Refining 3D Human Digitization Using Learnable Priors

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science in ***Computer Science and Engineering***
by Research

by

Routhu Snehith Goud
20171105
snehith.goud@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500032, INDIA
December, 2022

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled ***"Refining 3D Human Digitization Using Learnable Priors"*** by Routhu Snehith Goud, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Avinash Sharma

This paper is dedicated to my institution mentors under whose constant guidance I have completed this dissertation. They not only enlightened me with academic knowledge but also gave me valuable advice whenever needed.

Acknowledgments

This thesis has been possible because of the constant support of my adviser, family, and peers. In this journey, I have met new people, made new friendships, and would cherish the memories made during this process.

Firstly, I would like to thank my adviser, Dr. Avinash Sharma, for his guidance in my research and writing of this thesis. He has been extremely supportive and considerate towards me this matters more to an undergrad researcher like me having to handle both studies and research. Even though I doubted myself at times he trusted and believed in me. Whenever I wandered into the wrong path he motivated me through philosophical teachings. I respect his efforts in organizing occasional group outings for relaxation and bonding.

Next, I would like to thank my brother for his constant motivational support throughout my engineering life. He understands my problems and suggests various options I can opt for in those scenarios. He is my go-to person to share my happiness and sorrows. I feel a special bonding with my brother where I can share anything without hesitation. He mostly takes care of the family responsibilities enabling me to focus more on my studies. I am very grateful for his existence in my life and can never repay him.

Sai Sagar Jinka is an amazing person and has been a great mentor to me. The amount of knowledge he possesses astonishes me. His hunger for research is something I won't be able to match. He would always contribute ideas in various threads and was the go-to person for help. He always had my back and motivated me, especially during the conference deadlines, he would calm my nerves and give me confidence. I would also like to pay gratitude to my peers, Astitva Srivastava, who helped me get the hang of data generation pipelines, had brief discussions when I get stuck, and had some fun csgo nights together; and Chandradeep for helping me with technical assistance wish him all the luck and success; Abhinaba is an extremely friendly person.

This acknowledgment would be incomplete without mentioning my undergrad dual degree friends for making my life blissful here at IIIT-H. I would like to thank my friends Vijay, Aravind, Surendra, Vivek, and Sireesha for all the fun times, movies watched, late night talks that we had together. A special mention to my undergrad BTech friends Srikar, Ravi Teja, and Pavan Kalyan. I wish each of them success and happiness in their lives.

Abstract

The 3D reconstruction of a human from a monocular RGB image is an interesting yet very challenging research problem in the field of computer vision. It has various applications in the movie industry, the gaming industry, and AR/VR applications. Hence it is important to recover detailed geometric 3D reconstruction of humans in order to enhance the realism. The problem is ill-posed in nature owing to high freedom of human pose, self-occlusions, loose clothing, camera viewpoint, and illumination. Though it is possible to generate accurate geometric reconstructions using structured light scanners or multi-view cameras these are cost expensive and require a specific setup e.g., numerous cameras, controlled illumination, etc. With the current advancement of deep learning techniques, the focus of the community shifted to the 3D reconstruction of people from monocular RGB images. The goal of the thesis is toward 3D Human Digitization of people in loose clothing with accurate person-specific details.

The problem of reconstruction of people in loose clothing is difficult as the topology of clothing is different from the human body. A multi-layered shape representation called PeeledHuman was able to deal with loose clothing and occlusions but it suffered from discontinuity or distorted body parts in the occluded regions. To overcome this we propose peeled semantic segmentation maps which provide the semantic information of the body parts across multiple peeled layers. These Peeled semantic maps help the network in predicting consistent depth for the pixels belonging to the same body part across different peeled maps. Our proposed Peeled Segmentation maps help in improving reconstruction both quantitatively and qualitatively. Additionally, the 3D semantic segmentation labels have various applications, for example, can be used to extract the clothing from the reconstructed output.

The face plays an important role in 3D Human digitization, it gives a person their identity and the realism enhances with high-frequency facial details. However, the face appears in a smaller region of the image which captures the complete body and makes the task of high-fidelity face reconstruction along with the body even more challenging. We reconstruct person-specific facial geometry along with the complete body by incorporating a facial prior and refining it further using our proposed framework. Another common challenge faced by the existing methods is the surface noise i.e, false geometric edges generated because of the textural edges present in the image space. We address this problem by incorporating a wrinkle map prior that distinguishes the geometrical from the textural edges coming from image space. In summary, in this thesis, we address the problem of 3D human digitization from monocular images. We evaluate our proposed solution on various existing 3D human datasets and demonstrate that our proposed solutions outperform the existing state-of-the-art methods. The limitations present in the

proposed methods were mentioned and potential solutions on how to address them have been briefly discussed. In the end, some future directions that can be potentially explored based on the proposed solutions have been discussed. Finally, we proposed efficient and robust methods to recover accurate and personalized 3D humans from images.

Contents

List of Figures

Chapter 1

Introduction

3D computer vision is a challenging but interesting research domain that requires an understanding of geometry and depth which has applications in various areas such as Robotics, Autonomous Driving, 3D Scene Reconstruction from point cloud data, video gaming industry, entertainment industry, Virtual try-on, and in AR/VR industry. The advancements in hardware devices enabled the classical computer vision methods such as SfM, depth from the stereo images, bundle adjustment, and volumetric regression to run in real time. The deep learning era further revolutionized the field in terms of both real-time performance and accurate predictions.

Metrically accurate and precise reconstructions of humans are possible with calibrated multi-view systems that use RGB or structured light cameras. Artec Leo[1] can reconstruct high-quality dense textured mesh but fail in the scenario of non-rigid objects, like clothing, hair, etc. The RGBD sensors like Microsoft's Azure Kinect[2] enable us to capture multi-view RGB and depth data using a multi-camera setup at lower costs. The depth data and RGB images from different devices are combined to generate a sparse point cloud. Nevertheless, they suffer from sensor noise and changes in illumination. Further, post-processing is required to convert these point clouds into high-quality textured meshes. *The Relightables* [75] is a volumetric capture system that can capture the reflectance of 3D human performances using a Light Stage fitted with 331 programmable custom LED lights and a set of high-resolution depth sensors which generate depth maps of 4112×3008 pixels per viewpoint acquired at 60 Hz. It can easily integrate the performances into new environments using AR or into films and video games. Though it can capture high fidelity volumetric 4D data with ease the setup is highly expensive. Figure 1.1 depicts some of the existing technologies for human capture and digitization.

These methods remain inaccessible to a large part of the community owing to their strict environmental setup and expensive capture systems. it's very expensive for the general public to have a large number of cameras, structured illuminations, and hardware devices. Hence, the reconstruction of 3D humans from a monocular RGB image garnered interest.

[1] https://www.artec3d.com/portable-3d-scanners/artec-leo

[2] https://azure.microsoft.com/en-us/services/kinect-dk

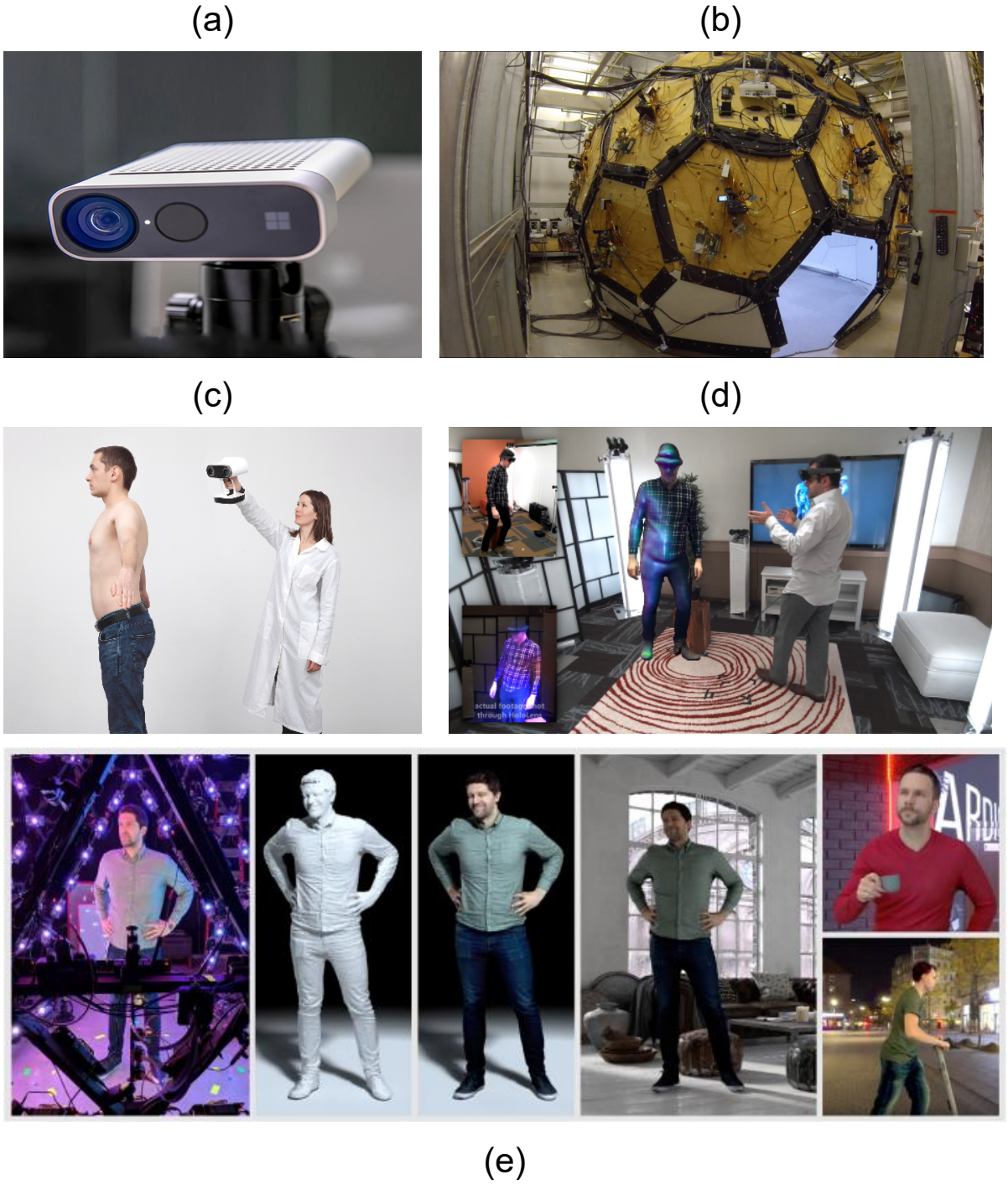

Figure 1.1 3D data digitization technologies: **(a)**Microsoft's Azure Kinect **(b)** CMU Panoptic Studio **(c)** Artec Leo **(d)** Microsoft Holoportation **(e)** *The Relightables* by Google

Therefore, in this thesis, we address the problem of 3D human digitization from monocular RGB images without using any expensive capture system. We aim to generate a high-fidelity 3D reconstruction of people in arbitrary clothing with person-specific details. By reconstructing highly Fidel reconstructions of the face we aim to enhance the realism and identity of the digitized human.

1.1 3D Digitization of Humans

1.1.1 Motivation

The field of 3D human digitization Figure 1.2 has gained high attention owing to the development of AR/VR technologies, virtual try-on, etc. Being able to generate video from multiple viewpoints enables the user to have an immersive experience in many applications such as AR/VR, gaming, virtual try-on, etc. as depicted in Figure 1.3. With the advent of the deep learning era, the trend of automated acquisition of detailed 3D human reconstruction from images has gained interest both in academic research and industrial research. A highly accurate, reliable, robust, and simple straightforward way mechanism to digitize the 3D human in arbitrary clothing has garnered immense attention.

3D human digitization has largely been enabled by the availability of complex, expensive 3D acquisition systems and highly sophisticated reconstruction algorithms. These complex capturing systems use expensive industry-level devices under a rig set up, with high-resolution synchronized cameras capturing information at high speeds. For example, the CMU Panoptic studio pipeline consists of 480 VGA cameras, 31 HD cameras, and 10 Kinect sensors in order to capture a multitude of human activities and perform the reconstruction. Industrial solutions such as Microsoft Holoportation and 8i[3] utilize infrared structured light for high-resolution capture consisting of a lot fewer cameras in usage. However, the huge setup of these high-quality reconstruction dome pipelines makes it infeasible to adapt to daily life scenarios. Most of these reconstruction pipelines use complex capture systems and equipment which involve highly sophisticated algorithms and processing to achieve the final reconstruction. The cost-effectiveness and huge setup, make it impractical for everyday users to adopt these systems. Hence the need for a simple and efficient setup that requires just a single image from a device such as a mobile or a camera to obtain a robust and photo-realistic 3D reconstruction of people in clothing.

A huge amount of effort has been invested in the development of 3D capture hardware devices and reconstruction algorithms. However, it has to be noticed that the traditional reconstruction pipelines are cost-ineffective and require following tedious capture routines. These complications make the reconstruction pipelines inaccessible to the general community. Moreover, the hand-crafted pipelines are prone to reconstruction artifacts, and noise from the capture devices, leading to a diminishment in the visual quality of the reconstruction.

Recent advances in image-based 3D human digitization have been driven by the significant improvement in representational power afforded by deep neural networks. Deep neural networks can be used

[3]https://8i.com/

Figure 1.2 3D digitization of humans.

to improve the reconstruction efficiency and robustness by learning the shape and structure of humans explicitly or implicitly from the existing 3D data. The major challenge is to achieve accurate, efficient, and photo-realistic reconstruction and re-rendering of humans, using devices that are portable and affordable. Although existing deep learning approaches have demonstrated significant potential in real-world scenarios, they often fail to produce person-specific photo-realistic reconstructions with the level of detail that is present in the input images.

Recently Facebook announced that it would focus its future and resources on the 'metaverse'. With this announcement, there is a lot of hype around the selling of 'metaverse' which is a VR social platform. Many other tech giants like Microsoft, Nvidia, Unity, etc. are investing in AR/VR which represents our physical world. The most common being of the physical world is humans and virtual human modeling is going to garner huge attention. For example, with the recent covid era many efforts have been put towards virtual reality meetings, classes, etc. Realism and user immersion in such applications can be achieved only when the virtual avatars are person specific and resemble their identity in the physical world. Hence the focus is to precisely capture the body shape, person-specific geometric facial details, along with other geometric details such as hairs, clothing deformations, etc. It is difficult to employ bulk complex capture systems or 3D acquisition devices as they are cost expensive and are not easily accessible to the general public. A practical approach is to use the images(single or multiple) to reconstruct personalized 3D human avatars. The existing state-of-the-art methods estimate reasonably well predictions in the scenario of tight clothing but fail terribly in the scenario of loose clothing such as robes, sarees, kurtas, frocks, gowns, etc. The existing methods also reconstruct flat faces with no facial details, as the face provides identity to a person it is important that we recover person-specific geometric facial details to enhance realism.

1.1.2 Research Challenges

The problem of 3D human digitization from monocular RGB images is a severely ill-posed problem in itself. It is very challenging to obtain high-frequency geometric details prevailing in various forms of clothing consisting of various textural patterns while also maintaining the person-specific shape. The face is another important aspect that gives identity to a person. A highly detailed face not only enhances realism but also captures the person-specific identity. A visual depiction of these challenges has been provided in Figure 1.4. We will now look into the challenges faced while digitizing 3D humans from monocular RGB images:

- The ill-posed nature of the problem arises due to various factors such as the viewpoint, self-occlusion with the body, and estimating the occluded parts such that they belong to the plausible pose of the human. The model should be able to understand what all shapes are feasible and also what poses are acceptable. Not being able to do so will lead to artifacts, irregular shapes, and noise in the reconstruction.

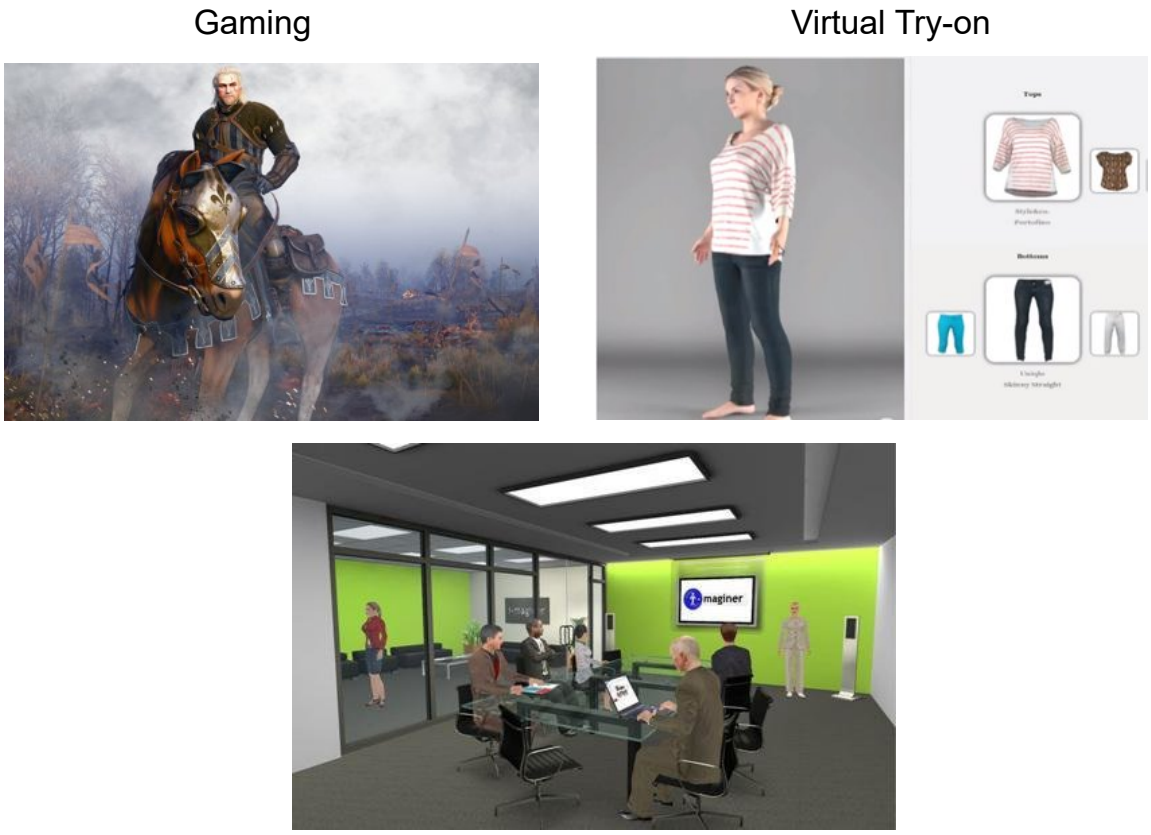

Figure 1.3 Applications of 3D human digitization

- Loose clothing has a topology that is entirely different from the human body. These can't be modeled using template-based methods. Recovering the geometry of loose clothing in occlude regions is an arduous task as the complex and skewed poses of the human, induce non-rigid deformations in the garment. The contextual information we get from the image is not enough to model these deformations.

- It is also difficult to obtain fine-grain facial details from the full body monocular RGB image owing to the lower resolution of the face in the image (approximately 50×50 in a 512×512 image). Nevertheless obtaining facial details not only enhances the realism but also gives a person-specific identity.

- It has been observed that various clothing has different textural patterns which get projected as false geometrical edges in the reconstructed geometry. It is difficult to distinguish whether the high-frequency details are due to the geometry or coming from the image space.

- Multiple layers of clothing on top of each other is an even more challenging task, where a particular garment is occluding another garment partially.

- Illumination plays a key role in the monocular reconstruction as the reconstruction is completely based on the RGB values of the image, without proper lighting, the quality of reconstruction is affected both in terms of geometrical details and color.
- The deep learning methods require huge amounts of data in order to generalize well on unseen in-the-wild images. Most of the existing 3D human data is either synthetic or captured using studio capture pipelines. These datasets often lack variations in garment styles. Hence, 3D human reconstruction in "in-the-wild" without proper large-scale real-world image-3D human datasets makes it challenging.

