network similar to Chibane et al. (2020a,b) that takes the segmented and voxelized human $\mathcal{S}^h$ and object $\mathcal{S}^o$ point cloud as inputs, and generates a voxel aligned grid of features $\mathbf{F} = f^{\text{enc}}_{\phi}(\mathcal{S}^h, \mathcal{S}^o)$. We then sample N 3D query points, $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}, \mathbf{p}_i \in \mathbb{R}^3$ and for each point $\mathbf{p}_i = (x,y,z)$ obtain the corresponding point feature $\mathbf{F}_i = \mathbf{F}(x,y,z)$. We pass this point feature through a decoder network f^{udf}_{ϕ}, to predict the unsigned distance to object and human surfaces, $u^o_i, u^h_i = f^{\text{udf}}_{\phi}(\mathbf{F}_i), u^o_i, u^h_i \in \mathbb{R}$, respectively. We use a second decoder network f^{corr}_{ϕ}, to predict the correspondence of point $\mathbf{p}_i$ to the SMPL model, $\mathbf{c}_i = f^{\text{corr}}_{\phi}(\mathbf{F}_i), \mathbf{c}_i \in \mathbb{R}^3$.

E^{corr} enforces that the distance between the input point $\mathbf{p}_i$ and the corresponding point $\mathbf{c}_i$ after transforming it with the SMPL model is same as the distance predicted by the network u_i^h. Under a slight abuse of notation we use $M(\mathbf{c}_i, \boldsymbol{\theta}, \boldsymbol{\beta})$ to transform $\mathbf{c}_i$ with the SMPL function.

$$E^{\text{corr}} = \sum_{i=1}^{N} ||\mathbf{p}_i - M(\mathbf{c}_i, \boldsymbol{\theta}, \boldsymbol{\beta})|_2 - u_i^h|. \tag{7.3}$$

If the correspondences predicted by the network $\mathbf{c}_i$ deviate from the SMPL surface, these cannot be skinned using the SMPL model as its function is only defined on the body surface. To alleviate this issue, we use the LoopReg (Bhatnagar et al., 2020b) formulation that allows us to pose and shape off-the-surface correspondences as well.

The final term $E^{\text{reg}} = E^{\text{J2D}} + E^{\boldsymbol{\theta}} + E^{\boldsymbol{\beta}}$, adds regularisation for SMPL joints. $E^{\text{J2D}} = \sum_{k=1}^{K} |\pi_k M^J(\boldsymbol{\theta}, \boldsymbol{\beta}) - \mathbf{J}_{2D}^k|_2$, where π_k is the camera projection matrix of camera k, $M^J(\cdot)$ are the 3D body joints and $\mathbf{J}_{2D}^k$ are the 2D joints detected in the k^{th} Kinect image. $E^{\boldsymbol{\theta}}$ and $E^{\boldsymbol{\beta}}$ are regularisation terms on SMPL pose and shape similar to Bogo et al. (2016).

7.3.2 *Fitting the object mesh to the object point cloud*

In order to fit the object mesh, we must ensure that distance from the input object point cloud to the object mesh is minimized. Minimizing this one-sided distance is necessary but not sufficient. Since severe occlusions are common in our interaction setting, large parts of object might be missing from the object point cloud, making fitting difficult. To alleviate this issue we must also ensure that all the vertices of the object mesh are correctly placed w.r.t. the input, even when the point cloud is incomplete. To do so, we take the object mesh vertices $\mathbf{v}_j^o \in \mathbf{W}^o, j \in \{1, \ldots, L\}$ and obtain the corresponding point feature $\mathbf{F}_j$, same as Sec. 7.3.1, where L is the number of object mesh vertices. We then obtain the unsigned distances to the object and human surfaces using the point feature $u_j^o, u_j^h = f_\phi^{\text{udf}}(\mathbf{F}_j)$. Since $\mathbf{v}_j^o$ is a vertex on the object mesh, its distance to the object surface u_j^o must be zero for a correct fit. This allows us to accurately fit the object vertices to the point data even when corresponding parts are missing from the object point cloud.

$$E^{\text{obj}} = \boldsymbol{d}(\mathcal{S}^o, \mathbf{W}^o) + \sum_{j=1}^{L} |u_j^o|, \tag{7.4}$$

where, $\boldsymbol{d}(\mathcal{S}^o, \mathbf{W}^o)$ minimizes the point-to-mesh distance between the object point cloud and the object mesh, and the term $\sum_{j=1}^{L} |u_j^o|$ uses implicit unsigned distance prediction to reason about missing object parts.

PREDICTING OBJECT ORIENTATION. Although the terms in Eq. 7.4 minimize the bi-directional distance between the object point cloud and the object mesh, they do not guarantee that parts of the object point cloud are explained by the semantically corresponding parts on the object mesh, e.g. in Fig. 7.5, the legs of the table are not aligned correctly. This issue can be fixed if we obtain the global object orientation during fitting. We represent the orientation of the object with the principal components obtained by running PCA on the object vertices.

We train a neural network f_ϕ^a, that uses the point feature $\mathbf{F}_j$ (same as Sec. 7.3.1) corresponding to each query point $\mathbf{p}_j$ and predicts the global orientation of the object $\mathbf{a}_j = f_\phi^a(\mathbf{F}_j), \mathbf{a}_j \in \mathbb{R}^9$. We find that orientation prediction is unreliable if the query point is far from the object surface, hence we filter out points whose unsigned distance from the object surface u_j^o, is greater than a threshold $\epsilon = 2cm$. The global orientation of the object is obtained by averaging the orientation predictions from the filtered points, $\mathbf{a}^o = \frac{1}{M}\sum_{j=1}^{M} \mathbf{a}_j$ where M is the number of filtered points.

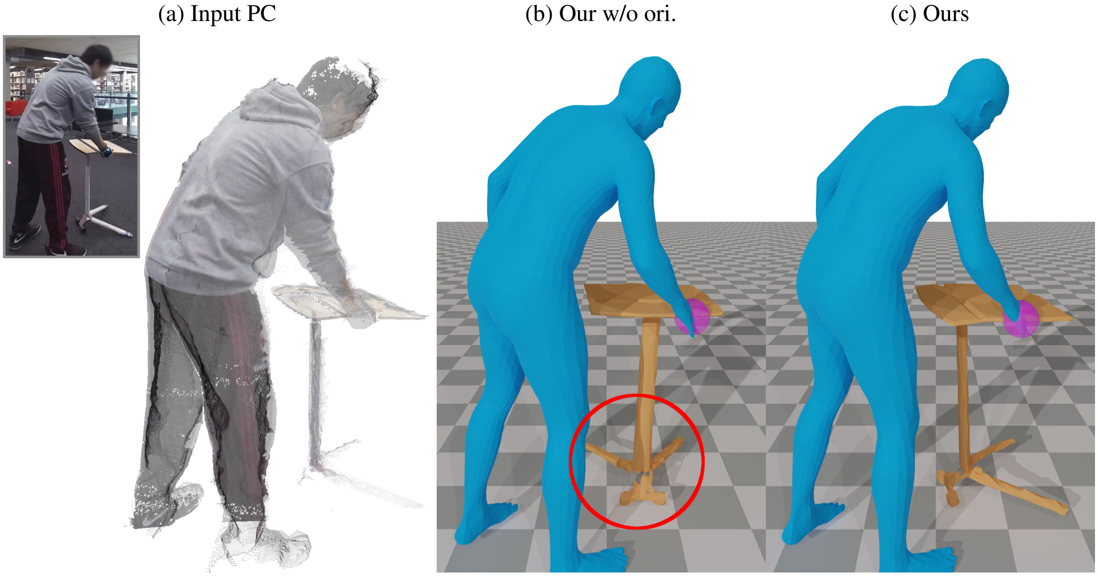


Figure 7.5: We show that our network predicted orientation is important for accurate object fitting. Without our orientation prediction the fitting gets stuck in a local minima.

Next, we compute the relative rotation between the current object orientation $\bar{\mathbf{a}}$ and the predicted object orientation $\mathbf{a}^o$, and use this to initialise the object rotation $\mathbf{R}^o = \mathbf{a}^o(\bar{\mathbf{a}}^T\bar{\mathbf{a}})^{-1}\bar{\mathbf{a}}^T$. We further run SVD on $\mathbf{R}^o$ and only keep the rotation matrix.
Initialising $\mathbf{R}^o$ with the network predicted object orientation is crucial to avoid local minima during object fitting, as can be seen in Fig. 7.5 and 7.3.

7.3.3 *Refining human & object models using contacts*

Our formulation above gives reasonably good human and object fits but does not ensure that human and object meshes satisfy the contacts predicted by the network. This often leads to floating objects and hovering hands see Fig. 7.6, as human and object models are not in contact. In this section we explicitly optimize the human and object meshes to fit the contacts predicted by the network. We model contacts as vertices in the registered object mesh $\mathbf{v}_j^o \in \mathbf{W}^o$, that are very close to the input human $u_j^h < \epsilon$ and object $u_j^o < \epsilon$ surface. Similarly to Sec. 7.3.2, we use f_ϕ^{udf} to obtain the unsigned distances u_j^o, u_j^h and f_ϕ^{corr} to obtain the correspondences $\mathbf{c}_j$ of these points to the SMPL model, respectively. In order to filter query points close to human and object surfaces we compute a binary indicator matrix $\mathbb{1}^c \in \mathbb{R}^N$ such that $\mathbb{1}_j^c = 1$ iff $u_j^o < \epsilon, u_j^h < \epsilon$.

$$E^{\text{contact}} = \sum_{j=1}^{N} \mathbb{1}_j^c |\mathbf{v}_j^o - M(\mathbf{c}_j, \boldsymbol{\theta}, \boldsymbol{\beta})|_2. \tag{7.5}$$

E^{contact} allows us to jointly optimise the SMPL model and the object parameters $\mathbf{R}^o, \mathbf{t}^o$ to satisfy the contacts predicted by the network.

7.3.4 *Network training*

In this section we elaborate on training our networks.