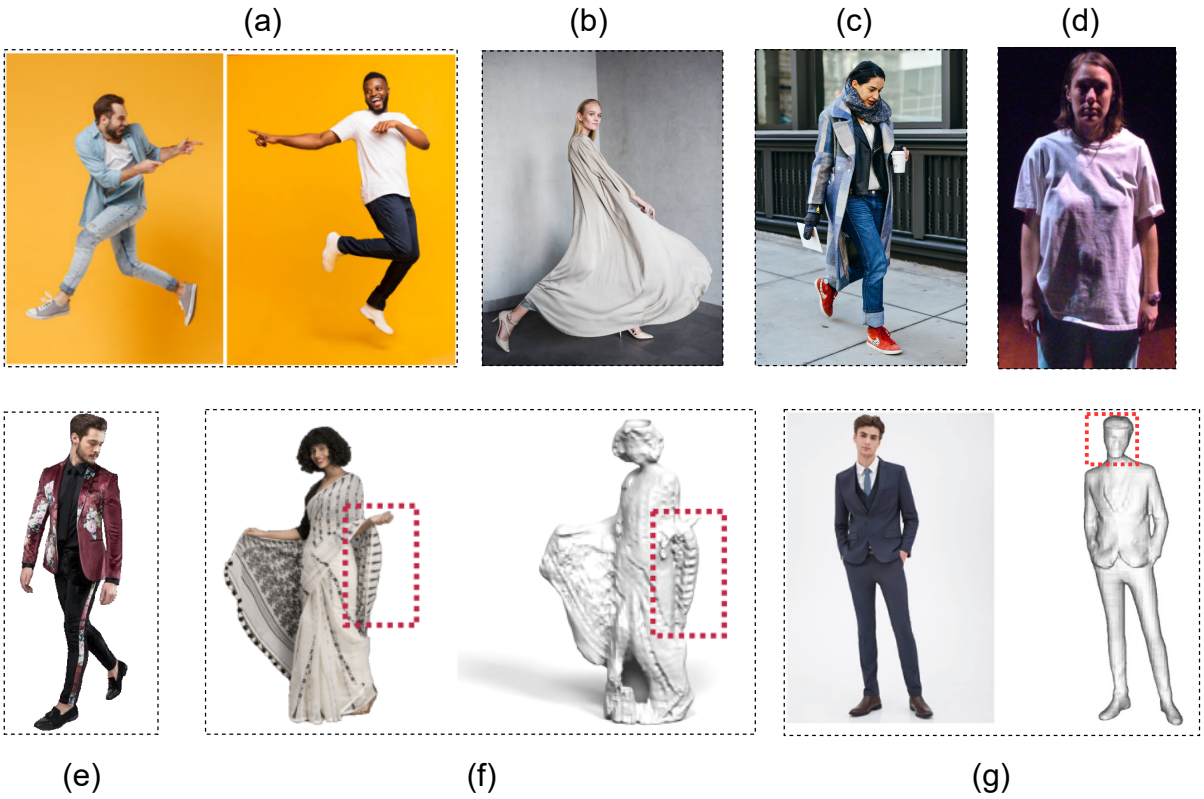

Figure 1.4 Challenges in 3D human digitization. **(a)**Skewed Complex Poses **(b)** Extreme Loose Clothing **(c)** Layered Clothing **(d)** Improper Illumination **(e)** Self Occlusion **(f)**False Geometric Details **(g)** Smooth Facial Details.

1.1.3 Problem Statement

In this thesis, we address the task of robust and high-fidelity 3D human digitization from monocular RGB images with the help of deep learning techniques in computer vision. Even though accurate reconstructions can be made using multi-view images assumptions about the calibration need to be

Figure 1.5 Problem Statement: Given a monocular RGB image as input, the final goal is to reconstruct robust photo-realisitc 3D humans.

made and this is not the case in in-the-wild scenarios. Hence we stick to 3D human reconstruction from 2D images which in general can be captured using a phone and are a more general case in the real world.

We mainly try to address the following challenges present in 3D human digitization: **(i)** Recovering the robust 3D geometry of the human under challenging clothing(loose clothing) with cloth-specific deformations using a single RGB image, and **(ii)** Realistic and person-specific reconstruction in terms of overall body shape, facial details, etc. Though the problem of estimating 3D geometry from an image is ill-posed, by using the representational power of deep learning it is plausible to model the underlying geometry of humans just using the images. In order to deal with the ambiguities in terms of occlusion, and loose clothing, it is always better to have statistical priors which give plausible pose and shape information. Though deep learning has the power to model the 3D geometry from images it still needs a huge amount of training data with various categories of clothing, poses, and shapes in order to generalize well in an in-the-wild setting and generate robust reconstructions.

1.2 Research Landscape

As mentioned in the previous section there was a need for a robust, reliable, and straightforward way to digitize 3D humans. The RGBD devices and complex capture rigs are not portable and affordable, hence there is a need for 3D human digitization from an easy-to-use setup as simple as reconstruction from monocular RGB images which are way cheaper to get than previously mentioned complex capture systems. The representational power of deep learning methods has made it feasible to reconstruct 3D humans from 2D RGB images. Thus learning-based geometry prediction methods can be employed to eliminate the need for complex 3D capture and reconstruction systems.

1. **Parametric Methods:** These methods[2, 73, 56, 77, 17, 117] build upon the statistical body model such as [62, 78, 10] and tries to optimize the pose and shape parameters of these models to match the input image. Most of the current model-based approaches optimize the pose and shape of SMPL to match image features, which are extracted with bottom-up predictors. The most popular image features are 2D joints [17], or 2D joints and silhouettes [5]. HMR[2] proposes to regress SMPL parameters when minimizing re-projection loss with the known 2D joints. These methods do well in recovering reasonably accurate body pose, but the reconstructed geometry is constrained to be within the model space which is in general low dimensional and can not capture the surface geometrical details and loose clothing. Other works [7, 6, 14] recover static body shape and clothing as displacements on top of the SMPL body model. Nevertheless, under this configuration, only tight garments can be modeled. Thus, loose garments like skirts, sarees, kurtas, and robes, which have a different topology from body shape, are beyond their representational range.

2. **Non-Parametric Methods:** The non-parametric class of monocular 3D reconstruction techniques [69, 91, 93, 101, 115] uses non-parametric representations to infer detailed geometry beyond basic body shape and pose. These non-parametric representations are either based on voxels or implicit functions and can recover arbitrary shapes. BodyNet [101] leverages intermediate tasks (e.g. segmentation, 2D/3D skeletons) and estimates low-resolution body voxel grids. Similar work to BodyNet is VRN [3], but it directly estimates occupancy voxels without solving any intermediate task. Voxel-based methods [102, 69] often produce errors at the limbs of the body and require fitting a model post-hoc [102]. However, voxel representation is of lower resolution due to memory limitations and thus fails to recover high-frequency surface geometric details.

   Whereas, deep implicit function learning techniques train deep neural networks to estimate dense, continuous occupancy fields from which meshes may be reconstructed e.g. via Marching Cubes [63]. These methods either extract either a global image feature [43, 21, 57], or pixel-aligned features [91, 93] to drive this estimation. However, these methods fail to take into account the fine-grained local shape patterns and do not seek to enforce global consistency to encourage physically plausible shapes and poses in the reconstructed mesh. This can lead to unnatural body shapes or poses, and loss of high-frequency surface details within the reconstructed mesh.

3. **Prior based Non-Parametric Methods:** DeepHuman [115] uses the SMPL as a prior to reconstruct the clothed body volume and tries to further refine the surface details using image features. However, their method fails to recover high-quality geometric surface details owing to the resolution limitation of the regular occupancy volume. Geo-PIFu [37] tries to resolve the feature ambiguity of PIFu by combining the volumetric features obtained from a 3D-UNet with 2D image local features to learn the surface as an implicit function. Both these approaches are computationally expensive and also require surface networks for surface and texture inference. ARCH

[40] proposed a semantic deformation filed(SemDF) based approach where the query points are sampled around the body in a canonical space (A -pose), an implicit surface is learned in the canonical space, and deformed using SemDF to match the pose in the input. However, it fails to generate accurate results, especially in the scenarios of loose clothing. PaMIR [114] proposes to condition the implicit field on the SMPL prior, they do this by combining the 2D image features with SMPL volume features while querying. SHARP [45] proposes to model the reconstruction as two tasks (1) deforming the SMPL body prior peeled maps in order to obtain fine-grained geometric surface details. (2) directly regress the loose clothing as auxiliary peeled depth maps and fuse both of them to obtain the reconstruction

4. **Face reconstruction methods** Similar to SMPL ([62]) for full body, 3D Morphable Models (3DMM) [28] is proposed to model face in parametric representation. Methods proposed in [61, 94, 34, 96] estimate parameters of 3DMM. Implicit functions are combined with 3D morphable models in [109, 85]. Full head models including hair are recovered in [60, 29, 23].

It can be observed that the main limitation of parametric methods is that they fail to model loose clothing due to the assumption of the available statistical templates. It is also not feasible to have multiple garment templates for the large spectrum of clothes available to model body and clothing separately. Non-parametric methods can deal with arbitrary clothing but fail to reconstruct high-frequency geometric details and noisy reconstructions in the scenarios of occlusion. The full body reconstruction methods either it is parametric, non-parametric, or prior-based non-parametric methods don't focus on recovering the facial details and are always smooth without minimal details such as eyes and nose. All of these methods are prone to noisy reconstructions when high-frequency details are present in the image space. In this thesis, we address the previously mentioned issues and provide a prior-based non-parametric solution to deal with them.

1.3 Key Contributions of Thesis

1. **Coarse-to-fine framework for 3D human digitization using peeled semantic segmentation context:** We address some of the problems existing in PeeledHuman[44] representation by proposing peeled semantic segmentation maps which help in refining the reconstruction in the occluded regions and loose clothing. The 3D segmentation can be useful in extracting the cloth out of the reconstructed 3D Human.

2. **Noise-free 3D human digitization using wrinkle map prior:** Even though peeled representation is sparse and can handle various poses it still suffers from severe occlusions. [45] Introduced a body prior such as SMPL to the peeled representation which solves the ambiguity of pose due to occlusion. However, like any other method, they also suffer from noisy reconstructions due to high-frequency details coming from texture in the image space. To address this issue, we

introduce a wrinkle map prior which distinguishes between textural and geometric high-frequency details. Thus, we obtain high-quality robust reconstructions without any false geometric noise due to textural patterns.

3. **Person-specific facial geometry prediction by incorporating facial prior:** The SMPL parametric model face doesn't capture person-specific details and is smooth owing to its low dimensional representational space. The alignment between the estimated geometry and texture is misaligned in most scenarios. We introduce a per-pixel depth-based facial prior that blends well with the peeled representation and is aligned with the facial texture in the image. We further enhance the facial details and refine the overall geometry using our proposed novel framework.

1.4 Thesis Roadmap

In this chapter, we briefly introduced the 3D human digitization problem, hardware and capture system adaptability, and challenges present in the monocular setup. A brief discussion on various classes of existing methods, their failure cases, and limitations has been made in the earlier sections. We also briefly discussed how we address these problems and challenges.

In *Chapter-2*, the background study required to have a smooth understanding of the various representations and priors used has been provided. Initially, we begin with various ways of representing 3D data, move on to SMPL representation of a naked human in detail, then the PeeledHuman [44] representation and its benefits over other representations. We also discuss how to integrate the SMPL body prior into the peeled representation as proposed in [45]

Chapter-3 addresses the challenges of noisy reconstruction due to occlusions, complex poses, and loose clothing. A novel semantic segmentation context-based approach was proposed to generate coherent shapes even in the scenarios of occlusion.

Chapter-4 focuses on reconstructing high fidelity person-specific 3D reconstruction of people in arbitrary clothing ,texture with high-frequency facial details. In order to make the reconstruction robust, body and face priors have been combined with non-parametric PeeledHuman representation.

Chapter-5 concludes the work presented in the thesis and future directions to explore.

Chapter 2

Background

In the following chapter, we build the basic knowledge required for the upcoming chapters. Initially, we discuss the various ways to represent 3D objects. Then we move on to the parametric body model SMPL which can be used as a geometric prior that guarantees shape and pose-aware 3D human reconstructions. Later on, we study PeeledHuman representation in detail which is important for modeling humans efficiently. Finally, we briefly go through SHARP which combines body prior SMPL with the PeeledHuman representation.

2.1 3D Representations

3D objects can be represented in various forms such as Raw Data (point clouds, voxels, range image), Solids (octree, BSP tree), Surfaces (Mesh, Implicit), and high-level structures. As only point cloud and mesh representations are used in this thesis we will do a background study of the same.

2.1.1 Point cloud

A point cloud can be viewed as an unstructured set of 3D point samples in a given reference coordinate system. These 3D spatial points are represented by the (x,y,z) coordinates in the given coordinate frame. You can also add RGB values to each 3D point to make it a colored point cloud. Point cloud representation captures the envelope of an object. It is possible to extract the point cloud of an object using capture devices. These can be commonly obtained using 3D acquisition devices such as LiDAR terrestrial laser scanners mounted on vehicles, depth and infrared sensing by Microsoft Azure Kinect, and recently deep learning has enabled this using Generative Adversarial Networks. A point cloud can be converted into a mesh using traditional methods such as Poisson Surface Reconstruction[51] or using the advanced deep learning approaches. The advantage of point cloud representation is that you can combine objects by merging the point lists, fast rendering, fast transformations,

Figure 2.1 3D geometry represented as point cloud and mesh.

2.1.2 Mesh

A polygonal mesh is a geometric representation of the surface which allows the subdivision operation consisting of a set of polygons. Meshes are widely used in the field of computer graphics in order to represent various objects or surfaces present in the scene. Mesh representation is also used to discretize the continuous implicit surfaces. A mesh consists of vertices (3D points in the space) linked together by faces consisting of edges between vertices to make a polygonal shape. A mesh is called triangular mesh when all the faces of the mesh are triangles. These are the meshes that one encounters in Graphics, Real World Capture systems, etc. We can say that meshes represent the geometry of a point cloud to a certain extent depending on the complexity of the mesh. Mesh representation allows us to perform various operations on mesh for e.g, simplification, decimation, subdivision, etc. One can increase or decrease the number of vertices using these operations.

The common properties of the mesh that are used to represent the polygonal meshes in various file formats like OBJ are listed below:

- Geometric Vertices: A vertex is represented by the (x,y,z) coordinates in the 3D space. To represent a vertex ”v x y z” is used. RGB values of the corresponding vertex can also be added and are represented using ”v x y z r g b”.

- Vertex Normals: The vertex normal represents the direction associated with the vertex. Vertex normals are represented as ”vn dx dy dz”.

- Texture coordinates: If the mesh is uv parameterized then for each 3D vertex there is a 2D uv coordinate. These coordinates are represented as ”vt u v”.

- Face: A face consisting of only vertices is represented by "f v1 v2 ... vn". Face with normals is represented by "f v1//n1 v2//n2 .... vn//nn". Face with texture coordinates is represented by "f v1/t1 v2/t2 .... vn/tn". Face with both normals and texture coordinates is represented using "f v1/t1/n1 v2/t2/n2 .... vn/tn/nn", where vi,ni, and ti are the indices of the corresponding vertex, texture coordinate and normals.

2.2 SMPL: Skinned Multi-Person Linear model

SMPL[62] is a parametric model of human that disentangles the surface of a body into shape(β) and pose(θ) spaces. SMPL initially starts with a triangulated template mesh $\bar{T}$ which is in rest pose and consists of 6890 vertices in total. Based on the shape(β) and pose(θ) parameters offsets are added to $\bar{T}$ corresponding to shape-dependent deformations $B_S(\beta)$ and pose dependent deformations $B_P(\theta)$ respectively. We obtain the final mesh by posing the offset deformed template using the skinning function W. Formally :

$$T(\beta, \theta) = \bar{T} + B_S(\beta) + B_P(\theta) \tag{2.1}$$

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \theta, \mathcal{W}) \tag{2.2}$$

where the rest pose vertices of the deformed mean template $\bar{T}$ are posed using the LBS function W around their respective parent joints which can be obtained using a joint regressor matrix that weighs vertices contribution differently for different joints. The blend weights smooth the vertices and return the final posed vertices. The pose $\theta \in R^{(3\times(23+1))}$ of the SMPL is represented using the axis angle representation consisting of relative 3D rotations for each of the 23 joints and one for the global rotation of the body.

The entire SMPL workflow has been demonstrated in the Figure 2.2. Initially, a handcrafted template mesh $\bar{T}$ with blend weights $\mathcal{W}$ by an artist as shown in column (a). These vertices are deformed by adding (b) 3D offsets due to shape-dependent deformations $B_S(\beta)$ resulting in updated joint locations which are regressed using a joint regressor matrix. In (c) pose dependent deformations are added to the mesh. In (d) the mesh is posed using the pose parameters θ using linear blend skinning function W and skinning weights $\mathcal{W}$.

Shape parameters: Shape parameters(β) consist of a vector of 10 scalar values. These 10 parameters control the amount of expansion or shrinkage of a human subject along the principal components which can represent physical features such as height, width, etc. The shape basis is obtained by performing Principal Component Analysis on a dataset of various human subjects.

Pose Parameters: Pose parameters(θ) is a vector consisting of 24x3 scalar values that stores the relative rotations of joints in axis angle representation relative to the parent joint.

Though SMPL can simulate a wide spectrum of shape variations it still cannot handle extreme shapes owing to its low dimensional representation. SMPL doesn't capture the facial expression and hand pose.

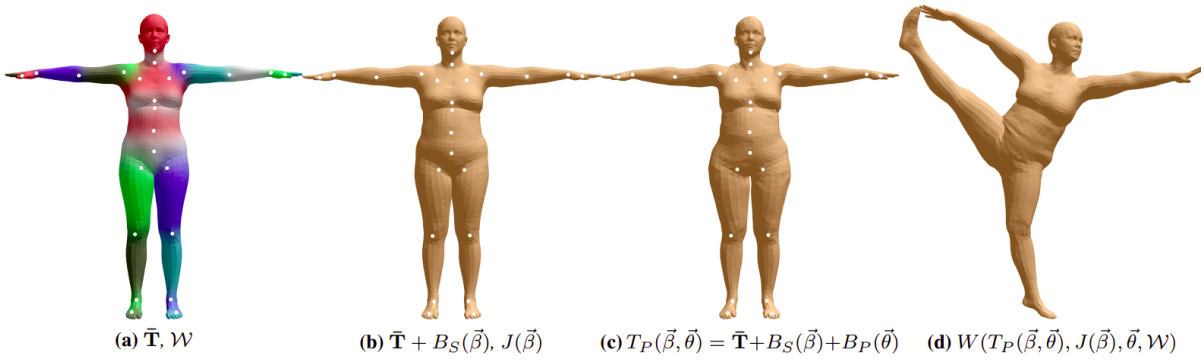

Figure 2.2 (a) Triangulated template mesh $\bar{T}$ (b) Shape-dependent deformations (c) Pose-dependent deformations; observe the change in shape of hips region. (d) Final mesh after reposing using blend weights. Figure borrowed from [62]

An extension of SMPL with different basis spaces for the face, hand, and the remaining body SMPL-X [78] was proposed with 300 shape parameters that have more control on the overall body shape, facial expressions, and hand pose.

2.3 PeeledHuman Representation

The PeeledHuman [44] representation modeled the human body as a set of multi-layered sparse peeled maps similar to the layered depth images in novel view synthesis. They estimate a fixed number of ray intersection points with the human body surface in the canonical view volume for every pixel in an image, yielding a multi-layered shape representation called PeeledHuman. It has been demonstrated that this representation can handle self-occlusions, generic poses, and loose clothing.

Under the assumption that the human body is a non-convex object placed in a virtual scene, a set of rays originating from the camera center are traced through each pixel to the 3D world. The first set of ray-intersections with the 3D body is recorded as depth map D^1 and RGB map R^1, capturing the visible surface details nearest to the camera. Subsequently, the rays are extended beyond the first bounce to hit the next intersecting surface. The corresponding depth and RGB values of the next layer are represented by D^i and R^i, respectively as illustrated in Figure 2.3. PeeledHuman [44] demonstrated that 4 intersections of each ray i.e., 4 *Peeled Depth & RGB maps* are sufficient to faithfully reconstruct a human body, which can handle self-occlusions caused by the most frequent body poses. Subsequently, a point cloud can be reconstructed from these depth maps using classical camera back projection. The RGB vertex colors are transferred to the point cloud from respective RGB peel maps to obtain a high-density colored point cloud as illustrated in Figure 2.4. The point cloud can be converted to mesh using Screened Poisson Surface Reconstruction [52]. This sparse layered 2D representation is more

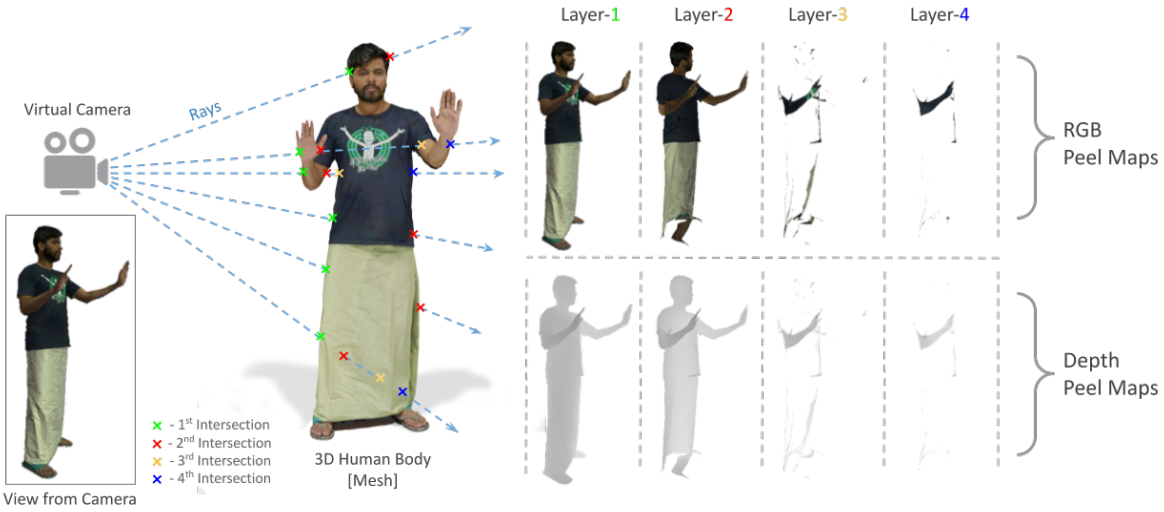

Figure 2.3 Encoding of 3D human body using PeeledHuman Representation. Courtesy:SHARP[45].

efficient to process than volumetric and implicit functions as it only stores ray-surface intersection in a 2D multi-layered layout.

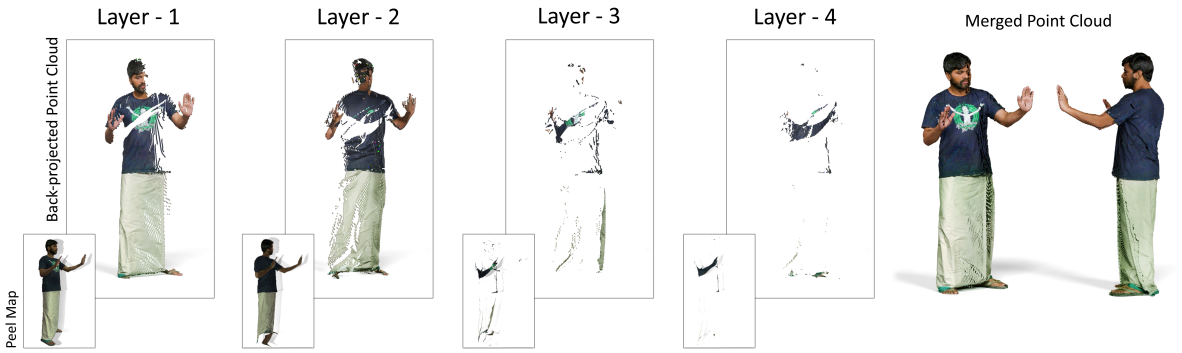

Figure 2.4 Back-projecting peelmaps to form a point cloud.

2.4 Combining Parametric Prior with PeeledHuman representation

SHARP[45] imposes the geometric shape constraints on the PeeledHuman[44] representation by incorporating a parametric body prior in the form of SMPL prior peeled maps $\mathcal{D}_{smpl}$. Sharp disentangles the reconstruction of 3D clothed body surface into two complementary reconstruction tasks (1) recovering the person-specific geometry and surface by deforming the SMPL prior peeled maps $\mathcal{D}_{smpl}$ by predicting the residual peeled maps $\hat{\mathcal{D}}_{rd}$ which is the difference between ground truth and SMPL peeled maps wherever SMPL peeled map exists. (2) recover the remaining surface details of the loose clothing in the form of auxiliary peeled maps $\hat{\mathcal{D}}_{aux}$ as illustrated in Figure 2.5. The SMPL prior peeled maps

$\mathcal{D}_{smpl}$, residual peeled maps $\hat{\mathcal{D}}_{rd}$ and auxiliary peeled maps $\hat{\mathcal{D}}_{aux}$ are further combined using SMPL mask to obtain the final fused peeled maps $\hat{\mathcal{D}}_{fused}$ as shown in the Figure 2.6.

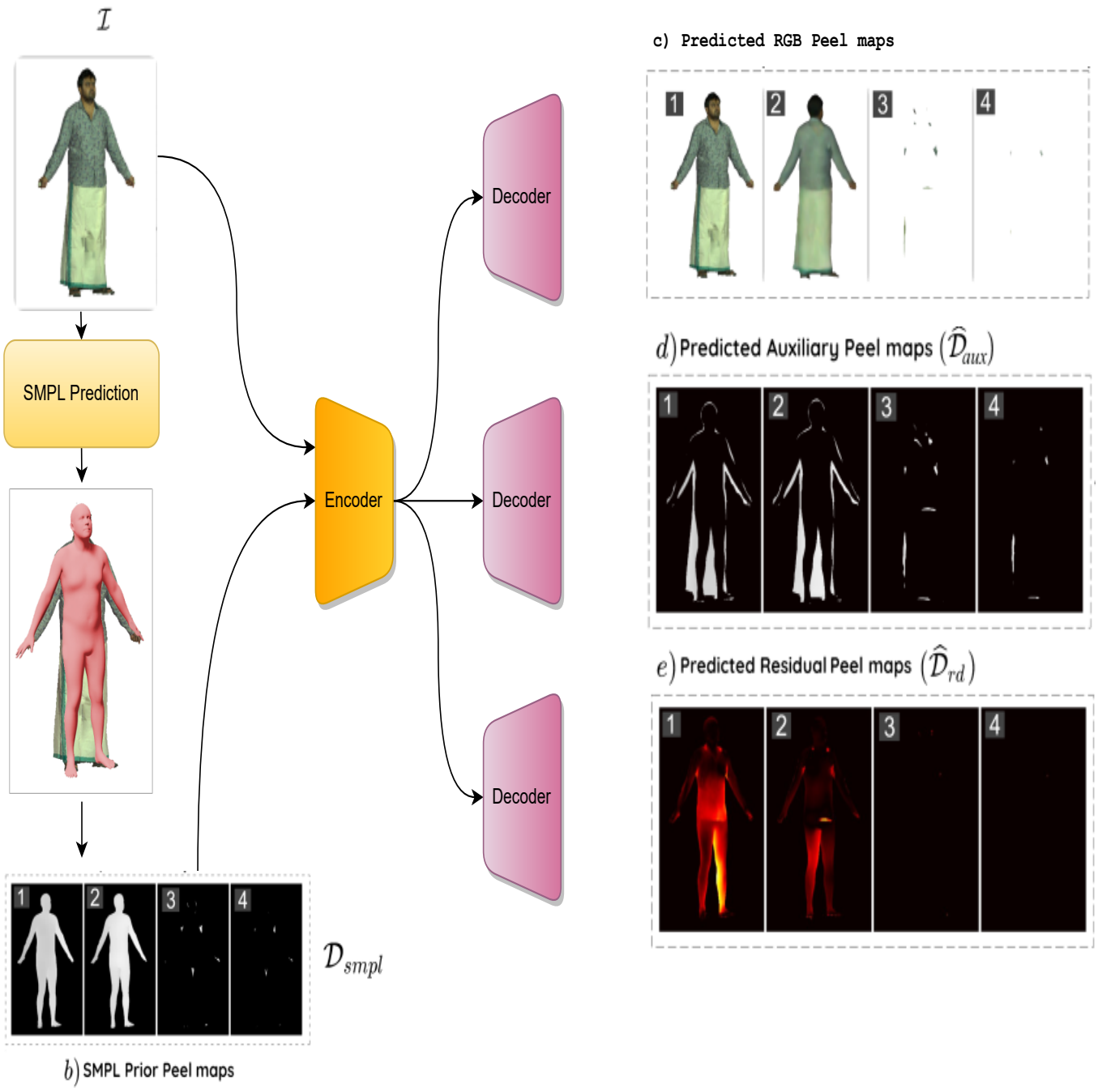

Figure 2.5 SHARP Pipeline

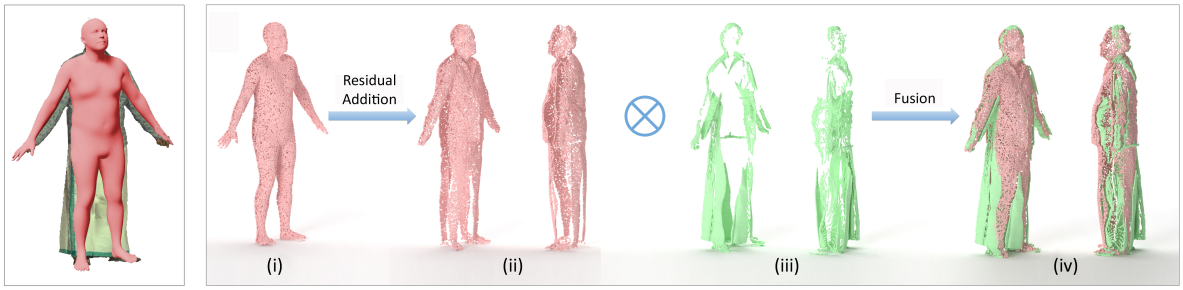

Figure 2.6 Fusion of residual and auxiliary peeled maps. (a) SMPL prior overlayed on the input image: The residual peeled maps recover depth along the pixels over which SMPL prior is present across all the layers. For the remaining pixels, auxiliary peeled maps are used to recover depth. (b) 3D representation of fusion: (i) Point cloud obtained D_{smpl} is illustrated in red. (ii) Point cloud obtained from $D_{smpl}+\hat{D_{rd}}$ is illustrated from two views in red. (iii) Point cloud obtained from $\hat{D}_{aux}$ is illustrated from two views in green. (iv) Final point cloud obtained after fusion. Courtesy:SHARP[45].

Chapter 3

Coarse-to-fine 3D Clothed Human Reconstruction using Peeled Semantic Segmentation Context

In the previous chapter, we discussed the background knowledge required for this dissertation. This chapter discusses the proposed coarse-to-fine reconstruction approach to deal with partial self-occlusion and loose clothing. Our proposed peeled semantic segmentation maps enable us to capture 3D shape context resulting in robust and coherent reconstructions. Along with 3D human reconstruction, we have additional 3D segmentation labels. An application of these 3D semantic labels is to extract the clothing from the reconstructed mesh.

3.1 Introduction

3D reconstruction of human body model is a very challenging yet desired research problem with applications in VR/AR domain, entertainment & gaming as well as marketing domains (e.g., virtual try-on). Metrically accurate and precise reconstructions of humans is now possible with calibrated multi-view systems [48, 87] that uses RGB or structured light cameras. Nevertheless, these techniques have remained largely inaccessible to the general community due to its reliance on professional capture systems with strict environmental constraints like large number of cameras, controlled illuminations etc., that are overly expensive.

On the other hand, reconstruction from a single view is an inherently under-constrained problem. However, recent advancement in deep learning domain [2, 32, 70, 99, 67, 50] have shown great promise in acquiring 3D body reconstructions from a monocular image. The first class of parametric body model based monocular shape reconstruction techniques infer the shape and pose parameter of a statistical body model, like SMPL [65, 30, 27]. These methods do well in recovering reasonably accurate body pose, but the reconstructed geometry is constrained to be within the model space, which can not capture the geometrical details including surface geometrical details and loose clothing. Other recent works [9, 99, 97, 12], recover static body shape, and clothing as displacements on top of the SMPL body model [65] (model-based). Semantic segmentation labels have been used to refine parametric body estimation

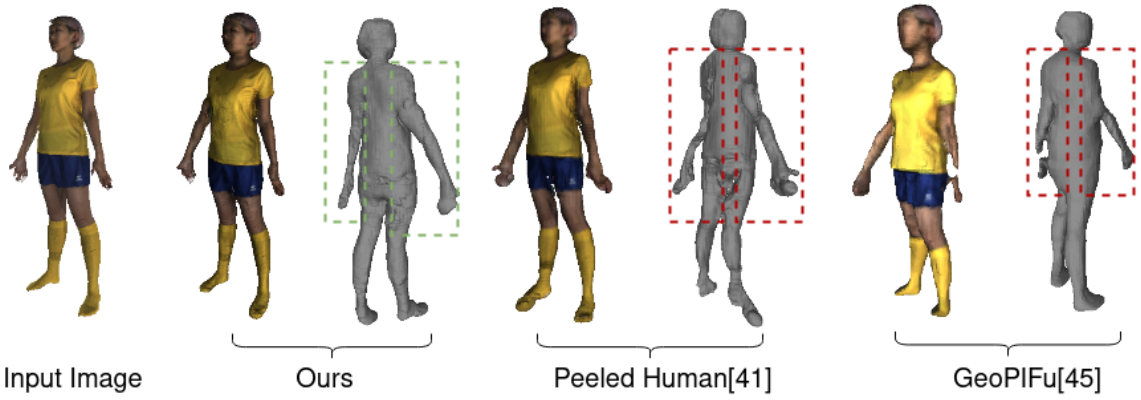

Figure 3.1 Our coarse-to-fine refinement approach using semantic segmentation context yields superior reconstruction of 3D human body from monocular input image compared to other SOTA methods.

[103]. Nevertheless, under this configuration only tight garments can be reconstructed. Thus, garments like skirts and robes, which have a different topology with body shape, are beyond their representation range.

The second class of monocular 3D reconstruction techniques [70, 92, 89, 32, 116] uses non-parametric representations to infer detailed geometry beyond basic body shape and pose. These non-parametric representations are either based on voxels or implicit function and can recover arbitrary shapes. BodyNet [32] leverages intermediate tasks (e.g. segmentation, 2D/3D skeletons) and estimates low-resolution body voxel grids. A similar work to BodyNet is VRN [3], but it directly estimates occupancy voxels without solving any intermediate task. Voxel-based methods [33, 70] often produce errors at the limbs of the body and require fitting a model post-hoc [33]. However, voxel representation are typically of lower resolution due to memory limitations and thus fails to recover high frequency surface geometry details. Whereas, deep implicit function learning techniques train deep neural networks to estimate dense, continuous occupancy fields from which meshes may be reconstructed e.g. via Marching Cubes [63]. These methods either extracts either a global image feature [43, 21, 57], or pixel-aligned features [92, 89] to drive this estimation. However, these methods fails to take into account the fine-grained local shape patterns and do not seek to enforce global consistency to encourage physically plausible shapes and poses in the reconstructed mesh. This can lead to unnatural body shapes or poses, and loss of high-frequency surface details within the reconstructed mesh. Recently proposed model GeoPiFu [98] attempt to combine both volumetric and implicit function learning by providing U-Net based volumetric features as prior for implicit function prior. However, their method is computationally intensive and perform dense inference at test time (similar to [92]).

An alternative set of non-parametric approaches attempt to model 3D objects/scenes as sparse layered representation. [100] attempted to predict 3D human body reconstruction by predicting the frontal

and backside depth maps from monocular image. However, they failed to model the scenarios of self-occlusions by body parts, clothing etc. [88] attempted to address the self-occlusion problem by predicting multiple peeled depth maps. They estimate a fixed number of ray intersection points with the human body surface in the canonical view volume for every pixel in an image, yielding a multi-layered shape representation called **PeeledHuman**. Their representation encodes a 3D shape as a set of layered depth maps called as *Peeled Depth maps* and the surface texture/color maps called as *Peeled RGB maps*. Subsequently, they learn to predict these peeled RGB and Depth Maps from monocular input image using their *PeelGAN* model. Their proposed shape representation allows to recover multiple 3D points that project to the same pixel in the 2D image plane thereby overcoming the problem of self-occlusion. Recently, [90] proposes to introduce parametric body prior to variant of PeelGAN. However, they require good estimate of SMPL prior fitted to input image.

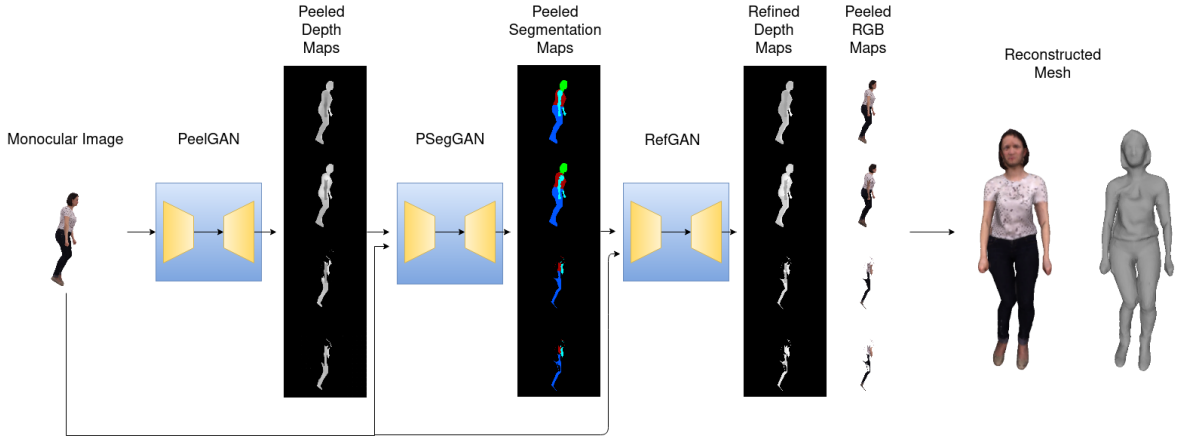

Figure 3.2 Coarse-to-fine refinement framework for 3D reconstruction of human body using peeled semantic segmentation.

In this paper, we address the problem of detailed full-body shape reconstruction from a single image. We adopt the PeeledHuman [88] representation as this encoding is sparse and robust to self-occlusions. However, the key limitation of their PeelGAN model is that predicted depth maps of self-occluded parts are sparse and noisy, and hence after back-projection lead to distorted body parts and sometimes discontinuity between them. We hypothesized that introducing coherence in semantic labels would alleviate such issues. Additionally, the semantic labels also helps in improving the generalisation of our model as shown in Figure 3.1 where we compare our method with SOTA. It is important to note that, we propose and use peeled semantic representation that enables capturing 3D shape context as compare to traditionally used segmentation map for monocular rgb image. To summarize, we propose to introduce *Peeled Segmentation map* representation in a coarse-to-fine refinement framework which consist of a cascade of three networks namely, PeelGAN, PSegGAN and RefGAN, as shown in Figure Figure 3.2. At first, we use original PeelGAN [88] as baseline model to predict initial coarse estimation of peeled depth maps from input RGB image. These peeled maps are subsequently fed as input along with monocular

RGB image to our novel PSegGAN which predict *Peeled Segmentation maps* in a generative fashion. These peeled segmentation maps are spatially aligned to peeled depth maps and thus we get semantic segmentation label corresponding to each pixel of the peeled depth maps. Finally, we fed these peeled segmentation maps as additional context along with monocular input image to our RefGAN to predict the refined peeled RBG and Depth maps. We observe that providing the semantic segmentation context improves the reconstructions both quantitatively and qualitatively especially in the case of self-occlusions. This also provide an additional output as 3D semantic segmentation of the reconstructed shape. We perform a thorough empirical evaluation over four publicly available datasets to demonstrate the superiority of our model over [88]. Additionally, we intend to release the semantic segmentation label data and proposed model code for the community to use.

3.2 Literature Survey

3.2.1 Parametric 3D Reconstruction

Many existing methods have attempted estimating a naked body shape (statistical models [27, 65]) from image [2, 59, 81, 67, 19, 26, 47, 54, 108]. Most of the current model based approaches optimize the pose and shape of SMPL[65, 31] to match image features, which are extracted with bottom-up predictors. The most popular image features are 2D joints [19], or 2D joints and silhouettes [8, 9, 36]. HMR proposes to regress SMPL parameters when minimizing re-projection loss with the known 2D joints. For better representation ability, a displacement vector is added for each vertex. [9, 99, 8] adopt this strategy to reconstruct clothed body with skin-tight garment. Alldieck et al. [97] estimate detailed normal and vector displacement on the UV map, which leads to finer-scale details. Zhu et al. [119] model fine-scale details by adding free-form 3D deformation on top of parametric model. Instead of using a single surface to represent both garment and body, [12] separates SMPL mesh to represent upper garment and pant independently, leading to more flexible control. However, binding garment vertices to the body model strictly restricts the topology of support garment categories, and it is hard to represent more loose garment types, such as skirts. [83] also uses separate body and garment templates to register clothed body motion sequences. Pixel2Mesh [34] generates water-tight meshes by progressively deforming a sphere template.

3.2.2 Non-parametric 3D Reconstruction

Some non-parametric methods based on voxel or implicit function have been proposed to address the complex topology of garments. Volumetric regression [32, 4, 111] directly infers a voxel representation of clothed bodies with a deep network. Due to the large memory cost for high resolution, high-frequency details are often missed. [116] infer clothed body volume representation with an initial aligned SMPL body, and combine image features to enhance reconstruction details. Natsume et al. [70] propose a reconstruction method based on a multi-view framework using synthesizing new silhouettes

from a single image. TSDFs [22] can represent the human surface implicitly, which is common in depth-fusion approaches [71, 72]. Occupancy Networks [57], DeepSDF [43] and LIF [21] proposed to use global representations of a single-view image input to learn deep implicit surface functions for mesh reconstruction. However, the global representation-based implicit function does not have dedicated query point encodings and thus lacks modeling power for articulated parts and fine-scale surface details.This motivates later works of PIFu [92] and DISN [84]. They utilize pixel-aligned 2D local features to encode each query point when estimating its occupancy value. The alignment is based on (weak) perspective projection from query points to the image plane, followed by bi-linear image feature interpolation. However, PIFu still suffers from the feature ambiguity problem and lacks global shape robustness. Another two PIFu variations are PIFuHD [89] and ARCH [41]. PIFuHD leverages higher resolution input then PIFu through patch-based feature extraction to accommodate GPU memory constraints. ARCH combines parametric human meshes (e.g. SMPL [65]) with implicit surface functions in order to assign skinning weights for the reconstructed mesh and enable animations. Both these methods require more input / annotations (e.g. 2×higher resolution color images, SMPL registrations) than PIFu. Geo-PIFu [98] is volumetric-regression based approach that incur high computational and memory costs. Unsupervised estimation of implicit functions has been addressed in [86, 66] . Peeled maps [88] proposed a sparse representation by estimating only surface intersections by posing the problem as an extension of ray tracing.