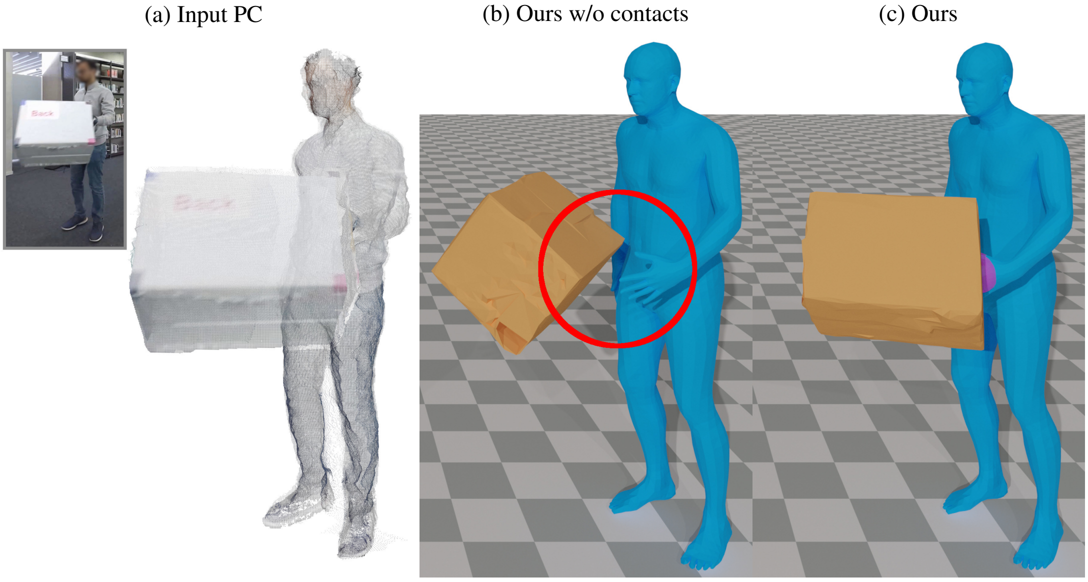


Figure 7.6: Without our network predicted contacts we observe artefacts like floating objects, leading to unrealistic tracking.

FEATURE ENCODING. We use a 3D CNN similar to IF-Net (Chibane et al., 2020a) to obtain a voxel aligned multi-scale grid of features $\mathbf{F} = f_\phi^{\text{enc}}(\mathcal{S}^h, \mathcal{S}^o)$.

UNSIGNED DISTANCE PREDICTION. To train the network f_ϕ^{udf}, we sample N query points $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ in 3D. For each query point $\mathbf{p}_j$ we obtain its point feature $\mathbf{F}_j$ (Sec. 7.3.1) and use this to predict the unsigned distance (Chibane et al., 2020b) to human and object surface $u_j^o, u_j^h = f_\phi^{\text{udf}}(\mathbf{F}_j)$.
We jointly train $f_\phi^{\text{enc}}, f_\phi^{\text{udf}}$ with standard L2 loss. The GT for u_j^o, u_j^h is easily available as our dataset contains GT SMPL and object fits allowing us to obtain GT distance of point $\mathbf{p}_j$ from the SMPL and object mesh.

SMPL CORRESPONDENCE PREDICTION. To train f_ϕ^{corr}, we use the point feature $\mathbf{F}_j$ of sampled query point $\mathbf{p}_j$ to predict its correspondence to the SMPL model $\mathbf{c}_j = f_\phi^{\text{corr}}(\mathbf{F}_j)$.
We jointly train $f_\phi^{\text{enc}}, f_\phi^{\text{corr}}$ using a standard L2 loss. Since we have the GT SMPL fit in our dataset we simply find the closest SMPL surface point for the query point $\mathbf{p}_j$ and use this as the GT correspondence.

OBJECT ORIENTATION PREDICTION. To train the network f_ϕ^a, we use the point feature $\mathbf{F}_j$ of a sampled query point $\mathbf{p}_j$ to predict the global object orientation $\mathbf{a}_j = f_\phi^a(\mathbf{F}_j)$.
We jointly train $f_\phi^{\text{enc}}, f_\phi^a$ with standard L2 loss. We find that points far away from the object surface are unreliable in predicting the object orientation. Hence we only apply this loss to points that are close to the object, i.e. the GT $u_j^o < \epsilon$. Since we have the GT object fit, we obtain the GT orientation by running PCA on the object mesh vertices and use the 3 principal axes in $\mathbb{R}^9$.

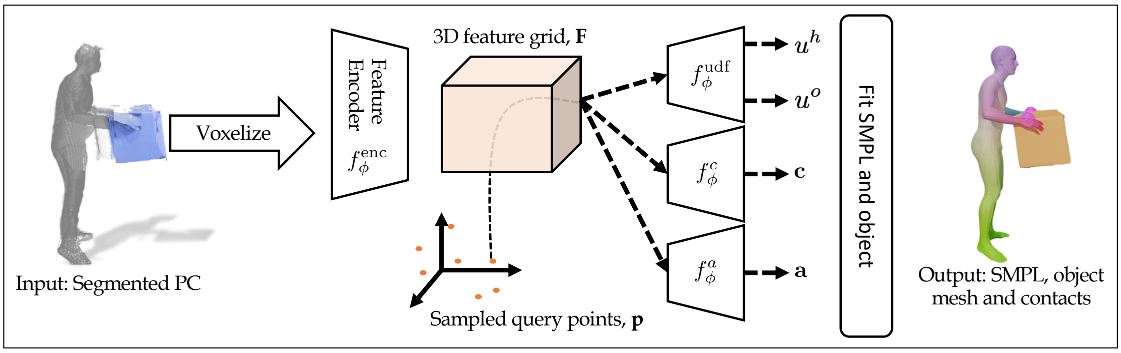


Figure 7.7: The BEHAVE network takes voxelized human and object point cloud as input and generates an input aligned 3D feature grid, $\mathbf{F}$. We then sample 3D query points and for each input point $\mathbf{p}$, we use the point feature $\mathbf{F}(\mathbf{p})$ to predict unsigned distance to human and object surfaces, u^h, u^o, correspondences to the SMPL model $\mathbf{c}$ and object orientation $\mathbf{a}$. We use these predictions to fit SMPL and object meshes to the input, explicitly taking the contacts between them into account.

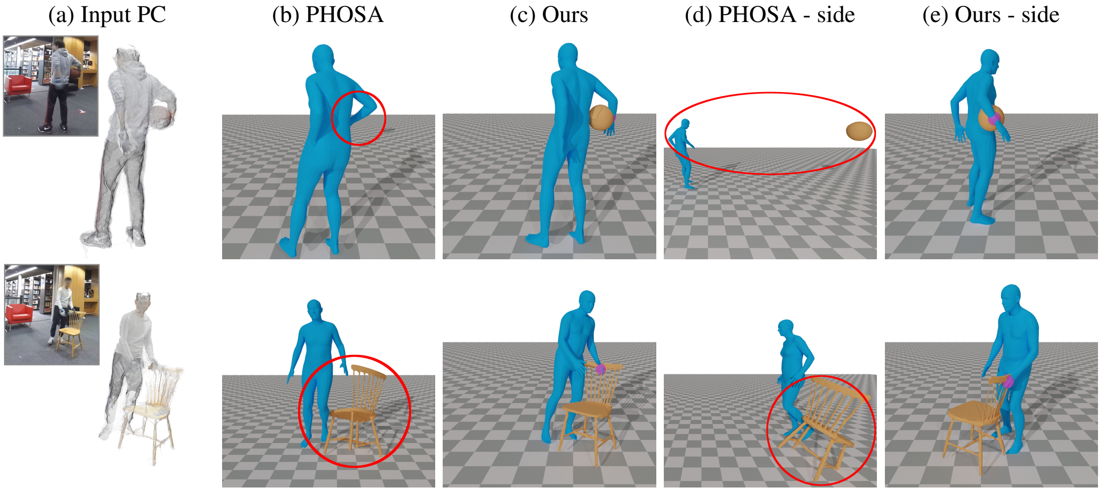


Figure 7.8: We compare our method to track human, object and contacts with PHOSA Zhang et al., 2020a. It can clearly be seen that our method can reason about the human-objects contacts and produces more accurate results.

7.3.5 *Implementation details*

The input to our method is multi-view segmented point cloud of the human and the object. We voxelize it and feed it to our feature encoder f_ϕ^{enc} to obtain a grid aligned set of features. The network f_ϕ^{enc} comprises of $4 \times \{2 \times \text{Conv3D} + \text{ReLU} + \text{BN}\}$ layers. We use a stride of 2 in our convolutional layers.
We then train three separate decoders $f_\phi^{\text{udf}}, f_\phi^{\text{corr}}$ and f_ϕ^{a} to predict (i) unsigned distances to the human u^h, and the object u^o, surfaces; (ii) correspondences to the SMPL model $\mathbf{c}$ and (iii) object orientation $\mathbf{a}$, respectively. We use the same architecture for all our decoders, $3 \times \{\text{Conv1D} + \text{ReLU}\}$, followed by a regression layer, Conv1D. Our network structure can be seen in Fig. 7.7.