3.3 Method

3.3.1 PeeledHuman Representation

Peeled representation is a sparse, non-parametric encoding of 3D shape proposed in [88] where 3D human body is modeled as a set of *Peeled Depth & RGB maps*. Under the assumption that human body is a non-convex object placed in a virtual scene, a set of rays originating from the camera center are traced through each pixel to the 3D world. The first set of ray-intersections with the 3D body are recorded as depth map d_1 and RGB map r_1, capturing visible surface details nearest to the camera. Subsequently, the rays are extended beyond the first bounce to hit the next intersecting surface. The corresponding depth and RGB values of the next layer are represented by d_i and r_i, respectively. Additionally, [88] claimed that 4 intersections of each ray i.e., 4 *Peeled Depth & RGB maps* are sufficient to faithfully reconstruct a human body, which can handle self-occlusions caused by the most frequent body poses. Subsequently, a point-cloud can be reconstructed from these maps using classical camera back-projection methods. Please refer to [88] for visualization and other details. This sparse 2D representation is more efficient to process than volumetric and implicit functions as it only stores ray-surface intersection in a 2D multi-layered layout.

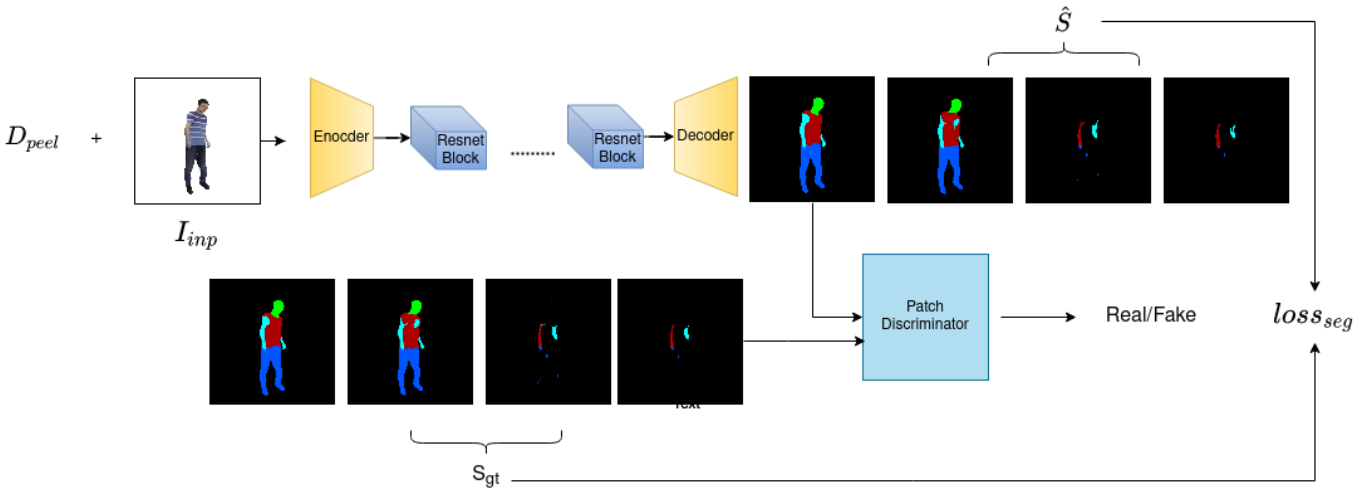

Figure 3.3 Architecture of the proposed PSegGAN that predicts Peeled Segmentation maps given input monocular image and coarse Peeled Depth maps predicted from PeelGAN.

3.3.2 Overview & Notations

Our objective is to reconstruct a 3D Human with clothing from a given monocular RGB Image I_{inp}. We employ a coarse-to-fine strategy with a cascade of multiple GANs. Initially, pre-trained **PeelGAN** [88] is used to predict the coarse peeled depth maps D_{peel} from monocular image I_{inp}. Subsequently, these predicted D_{peel} along with input RGB image are fed to our proposed **PSegGAN** which outputs *Peeled Segmentation maps* $\hat{S}$ in a generative fashion using the supervision of ground-truth *Peeled Segmentation maps* S_{gt}. Later, we feed these predicted segmentation maps $\hat{S}$ as a semantic context along with the monocular RGB image I_{inp} to the proposed **RefGAN** and predict the final *Peeled Depth maps* $\hat{D}$ *and RGB maps* $\hat{R}$. We propose **RefGAN** as a conditional generative adversarial network to refine the coarse peeled maps. We train the network in supervised manner using the ground-truth *Peeled Depth* (D_{gt}) *and RGB* (R_{gt}) *maps*.

In the case of self-occlusion, the peeled semantic segmentation maps provides refined information about the boundary of occluded parts. Even in case of poses without self-occlusion the semantic information provides the spatial homogeneity context for pixels belonging to the same part. Finally, we can back-project the predicted $\hat{D}$ and $\hat{R}$ into a 3D point-cloud using the known camera intrinsic parameters, which are dependent only on the image resolution. We further extract the mesh from the resulting point-cloud using the Poisson surface reconstruction [51] Note that the camera intrinsic parameters is dependent on input resolution and is fixed across the datasets.

3.3.3 Peeled Segmentation Network (PSegGAN)

In order to generate the *Peeled Segmentation maps ($\hat{S}$)* from *Peeled Depth maps* D_{peel} and RGB image I_{inp} in a generative fashion, we propose PSegGAN, a conditional GAN. The input to our PSegGAN is D_{peel} & I_{inp} and it generates *Peeled Segmentation maps* $\hat{S}$ as output as shown in Figure 3.3. We use cross-entropy loss over the predicted segmentation maps $\hat{S}$ and ground-truth segmentation maps S_{gt} as it is a classification task along with the GAN loss. The final loss function is shown below:

$$loss_{total} = w_g * loss_{GAN} + w_s * loss_{seg} \tag{3.1}$$

where w_g and w_s are the weighting factors for $loss_{GAN}$ and $loss_{seg}$.

$$loss_{GAN} = E_{I_{inp},S_{gt}}[logD(I_{inp}, S_{gt})] + E_{I_{inp},\hat{S}}[log(1 - D(I_{inp}, \hat{S}))] \tag{3.2}$$

and

$$loss_{seg} = \sum_{i=1}^{n}\sum_{k=1}^{c} -t_k^i * log(\hat{y}^i) \tag{3.3}$$

where n is the total number of pixels in the image,c is the total number of part labels of human, t_k^i is 1 when $S_{gt}^i = c$ other wise it is 0 and $\hat{y}$ is the last layer of PSegGan i.e $argmax(\hat{y}) = \hat{S}$

3.3.4 Refined Peeled Map Prediction (RefGAN)

We predict the refined *Peeled Depth and RGB maps* using a generative network, named **RefGAN**. The RefGAN takes a set of peeled segmentation maps $\hat{S}$ and monocular RGB image I as input and generates refined *Peeled Depth $\hat{D}$ and RGB Maps $\hat{D}$*. The peeled segmentation maps provide dense semantic context to our generator network RefGAN. Refer Figure 3.4 for the architecture. We use L_1 loss over the predicted and ground-truth depth maps and RGB maps. In order to maintain 3D structure consistency, we add Chamfer loss over point-cloud obtained by back projecting the depth maps. Additionally, in order to enforce smoothness over predicted depth maps, we make first order gradients of generated depth maps $\hat{D}$ to be similar with ground-truth depth maps D_{gt}. So the combined loss function is shown as below:

$$L = w_g * l_{GAN} + w_l * l_{L_1} + w_c * l_c + w_{sm} * l_{sm} + w_r * l_{rgb} \tag{3.4}$$

where $w_g, w_l, w_r, w_c, w_{sm}$ are the respective weights for
l_{GAN}, l_{L1}, l_{rgb}, l_c and l_{sm}.

Below, we define each loss terms:

$$\begin{aligned} l_{GAN} = E_{I_{inp},D_{gt}}[log(D(I_{inp}, D_{gt})] + E_{I_{inp},\hat{D}}[log(1 - D(I_{inp}, \hat{D}))] \\ + E_{I_{inp},R_{gt}}[log(D(I_{inp}, R_{gt}))] + E_{I_{inp},\hat{R}}[log(1 - D(I_{inp}, \hat{R}))] \end{aligned} \tag{3.5}$$

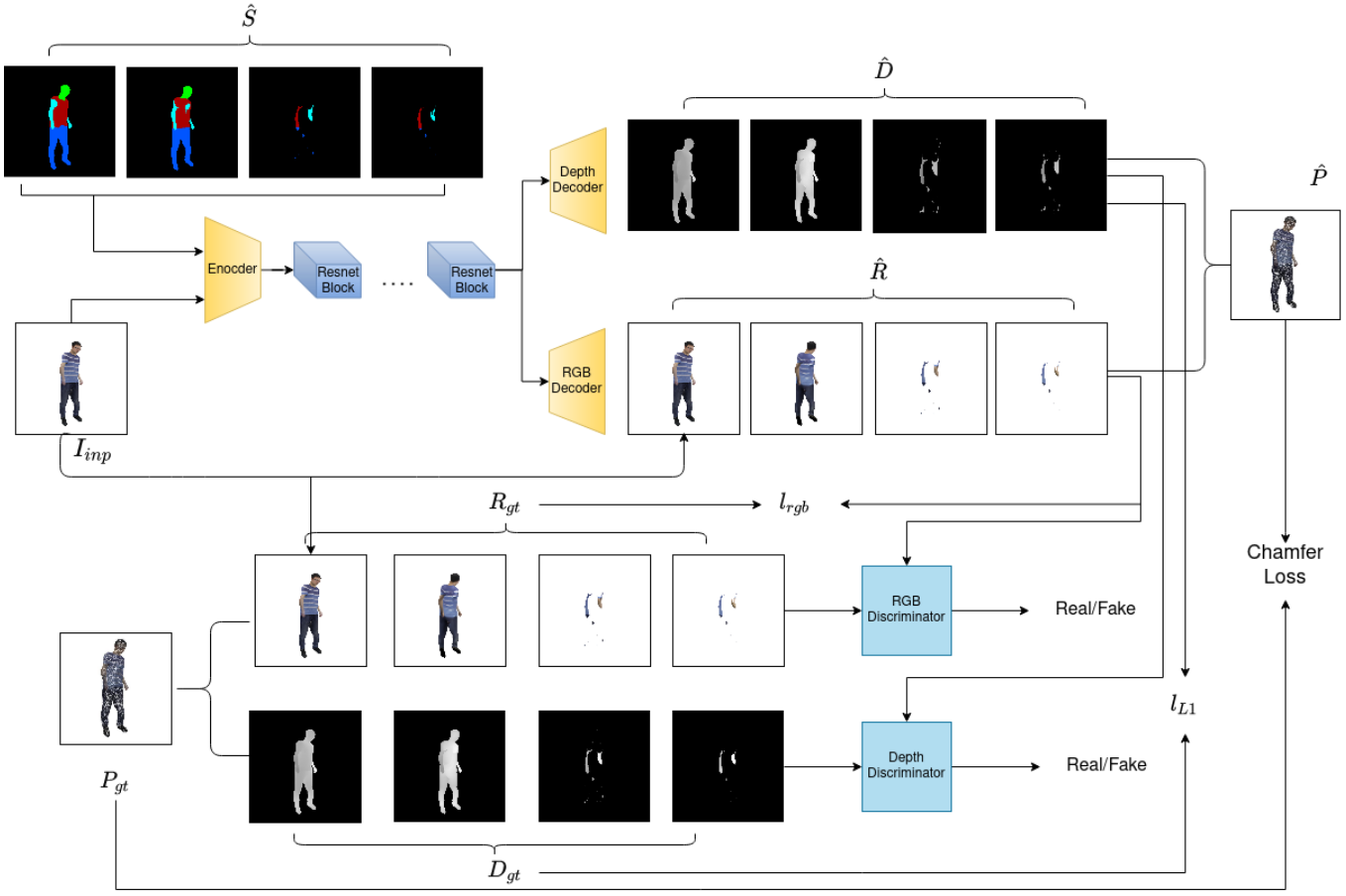

Figure 3.4 Architecture of the proposed RefGAN that predicts refined *Peeled Depth and RGB maps* given input monocular image and *Peeled Segmentation maps* predicted by PSegGAN.

Depth Pipeline

We define depth loss (l_{L_1}) as:

$$l_{L1} = \sum_{i=1}^{4} m_i \left| \hat{D}^i - D^i_{gt} \right| \tag{3.6}$$

where i is the i^{th} layer in peeled maps, m_i is 1 for occupied pixels (segmentation label at the pixel is not background), else 0 for depth maps ($\hat{D}^1, \hat{D}^2$) and m_i is 10 for occupied pixels, else 0 for depth maps ($\hat{D}^3, \hat{D}^4$), as these are sparse and hence more weight is given to them.

Smoothness loss (l_{sm}) is defined as :

$$l_{sm} = \sum_{i=1}^{4} \left| \nabla \hat{D}^i - \nabla D^i_{gt} \right| \tag{3.7}$$

To further improve the RGB peel maps conditioned on generated segmentation peel maps ($\hat{S}$), we use l_{rgb}

$$l_{rgb} = \left| \hat{R} - R_{gt} \right| \tag{3.8}$$

To ensure 3D consistency, we further use the chamfer loss l_c between predicted and ground truth point clouds.

$$l_c = \sum_{\hat{p} \epsilon \hat{P}} \min_{p \epsilon P_{gt}} ||\hat{p} - p||_2^2 + \sum_{p \epsilon P_{gt}} \min_{\hat{p} \epsilon \hat{P}} ||p - \hat{p}||_2^2 \tag{3.9}$$

3.3.5 Network Architecture

The peeled segmentation network (PSegGAN) consists of an encoder block, followed by residual blocks and a decoder block. The input to the network is 512x512 monocular RGB image and four peeled depth maps predicted from PeelGAN network. Dropout and Batch Normalisation are used for regularisation. The activation function used is ReLU. The RefGAN network also use the same architecture of encoder, residual blocks but different decoder branches for depth and RGB maps prediction. The activation function used is ReLU except for the last convolution layer of decoders where sigmoid is used in depth decoder and Tanh is used in RGB decoder. The architecture of the networks is similar to [88]. The discriminator for both the networks is a patch-based discriminator similar to *PatchGAN* proposed in [76].

3.4 Experiments and Results

3.4.1 Datasets

3.4.1.1 Thumans

The dataset [115] consists of 6800 human meshes registered with SMPL body in varying poses and garments. We manually do semantic segmentation on the SMPL body. For each vertex in the ground-truth mesh, we find its nearest neighbour in the registered SMPL mesh and assign its label.

3.4.1.2 MonoPerfCap

The dataset [107] consists of 13 daily human motion sequences in relatively tight clothing. It has approximately 40,000 3D human body models. We manually segment the meshes of human subjects.

3.4.1.3 Cloth3D

The dataset [35] comprises of 6500 sequences of draped SMPL meshes simulated with MoCap data. Each frame of a sequence contains a garment and an SMPL body. We augment this data by capturing SMPLtexture maps with minimal clothing to simulate realistic body textures. For each sequence, five frames are randomly sampled and the naked body is subtracted from the garment. Thus, we obtain clothing occluded body as ground-truth for training. For semantic segmentation, we assign labels to clothing and manually label SMPL body. While we perform the peeling of the mesh as suggested in [88], we assign the cloth label if the naked body is occluded by cloth, else the semantic labels of SMPL are assigned to a pixel in the peeled segmentation map.

3.4.1.4 Buff

The dataset consists of 5 subjects with tight clothing performing simple motions. For generating segmentation labels we register an SMPL-D mesh with the original meshes. We have a segmented template SMPL mesh as prior, using this while performing peeling of the mesh we assign label to a particular vertex the label of its closest vertex in registered SMPL-D mesh. Figure 3.5 shows the sample mesh which is registered with SMPL and transferred labels.

3.4.2 Implementation Details

Our model is implemented in PyTorch and trained on 4 Nvidia RTX GPU's in parallel. In Cloth3D, BUFF and MonoPerfCap datasets, we keep subjects (which are not in training) for testing where we follow 80-20 training and test split. For THumans, we adopted test/train split from [98]. We initially train PeelGAN and freeze the network. Our segmentation network is trained for around 30 epochs with a batch size of 8 with an image resolution of 512x512 with w_g, w_{seg} set as 1 and 100. We freeze the

Table 3.1 Quantitative results on Thuman Dataset

Method	ChamferDistance
DeepHuman [116]	0.0011
PeeledHuman [88]	0.00051
GeoPiFu [98]	0.00017
Our Method	0.00037

segmentation network and pass its output to the depth refinement network. Our depth refinement network is trained for 50 epochs with a batch size of 8 and same image resolution with $w_g, w_l, w_r, w_c, w_{sm}$ as 1,500,500,100 and 10. Adam optimiser was used to train both networks, initialised with a learning rate of 0.0002, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We train our segmentation network and depth refinement network in a sequential fashion. The ground-truth segmentation, depth and RGB maps are produced by performing ray tracing using Trimesh library. The reconstructed point-cloud has around 50,000 points.

3.4.3 Results

We performed both quantitative and qualitative evaluations on Thuman, MonoPerfCap, Buff and Cloth3D dataset. Figure 3.6 shows the qualitative results of our method on all four datasets where our method seems to perform well on variations in body shape, pose and (loose) clothing.

We also compared our method with other state-of-the-art (SOTA) methods. Table 3.1 report quantitative results using Chamfer distance on Thuman dataset. As we can see that our method achieves lower Chamfer distance on the test set as compared to DeepHuman [116] and PeeledHuman [88]. Although GeoPiFu [98] seems to perform superior in terms of Chamfer distance, we compare the qualitative results with some of the test samples where our method reconstructs more consistent 3D body models as compare to GeoPiFu, as shown in Figure 3.9. Also note that in Figure 3.1, we show the result on BUFF dataset where all the models are trained on THumans and tested on BUFF. This experiment proves the generalizability of our model when tested on other datasets. Additionally, GeoPIFu training takes 7 days on six GPUs while our training takes less than three days on four GPUs. Please refer to supplementary for more results.

We also compared with the most relevant PeeledHuman [88] method on Cloth3D and MonoPerfCap dataset where our method achieves lower Chamfer distance, as reported in Table 3.2. This implies that the PeeledHuman representation require more semantic context to reconstruct plausible shapes. Figure 3.8 shows the comparison with [88] where we render predicted point clouds (raw output) on all four datasets where our method consistently achieves superior quality reconstruction results. We are

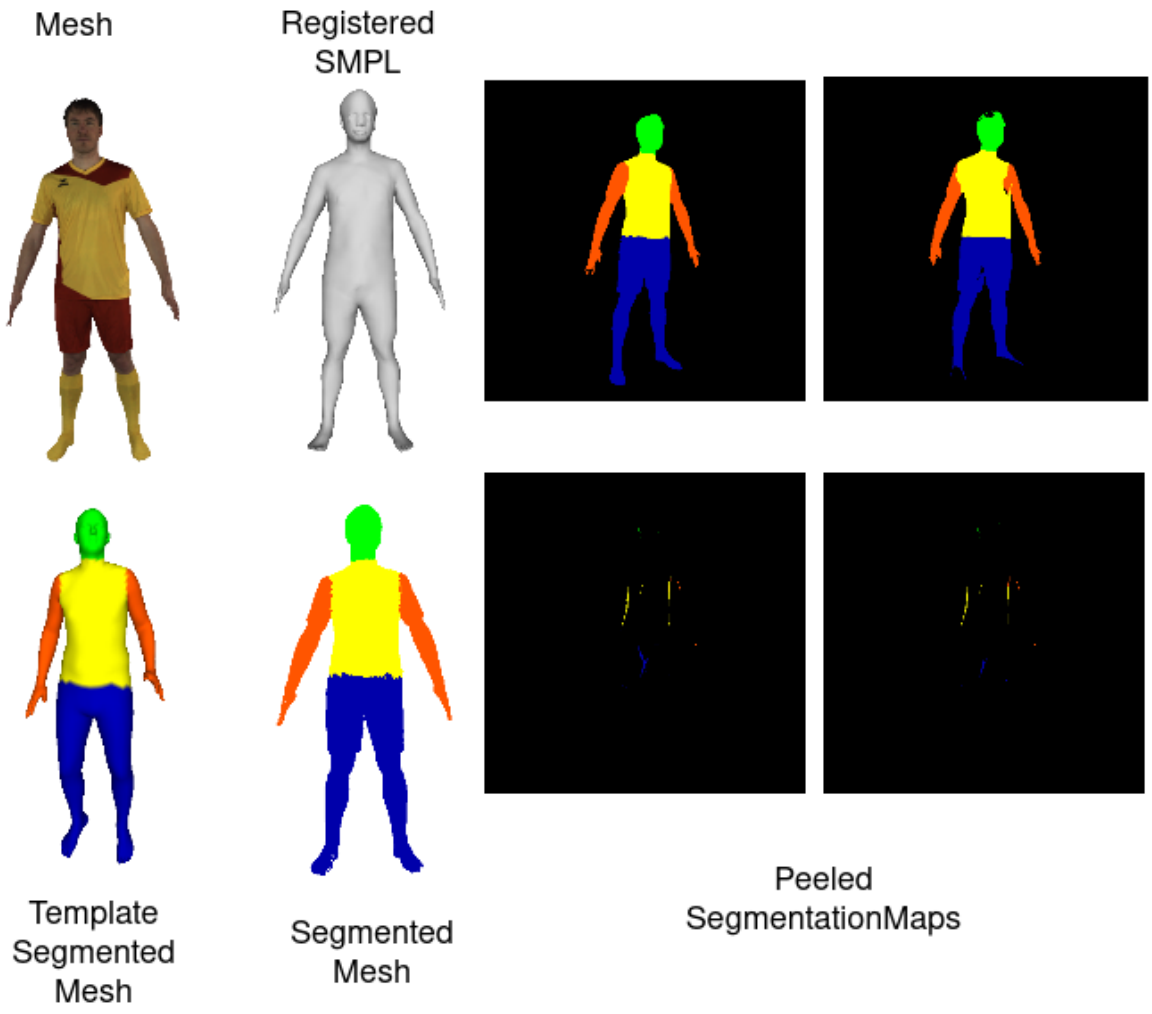

Figure 3.5 Generating semantic segmentation maps.

Buff Dataset

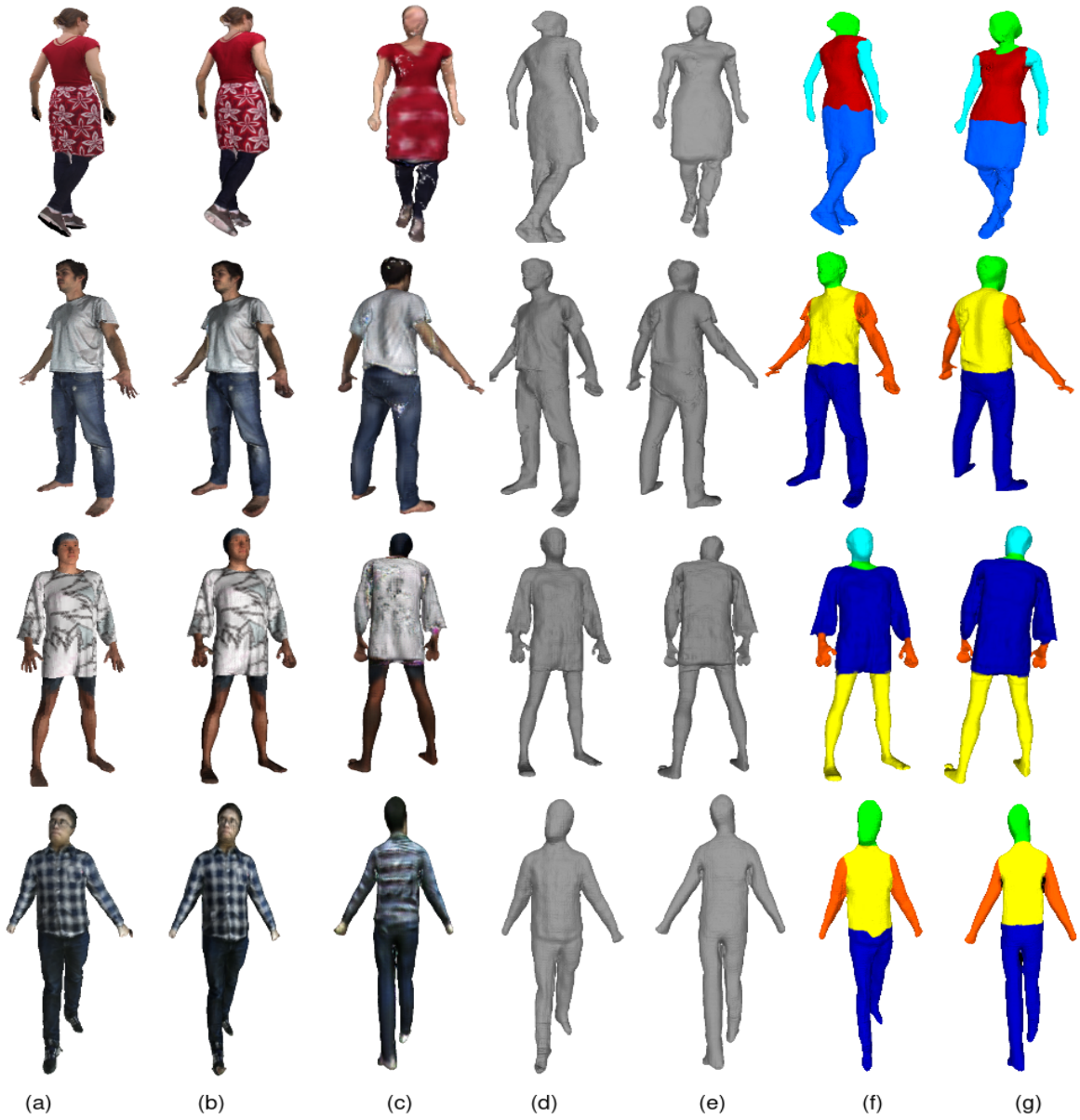

Figure 3.6 Qualitative results (meshes) of our method on MonoPerfCap, BUFF, Cloth3D and Thuman Datasets.(a) input RGB image, (b) & (c) predicted 3D output in different views, (d) & (e) 3D mesh without texture in different views, (f) & (g) our predicted semantic segmentation in different views.

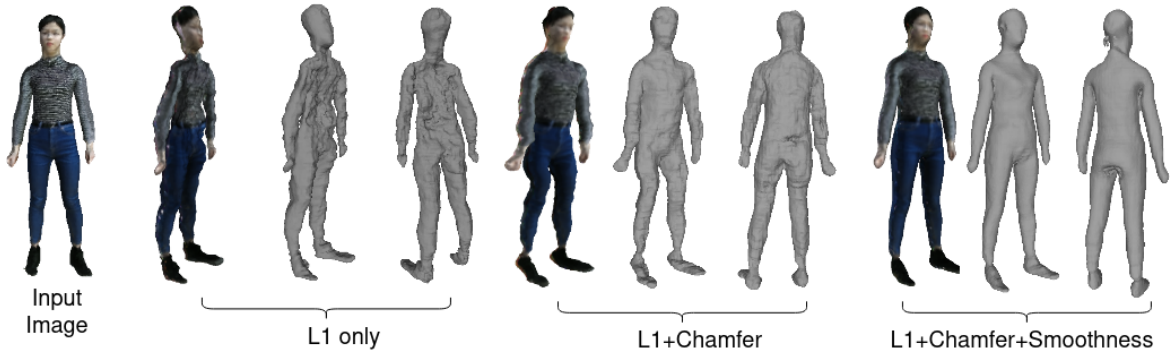

Figure 3.7 Qualitative visualization of ablative study on various loss terms.

Table 3.2 Quantitative results (Chamfer distance) of our method with [88]

Dataset	PeeledHuman	Our Method
Cloth3D	0.00147	0.00135
MonoPerfCap	0.00095	0.00086
Buff	0.00030	0.00025

able to reconstruct plausible body shapes because of the inclusion of semantic context proposed in our method. Please refer to our supplementary material for video rendering of qualitative results.

We performed ablation study on the depth of network for RefGAN by varying the number of ResNet blocks. As reported in Table 3.3, we can conclude that a deeper network with more ResNet blocks helps improve the performance of our model on Cloth3D dataset. Additionally, we also performed ablation study of the effect of various loss terms on the final prediction. The effect of loss terms on the chamfer distance between prediction and ground truth has been reported in Table 3.4 and Figure 3.7. Video results can be found at [1]

3.5 Real World Results

We demonstrate the performance of our model on real world images in the Figure 3.10.

[1] https://iiitaphyd-my.sharepoint.com/:f:/g/personal/snehith_goud_research_iiit_ac_in/ErwMD1BwrzJBjeL0uyIVLgMBicUMkxibyAEcjSl2iHC92w?e=byOihG

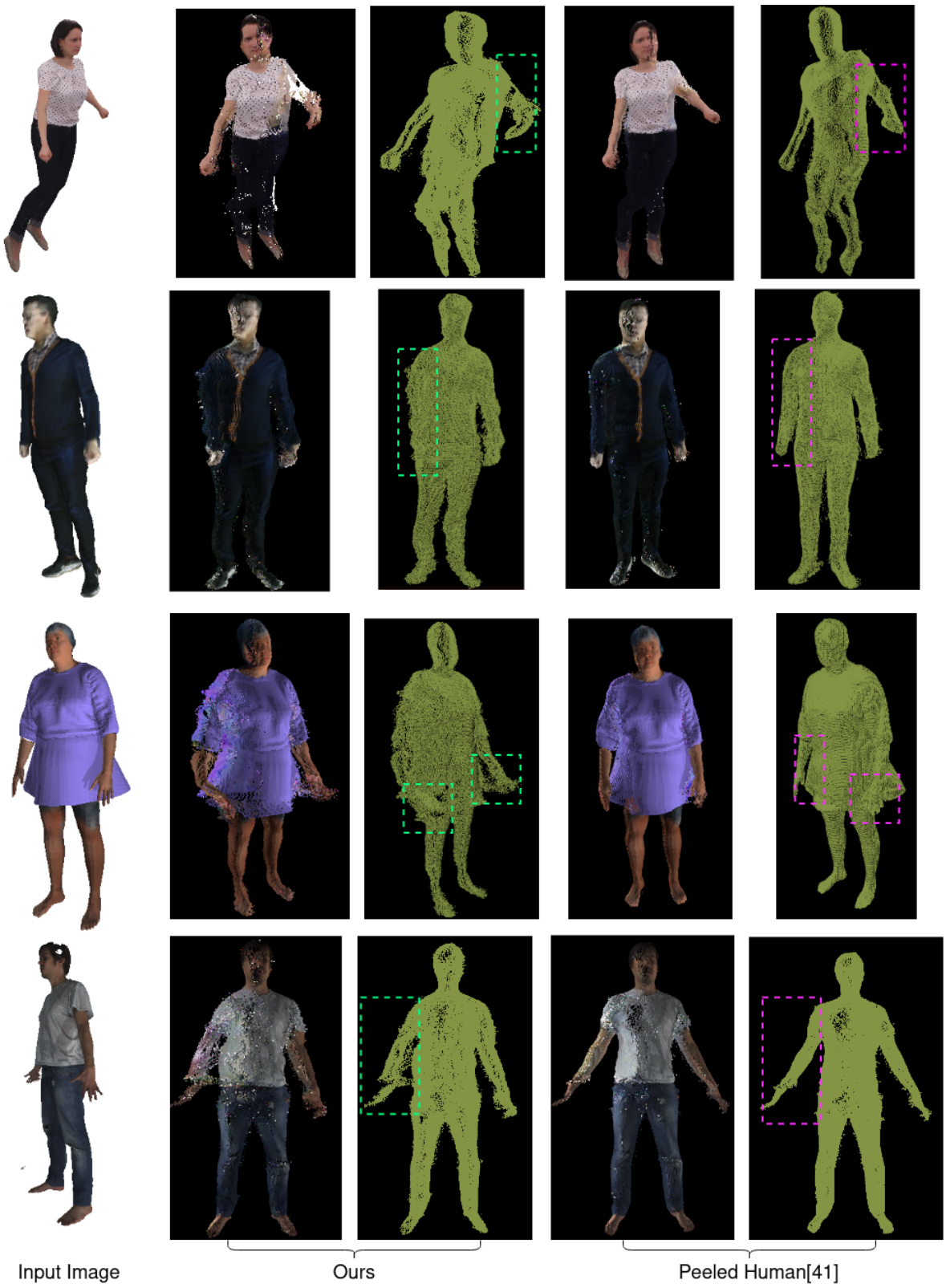

Figure 3.8 Qualitative comparison of generated point cloud (colored and without color) of our method (d,e) and PeeledHuman (b,c) [88].

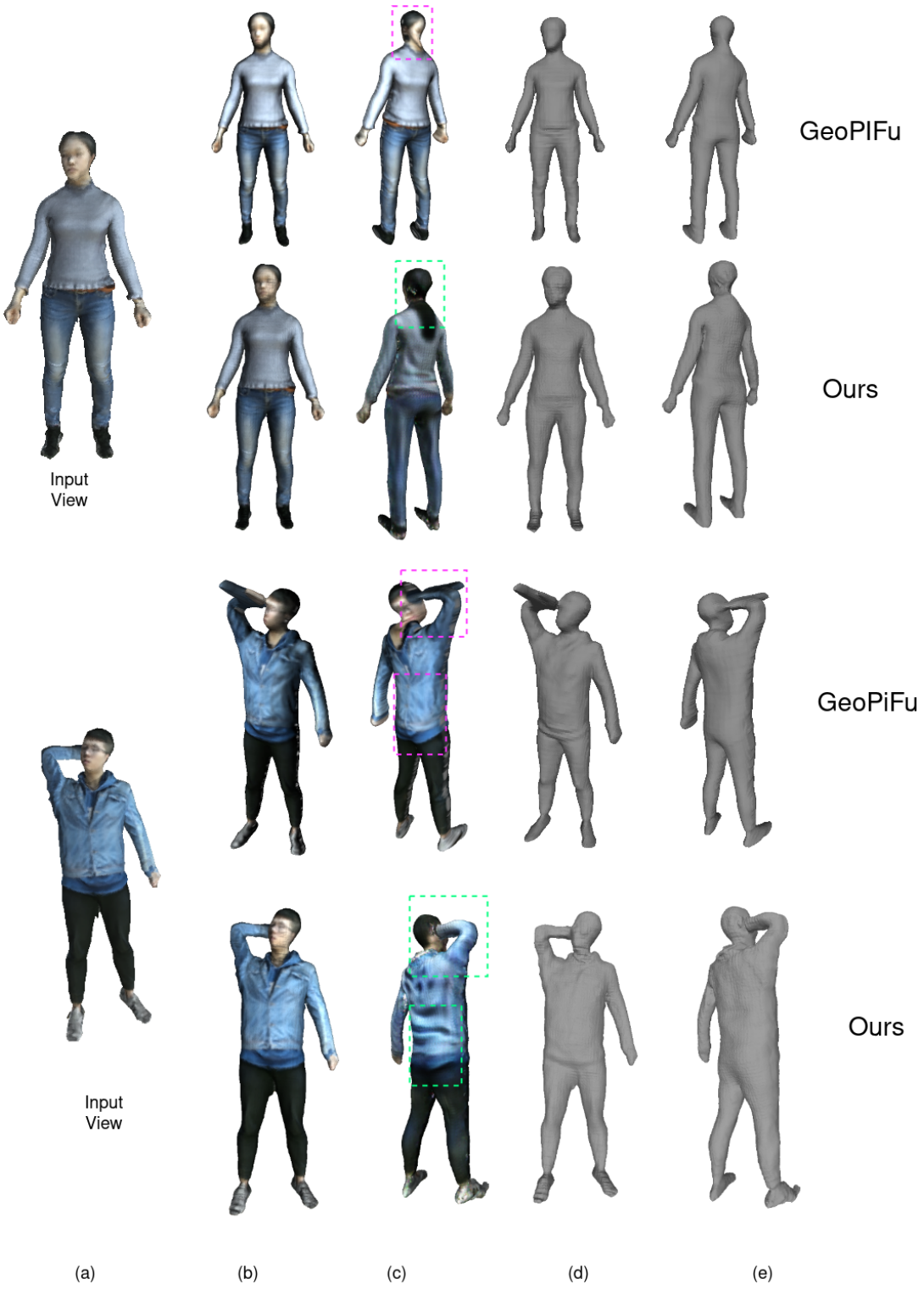

Figure 3.9 Comparison with GeoPIFu on THUman dataset

Table 3.3 Effect of varying number of ResNet blocks for ResGAN.

ResNetBlocks	Chamfer Distance
6	0.00154
9	0.00141
18	0.00135

Table 3.4 Qualitaive results of ablation study performed on loss terms on the THumans Dataset

Loss function	Chamfer Distance
L1	0.000456
L1+Chamfer	0.000411
L1+Chamfer+smoothness	0.000370

3.6 Generalisation

We evaluate generalisation of our model and GeoPIFu both trained on THUman dataset and tested on unseen images. This has been qualitatively shown in Figure 3.11.

3.7 RGB + Depth as input

We trained our model by giving first layer depth map D1 as input along with the monocular rgb image and generated Segmentation. We calculated the chamfer loss for the same on Thuman dataset reported in table 3.5.

We show the qualitative results with depth D^1 as additional input to our network in the Figure 3.12.

Table 3.5 RGB + Segmentation + Depth D^1 as input

Method	Chamfer Loss
Ours(RGB + SegMaps)	0.00037
Ours(RGB + SegMaps + D^1)	0.00010

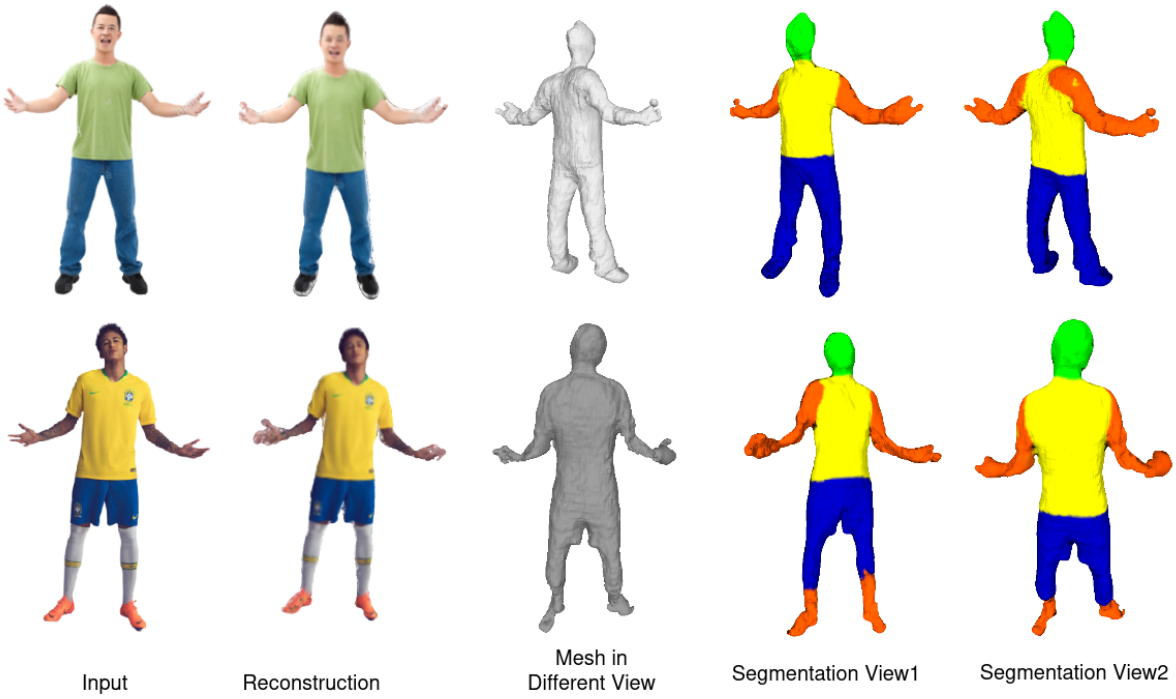

Figure 3.10 Real World Results

3.8 Conclusion

We introduced *Peeled Segmentation map* representation in a coarse-to-fine refinement framework that reconstructs the 3D human body with clothing from a monocular image. Along with the reconstruction, we also provide 3D semantic segmentation of the human from monocular image. We use the semantic segmentation information as a prior to our refinement network which provides a global context of human shape. Our method performs well even in the case of partial occlusions. We evaluate our method on various datasets and report impressive results on par with state-of-the-art methods. As part of future work, we intend to extend the work for the temporal reconstruction of human bodies in action.