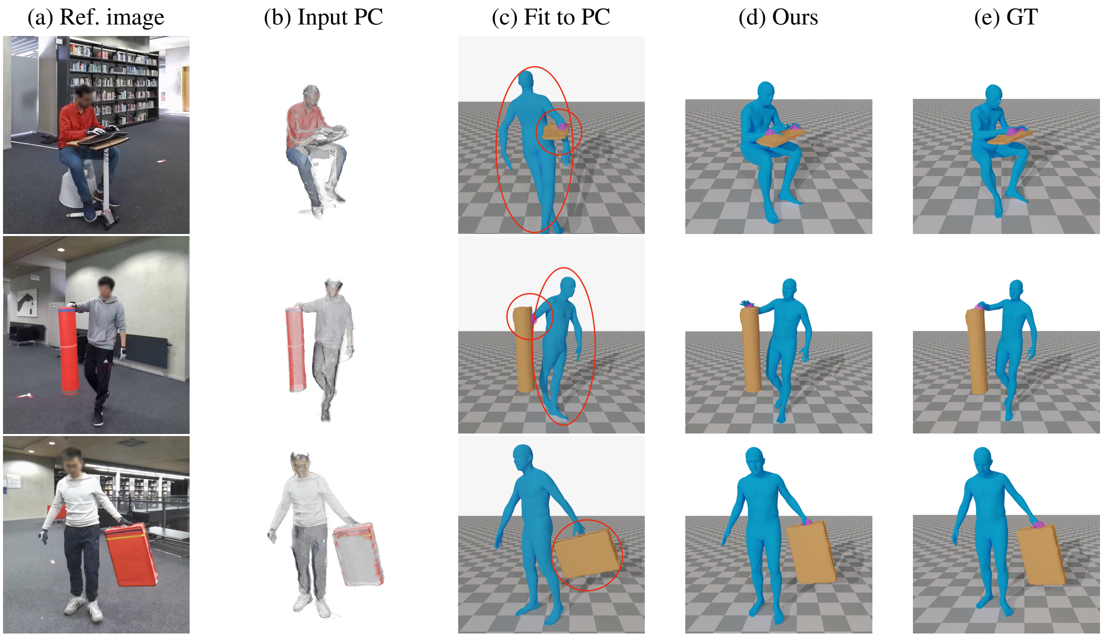


Figure 7.9: We show that registering SMPL and object meshes directly to input point cloud results in inaccurate fitting. Our neural predictions corresponding to human and object unsigned distance fields, SMPL correspondence field and object orientation field are key to good fitting.

Method	**SMPL v2v** (cm)	**Obj. v2v** (cm)
IP-Net (Bhatnagar et al., 2020a)	6.61	NA
LoopReg (Bhatnagar et al., 2020b)	9.12	NA
Fit to input	16.15	26.09
PHOSA (Zhang et al., 2020a)	13.73	34.73
Ours	**4.99**	**21.20**

Table 7.2: We compare our method to obtain SMPL and object fits with IP-Net, LoopReg and PHOSA. We also show that directly fitting SMPL and object meshes to the input leads to sub-optimal performance. Our method not only obtains better fits, but unlike LoopReg and IP-Net, we can also fit the object.

7.4 DATASET AND EXPERIMENTS

In this section we compare our approach with existing methods. Our experiments show that we clearly outperform existing baselines. Next, we ablate our design choices and highlight the importance of contact and object orientation prediction in capturing human-object interactions.

7.4.1 *Comparing with PHOSA*

We find PHOSA (Zhang et al., 2020a), a method to reconstruct humans and objects from a single image, quite relevant to our work. Although PHOSA uses only a single image whereas we use multi-view images, thus giving our method an advantage, it is still the closest competing method. We run Procrustes alignment on PHOSA results to remove depth ambiguity.

It should be noted that PHOSA depends on pre-defined fixed contact regions whereas our

Method	**SMPL v2v** (cm)	**Obj. v2v** (cm)
A) Ours w/o ori.	4.98	24.02
B) Ours w/o cont.	**4.96**	21.28
C) Ours	4.99	**21.20**

Table 7.3: We analyse the importance of (A) object orientation prediction and (B) contact prediction for our method. It can be seen that object orientation prediction noticeably improves object localisation error. The effect of contact loss is not significant quantitatively but makes noticeable difference qualitatively see Fig. 7.6.

approach can freely predict full-body contacts and PHOSA uses hand crafted heuristics to model contacts whereas our approach learns contact modelling from data, making our method more scalable. We compare our method with PHOSA in Fig. 7.8 and Table 7.2, and clearly outperform it.

7.4.2 *Why not fit human and object models directly to point clouds?*

Since there are no existing methods that can jointly track humans, objects and the contacts from a multi-view input, we create an obvious baseline where we fit the SMPL and object meshes directly to the input point cloud. We show (Table 7.2 and Fig. 7.9) that direct fitting easily gets stuck in local minima. This is because the point clouds are very noisy and large parts are missing due to heavy occlusion between the person and the object during interactions. Our network, on the other hand, can implicitly reason about missing parts, thus generating more accurate results.

7.4.3 *Why can't existing human registration approaches be extended to our setting?*

There are no direct baselines that can jointly track humans, objects, and contacts from multi-view input. There are works (Bhatnagar et al., 2020a,b) that pursue similar ideas of predicting correspondences and fitting SMPL to the human point cloud. In this subsection we explore their suitability in our setting.

COMPARISON TO IPNET. IPNet (Bhatnagar et al., 2020a) takes as input a human point cloud and predicts an implicit reconstruction of the human and sparse correspondences to the SMPL model, which enables its fitting to the implicit reconstruction.
This approach has three major disadvantages. First, querying occupancies for a 128^3 grid to obtain implicit reconstruction is expensive. Second, it predicts occupancies which requires water-tight surfaces. And third, running traditional Marching Cubes makes occupancy prediction non-differentiable w.r.t. SMPL fitting.
Our formulation in Eqs. 7.2 and 7.3 alleviates these problem as we can fit SMPL by only querying $N = 30k$ points instead of $128^3 (\sim 2M)$ points. Since we use unsigned distance prediction, our method can work with non-water tight surfaces. We can also fit SMPL directly to unsigned distance predictions, thus removing the requirement for Marching Cubes. We compare our approach with IP-Net (Bhatnagar et al., 2020a) (trained on our dataset) in Table 7.2 and Fig. 7.10. We show that we obtain better performance than IP-Net at much lower cost ($30k$ (ours) vs. $\sim 2M$ (IP-Net) query points and no Marching Cubes). This shows that our formulation

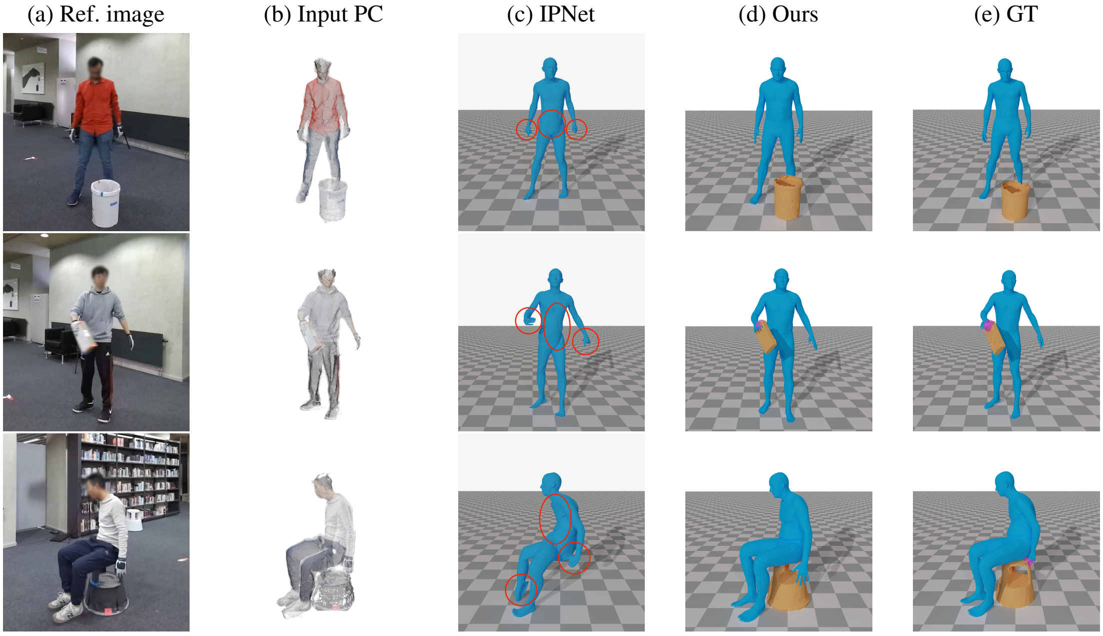


Figure 7.10: We compare our SMPL registration with IPNet (Bhatnagar et al., 2020a) and show superior results. IPNet cannot register objects that our approach can.

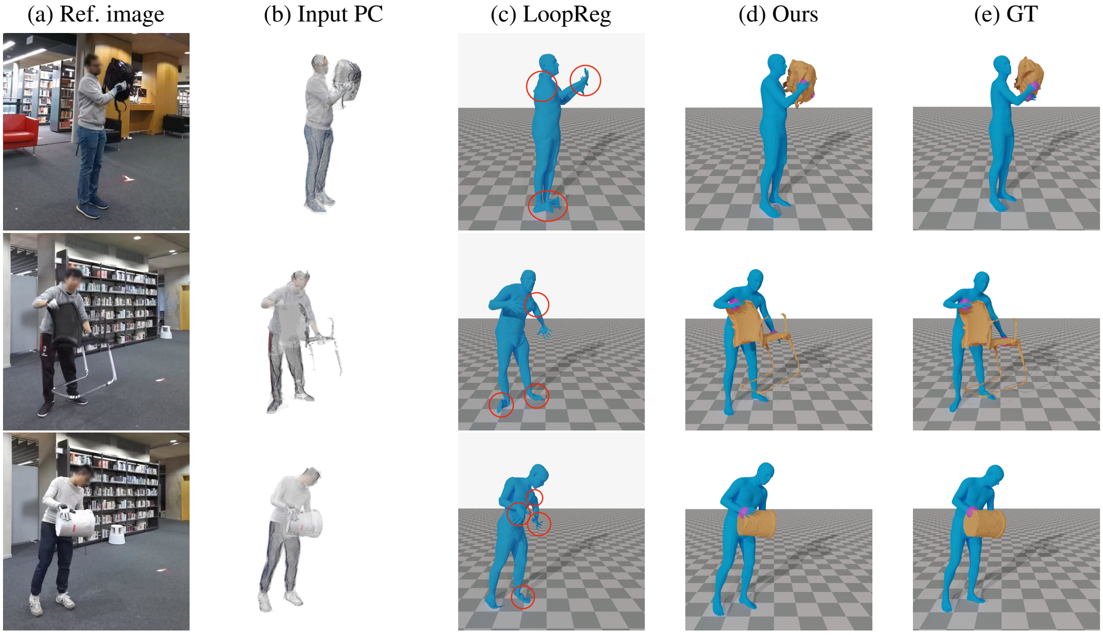


Figure 7.11: We compare our SMPL registration with LoopReg (Bhatnagar et al., 2020b) and show superior results. LoopReg cannot register objects that our approach can.

is superior than IPNet even for human registration. We can additionally handle objects and interactions. Qualitative comparisons are given in the supplementary material.