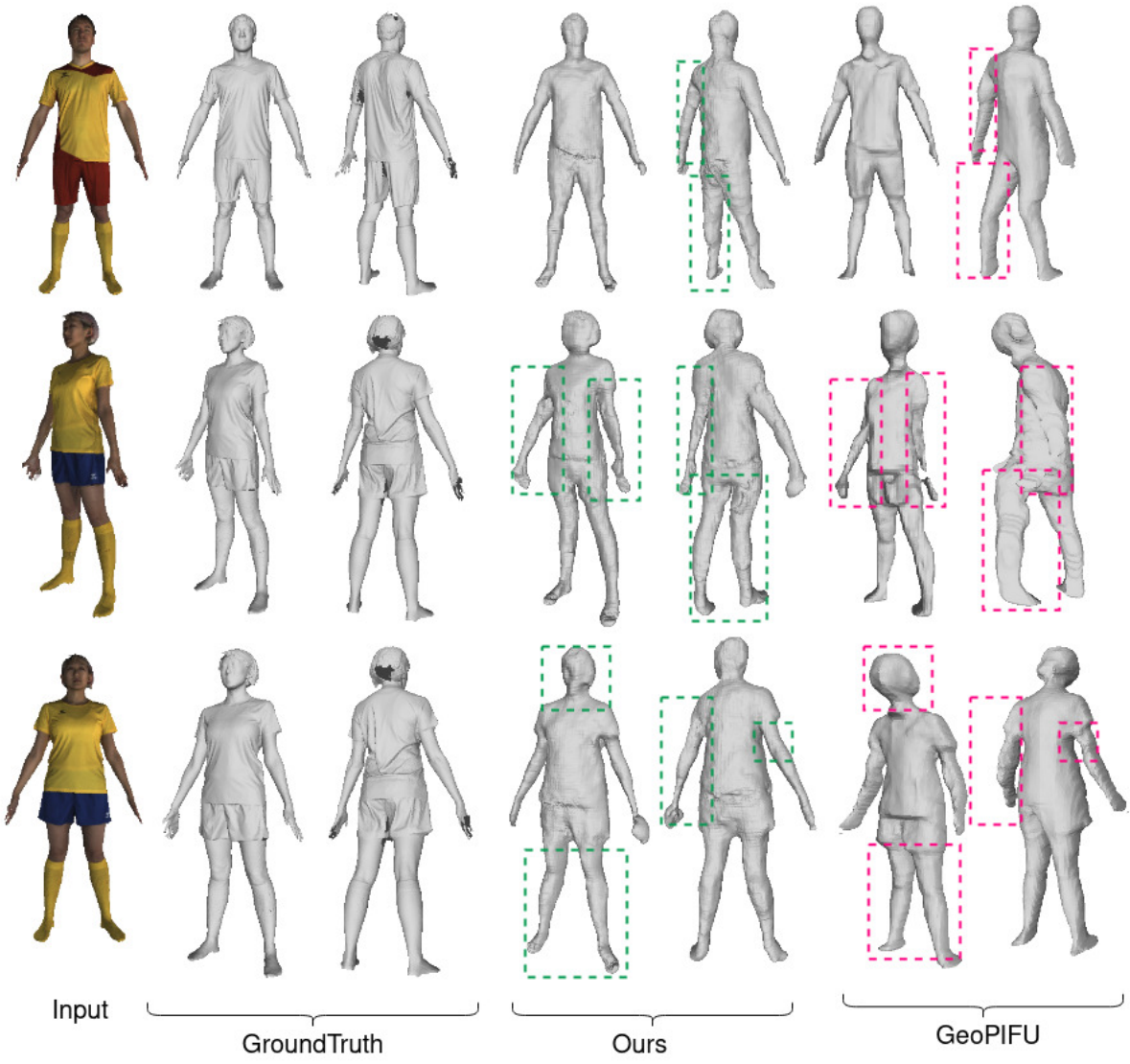

Figure 3.11 Generalisation of our model and GeoPIFu on unseen data

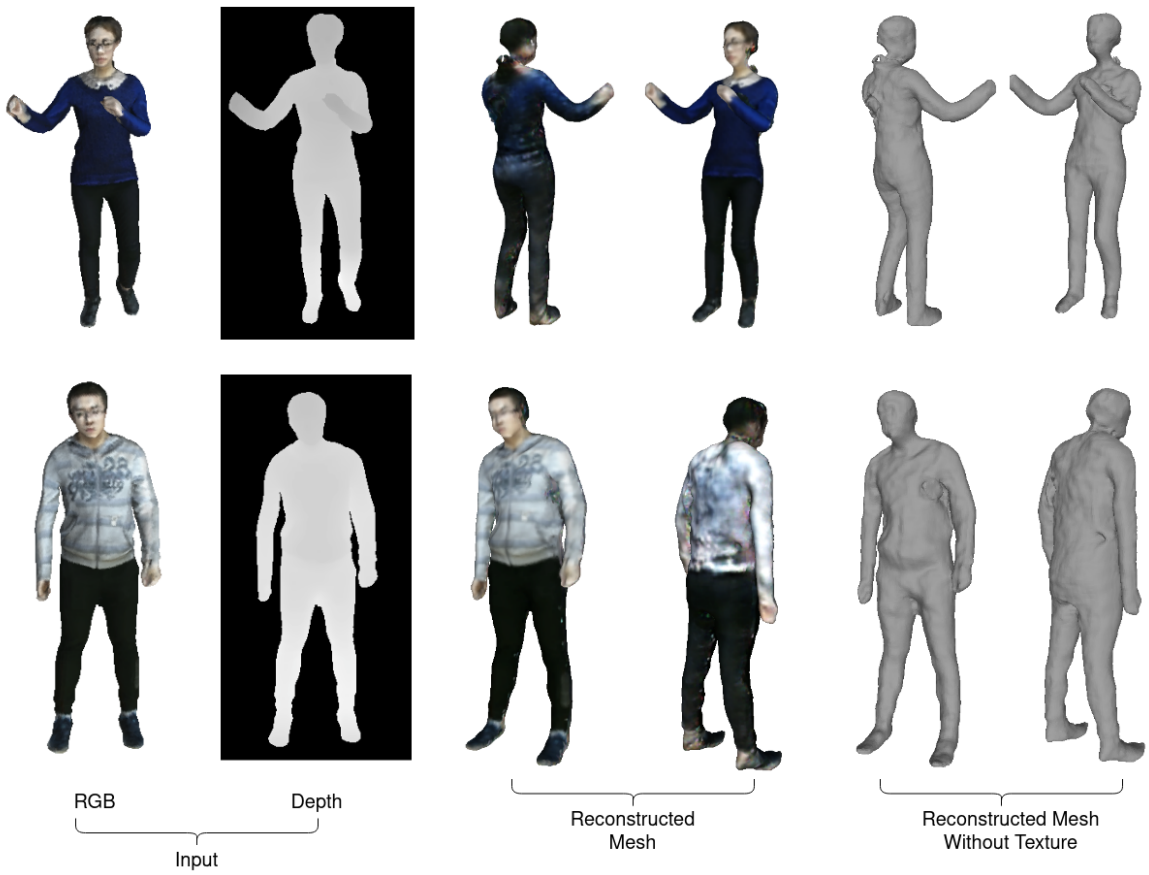

Figure 3.12 Qualitative results with D^1 as additional input

Chapter 4

REF-SHARP: REFined face and geometry reconstruction of people in loose clothing

In the previous chapter, we used peeled semantic segmentation in a coarse-to-fine framework to reconstruct 3D humans. Nevertheless, it suffers from false geometric noise coming from the textural edges and patterns present in the image space. Additionally, the reconstructed face lacks realism and person-specific geometry. To address the aforementioned problems in this chapter we introduce wrinkle map prior, facial prior, and body prior into the PeeledHuman representation. The wrinkle map prior helps in distinguishing the geometrical edges from the textural edges leading to noise-free reconstruction. The facial provides an overall coarse person-specific shape which is smooth but is better compared to SMPL face. We further refine the overall geometry and enhance the high-frequency details of the reconstructed face using our proposed novel framework.

4.1 Introduction

The 3D modeling of humans is interesting and active area of research in computer vision which has tremendous applications in VR/AR, gaming, image, and video editing, tele-presence, virtual try-on, to name a few. With the advent of deep learning, 3D human body reconstruction from monocular RGB images [49, 90, 92, 93, 113] is feasible eliminating the requirement of expensive multi-camera calibrated setups [75, 95]. However, the problem is ill-posed in nature because of challenges which include self-occlusion, loose clothing, skewed viewpoints, pose and shape variation, etc. The two key challenges that are largely remain unaddressed by existing literature are fidel reconstruction of high frequency geometrical detail in the facial region as well as complex loose clothing covering large parts of the body. High fidelity reconstruction of faces largely enhances the realism and identity of the digitized human models. However, it appears in a smaller area of the input image making the reconstruction more difficult for existing methods. On the other hand, although clothing covers significant portion of the human body, it is difficult to model large space of garment designs, specifically in case of loose clothing where high frequency geometrical details (owing to folds and curls) prevails in a highly unstructured

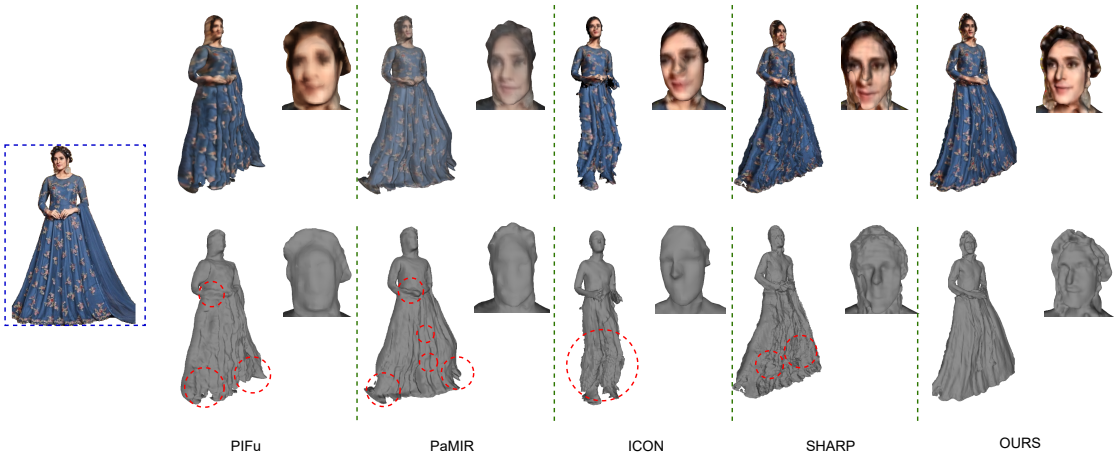

Figure 4.1 Reconstruction from in-the-wild images using PIFu [92], PaMIR [113], ICON [105] and SHARP [90] and ours. Our method predicts high fidelity geometry reconstruction along with consistent appearance in face and loose clothing regions.

manner. More importantly, cloths also consist of high frequency textural details, making it difficult to distinguish textural edges with geometrical edges in a monocular reconstruction setup.

Existing monocular 3D human body reconstruction methods can be broadly classified as parametric and non-parametric methods. The first class of methods [79, 74, 49, 56] regress pose and shape parameters of statistical model (SMPL) [62]. Nevertheless, they fail to capture fine geometrical details as the reconstructed geometry is bound within the parametric model space. Note that most of the parametric templates are modeled as naked human body and hence fail to deal with loose clothing scenario.

The other class of non-parametric methods [70, 92, 101, 13, 104] are not constrained by the body prior and use non-parametric representations to infer detailed geometry beyond basic body shape and pose. Recently, deep implicit function learning [92, 93, 38] techniques witnessed increased attention. These methods train multi-layer perceptrons to estimate dense, continuous signed distance fields. from which 3D mesh is reconstructed via Marching Cubes [64]. Another work, [44] proposed to represent the human body by predicting multiple peeled depth maps. These methods fail to ensure the prediction of physically plausible body shapes/poses (like consistent geometry of arms) in the reconstructed mesh as there are no global body shape/pose constraints. Some of the very recent works [113, 41, 39, 90, 116] have addressed this problem by incorporating a body prior i.e. SMPL into the reconstruction framework. Nevertheless, these methods typically yield reconstruction with facial geometrical details incorporated from the input SMPL prior while the person-specific facial details are not retained. This leads to loss of identity as well as misalignment in geometry and appearance of facial features. There are few methods that specifically focus on accurate 3D face reconstruction [28, 112, 24, 34, 61] but does not model the complete human body. Very recently, [20] proposed to integrate facial features from 3DMM model [28] into implicit function learning approaches. However, the method is not able to

perform better on the remaining parts of body as there is no explicit body prior. Additionally, all these methods are prone to interpret high frequency textural details present in the appearance space of clothes as false geometrical edges/details thereby yielding inconsistent and noisy geometrical reconstruction specifically in the loose clothing. Although, disambiguate textural edges with geometrical edges is very hard problem but these methods entirely neglect this challenge during reconstruction. Hence, there is an acute need of framework which recover accurate and consistent geometrical/appearance reconstruction of both face and complex in-the-wild loose clothing while modelling 3D human body.

In this paper, we propose REF-SHARP, a novel 3D reconstruction framework for recovering people in loose clothing from a monocular image while recovering fine 3D geometrical details of the face. Additionally, we also attempt to alleviate false geometrical edges caused by textural details in clothes. We improve upon SHARP [90], a SOTA method which uses PeeledHuman representation [44] along with SMPL prior for the prediction of human in loose clothing. However, similar to other methods Figure 4.1, SHARP also yields low-quality facial geometry as well as false geometrical details misled by textural edges. Thus, we propose to provide high quality pixel-aligned depth prior for face region along with prediction of wrinkle maps capturing geometrical edges to improve upon the aforementioned limitations of SHARP. More specifically, we attempt to recover finer facial details by predicting high resolution pixel-aligned depth in the face region and integrate this with SMPL peeled depth prior. This ensures that person-specific facial geometrical details will be preserved in the final reconstruction after learning-based fusion. Additionally, we propose to learn and predict wrinkle map prior by regressing over geometrical edge maps and concatenate this with face modified SMPL prior to and use suppression and normal loss to ensure suppression of false geometrical edges.

We evaluate our method on publicly available THUman2.0 [110] and 3DHumans [90] datasets and report superior performance over the SOTA methods. To summarize our contributions are: we propose REF-SHARP where we reconstruct high-fidelity geometrical details in the face while recovering full body reconstruction under loose clothing scenarios.

4.2 Related Work

4.2.1 Parametric Reconstruction Methods

With the emergence of statistical human models such as SMPL[62] and SCAPE [11] interest has shifted to estimating these models pose and shape from a single image using deep learning methods [49, 25, 46, 18, 58, 74, 80]. [79, 49] tries to optimize the pose and shape parameters of statistical human models for *e.g.*(SMPL) by matching with image features obtained from CNN. The commonly used features are 2D joints [18, 49, 74], 2D joints with silhouettes [58, 25].

Methods proposed in [7, 15] deform the statistical models by adding displacements over the surface to obtain geometric details and clothing to a certain extent. It is to be noted that under this SMPL[62] plus displacement setup only tight clothing can be modeled as the loose garments such as skirts and

robes have a surface topology different from the body and are beyond the representation range. [118] models the fine geometric details as free-form 3D deformations applied on the parametric body model. [55] improves the estimation of the prediction by introducing model optimization in the training loop, [68] combines local and global features to estimate fine body poses. Nevertheless parametric models can only capture minimalistic clothing where the clothing is tight fitted with the body and fails to represent humans with loose clothing.

[82, 16] use separate templates for body and garments and bind garment vertices to the parametric body model which becomes difficult to represent very loose clothing as sarees, robes, and skirts as the topology of the garment is constrained by the binding with the body model.

4.2.2 Non-Parametric Reconstruction Methods

Volumetric regression based methods [101, 104] estimate the occupancy of the voxels in the volumetric space using deep neural networks. These methods are computationally costly for higher resolution as voxel representation is memory intensive, high frequency details are not captured due to lower resolution (typically 128). [92, 106] combines the pixel aligned 2D local features extracted using deep convolutional networks with the implicit representation. Nevertheless PIFu [92] suffers from feature ambiguity problem due to multiple query points mapping to the same 2D image features upon weak perspective projection and lacks global shape robustness. PIFuHD [93] is another variation of PIFu which generates human meshes from high resolution images. GeoPiFu [38] attempted to resolve the feature ambiguity of 3d points projecting to the same image feature by combining U-Net based volumetric features with the pixel aligned features. However, this method is computationally intensive during training and inference. An alternative set of non-parametric approaches attempt to model 3D objects/scenes as sparse layered representation. PeeledHuman [44] proposes a sparse2D representation by posing the problem as an extension to ray tracing, they model the 3D surface by performing ray intersection with the surface and storing them as the peeled depth and rgb maps.

4.2.3 Prior based non-parametric Methods

DeepHuman [116] uses the SMPL as a prior to reconstruct the clothed body volume and tries to further refine the surface details using image features. However, their method fails to recover high-quality geometric surface details owing to the resolution limitation of the regular occupancy volume. ARCH [41] proposed a Semantic Deformation Fields (SemDF) based approach where the query points are sampled around the body in a canonical space (A-pose), an implicit surface is learned in the canonical space, and deformed using SemDF to match the pose in the input. However, it fails to generate accurate results, especially in the scenarios of loose clothing. PaMIR [113] proposes to condition the implicit field on the SMPL prior, they do this by combining the 2D image features with SMPL volume features while querying. SHARP [90] proposes to model the reconstruction as two tasks (1) deform the SMPL

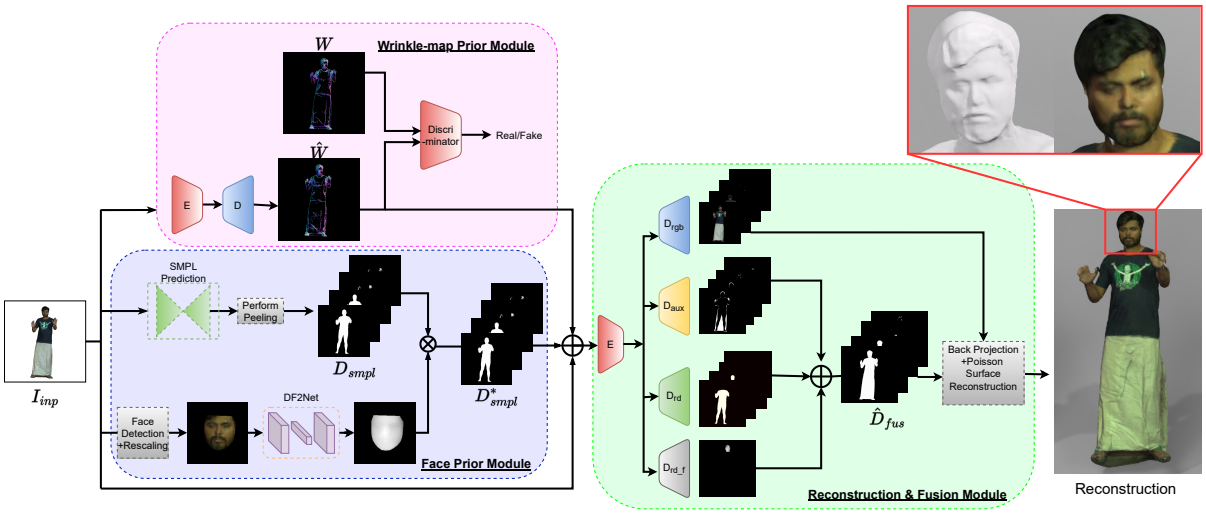

Figure 4.2 Architecture of the proposed framework.

body prior peeled maps in order to obtain fine-grained geometric surface details (2) directly regress the loose clothing as auxiliary peeled depth maps and combine both of them to obtain 3D reconstruction.

4.2.4 Face reconstruction methods

Similar to SMPL ([62]) for full body, 3D Morphable Models (3DMM) [28] is proposed to model face in parametric representation. Methods proposed in [61, 94, 34, 96] estimate parameters of 3DMM. Implicit functions are combined with 3D morphable models in [109, 85]. Full head models including hair are recovered in [60, 29, 23].

4.3 BACKGROUND

PEELED-HUMAN representation This is a sparse 2D representation of 3D objects modeled as Peeled Depth and RGB maps. The 3D Human mesh is placed in a virtual world and rays are passed from the camera to intersect with the mesh. The primary set of rays intersecting with the surface is captured as depth map d_1 and RGB map r_1 depicting the visible surface details nearest to the camera. The rays are then further extended beyond the first intersection to hit the next intersecting surface. The corresponding depth and RGB values of the i_{th} layer are represented by d_i and r_i, refer to [44] for further understanding. [44] demonstrate that 4 layers are sufficient to handle self-occlusions and common body poses though the method can be extended to multiple layers.

4.4 Method

Our proposed REF-SHARP is divided into three key modules as shown in Figure 4.2. The input monocular RGB image is fed to the *"Face Prior Module"* where we predict pixel-aligned high resolution depth map of the cropped face region and overlay it over SMPL peeled prior. In parallel, we predict wrinkle map that models prominent geometrical edges over the surface as outlined in *"Wrinkle-map Prior Module"*. The generated face+SMPL prior and wrinkle maps along with input RGB image are then fed to the *"Reconstruction & Fusion Module"* which reconstructs full body. This module predicts auxiliary and residual depth peel maps, RGB peel maps for the full body, and single layer face residual peel map. These maps are subsequently *fused* to get final peeled maps that are jointly back-projected to get high resolution vertex colored point cloud representing full body. This dense surface point cloud is further converted to mesh representation using Poisson surface reconstruction[53].

4.4.1 Face Prior Module

Similar to [90], we initially predict SMPL peeled prior $D^i_{smpl} \in D_{smpl}$. We modify the SMPL peeled prior to obtain D^*_{smpl} by replacing the face region of SMPL (which has inconsistency in the face) in the first layer (D^1_{smpl}) with face predicted from [112].

Since SMPL face is not exactly aligned to actual face, we use off-the-shelf face estimation [112] to achieve pixel-to-pixel consistency with the input image. We perform face detection using an off-the-shelf method [1] to detect face in the image and crop it using the detected bounding box M_b which contains the complete head region (hair to chin). We rescale (bicubic interpolation) the cropped region to 512x512 and mask the cropped image using 68 landmarks detected using the face detection library [1] and feed it to [112] for pixel-wise depth prediction. We rescale the depth prediction back to the original resolution of the face in the image. It is important to note that although we re-scale the face region, we are not using any information from any external source. Finally, we replace the SMPL face in first layer of SMPL prior (D_{smpl}) with the rescaled person-specific face prediction to obtain D^*_{smpl}.

However, [112] predicts depth in an orthogonal fashion, in order to adopt this to the perspective setup and compensate for the details lost during the up-scaling we predict per pixel offsets for the bounding box region(M_b) of D^{*1}_{smpl}. In the next module, we use a separate decoder branch for predicting these residual deformations to focus on the region containing the face.

4.4.2 Wrinkle-map Prior Module

High frequency detail in the input RGB image can be from both geometrical or textural variations. Our wrinkle map representation aims to capture only the high frequency (edges) details that are caused by variation in surface geometry, thereby decoupling these details from textural high frequency details present in the appearance space. However, such prediction can be done reliably only for the first (and visible) peel layer as learning for other layers is hard since the respective body surface is not observed

in the input image. We train an image-to-image translation GAN [42] to generate wrinkle map from the input image using ground truth wrinkle map as supervision. We use L1 loss over the predicted wrinkle map w and ground-truth wrinkle map $\hat{w}$ along with the adversarial GAN loss. The final loss function is:

$$loss = w_g * l_{GAN} + w_{L1} * l_{L1} \tag{4.1}$$

where w_g and w_{L1} are the weighting factors for l_{GAN} and l_{L1}

$$l_{GAN} = E_{I_{inp},\hat{w}}[logD(I_{inp}, \hat{w})] + E_{I_{inp},w}[log(1 - D(I_{inp}, w))] \tag{4.2}$$

$$l_{L1} = |\hat{w} - w|\,. \tag{4.3}$$

For generating ground truth wrinkle map $\hat{w}$ we smooth the normal map using a bilateral filter and then calculate the change in normal using Laplacian filter and threshold it to obtain geometric regions which are not smooth.

4.4.3 Reconstruction & Fusion Module

The generated wrinkle maps ($\hat{w}$) along with D^*_{smpl} are fed as prior to the encoder which predicts auxiliary peel maps $\hat{D}_{aux}$, RGB peel maps $\hat{D}_{rgb}$, residual peel maps $\hat{D}_{rd}$ similar to [90]. We additionally predict residual deformation for face $\hat{D}_{rd\_f}$. The predicted residual deformation $\hat{D}_{rd\_f}$ captures the face region and the hair region is obtained from auxiliary peel maps $\hat{D}_{aux}$. Given the face bounding box obtained from the face detector M_b (obtained in the previous module), we estimate the complete face region peel map $\hat{D}_f$ as a fusion of $\hat{D}_{rd\_f}$, and $\hat{D}^1_{aux}$ as follows:

$$\hat{D}_f = M_b * \hat{D}^1_{aux} + M_b * (\hat{D}_{rd\_f} + D^{*1}_{smpl}). \tag{4.4}$$

The residual deformation maps are added to D^*_{smpl} to get deformation maps $\hat{D}_{def}$. We finally fuse all the peel maps to obtain the final fused peel maps ($\hat{D}_{fus}$). The final depth peel maps are obtained by fusion of auxiliary ($\hat{D}_{aux}$), face peel map ($\hat{D}_f$), and body peel map ($\hat{D}_{rd} + D^*_{smpl}$). The first layer fused peel map ($\hat{D}^1_{fus}$) (which consists of face region) and remaining layer fuse peel maps ($\hat{D}_{fus}$) can be expressed as:

$$\begin{aligned} \hat{D}^1_{fus} &= M^1_s(1 - M_b) * \hat{D}^1_{def} + (1 - M^1_s) * \hat{D}^1_{aux} + \hat{D}_f \\ \hat{D}^i_{fus} &= M^i_s * \hat{D}^i_{def} + (1 - M^i_s) * \hat{D}^i_{aux} \forall i \in 2, 3, 4 \end{aligned} \tag{4.5}$$

,where M_s is the mask for $SMPL^*$ peeled prior defined as

$$M^i_s = \begin{cases} 1, & \text{if } D^{*i}_{smpl} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.6}$$

4.4.4 Loss Functions

We propose to use the following loss functions: face loss (l_{face}), residual deformation loss (l_{rd}), RGB loss (l_{rgb}) and fusion loss (l_{fus}). Additionally, we use smoothness loss (l_{sm}), normal loss (l_{nl}) for regularization. For normal loss, we back-project the predicted and ground truth depth maps. The overall loss function used is :

$$L = w_{fus} * l_{fus} + w_{rd} * l_{rd} + w_{sm} * l_{sm} + w_{rgb} * l_{rgb} + w_n * l_{nl} + w_{sup} * l_{sup} + l_{face} \quad (4.7)$$

, where w_{fus}, w_{rd}, w_{sm}, w_{rgb}, w_n and w_{sup} are the respective weights for l_{fus}, l_{rd},l_{sm}, l_{rgb}, l_{nl} and l_{sup}.

The fusion (l_{fus}), RGB (l_{rgb}), residual deformation (l_{rd}) and smoothness (l_{sm}) losses are similar to [90], i.e. L_1 loss between respective ground truth and predicted peel maps. The remaining loss terms are defined as follows:

$$\begin{aligned} l_{face} = w_f * \left| M_b * D^1_{fus} - \hat{D}_f \right| + \\ w_{f\_rd} * \left| \hat{D}_{rd\_f} - D_{rd\_f} \right| + \\ w_{f\_n} * \left| \hat{N}_f - N_f \right|. \end{aligned} \quad (4.8)$$

We employ L_1 loss over the $\hat{D}_f$ and the ground truth depth in the face region. We calculate L_1 loss between ground truth residual depth D_{rd_f} and predicted $\hat{D}_{rd\_f}$. We use L_1 loss over ground truth and predicted normals N_f and $\hat{N}_f$ respectively.

In order to compensate for the over smoothing we apply the L1 loss between normal maps of predicted and ground truth depth of the first fused layer. Let N be the normals obtained from D^1_{fus} and $\hat{N}$ be the normals obtained from $\hat{D}^1_{fus}$ then the loss is defined as below.

$$l_{nl} = \left| \hat{N} - N \right|. \quad (4.9)$$

Based on the wrinkle map we penalise the change in gradients of the depth of non-wrinkle regions as regularisation term to ensure that empty regions in the wrinkle map are locally smooth.

$$M_w = \begin{cases} 1, & \text{if } \hat{w} > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (4.10)$$

$$l_{sup} = \left| (1 - M_w) \nabla \hat{D}^1_{fus} \right|. \quad (4.11)$$

We back-project the $\hat{\mathcal{D}}_{fus}$ and $\hat{R}$ to camera co-ordinate frame assuming the projection is weak perspective in order to obtain the 3D point cloud of the reconstruction. The point-cloud is then post-processed, and further meshified using Poisson Surface Reconstruction (PSR)[53] to generate the final 3D body mesh.

	3DHumans [90]		THUman2.0 [110]	
Method	Chamfer($*10^{-5}$)	P2S	Chamfer($*10^{-5}$)	P2S
PIFu	13.9	0.0071	18.1	0.0074
PaMIR	5.9	0.0068	3.6	0.0033
ICON	4.1	0.0045	3.1	0.0039
SHARP	3.1	0.0033	2.7	0.0031
OURS	**2.5**	**0.0028**	**2.4**	**0.0027**

Table 4.1 Quantitative analysis of our method in the head-only region.

	3DHumans [90]	THUman2.0 [110]
Method	P2S	P2S
PIFu	0.00826	0.0091
ICON	0.00822	0.0064
PaMIR	0.00714	0.0049
SHARP	0.00514	0.0055
OURS	**0.00508**	**0.00546**

Table 4.2 Quantitative analysis of our method in the full body region.

4.5 Experiments and Results

4.5.1 Datasets

3DHumans [90]: The dataset is a collection of 200 subjects with diverse body shapes and various clothing styles. This dataset consists of relatively loose clothing (South Asian styles), and also tight clothing such as shirts and pants. The dataset consists of around 150 male and 50 unique female subjects with a database of 200 scans.

THuman2.0 [110]: The dataset is a collection of 500 subjects with high quality 3D scans captured with a DSLR rig. Each subject has around 3-4 poses hence providing us with various poses in the dataset. However, the dataset lacks very loose clothing like long skirts etc.

4.5.2 Implementation Details

For wrinkle map generation we employ a GAN [42] with our generator being a ResNet generator with a set of down convolution blocks followed by 18 residual blocks then a set of up convolution

Figure 4.3 Qualitative results of our method on 3DHumans (columns 1 and 2) and THuman2 (columns 3 and 4) datasets. Top row: input image, 2nd and 4th rows: full-body and head region reconstruction of our method, 3rd and 5th rows: ground truth full body and head only region scans.

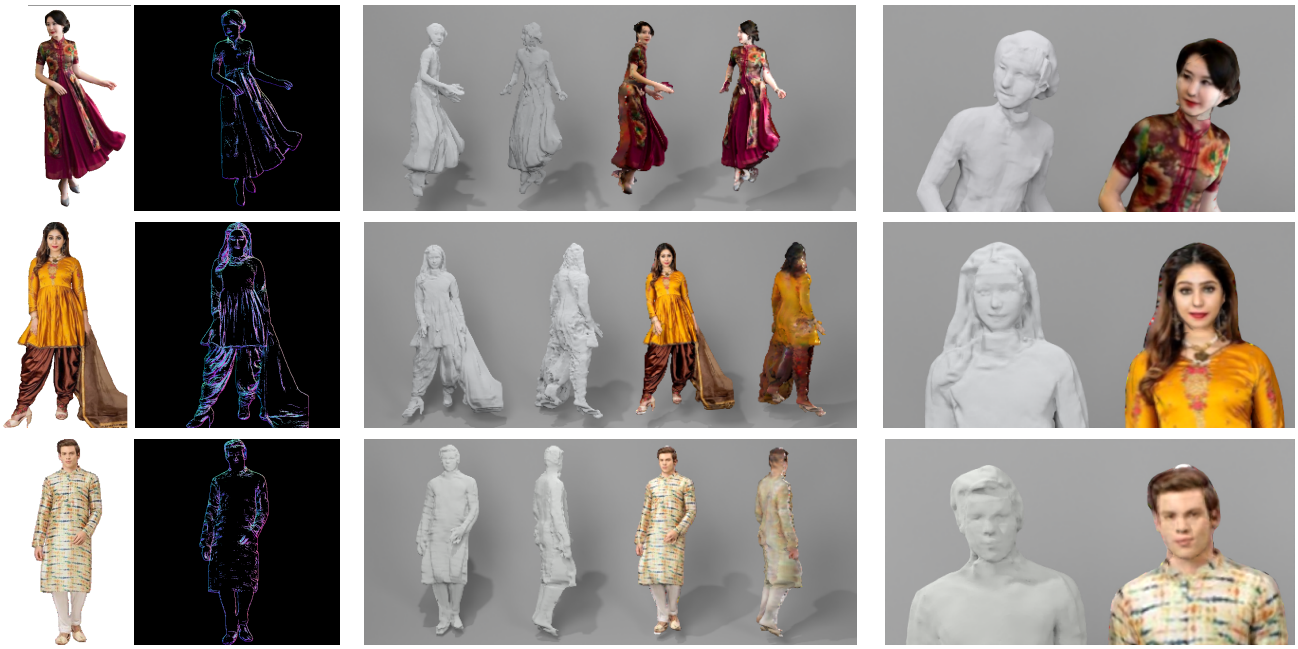

Figure 4.4 Qualitative results of our method on in-the-wild internet images.

blocks and our discriminator is a patch-based discriminator. The peeled depth estimation network is an encoder-decoder network. The input to the network is a concatenation of RGB image, SMPL peeled prior and generated wrinkle map. The shared encoder consists of an initial convolution layer of 64 filters of size 7×7 followed by a couple of down sampling layers of filter size 3×3 with stride 2 and respective filters of 128 and 256, in each layer. The down sampled output of $256 \times 128 \times 128$ is then passed through a series of (18) residual blocks. The encoded shared features are then passed to the decoders which predict different outcomes based on the task. The decoders D_{aux}, D_{rd} ,D_{rgb} and D_{rd_f}, consists of 2 upsampling layers of filter sizes 3×3 and channels 128 and 64, respectively. This is followed by a convolutional layer of filter size 7×7. Sigmoid activation is used in D_{aux} ,D_{rd} andD_{rd_f} decoder branches, whereas a Tanh activation is used for the D_{rgb} decoder branch. The D_{rd} output values are scaled to a range of [- 1, 0.5] and D_{rd_f} to [-0.025,0.025] which can be found empirically.

We use Adam optimiser with an initial learning rate of 0.0005. Our network takes around 30 hrs to train for 30 epochs on 4 Nvidia GTX 1080Ti GPUs with a batch size of 4 and w_{fus}, w_{rd}, w_{sm}, w_{rgb}, w_n and w_{sup} are set to $1, 1, 0.1, 0.001, 0.2 and 0.001$ respectively.

4.5.3 Qualitative & Quantitative Evaluation

We evaluate our model by comparing it with following SOTA methods: PIFu[92], PaMIR[113], ICON[105] and SHARP[90].

Figure 4.5 Qualitative comparison on 3DHumans (top row), THuman 2.0 (middle row) and in-the-wild (bottom row) images.

Qualitative Results: We present detailed qualitative results generated by our method and comparisons.First, Figure 4.3 we refer to the results generated by our method on the subjects from 3DHumans and THUman2.0 in Figure 4.3. Here, the first two columns are on 3DHumans dataset and last two columns are on THUman2.0 dataset. It can be observed that our model can deal with various styles of clothing while predicting high geometric details in the face. Next, we also test the generalization of our model on unseen internet images and report them in Figure 4.4. It can be observed that our wrinkle map enhances the results qualitatively. It is also observed that our method produces highly detailed faces which can be observed in the third column of Figure 4.4.

Finally, we qualitatively compare the aforementioned SOTA methods in Figure 4.10, Figure 4.1. For each input image, we show full body reconstruction along with face geometry and texture. Accurate face prediction enhances the quality of reconstruction after back projecting the texture. It is observed that in all the SOTA methods, when the texture is back-projected, misalignment is clearly visible. Our method reconstructs face to actual geometry thereby preserving back-projected texture and enhancing the realism.

Quantitative Results: We also evaluate our model quantitatively by comparing it with the aforementioned SOTA methods. In Table 4.1, we show quantitative results on head region where we trained the models on both 3DHUmans and THUman2 .0 datasets using the same train and test split. We use Chamfer distance (lower the better) and point to surface distance (P2S)(lower the better) as evaluation metrics. We can infer that our method consistently outperform all the SOTA methods. Our framework which integrates face prior reconstructs accurate facial geometry close to ground truth face while achieving consistent fusion with face and head region. We also perform quantitative evaluation of full body region in Table 4.2. In full body reconstruction we are close to SHARP. Nevertheless, we can observe in Figure 4.1, Figure 4.4, and Figure 4.10 (bottom row) our method yield qualitatively far superior results, while generalizing to in-the-wild internet images. The inference time of the network is 0.1 seconds.

loss terms	P2S
$w/o\ l_{sup}$ & $w/o\ l_{nl}$	0.0055
$w\ l_{sup}$ & $w/o\ l_{nl}$	0.0054
$w/o\ l_{sup}$ & $w\ l_{nl}$	00518
$w\ l_{sup}$ & $w\ l_{nl}$	**0.00508**

Table 4.3 Ablation study of l_{nl} and l_{sup}.

4.5.4 Ablation

Our model predicts accurate geometry compared to directly overlaying the face prior with the body, as shown in Figure 4.6. It is to be noted that directly overlaying face prior produces artifacts in the head

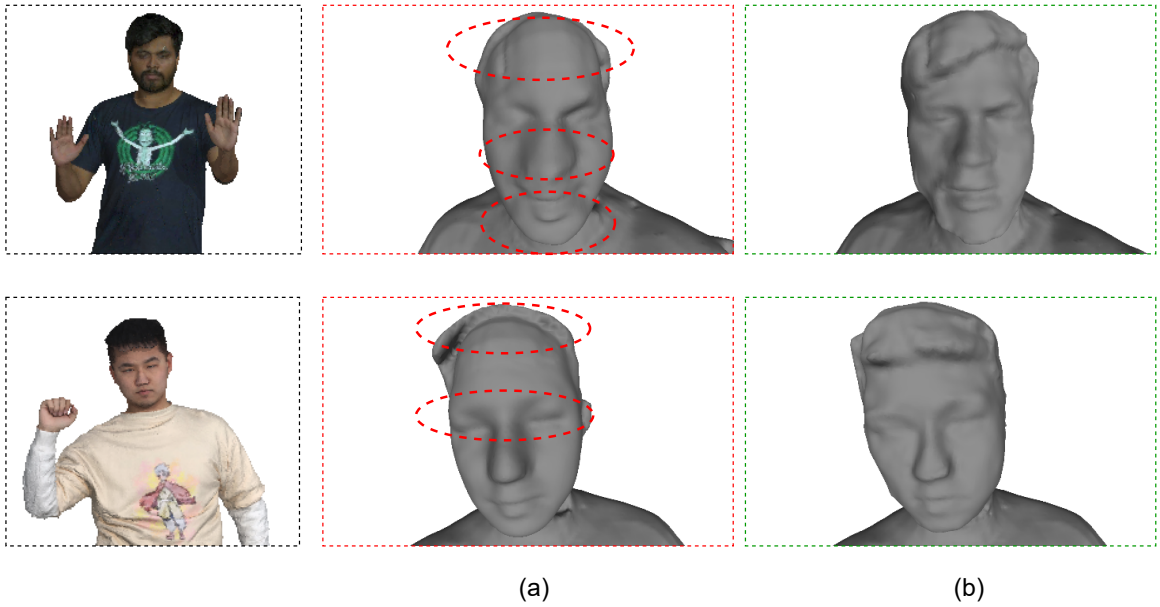

Figure 4.6 Effect of our networks refinement over face. (a) Directly overlaying fae prior; (b) Our fused reconstruction.

region. Our proposed residual deformation on the face (D_{rd_f}) seamlessly fuse the face prior with the head region (e.g., near front hairline), as illustrated. Video results can be found at [1] In Table 4.3, we provide a study on the impact of suppression loss (l_{sup}) and normal loss (l_{nl}) on the model's performance. We show P2S estimation on full body (which includes face) on 3DHumans [90] dataset. We observe when we train our model without l_{sup} and l_{nl} losses, we obtain inferior P2S values. Subsequently, we add the suppression loss (l_{sup}) and get slightly improved performance over the previous setup. Further, only adding normal loss (l_{nl}) also significantly boosts the performance where we achieve P2S value of 0.00518. Finally, when trained the model by using both normal and suppression loss to achieve further improvement in P2S, i.e, 0.00508. Hence, the proposed losses contribute to improved performance of our framework. Additionally, we quantitatively evaluate the performance of our method by removing the wrinkle map prior in Table 4.4. As we can observe, the performance deteriorates in absence of this prior. Finally, we also perform ablative study on the face prior in Table 4.5 where we can infer that both the P2S error as well as Chamfer distance around head region increases in absence of face prior. We also show qualitatively that there is significant misalignment between the predicted geometry and texture of the face while using SMPL face as prior in our framework whereas our proposed method with pixel-aligned depth face prior achieve much superior texture to geometry alignment as shown in Figure 4.7.

[1] https://iiitaphyd-my.sharepoint.com/:f:/g/personal/snehith_goud_research_iiit_ac_in/EiGJ-0j08oRBqIzAkDkX0fQBAjHknRgoPrnLfgYlaWvRng?e=DfnLBX

	P2S
Without wrinkle map prior	0.00517
With wrinkle map prior	**0.00508**

Table 4.4 Effect of wrinkle map prior.

Prior	Chamfer($*10^{-5}$)	P2S
w/o face prior	2.8	0.0031
w face prior	**2.5**	**0.0028**

Table 4.5 Ablation study of facial prior on head region.

4.5.5 Limitations

Majority of existing pixel-aligned face depth prediction methods predicts only the frontal faces reliably. Thus, our method can't recover good geometrical reconstruction where the face has skewed pose with large portion of face self-occluded. In this scenario, our proposed model predicts a smooth face in the occluded region as illustrated in the Figure 4.8. This can be resolved partially if we use a full head parametric model which provides prior for the occluded regions of the face. Nevertheless, full head models also suffers from recovering accurate person-specific face in extreme self-occlusion cases. Alternatively, we can also model the occluded regions in a generative fashion to recover hidden region which can be explored as part of the future work.

It is important to note that in monocular reconstruction setup (with fixed illumination), it is an ill-posed problem to absolutely discriminate between geometrical and textural edges. Hence, our proposed wrinkle map formulation is also susceptible to failure as it is largely dependent on training data distribution. Thus, it will be interesting to explore a solution in the multi-view (varying illumination) or temporal learning setup where it can be easy to differentiate between these geometrical and textural edges.

4.6 Conclusion

Predicting accurate 3D face and body largely enhances the realism of 3D human body models. In this paper, we proposed a novel reconstruction framework where we incorporate facial prior and wrinkle map prior to recover detailed geometry of face and body in people wearing loose clothing. We demonstrated results on in-the-wild settings by training our model with publicly available datasets.