COMPARISON WITH LOOPREG. LoopReg (Bhatnagar et al., 2020b) fits SMPL to the input point cloud by explicitly predicting correspondences. We find the idea interesting and use their diffused SMPL formulation in our method. LoopReg is, however, not directly applicable in our setting as it assumes a noise free and complete human point cloud. When the point cloud is incomplete due to occlusions, no correspondences are predicted for missing parts. Since LoopReg can only use surface points for fitting, this makes registration inaccurate. BEHAVE handles this case by using distances to the SMPL surface(Eq. 7.3) predicted for each of the sampled query points to fit the body model, thus allowing the use of non-surface points for fitting. This is important as the Kinect point cloud is noisy. We outperform LoopReg (Bhatnagar et al., 2020b) (trained on our dataset) and show (Table 7.2 and Fig. 7.11) that our formulation is robust to missing parts and noisy input.

7.4.4 *Importance of contacts*

In this experiment we show that our network predicted contacts are key for physically plausible tracking. Even though quantitative difference is not significant (Table 7.3), it can be seen in Fig. 7.6 that without contact information, the human and the object do not lock into the correct location. Hence, we notice unnatural results like floating objects. Using our contact prediction alleviates such issues.

7.5 LIMITATIONS AND FUTURE WORKS

Although we outperform existing baselines, the research in capturing human-object interactions is still nascent. Our approach uses a set of predefined object templates, but a more flexible method would build models ‘on the fly’. Current research on fusion methods is promising and integrating it with our trackers is an interesting future direction.

Improving tracking performance is also important. Our object orientation prediction is reliable, but predicting object orientation for symmetric objects is ambiguous, leading to erroneous fitting sometimes. We also observe that since Kinect data is noisy, we cannot model fine grained hand interactions, this results in interpenetrations between the hand and the object and sometimes unrealistic grasps. More interesting limitation arises from the fact that we assume the objects to be rigid which is not the case in reality. Objects like backpacks when grabbed from the straps are not accurately registered as the deformations in this case are non-rigid. All these are challenging scenarios with no straightforward solution and these directions warrant further research.

It would be also useful to build single camera trackers for wider applicability. Our BEHAVE dataset can be a starting point in this quest.

7.6 CONCLUSIONS

We have presented BEHAVE, the first methodology to jointly track humans, objects and explicit contacts in natural environments. By introducing neural networks to predict correspondences to a 3D human body model along with unsigned distance fields defined over human and object surfaces, we are able to accurately model human-object contacts. We further integrate such neural predictions into a proposed joint registration method resulting in the robust 3D tracking of human-object interactions.

Along with our proposed method we also provide BEHAVE, the *largest* dataset of RGBD sequences and annotated humans, objects, and contacts to date. BEHAVE dataset is the *first* benchmark for the part of the research community interested in modelling human-object interactions. We propose real-world challenges like reconstructing humans and object from a single RGB image, tracking human-object interactions from multiple and single-view RGB(D) input, pose estimation etc. Our dataset together with our code is released in order to stimulate future research in this important emerging domain.

8 CONCLUSION AND FUTURE WORK

Digital models of humans are becoming increasingly popular due to real world applications such as gaming, telepresence, animated content creation and Metaverse to name a few. These models must not only look real but should also interact naturally with their surroundings. This goal requires modelling several characteristics of real humans such as our body shape, pose, clothing and also interactions. To this end, this thesis proposes several advancements in the form of Multi-Garment Network (MGN), Implicit Part Network (IPNet), LoopReg and BEHAVE.

In chapter 4, we propose the first method to reconstruct 3D clothing and body shape as separate meshes from a few RGB images. To train MGN we need data of people dressed in registered 3D garments. We propose a novel method to extract 3D garments from a 3D scan and register them to a common set of templates. SMPL model opened up several frontiers in modelling 3D humans but is limited to undressed body shape. Our garment registration allows extending the SMPL model (SMPL+G) to be easily dressed with garments, thus modelling garments on top of the SMPL model. We demonstrate this with our released digital wardrobe. Our SMPL+G garments can be easily posed and shaped with the SMPL model. Our SMPL+G model allows reconstructing garments on top of SMPL for the first time using just images. Our SMPL+G formulation is currently being used in several other tasks like garment texture generation (Mir et al., 2020), modelling pose dependent deformations (Patel et al., 2020), modelling garment size dependent deformations (Tiwari et al., 2020) etc.

Although quite powerful, the mesh based SMPL+G model is limited by mesh resolution and unable to capture high frequency garment details. Recently, implicit functions have become quite useful to produce high quality surfaces. Unfortunately they typically produce static surfaces that cannot be reposed or reshaped. This control is a requirement for a lot of downstream tasks such as animation. Therefore we propose Implicit Part Network (IPNet), in chapter 5, that not only leverages implicit functions for detailed 3D reconstruction but also registers them with the SMPL model. This makes our implicit reconstructions both *detailed* and *controllable* for the first time.

It is evident that, registering human meshes with the SMPL model has a lot of utilities, making 3D registration a work horse of several graphics and vision applications. This is further demonstrated in our work IPNet (chapter 5). In chapter 6, we further explore this decade old task of 3D registration and highlight that existing methods on 3D registration are not end-to-end differentiable with respect to the correspondence prediction. We propose a novel representation of SMPL model by diffusing the mesh based SMPL function to entire $\mathbb{R}^3$. This allows us to create a self-supervised registration loop and our experiments demonstrate that our semi-supervised approach outperforms existing supervised methods using 100x less supervised data.

In chapter 7 we present BEHAVE, the first dataset and method to track human-object interactions. BEHAVE dataset contains $\sim$15K multi-view RGBD frames of people interacting with various daily use objects. We also provide segmentation masks, corresponding SMPL and object fits in 3D as well as annotated contacts. We also release an extended BEHAVE dataset with $\sim$450K frames with pseudo-ground truth annotations. BEHAVE dataset can be used for tasks like tracking human-object interaction, reconstructing 3D human and object from a single image, pose and shape estimation under heavy occlusion, interaction motion synthesis etc. We use the BEHAVE dataset to learn the BEHAVE model, which is the first method that can track the human, the object and the contacts between them using multi-view input. Our BEHAVE dataset is publicly available and it has already been used for single image human-object interaction reconstruction (Xie et al., 2022).

8.1 KEY INSIGHTS

In this subsection, we discuss some high level insights that have shown to be quite useful, sometimes even beyond the scope in which they were proposed. This includes everything ranging from practical tips, low level details to high level ideas. We have tried to pool together insights from other works as well, that also support our claims.
Note: This section might contain subjective interpretation of some experimental results. We point to the relevant experiment in all the cases and the reader is encouraged to check out the detailed experiment in the main chapters.

8.1.1 *MGN (Bhatnagar et al., 2019)*

CHOICE OF REPRESENTATION. How we choose to represent garments is key and heavily application dependent. Mesh based representation, as used in MGN (Bhatnagar et al., 2019) is really powerful as it allows us to register garments with diverse poses and shapes to a common template. This template can be rigged, for instance with SMPL model, to allow animation and control of all the registered garments. Mesh based representation is also suitable for texture and geometry transfer as well as rendering. However, we found that predicting garments as meshes does not preserve high frequency details. We partially address the problem by decomposing the meshes into a ‘low frequency coarse shape’ and ‘high frequency displacements’ on top. This approach has shown to work well for other works on modelling pose dependent garment deformation (Patel et al., 2020) and reconstructing 3D clothes from images (Jiang et al., 2020). Implicit function based representation has shown to preserve better details although they are typically not controllable.

TEST TIME REFINEMENT. Given an input image, MGN would predict 3D garments in a forward pass. To improve the details during inference, we froze the network and optimized the latent representation to minimise the loss between the input image and the projection of our 3D prediction. This significantly improved the final results. This test time refinement (and its variants) have proven to useful in other works such as estimating pose and shape ‘in the wild’ (Kolotouros et al., 2019), reconstructing detailed 3D shape from images (Alldieck et al., 2019a), registering point clouds and scans (Bhatnagar et al., 2020a,b).

RECONSTRUCTION FROM SINGLE IMAGE VS VIDEO. MGN is flexible in its input and works with a single image or a sequence of images as input. Intuitively, more images should allow us to reconstruct better details but in practice we found that network would smooth out the details across images. The challenge was fusing details from different images (different poses and deformations) together. This is an interesting future direction.

8.1.2 *IPNet (Bhatnagar et al., 2020a)*

MAPPING BETWEEN IMPLICIT AND PARAMETRIC REPRESENTATION. As discussed above, the choice of representation for 3D shapes is critical and application dependent. Two of the most popular representations are mesh based and implicit, with each having well studied advantages/ disadvantages. Techniques described in chapter 5 can be seen as a way of learning a continuous mapping between implicit and mesh based representations. This can be very powerful when different use cases prefer different representations eg. implicit representations are more suitable for details but meshes are easier to render. Our formulation can allow a user to freely switch between representations depending on the use case while still maintaining a mapping between the two.

USING IMPLICIT FUNCTIONS TO LEARN CONTINUOUS FIELDS. Prior work has shown that implicit functions can represent 3D shapes using using occupancies/SDFs/UDFs etc. But implicit functions can be used to model continuous fields in general. IPNet uses implicit functions to learn a part correspondence field. This formulation is also used for model based fitting (Bhatnagar et al., 2020b). Similar ideas have been used to also model contact fields (Bhatnagar et al., 2022; Karunratanakul et al., 2020; Xie et al., 2022) and interaction fields (Zhou et al., 2022a). Very recently, Tiwari et al. (2022) showed that implicit functions can even be used to model the high dimensional manifold of 3D poses.

8.1.3 *LoopReg (Bhatnagar et al., 2020b)*

IMPORTANCE OF LOOP CLOSURE. LoopReg formulates registration, the task to fit a 'model' (eg. SMPL) to 'data' (input scans/ point clouds), as a mapping task. We learn a differentiable map from data to model and back to data. This loop closure, enabled by our novel differentiable map, allows us to train LoopReg primarily with self-supervision. Compared to prior work, Groueix et al. (2018), LoopReg requires $\sim$ 100x less annotated data and can register detailed dress shape, unlike Groueix et al. (2018).

IMPORTANCE OF CORRESPONDENCE PREDICTION FOR FITTING SMPL. There are two main schools of thought on how to fit SMPL to an input (image/video/point cloud/scan). First is to directly regress SMPL parameters (Bhatnagar et al., 2019; Kanazawa et al., 2018; Kolotouros et al., 2019; Omran et al., 2018) and second is to predict intermediate correspondences from the input to the SMPL model and then optimize SMPL parameters to fit the correspondences (Bhatnagar et al., 2020a,b, 2022; Xie et al., 2022). Our findings support the latter for several reasons. Neural Networks, especially CNNs, are good at making localised predictions. Therefore learning to predict input aligned correspondences is an easier task for CNNs by design. Moreover, given decent correspondences, instance specific optimization fits the SMPL model accurately. *Learning correspondences* and *optimizing SMPL* leverages best of 'bottom-up prediction and top-down optimization'. Secondly, existing work can directly regress SMPL pose and shape parameters when trained on large amounts of data (Kanazawa et al., 2018; Kolotouros et al., 2019; Omran et al., 2018) but cannot regress non-rigid deformations (Bhatnagar et al., 2020a). Our approach of 'correspondence prediction followed by SMPL optimization', captures non-rigid deformations well.

A very important point to keep in mind, when deciding between the two approaches is to also consider the input. With 3D inputs like point clouds and scans, we already have a lot of information about the 3D shape and 'correspondence + optimization' works quite well. When predicting from images, we will need additional prediction about the 3D shape. Our recent work on modelling human-object interaction from a monocular RGB camera (Xie et al., 2022) learns to predict 3D human shape as an unsigned distance field. And uses 'correspondence + optimization' strategy to fit SMPL to the image.

DIFFUSION AS ALTERNATIVE TO UV MAPS. One of the key challenges in LoopReg was to ensure that the network predicted correspondences to the SMPL model lie on the surface of SMPL mesh. To ensure this, prior work would predict correspondences to the SMPL UV-map. UV-maps are powerful representations of 2D manifolds living in a 3D space but require cutting and flattening the mesh. This makes the surface distorted and discontinuous at the location of the cuts.
To alleviate these problems we propose to diffuse the SMPL model to $\mathbb{R}^3$ instead. This diffused representation is continuous, differentiable and well defined throughout $\mathbb{R}^3$, thus making it highly amenable to neural network predictions. We think that the diffused model representation can be a powerful alternative to UV maps.

8.1.4 *BEHAVE (Bhatnagar et al., 2022)*

HOW TO MODEL CONTACTS? Our work LoopReg (Bhatnagar et al., 2020b) proposed a formulation to use a neural correspondence field to SMPL model for registering 3D human point clouds/ scans. BEHAVE extends this idea further by learning a correspondence field between the human and the object. This is easily implemented by learning two unsigned distance fields for the human and the object as well as a correspondence filed to the SMPL model. The points where the two distance fields are zero give us the contacts (Karunratanakul et al., 2020) and the SMPL correspondence tells us exactly which point on the body is in contact with the object. This formulation leverages existing distance fields (Chibane et al., 2020b; Park et al., 2019) and correspondence field to SMPL (Bhatnagar et al., 2020a,b) that have shown to work well and models a novel *contact field*. This formulation was extended by Xie et al. (2022) to reconstruct human-object reconstruction from a single image and Zhou et al. (2022a) to model hand-object interaction.

IMPLICIT FUNCTION TO MODEL OBJECT ORIENTATION FIELD. We discussed earlier how implicit functions can be used to model general continuous fields. We explore this idea further and learn a global object orientation field to obtain the 6D pose of an object. The ability of neural networks to learn such a field further corroborates our intuition that neural network based implicit representation can learn not just surfaces but general fields.

FITTING SMPL TO IMPLICIT NEURAL FIELDS. Our work Bhatnagar et al. (2020a) presented one of the first ideas to fit parametric model to implicit neural field predictions. IPNet would predict occupancy field that would reconstruct the 3D mesh using marching cubes. IPNet also predicts correspondences to the SMPL model that allow fitting SMPL model. Although powerful, this formulation has some limitations. First, marching cube requires predicting occupancy at dense grid points ($\sim$ 16M for 256^3 grid). This is very expensive and slow. Second marching cubes is typically non-differentiable, therefore network predictions do not receive any gradient from SMPL fitting.
We significantly improve on SMPL fitting to neural field predictions in BEHAVE by directly

fitting the SMPL model to neural fields without reconstructing the surface through marching cubes. Our formulation requires much less query points ($\sim$ 30k vs $\sim$ 16M) making it much faster and also differentiable. This way of fitting SMPL to implicit fields can be more widely used in registration tasks.

8.2 FUTURE WORK

In this section, we discuss some important research ideas and directions that this thesis opens up. We discuss future extensions specific to each chapter at the end of the chapter and present the broader research directions here.

LEARNING FROM 2D DATA. As mentioned earlier, the key focus of this thesis is data drive modelling of 3D humans. One of the key challenges here, is the lack of available 3D data. Such data is hard to collect as 3D and 4D scanners, marker based recording studios are very expensive and hard to set up. This was one of the motivations behind BEHAVE capture setup that greatly simplified 4D recording using just Kinects. Nonetheless, the 3D and 4D data is hard to get. On the other hand, there is lot of 2D data available in the form of annotated datasets and more importantly in the form of unlabeled images and videos on the internet. A very promising research direction is to explore using the 2D data to learn in 3D. For instance, we have hundreds of videos of people using a hammer on the internet, "can we use these videos to learn how to interact with a hammer in 3D?"

CAPTURING HUMAN-OBJECT INTERACTIONS. Over the last few years we have seen a lot of progress in modelling the appearance of digital humans but modelling the human actions/behaviour is only recently picking up. One of the main roadblocks is the availability of large scale annotated data of human-scene interaction. Capturing such data is hard as current 4D scanners and marker based capture studios cannot scale to record variety of interactions at different locations. For instance, we cannot relocate our 4D scanner from lab to kitchen, just to record cooking, cleaning, operating different appliances etc. as these scanners are huge and difficult to set up. Moreover, we cannot record long term motion like 'person walking from a room to kitchen to get water', as the recording volume is limited.

We provide an alternative using Kinects in BEHAVE (Bhatnagar et al., 2022), which is easier to scale, but we found several limitations. The recording area is for Kinects is quite small and does not allow long range capture. Kinect point clouds are noisy so it is difficult to capture detailed appearance and fine-grained contacts with hands.

This is a challenging task that will require several advancements at hardware and software level so that we can reliably capture 3D data, using a consumer grade setup. This is essential if concepts like Metaverse are to become a reality, where it is essential to capture a person and their actions and faithfully reproduce them in mixed reality.

AUTONOMOUS DIGITAL HUMANS. Modelling the appearance of digital humans is an active area of research and more recently we have been looking into not just how humans look but also interact with their surroundings. This is a step in the right direction but human beings can do much more than interact with their environment 'in the moment'. We can plan long term activities, explore and understand various elements of environments. This ability for long term understanding needs to be modelled in future models of digital humans.

There is active research in the area of embodied AI (Manolis Savva* et al., 2019; Ramakrishnan et al., 2021; Srivastava et al., 2022; Szot et al., 2021) and we are seeing great progress in robotic agents exploring virtual environment. A great future direction is to combine the research in modelling humans and embodied AI to build the futuristic autonomous digital humans.

8.3 SOCIAL IMPACT

A lot of research presented in this thesis is directly linked to real world applications. For instance, in Chapter 4 we show that given just a few images of a person we can reconstruct the body shape and clothing of a person in 3D, which is then used for virtual try-on. In Chapter 5, we show that we can animate and edit a 3D human from just an input point cloud. This has applications in creating digital content. In Chapter 7 we release the first method and dataset to model dynamic human-object interactions in *real* world settings. This has applications in robotics, animation, gaming etc. We have released our code and datasets curated by us, along with all our published research. We hope that this will allow the community to reliably use and build on our work.

PRIVACY CONCERNS. This thesis addresses the privacy sensitive area of modelling 3D humans, including details of how they look and behave. In this section we discuss potential concerns arising from our work, both from the methodology as well as data perspective. Our work in Chapter 4, can reconstruct the 3D clothing of a person but does not provide detailed identity markers like precise body shape and facial details. Thus posing only a small privacy risk. Our work in Chapters 5 and 6 is able to reconstruct detailed 3D appearance including facial details from a point cloud. Given that some of the mobile phones now provide depth sensors, it might be possible to obtain 3D models of people relatively easily. The privacy concern is similar to having your picture taken without consent, but instead of 2D image, 3D model of a person is at risk. In Chapter 7, we propose an approach to track a 3D human and it's interaction with objects. At the moment the method does not capture unique identity markers like face, but in future it might be able to track a person's actions along with detailed appearance. We should be building such systems with care as they can capture two of the most defining characteristics of a person, namely the appearance and behaviour. These models can be used in conjunction with 'Deep Fake' models to put the privacy of people at risk and spread misinformation. This calls for more research into privacy preserving 3D vision to ensure that advancements in 3D modelling are not misused.

This thesis presents several advancements in data driven 3D human modelling. This requires collecting data in the form of images, videos and 3D scans of people. Throughout the process, we ensure that the subjects are aware of how their data will be captured and used. We obtain written consent from all our subjects. Subjects can further choose to anonymise their identity (face) both in images and scans. Our data is free to use for research purposes but we do not allow commercial usage to ensure that people's data is not monetised without consent.