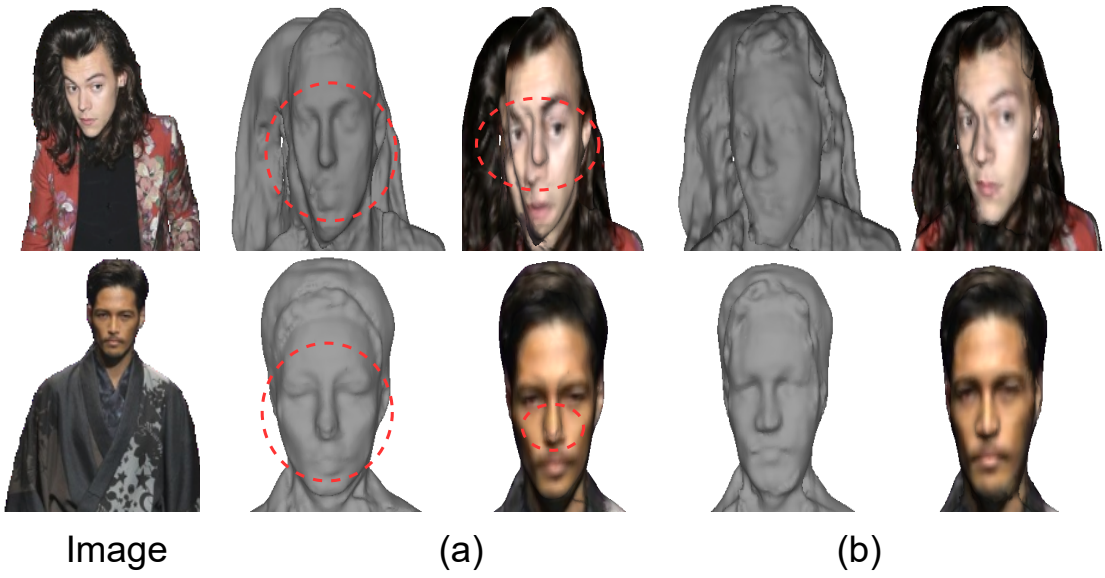

Figure 4.7 Qualitative ablative analysis of facial prior on internet images. (a) with SMPL face prior (b) with our face prior.

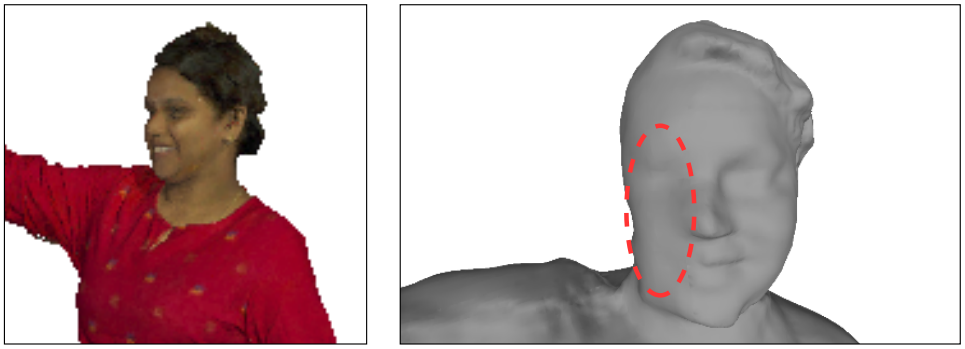

Figure 4.8 Limitation of our method.

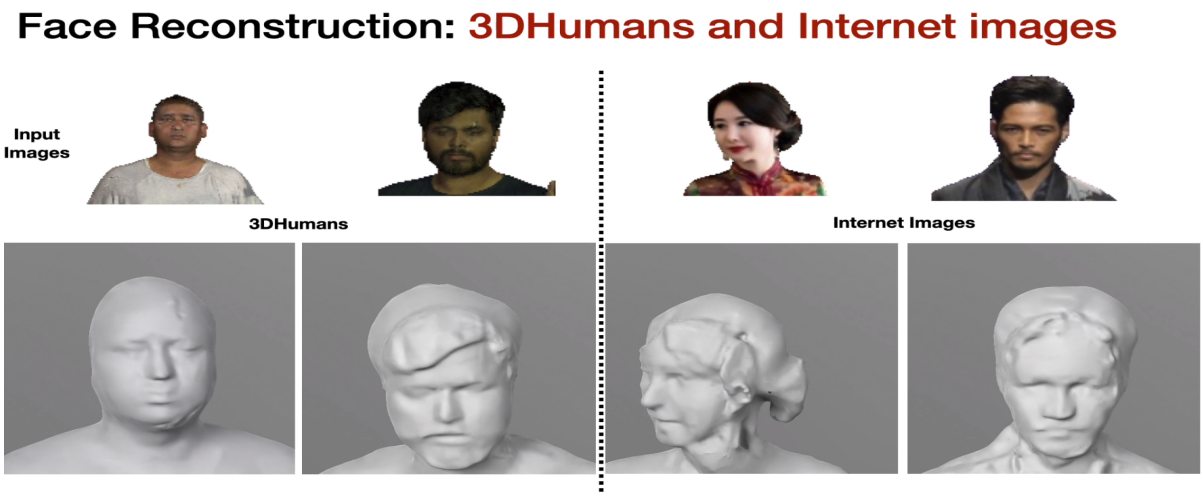

Figure 4.9 Additional face results on 3DHumans and real-world internet images.

Figure 4.10 Additional face results on THuman 2.0.

Chapter 5

Conclusion

We conclude the thesis by listing out the efforts and contributions made in digitizing humans. Further future directions to explore and look out for are discussed in the end.

5.1 Summary

In this thesis, we addressed the problem of efficient and realistic personalized 3D human digitization of people in arbitrary clothing from monocular images. We provided efficient solutions by incorporating various priors like peeled segmentation context prior, wrinkle map prior, SMPL body prior, and DF2net facial prior in the peeled representation of humans. We exploit deep learning techniques to achieve accurate high-fidelity person-specific reconstruction.

Initially, we propose a peeled semantic segmentation based prior to improve the reconstruction of peeled representation in the scenarios of partial occlusion due to body and clothing. Peeled semantic representation enables to capture 3D shape context unlike the traditionally used segmentation map of a monocular RGB image. Thus a coarse-to-fine refinement framework was proposed using the peeled segmentation maps as prior. We observed a significant improvement both quantitatively and qualitatively, especially in the scenarios of partial occlusion due to body and clothing. Additionally, these 3D semantic segmentation labels can be leveraged to extract the cloth from the reconstruction. Though this can handle partial occlusions it still suffers from complete occlusions and lacks high-frequency facial details. To eliminate the ambiguity due to severe occlusion we adopt SHARP which uses SMPL body prior to guarantee pose and shape-aware reconstruction. Even with SMPL body prior SHARP still lacks person-specific geometric details and is prone to false geometric edges coming from texture patterns present in the image space. To address these issues we present a novel framework consisting of a wrinkle map prior that distinguishes the geometrical edges from the textural edges and a facial prior was incorporated into the framework to extract person-specific facial details. It is to be noted that the facial prior provides a relatively smooth geometry of the person without any high-frequency details and it is through our network we obtain high-fidelity face reconstruction. We also observe that our wrinkle map prior enables perfect alignment between the reconstructed geometry and texture which is important in

applications such as AR/VR, the entertainment industry, etc. We demonstrate the generalizability of our proposed solutions by evaluating them on real-world images while being solely trained on publicly available datasets.

5.2 Limitations

Though our initially proposed peeled semantic segmentation-based reconstruction framework(chapter 3) can deal with partial self-occlusions it still fails in the scenarios of extreme occlusions. It also suffers from geometric noise due to textural patterns and lacks person-specific facial details and realism. We address these limitations in chapter 4 by incorporating body, face, and wrinkle map priors. However, the reconstruction is still affected by the quality of the wrinkle map. The way we generated wrinkle map data for training it is possible that the wrinkle map might miss capturing small-scale minute wrinkle details. Also, we used a frontal face prior thus leading to smooth output on the occlude face regions. This can be dealt with by incorporating a full head parametric prior and synthesizing high-frequency details and geometry of the occluded regions in a generative modeling fashion. This is a future direction to explore.

5.3 Impact

This thesis advances the field of high-fidelity 3D human digitization from monocular images in real time. A significant improvement both quantitatively and qualitatively has been achieved over multiple publicly available datasets. Though we built upon the existing representations the proposed solutions have the capacity to generalize to the in-the-wild scenarios. We addressed the problem of reconstruction noise due to textural patterns which has never been tended to by the existing methods. We are also the first to demonstrate full body reconstruction with high-frequency facial geometric details on in-the-wild internet images.

5.4 Future Research Directions

Even though a lot of progress has been made in this field there are a lot of unexplored problems. New ideas and solutions are required to address these problems. We point out a few directions to explore based on this thesis below:

- **Multiview 3D human digitization:** Extending the proposed solutions to multi-view setup both the quality of the reconstruction and texture can be significantly improved as statistical body priors generate plausible pose in case occlusion in single view setup whereas an accurate pose can be estimated using multi-view images. It is interesting to see how the challenge of dealing with calibration between in-the-wild images will be addressed.

- **Multiple Humans in the image:** Until now we assumed only a single human in the image. In real life, we come across many instances where there are multiple people in the image. This is a challenging task as now there are self-occlusions and occlusions due to other subjects. Also modeling the interactions between various subjects is challenging.
- **Temporal:** Animating non-parametric objects over time is a challenging task. As loose clothing is not tightly coupled with the body generating temporally coherent deformations over time could be an interesting direction to look out for.
- **Accurate Hand Pose:** In this thesis we addressed high-fidelity facial reconstruction. Similarly one can tackle the problem of accurate hand pose while reconstructing a complete body. There is a vast range of complex hand poses and this could also help with simulating human interaction with small objects.

Publications

Thesis Publications

- **Snehith Goud Routhu**, Sai Sagar Jinka and Avinash Sharma; *Coarse-to-Fine 3D Clothed Human Reconstruction Using Peeled Semantic Segmentation Context*; **Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing.** Association for Computing Machinery, New York, NY, USA, Article 34, 1–9. https://doi.org/10.1145/3490035.3490293.

- **Snehith Goud Routhu**, Sai Sagar Jinka and Avinash Sharma; *REF-SHARP: REFined face and geometry reconstruction of people in loose clothing*; **In Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'22).** ACM, New York, NY, USA, Article 22, 10 pages.

Other Publications

- Astitva Srivastava,**Snehith Routhu**, Chandradeep Pokhariya, Sai Sagar Jinka, and Avinash Sharma; *Template-free textured 3D garment reconstruction from a monocular image*; Patent application under submission.

Bibliography

[1] Face recognition. `https://github.com/ageitgey/face_recognition`.

[2] D. W. J. a. J. M. A. Kanazawa, M. J Black. End-to-end recovery of human shape and pose. *CVPR.*, 2018.

[3] C. M. A. S Jackson and G. Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. *ECCV*, 2018.

[4] S. J. A. Venkat and A. Sharma. Deep textured 3d reconstruction of human bodies. *BMVC*, 2018.

[5] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[6] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. mu reconstruction of 3d people models. *CoRR*, abs/1803.04758, 2018.

[7] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[8] X. W. T. C. P.-M. G. Alldieck T., Magnor M. Detailed human avatars from monocular video. *3D Vision*, 2018.

[9] X. W. T. C. P.-M. G. Alldieck T., Magnor M. Video based reconstruction of 3d people models. *CVPR*, 2018.

[10] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Transaction on Graphics*, 24:408–416, 2005.

[11] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM Trans. Graph.*, 24:408–416, 2005.

[12] C. T. B. L. Bhatnagar, G. Tiwari and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. *ICCV*, 2019.

[13] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Neural Information Processing Systems (NeurIPS)*, December 2020.

[14] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-Garment Net: Learning to dress 3D people from images. In *ICCV*, 2019.

[15] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct 2019.

[16] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5419–5429, 2019.

[17] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[18] F. Bogo, A. Kanazawa, C. Lassner, P. V. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[19] L. C. G. P. R. J. B. M. Bogo F., Kanazawa A. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. *ECCV*, 2016.

[20] Y. Cao, G. Chen, K. Han, W. Yang, and K.-Y. K. Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[21] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. *CVPR*, 2019.

[22] L. M. Curless B. A volumetric method for building complex models from range images. *Computer Graphics and Interactive Techniques, SIGGRAPH*, 1996.

[23] H. Dai, N. Pears, W. A. Smith, and C. Duncan. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*, pages 3085–3093, 2017.

[24] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[25] E. Dibra, H. P. Jain, A. C. Öztireli, R. Ziegler, and M. H. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5504–5514, 2017.

[26] O. C. Z. R. G. M. Dibra E., Jain H. Human shape from silhouettes using generative hks descriptors and cross-modal neural network. *CVPR*, 2017.

[27] D. K. S. T. J. R. Dragomir Anguelov, Praveen Srinivasan and J. Davis. Expressive body capture: 3d hands, face, and body from a single image. *In ACM Transactions on Graphics(ToG)*, 2005.

[28] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.

[29] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[30] N. G. T. B. A. A. O. D. T. G. Pavlakos, V. Choutas and M. J. Black. Expressive body capture:3d hands, face, and body from a single image. *CVPR*, 2019.

[31] N. G. T. B. A. O. D. T. G. Pavlakos, V. Choutas and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. *CVPR*, 2019.

[32] B. R. J. Y. E. Y. I. L. G. Varol, D. Ceylan and C. Schmid. Bodynet volumetric inference of 3d human body shapes. *ECCV.*, 2018.

[33] X. M. N. M. M. J. B. I. L. G. Varol, J. Romero and C. Schmid. Learning from synthetic humans. *CVPR*, 2017.

[34] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020.

[35] M. M. H. Bertiche and S. Escalera. Cloth3d: Clothed 3d humans. *ECCV*, 2020.

[36] Z. M. P.-M. G. T. C. Habermann M., Xu W. Livecap: Real-time human performance capture from monocular video. *In ACM Transactions on Graphics(ToG)*, 2019.

[37] T. He, J. Collomosse, H. Jin, and S. Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction, 2020.

[38] T. He, J. P. Collomosse, H. Jin, and S. Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *ArXiv*, abs/2006.08072, 2020.

[39] T. He, Y. Xu, S. Saito, S. Soatto, and T. Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021.

[40] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. ARCH: Animatable reconstruction of clothed humans. In *CVPR*, 2020.

[41] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. Arch: Animatable reconstruction of clothed humans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3090–3099, 2020.

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[43] J. S.-R. N. J. Joon Park, P. Florence and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *CVPR*, 2019.

[44] S. Jinka, R. Chacko, A. Sharma, and P. Narayanan. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. In *2020 International Conference on 3D Vision (3DV)*, pages 879–888. IEEE Computer Society, 2020.

[45] S. S. Jinka, R. Chacko, A. Srivastava, A. Sharma, and P. Narayanan. Sharp: Shape-aware reconstruction of people in loose clothing. *arXiv preprint arXiv:2106.04778*, 2021.

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

[47] S. Y. Joo H., Simon T. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CVPR.*, 2018.

[48] P. D. J. B. X. Y. M. W. G. H. S. O.-E. R. P. J. D. e. a. K. Guo, P. Lincoln. The relightables: volumetric performance capture of humans with realistic relighting. *In ACM Transactions on Graphics(ToG)*, 2019.

[49] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[50] Z. J. F. P. M. J. Kanazawa, A. Learning 3d human dynamics from video. *CVPR*, 2019.

[51] M. Kazhdan, M. Bolitho, and M. Hoppe. Poisson surface reconstruction. In *SGP*, volume 7, 2006.

[52] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29, 2013.

[53] M. M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29:1–29:13, 2013.

[54] P.-M. G. . Keyang Z., Bhatnagar B.L. Unsupervised shape and pose disentanglement for 3d meshes. *ECCV*, 2020.

[55] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019.

[56] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021.

[57] M. N. S. N. L. Mescheder, M. Oechsle and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. *CVPR*, 2019.

[58] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4704–4713, 2017.

[59] K. M. B. F. B. M. G. P. Lassner C., Romero J. Unite the people: Closing the loop between 3d and 2d human representation. *CVPR.*, 2017.

[60] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[61] J. Lin, Y. Yuan, T. Shao, and K. Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 5891–5900, 2020.

[62] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[63] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 1987.

[64] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *COMPUTER GRAPHICS*, 21(4):163–169, 1987.

[65] J. R. G. P.-M. M. Loper, N. Mahmood and M. J.Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (ToG)*, 2015.

[66] M. O. M. Niemeyer, L. Mescheder and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *CVPR*, 2020.

[67] G. P.-M. P. G. M. Omran, C. Lassner and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *3D Vision*, 2018.

[68] G. Moon and K. M. Lee. Pose2pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation. *ArXiv*, abs/2011.11534, 2020.

[69] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima. SiCloPe: Silhouette-based clothed people. In *CVPR*, 2019.

[70] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima. Siclope: Silhouette-based clothed people. In *CVPR*, pages 4480–4490. Computer Vision Foundation / IEEE, 2019.

[71] S. S. Newcombe R.A., Fox D. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. *CVPR*, 2015.

[72] S. S. Newcombe R.A., Fox D. Killingfusion: Non-rigid 3d reconstruction without correspondences. *CVPR*, 2017.

[73] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018.

[74] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. pages 484–494, 2018.

[75] X. Y.-M. W. G. H. S. O.-E. R. P. J. D. e. a. K. G. P. L. P. Davidson, J. Busch. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics(ToG)*, 34, 2019.

[76] T. Z. P. Isola, J.-Y. Zhu and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR,*, 2017.

[77] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[78] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[79] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[80] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[81] Z. X. D. K. Pavlakos G., Zhu L. Learning to estimate 3d human pose and shape from a single color image. *CVPR.*, 2018.

[82] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: seamless 4d clothing capture and retargeting. *ACM Trans. Graph.*, 36:73:1–73:15, 2017.

[83] H. S. B. M. Pons-Moll G., Pujades S. Clothcap: Seamless 4d clothing capture and retargeting. *In ACM Transactions on Graphics(ToG).*, 2017.

[84] D. C. R. M. Q. Xu, W. Wang and U. Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *NeurIPS*, 2019.

[85] E. Ramon, G. Triginer, J. Escur, A. Pumarola, J. Garcia, X. Giro-i Nieto, and F. Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021.

[86] W. C. S. Liu, S. Saito and H. Li. Learning to infer implicit surfaces without 3d supervision. *NeurIPS*, 2019.

[87] T. S. S. Lombardi, J. Saragih and Y. Sheikh. Deep appearance models for face rendering. *In ACM Transactions on Graphics(ToG)*, 2018.

[88] A. S. S. S. Jinka, R. Chacko and P. Narayanan. Peeledhuman: Robust shape representation for textured 3d human body reconstruction. *3DVision*, 2020.

[89] J. S. S. Saito, S. Simon and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. *CVPR*, 2020.

[90] S. Sagar Jinka, A. Srivastava, C. Pokhariya, A. Sharma, and P. J. Narayanan. SHARP: Shape-Aware Reconstruction of People in Loose Clothing. *arXiv e-prints*, page arXiv:2205.11948, May 2022.

[91] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019.

[92] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, and A. Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019.

[93] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

[94] E. Sariyanidi, C. J. Zampella, R. T. Schultz, and B. Tunc. Inequality-constrained and robust 3d face model fitting. In *European Conference on Computer Vision*, pages 433–449. Springer, 2020.

[95] T. Simon, S. Lombardi, J. Saragih, and Y. Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics(ToG)*, 6, 2018.

[96] J.-H. Song and H.-J. Shin. On parameterizing of human expression using ica. *Journal of the Korea Computer Graphics Society*, 15(1):7–15, 2009.

[97] C. T. T. Alldieck, G. Pons-Moll and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. *ICCV*, 2019.

[98] H. J. T. He, J. Collomosse and S. Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *NeurIPS*, 2020.

[99] B. L. B. C. T. Thiemo Alldieck, Marcus Magnor and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. *CVPR*, 2019.

[100] X. M. C. S. V. Gabeur, J.-S. Franco and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. *ICCV*, 2019.

[101] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018.

[102] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017.

[103] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[104] A. Venkat, S. S. Jinka, and A. Sharma. Deep textured 3d reconstruction of human bodies. In *BMVC*, 2018.

[105] Y. Xiu, J. Yang, D. Tzionas, and M. J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022.

[106] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *NeurIPS*, 2019.

[107] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. M., H. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, 2018.

[108] Z. A. F. W. S. R. S. C. Xu H., Bazavan E.G. Ghum : Generative 3d human shape and articulated pose models. *CVPR*, 2020.

[109] T. Yenamandra, A. Tewari, F. Bernard, H.-P. Seidel, M. Elgharib, D. Cremers, and C. Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12803–12813, 2021.

[110] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021.

[111] W. C. Y. Z. J. X. C. L.-L. L. C. M. Z. Huang, T. Li and H. Li. Deep volumetric video from very sparse multi-view performance capture. *ECCV*, 2018.

[112] X. Zeng, X. Peng, and Y. Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2315–2324, 2019.

[113] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[114] Z. Zheng, T. Yu, Y. Liu, and Q. Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[115] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, October 2019.

[116] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019.

[117] K. Zhou, B. L. Bhatnagar, and G. Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision (ECCV)*, August 2020.

[118] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4486–4495, 2019.

[119] W. S. C. X. Y. R. Zhu H., Zuo X. Detailed human shape estimation from a single image by hierarchical mesh deformation. *CVPR*, 2019.