BIBLIOGRAPHY

Alldieck, Thiemo, Marc Kassubeck, Bastian Wandt, Bodo Rosenhahn, and Marcus Magnor (2017). “Optical Flow-based 3D Human Motion Estimation from Monocular Video.” In: *German Conf. on Pattern Recognition*, pp. 347–360. ISBN: 978-3-319-66709-6.

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

Alldieck, Thiemo, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll (2018a). “Detailed Human Avatars from Monocular Video.” In: *International Conf. on 3D Vision*.

– (2018b). “Video based reconstruction of 3d people models.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8387–8397.

Alldieck, Thiemo, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor (2019b). “Tex2Shape: Detailed Full Human Body Geometry from a Single Image.” In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Alp Güler, Rıza, Natalia Neverova, and Iasonas Kokkinos (2018). “Densepose: Dense human pose estimation in the wild.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306.

Andriluka, Mykhaylo, Umar Iqbal, Anton Milan, Eldar Insafutdinov, Leonid Pishchulin, Juergen Gall, and Bernt Schiele (2018). “PoseTrack: A Benchmark for Human Pose Estimation and Tracking.” In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anguelov, Dragomir, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis (2005). “SCAPE: shape completion and animation of people.” In: *ACM Transactions on Graphics*. Vol. 24. 3. ACM, pp. 408–416.

Bhatnagar, Bharat Lal, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll (2020a). “Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction.” In: *European Conference on Computer Vision (ECCV)*.

– (2020b). “LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration.” In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Bhatnagar, Bharat Lal, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll (2019). “Multi-Garment Net: Learning to Dress 3D People from Images.” In: *IEEE International Conference on Computer Vision (ICCV)*.

Bhatnagar, Bharat Lal, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll (2022). “BEHAVE: Dataset and Method for Tracking Human Object Interactions.” In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Blanz, Volker and Thomas Vetter (1999). “A morphable model for the synthesis of 3D faces.” In: *Conf. on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., pp. 187–194.

Bodla, Navaneeth, Gaurav Shrivastava, Rama Chellappa, and Abhinav Shrivastava (2021). “Hierarchical Video Prediction Using Relational Layouts for Human-Object Interactions.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bogo, Federica, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black (2016). "Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image." In: *European Conf. on Computer Vision*. Springer International Publishing.

Bogo, Federica, Javier Romero, Matthew Loper, and Michael J. Black (2014). "FAUST: Dataset and evaluation for 3D mesh registration." In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Bogo, Federica, Javier Romero, Gerard Pons-Moll, and Michael J. Black (2017). "Dynamic FAUST: Registering Human Bodies in Motion." In: *IEEE Conf. on Computer Vision and Pattern Recognition*.

Brahmbhatt, Samarth, Ankur Handa, James Hays, and Dieter Fox (Apr. 2019). "ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact." In: *IROS*.

Brahmbhatt, Samarth, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays (2020). "ContactPose: A Dataset of Grasps with Object Contact and Hand Pose." In: *The European Conference on Computer Vision (ECCV)*.

Cao, Zhe, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik (2020). "Long-term Human Motion Prediction with Scene Context." In: *ArXiv* abs/2007.03672.

Chen, Xu, Yufeng Zheng, Michael J. Black, Otmar Hilliges, and Andreas Geiger (Oct. 2021). "SNARF: Differentiable Forward Skinning for Animating Non-Rigid Neural Implicit Shapes." In: *Proc. International Conference on Computer Vision (ICCV)*, pp. 11594–11604.

Chen, Yixin, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu (2019). "Holistic++ Scene Understanding: Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-Object Interaction and Physical Commonsense." In: *The IEEE International Conference on Computer Vision (ICCV)*.

Chen, Zhiqin and Hao Zhang (2019). "Learning Implicit Fields for Generative Shape Modeling." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5939–5948.

Chibane, Julian, Thiemo Alldieck, and Gerard Pons-Moll (2020a). "Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Chibane, Julian, Aymen Mir, and Gerard Pons-Moll (2020b). "Neural Unsigned Distance Fields for Implicit Function Learning." In: *Neural Information Processing Systems (NeurIPS)*.

Choutas, Vasileios, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black (2020). "Monocular expressive body regression through body-driven attention." In: *European Conference on Computer Vision*. Springer, pp. 20–40.

Choy, Christopher B, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese (2016). "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction." In: *European conference on computer vision*. Springer, pp. 628–644.

Corona, Enric, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Gregory Rogez (2020). "GanHand: Predicting Human Grasp Affordances in Multi-Object Scenes." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Crane, Keenan, Clarisse Weischedel, and Max Wardetzky (2013). "Geodesics in heat: A new approach to computing distance based on heat flow." In: *ACM Transactions on Graphics (TOG)* 32.5, p. 152.

Curless, Brian and Marc Levoy (1996). "A Volumetric Method for Building Complex Models from Range Images." In: *Proceedings of the 23rd Annual Conference on Computer Graphics* *and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pp. 303–312.

Dabral, Rishabh, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik (2021). "Gravity-Aware Monocular 3D Human-Object Reconstruction." In: *arXiv preprint arXiv:2108.08844*.

Deng, Boyang, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi (2020). "NASA: Neural Articulated Shape Approximation." In: *The European Conference on Computer Vision (ECCV)*.

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

Dyke, Roberto M., Yu-Kun Lai, Paul L. Rosin, and Gary K.L. Tam (2019). "Non-rigid registration under anisotropic deformations." In: *Computer Aided Gxeometric Design*.

Ehsani, Kiana, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta (2020). "Use the Force, Luke! Learning to Predict Physical Forces by Simulating Effects." In: *CVPR*.

Feng, Yao, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black (Dec. 2021). "Collaborative Regression of Expressive Bodies using Moderation." In: *International Conference on 3D Vision (3DV)*.

Fieraru, Mihai, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu (2020). "Three-dimensional reconstruction of human interactions." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7214–7223.

– (2021). "Learning Complex 3D Human Self-Contact." In: *Thirty-Fifth AAAI Conf. on Artificial Intelligence (AAAI'21)*.

Gabeur, Valentin, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Grégory Rogez (2019). "Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images." In: *CoRR*.

Gilbert, Andrew, Marco Volino, John P. Collomosse, and Adrian Hilton (2018). "Volumetric Performance Capture from Minimal Camera Viewpoints." In: *European Conference on Computer Vision (ECCV)*.

Gong, Ke, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin (2018). "Instance-level Human Parsing via Part Grouping Network." In: *European Conf. on Computer Vision*.

Grady, Patrick, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp (2021). "ContactOpt: Optimizing Contact to Improve Grasps." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1471–1481.

Groueix, Thibault, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry (2018). "3D-CODED : 3D Correspondences by Deep Deformation." In: *ECCV*.

Guler, Riza Alp and Iasonas Kokkinos (2019). "HoloPose: Holistic 3D Human Reconstruction In-The-Wild." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guzov, Vladimir, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll (2021). "Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Habermann, Marc, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt (2019). "Livecap: Real-time human performance capture from monocular video." In: *ACM Transactions On Graphics (TOG)* 38.2, pp. 1–17.

Habermann, Marc, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt (2020). “Deepcap: Monocular human performance capture using weak supervision.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5052–5063.

Hasler, N., C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel (2009). “A Statistical Model of Human Pose and Body Shape.” In: *Computer Graphics Forum* 28.2, pp. 337–346.

Hassan, Mohamed, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black (Oct. 2021a). “Stochastic Scene-Aware Motion Prediction.” In: *Proc. International Conference on Computer Vision (ICCV)*.

Hassan, Mohamed, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black (2019). “Resolving 3D Human Pose Ambiguities with 3D Scene Constraints.” In: *International Conference on Computer Vision*.

Hassan, Mohamed, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black (June 2021b). “Populating 3D Scenes by Learning Human-Scene Interaction.” In: *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 14708–14718.

Hasson, Yana, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid (2019). “Learning joint reconstruction of hands and manipulated objects.” In: *CVPR*.

He, Tong, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung (2021). “ARCH++: Animation-Ready Clothed Human Reconstruction Revisited.” In: *The IEEE International Conference on Computer Vision (ICCV)*.

Henderson, Paul and Vittorio Ferrari (2018). “Learning to Generate and Reconstruct 3D Meshes with only 2D Supervision.” In: *British Machine Vision Conference (BMVC)*.

Hirshberg, D., M. Loper, E. Rachlin, and M.J. Black (2012). “Coregistration: Simultaneous alignment and modeling of articulated 3D shape.” In: *European Conf. on Computer Vision (ECCV)*.

Hormann, Kai, Bruno Lévy, and Alla Sheffer (2007). “Mesh parameterization: Theory and practice.” In.

Hu, JF, WS Zheng, J Lai, and J Zhang (2017). “Jointly Learning Heterogeneous Features for RGB-D Activity Recognition.” In: *IEEE transactions on pattern analysis and machine intelligence* 39.11, pp. 2186–2200.

Huang, Yinghao, Federica Bogo, Christoph Classner, Angjoo Kanazawa, Peter V Gehler, Ijaz Akhter, and Michael J Black (2017). “Towards Accurate Markerless Human Shape and Pose Estimation over Time.” In: *International Conf. on 3D Vision*.

Huang, Yinghao, Omid Taheri, Michael J. Black, and Dimitrios Tzionas (2022). “InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction.” In: *German Conference on Pattern Recognition (GCPR)*.

Huang, Zeng, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li (2018). “Deep volumetric video from very sparse multi-view performance capture.” In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 336–354.

Huang, Zeng, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung (2020). “ARCH: Animatable Reconstruction of Clothed Humans.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Insafutdinov, Eldar, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele (2016). “DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model.” In: *European Conference on Computer Vision (ECCV)*.

Ionescu, Catalin, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu (2014). “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Iskakov, Karim, Egor Burkov, Victor Lempitsky, and Yury Malkov (2019). “Learnable triangulation of human pose.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7718–7727.

Jiang, Boyi, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao (2020). “BCNet: Learning Body and Cloth Shape from A Single Image.” In: *European Conference on Computer Vision*. Springer.

Joo, Hanbyul, Tomas Simon, and Yaser Sheikh (2018a). “Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies.” In: *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 8320–8329.

– (2018b). “Total capture: A 3d deformation model for tracking faces, hands, and bodies.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8320–8329.

Kanazawa, Angjoo, Michael J. Black, David W. Jacobs, and Jitendra Malik (2018). “End-to-end Recovery of Human Shape and Pose.” In: *IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE Computer Society.

Karunratanakul, Korrawe, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang (2021). “A Skeleton-Driven Neural Occupancy Representation for Articulated Hands.” In: *International Conference on 3D Vision (3DV)*.

Karunratanakul, Korrawe, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang (Nov. 2020). “Grasping Field: Learning Implicit Representations for Human Grasps.” In: *8th International Conference on 3D Vision*. IEEE, pp. 333–344.

Kocabas, Muhammed, Nikos Athanasiou, and Michael J. Black (June 2020). “VIBE: Video Inference for Human Body Pose and Shape Estimation.” In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5252–5262.

Kolotouros, Nikos, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis (2019). “Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop.” In: *ICCV*.

Kundu, Abhijit, Yin Li, and James M Rehg (2018). “3d-rcnn: Instance-level 3d object reconstruction via render-and-compare.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3559–3568.

Lazova, Verica, Eldar Insafutdinov, and Gerard Pons-Moll (2019). “360-Degree Textures of People in Clothing from a Single Image.” In: *International Conference on 3D Vision (3DV)*.

Lei, Jiahui, Srinath Sridhar, Paul Guerrero, Minhyuk Sung, Niloy Mitra, and Leonidas J. Guibas (2020). “Pix2Surf: Learning Parametric 3D Surface Models of Objects from Images.” In: *Proceedings of European Conference on Computer Vision (ECCV)*.

Leroy, Vincent, Jean-Sébastien Franco, and Edmond Boyer (2018). “Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency.” In: *European Conference on Computer Vision (ECCV)*.

Li, Chun-Liang, Tomas Simon, Jason M. Saragih, Barnabás Póczos, and Yaser Sheikh (2019a). “LBS Autoencoder: Self-Supervised Fitting of Articulated Meshes to Point Clouds.” In: *CVPR*, pp. 11967–11976. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Li_LBS_Autoencoder_Self-Supervised_Fitting_of_Articulated_Meshes_to_Point_Clouds_CVPR_2019_paper.html.

Li, Ruilong, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li (2020a). “Volumetric Human Teleportation.” In: *ACM SIGGRAPH 2020 Real-Time Live!* SIGGRAPH 2020. Association for Computing Machinery.

Li, Ruilong, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li (2020b). “Monocular Real-Time Volumetric Performance Capture.” In: *arXiv preprint arXiv:2007.13988*.

Li, Tianye, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero (2017). “Learning a model of facial shape and expression from 4D scans.” In: *ACM Transactions on Graphics* 36.6. Two first authors contributed equally, 194:1–194:17.

Li, Xueting, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz (2019b). “Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments.” In: *IEEE Conference on Computer Vision and Pattern Recognition*.

Ling, Hung Yu, Fabio Zinno, George Cheng, and Michiel van de Panne (2020). “Character controllers using motion VAEs.” In: *ACM Trans. Graph.* 39.4, p. 40.

Liu, Jun, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot (2019). “NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Rosanne, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski (2018). “An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution.” In: *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*. NIPS’18. USA: Curran Associates Inc., pp. 9628–9639.

Loper, Matthew, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black (2015). “SMPL: A Skinned Multi-Person Linear Model.” In: ACM.

Lorensen, William E and Harvey E Cline (1987). “Marching cubes: A high resolution 3D surface construction algorithm.” In: *ACM siggraph computer graphics* 21.4, pp. 163–169.

Luvizon, Diogo C., David Picard, and Hedi Tabia (2018). “2D/3D Pose Estimation and Action Recognition Using Multitask Deep Learning.” In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ma, Qianli, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black (June 2021a). “SCALE: Modeling Clothed Humans with a Surface Codec of Articulated Local Elements.” In: *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Ma, Qianli, Jinlong Yang, Michael J. Black, and Siyu Tang (2022). “Neural Point-based Shape Modeling of Humans in Challenging Clothing.” In: *2022 International Conference on 3D Vision (3DV)*.

Ma, Qianli, Jinlong Yang, Siyu Tang, and Michael J. Black (Oct. 2021b). “The Power of Points for Modeling Humans in Clothing.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mahmood, Naureen, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black (Oct. 2019). “AMASS: Archive of Motion Capture as Surface Shapes.” In: *International Conference on Computer Vision*, pp. 5442–5451.

Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra (2019). “Habitat: A Platform for Embodied AI Research.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Marcard, Timo von, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll (2018). “Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera.” In: *European Conference on Computer Vision (ECCV)*.

Marin, Riccardo, Simone Melzi, Emanuele Rodolà, and Umberto Castellani (2020). “FARM: Functional Automatic Registration Method for 3D Human Bodies.” In: *Comput. Graph. Forum* 39, pp. 160–173.

Mescheder, Lars, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger (2019). "Occupancy Networks: Learning 3D Reconstruction in Function Space." In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Michalkiewicz, Mateusz, Jhony K. Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders P. Eriksson (2019). "Deep Level Sets: Implicit Surface Representations for 3D Shape Inference." In: *CoRR.*

Mihajlovic, Marko, Yan Zhang, Michael J Black, and Siyu Tang (2021). "LEAP: Learning Articulated Occupancy of People." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10461–10471.

Mir, Aymen, Thiemo Alldieck, and Gerard Pons-Moll (2020). "Learning to Transfer Texture from Clothing Images to 3D Humans." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

Muller, Lea, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black (2021a). "On Self-Contact and Human Pose." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9990–9999.

Muller, Norman, Yu-Shiang Wong, Niloy J Mitra, Angela Dai, and Matthias Nießner (2021b). "Seeing Behind Objects for 3D Multi-Object Tracking in RGB-D Sequences." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6071–6080.

Natsume, Ryota, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima (2019). "Siclope: Silhouette-based clothed people." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4480–4490.

Newcombe, Richard A., Dieter Fox, and Steven M. Seitz (2015). "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 343–352.

Omran, Mohamed, Christop Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele (2018). "Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation." In: *International Conf. on 3D Vision.*

Or, Litany, Remez Tal, Rodolà Emanuele, Bronstein Alex M., and Bronstein Michael M. (2017). "Deep Functional Maps: Structured Prediction for Dense Shape Correspondence." In: *International Conference on Computer Vision (ICCV).*

Pandey, Rohit, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, et al. (2019). "Volumetric capture of humans with a single rgbd camera via semi-parametric learning." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9709–9718.

Park, Jeong Joon, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove (2019). "DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 165–174.

Patel, Chaitanya, Zhouyincheng Liao, and Gerard Pons-Moll (2020). "The Virtual Tailor: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

Pavlakos, Georgios, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black (2019). "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image." In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).*

Pavlakos, Georgios, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis (2018). “Learning to Estimate 3D Human Pose and Shape from a Single Color Image.” In: *IEEE Conf. on Computer Vision and Pattern Recognition*.

Pishchulin, Leonid, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele (2016). “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation.” In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pishchulin, Leonid, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele (2017). “Building statistical shape spaces for 3d human modeling.” In: *Pattern Recognition* 67, pp. 276–286.

Plänkers, Ralf and Pascal Fua (2003). “Articulated soft objects for multiview shape and motion capture.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Pons-Moll, Gerard, Sergi Pujades, Sonny Hu, and Michael Black (2017). “ClothCap: Seamless 4D Clothing Capture and Retargeting.” In: *ACM Transactions on Graphics* 36.4.

Pons-Moll, Gerard, Javier Romero, Naureen Mahmood, and Michael J Black (2015). “Dyna: a model of dynamic human shape in motion.” In: *ACM Transactions on Graphics* 34, p. 120.

Pons-Moll, Gerard and Bodo Rosenhahn (2011). “Model-Based Pose Estimation.” In: *Visual Analysis of Humans: Looking at People*. Springer. Chap. 9, pp. 139–170.

Pons-Moll, Gerard, Jonathan Taylor, Jamie Shotton, Aaron Hertzmann, and Andrew Fitzgibbon (2013). “Metric Regression Forests for Human Pose Estimation.” In: *British Machine Vision Conference (BMVC)*. BMVA Press.

Qi, Charles R, Li Yi, Hao Su, and Leonidas J Guibas (2017). “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space.” In.

Ramakrishnan, Santhosh Kumar, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra (2021). “Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI.” In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ranjan, Anurag, Timo Bolkart, Soubhik Sanyal, and Michael J. Black (2018). “Generating 3D faces using Convolutional Mesh Autoencoders.” In: *European Conf. on Computer Vision*, pp. 725–741.

Reddy, N Dinesh, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan (2021). “Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking.” In: *IEEE Conference on Computer Vision and Pattern Recognition*.

Rhodin, Helge, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt (2016a). “General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues.” In: *European Conference on Computer Vision (ECCV)*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling.

– (2016b). “General automatic human shape and motion capture using volumetric contour cues.” In: *European conference on computer vision*. Springer, pp. 509–526.

Rhodin, Helge, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua (2018). “Learning monocular 3d human pose estimation from multi-view images.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8437–8446.

Riza Alp Güler Natalia Neverova, Iasonas Kokkinos (2018). “DensePose: Dense Human Pose Estimation In The Wild.” In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Romero, Javier, Dimitrios Tzionas, and Michael J. Black (Nov. 2017). “Embodied Hands: Modeling and Capturing Hands and Bodies Together.” In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6.

Rong, Yu, Takaaki Shiratori, and Hanbyul Joo (2021). “FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration.” In: *IEEE International Conference on Computer Vision Workshops*.

Saito, Shunsuke, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li (2019). “PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization.” In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Saito, Shunsuke, Tomas Simon, Jason Saragih, and Hanbyul Joo (2020). “PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Saito, Shunsuke, Jinlong Yang, Qianli Ma, and Michael J. Black (June 2021). “SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks.” In: *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Savva, Manolis, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner (2016). “PiGraphs: Learning Interaction Snapshots from Observations.” In: *ACM Transactions on Graphics (TOG)* 35.4.

Slavcheva, Miroslava, Maximilian Baust, Daniel Cremers, and Slobodan Ilic (2017). “KillingFusion: Non-rigid 3D Reconstruction without Correspondences.” In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Sminchisescu, C. and A. Telea (2002). “Human pose estimation from silhouettes. A consistent approach using distance level sets.” In: *WSCG International Conference on Computer Graphics, Visualization and Computer Vision*.

Smith, David, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero (2019). “FACSIMILE: Fast and Accurate Scans From an Image in Less Than a Second.” In: *CoRR*.

Sofiiuk, Konstantin, Ilia Petrov, Olga Barinova, and Anton Konushin (2020). “f-brs: Rethinking backpropagating refinement for interactive segmentation.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8623–8632.

Song, Liangchen, Gang Yu, Junsong Yuan, and Zicheng Liu (2021). “Human pose estimation and its application to action recognition: A survey.” In: *Journal of Visual Communication and Image Representation*.

Sorkine, Olga (2005). “Laplacian mesh processing.” In: *Eurographics (STARs)*, pp. 53–70.

Srivastava, Sanjana, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei (2022). “BEHAVIOR: Benchmark for Everyday Household Activities in Virtual, Interactive, and Ecological Environments.” In: *Proceedings of the 5th Conference on Robot Learning*. PMLR, pp. 477–490.

Starke, Sebastian, He Zhang, Taku Komura, and Jun Saito (2019). “Neural state machine for character-scene interactions.” In: *ACM Trans. Graph.* 38.6, 209:1–209:14.

Stoll, Carsten, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt (n.d.). “Fast articulated motion tracking using a sums of Gaussians body model.” In: *IEEE International Conference on Computer Vision, ICCV 2011*.

Stoll, Carsten, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt (2011). “Fast articulated motion tracking using a sums of Gaussians body model.” In.

Su, Zhuo, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang (2020). “Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera.” In: *Com-*

puter Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer, pp. 246–264.

Su, Zhuo, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang (2021). "RobustFusion: Robust Volumetric Performance Reconstruction under Human-object Interactions from Monocular RGBD Stream." In: *arXiv preprint arXiv:2104.14837*.

Sun, Guoxing, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu (2021). "Neural Free-Viewpoint Performance Rendering under ComplexHuman-object Interactions." In: *arXiv preprint arXiv:2108.00362*.

Sun, Xingyuan, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman (2018). "Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Szot, Andrew, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra (2021). "Habitat 2.0: Training Home Assistants to Rearrange their Habitat." In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Taheri, Omid, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas (2020). "GRAB: A Dataset of Whole-Body Human Grasping of Objects." In: *European Conference on Computer Vision (ECCV)*.

Tao, Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Dai Quionhai, Hao Li, G. Pons-Moll, and Yebin Liu (2018). "DoubleFusion: Real-time Capture of Human Performance with Inner Body Shape from a Depth Sensor." In: *IEEE Conf. on Computer Vision and Pattern Recognition*.

Taylor, Jonathan, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. (2016). "Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences." In: *ACM Transactions on Graphics* 35.4, p. 143.

Taylor, Jonathan, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon (2012). "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation." In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 103–110.

Tewari, Ayush, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian (2017). "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction." In: *IEEE International Conf. on Computer Vision*.

Tiwari, Garvita, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll (2022). "Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields." In: *European Conference on Computer Vision (ECCV)*. Springer.

Tiwari, Garvita, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll (2020). "SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing." In: *European Conference on Computer Vision (ECCV)*. Springer.

Tiwari, Garvita, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll (2021). "Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing." In: *International Conference on Computer Vision (ICCV)*.

Tzionas, Dimitrios and Juergen Gall (2015). "3d object reconstruction from hand-object interactions." In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 729–737.

Varol, Gül, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid (2018). "BodyNet: Volumetric Inference of 3D Human Body Shapes." In: *European Conf. on Computer Vision*.

Varol, Gül, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid (2017). "Learning from Synthetic Humans." In: *IEEE Conf. on Computer Vision and Pattern Recognition*.

Walker, Jacob, Kenneth Marino, Abhinav Gupta, and Martial Hebert (2017). "The Pose Knows: Video Forecasting by Generating Pose Futures." In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society.

Wang, Lizhen, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu (2020). "NormalGAN: Learning detailed 3D human from a single RGB-D image." In: *European Conference on Computer Vision*. Springer, pp. 430–446.

Wang, Shaofei, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang (2021). "MetaAvatar: Learning Animatable Clothed Human Models from Few Depth Images." In: *Advances in Neural Information Processing Systems*.

Wei, Lingyu, Qixing Huang, Duygu Ceylan, Etienne Vouga, and Hao Li (2016). "Dense Human Body Correspondences Using Convolutional Networks." In: *Computer Vision and Pattern Recognition (CVPR)*.

Weng, Zhenzhen and Serena Yeung (2020). "Holistic 3D Human and Scene Mesh Estimation from Single View Images." In: *arXiv preprint arXiv:2012.01591*.

Wu, Jiajun, Yifan Wang, Tianfan Xue, Xingyuan Sun, William T Freeman, and Joshua B Tenenbaum (2017). "MarrNet: 3D Shape Reconstruction via 2.5D Sketches." In: *Advances In Neural Information Processing Systems*.

Wu, Jiajun, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum (2018). "Learning 3D Shape Priors for Shape Completion and Reconstruction." In: *European Conference on Computer Vision (ECCV)*.

Wu, Yan, Jiahao Wang, Yan Zhang, Siwei Zhang, Otmar Hilliges, Fisher Yu, and Siyu Tang (2022). "SAGA: Stochastic Whole-Body Grasping with Contact." In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wu, Yuxin, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick (2019). *Detectron2*. https://github.com/facebookresearch/detectron2.

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

Xie, Xianghui, Bharat Lal Bhatnagar, and Gerard Pons-Moll (2022). "CHORE: Contact, Human and Object Reconstruction from a single RGB image." In: *European Conference on Computer Vision (ECCV)*. Springer.

Xu, Hongyi, Thiemo Alldieck, and Cristian Sminchisescu (2021). "H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion." In: *Neural Information Processing Systems*.

Xu, Hongyi, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu (2020). "GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models." In: *IEEE International Conf. on Computer Vision and Pattern Recognition*.

Yang, Bo, Hongkai Wen, Sen Wang, Ronald Clark, Andrew Markham, and Niki Trigoni (2017). "3d object reconstruction from a single depth view with adversarial learning." In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 679–688.

Yang, Jinlong, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer (2016). "Estimation of Human Body Shape in Motion with Wide Clothing." In: *European Conference on Computer Vision*.

Yang, Lixin, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu (2021). "CPF: Learning a Contact Potential Field to Model the Hand-Object Interaction." In: *ICCV*.

Yu, Tao, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu (2021). "Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

Yu, Tao, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu (2018). "DoubleFusion: Real-Time Capture of Human Performances With Inner Body Shapes From a Single Depth Sensor." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7287–7296.

Zanfir, Andrei, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu (2020). "Neural Descent for Visual 3D Human Pose and Shape." In: *IEEE International Conf. on Computer Vision and Pattern Recognition*.

Zanfir, Andrei, Elisabeta Marinoiu, and Cristian Sminchisescu (2018). "Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes–The Importance of Multiple Scene Constraints." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2148–2157.

Zanfir, Mihai, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu (2021). "THUNDR: Transformer-Based 3D Human Reconstruction With Markers." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12971–12980.

Zhang, Chao, Sergi Pujades, Michael Black, and Gerard Pons-Moll (2017a). "Detailed, accurate, human shape estimation from clothed 3D scan sequences." In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

– (2017b). "Detailed, accurate, human shape estimation from clothed 3D scan sequences." In: *IEEE Conf. on Computer Vision and Pattern Recognition*.

Zhang, Jason Y., Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa (2020a). "Perceiving 3D Human-Object Spatial Arrangements from a Single Image in the Wild." In: *European Conference on Computer Vision (ECCV)*.

Zhang, Siwei, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang (Oct. 2022a). "EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices." In: *European conference on computer vision (ECCV)*.

Zhang, Siwei, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang (2021). "Learning motion priors for 4d human body capture in 3d scenes." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11343–11353.

Zhang, Siwei, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang (Nov. 2020b). "PLACE: Proximity Learning of Articulation and Contact in 3D Environments." In: *International Conference on 3D Vision (3DV)*.

Zhang, Xiaohan, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll (2022b). "COUCH: Towards Controllable Human-Chair Interactions." In: *European Conference on Computer Vision (ECCV)*. Springer.

Zhao, Kaifeng, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang (Oct. 2022). "Compositional Human-Scene Interaction Synthesis with Semantic Control." In: *European conference on computer vision (ECCV)*.

Zheng, Zerong, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu (2019). "Deephuman: 3d human reconstruction from a single image." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7739–7749.

Zhou, Keyang, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll (2022a). "TOCH: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement." In: *European Conference on Computer Vision (ECCV)*. Springer.

Zhou, Keyang, Bharat Lal Bhatnagar, and Gerard Pons-Moll (2020). "Unsupervised Shape and Pose Disentanglement for 3D Meshes." In: *The European Conference on Computer Vision (ECCV)*.

Zhou, Keyang, Bharat Lal Bhatnagar, Bernt Schiele, and Gerard Pons-Moll (2022b). "Adjoint Rigid Transform Network: Task-conditioned Alignment of 3D Shapes." In: *2022 International Conference on 3D Vision (3DV)*. IEEE.

Zhou, Linqi, Yilun Du, and Jiajun Wu (2021a). "3D Shape Generation and Completion Through Point-Voxel Diffusion." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5826–5835.

Zhou, Yuxiao, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu (2021b). "Monocular Real-time Full Body Capture with Inter-part Correlations." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4811–4822.