SHARP: Shape-Aware Reconstruction of People in Loose Clothing

Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma and P.J. Narayanan

[1*]Centre for Visual Information Technology, IIIT Hyderabad, Gachibowli, Hyderabad, 500032, Telangana, India.

Contributing authors: jinka.sagar@research.iiit.ac.in; astitva.srivastava@research.iiit.ac.in; chandradeep.pokhariya@research.iiit.ac.in; asharma@iiit.ac.in; pjn@iiit.ac.in;

Abstract

Recent advancements in deep learning have enabled 3D human body reconstruction from a monocular image, which has broad applications in multiple domains. In this paper, we propose SHARP (**SH**ape **A**ware **R**econstruction of **P**eople in loose clothing), a novel end-to-end trainable network that accurately recovers the 3D geometry and appearance of humans in loose clothing from a monocular image. SHARP uses a sparse and efficient fusion strategy to combine parametric body prior with a non-parametric 2D representation of clothed humans. The parametric body prior enforces geometrical consistency on the body shape and pose, while the non-parametric representation models loose clothing and handles self-occlusions as well. We also leverage the sparseness of the non-parametric representation for faster training of our network while using losses on 2D maps. Another key contribution is *3DHumans*, our new life-like dataset of 3D human body scans with rich geometrical and textural details. We evaluate SHARP on 3DHumans and other publicly available datasets, and show superior qualitative and quantitative performance than existing state-of-the-art methods.

Keywords: 3D human body reconstruction, parametric and non-parametric methods, monocular image, deep learning

1 Introduction

Image-based 3D reconstruction of humans in loose clothing is an interesting and challenging open problem in computer vision. It has several applications in the domains of fashion, AR/VR, sports and healthcare. Traditional stereo/multi-view (including RGB and depth sensor) based reconstruction solutions (Gall et al, 2009; Shotton et al, 2011; Wei et al, 2012; Baak et al, 2011; Newcombe et al, 2015; Dou et al, 2016; Bogo et al, 2017) typically require studio environments with controlled lighting and multiple synchronized and calibrated cameras. Thus, recent approaches have shifted their focus on in-the-wild 3D reconstruction of humans.

With the advent of deep learning models, significant interest has garnered around 3D reconstruction from a monocular image (Kanazawa et al, 2018; Varol et al, 2018; Habermann et al, 2020), which is an ill-posed problem. Challenges like self-occlusions, arbitrary viewpoints and clothing occlusions make the scenario more complicated. One class of existing deep learning solutions attempts to fit a parametric body

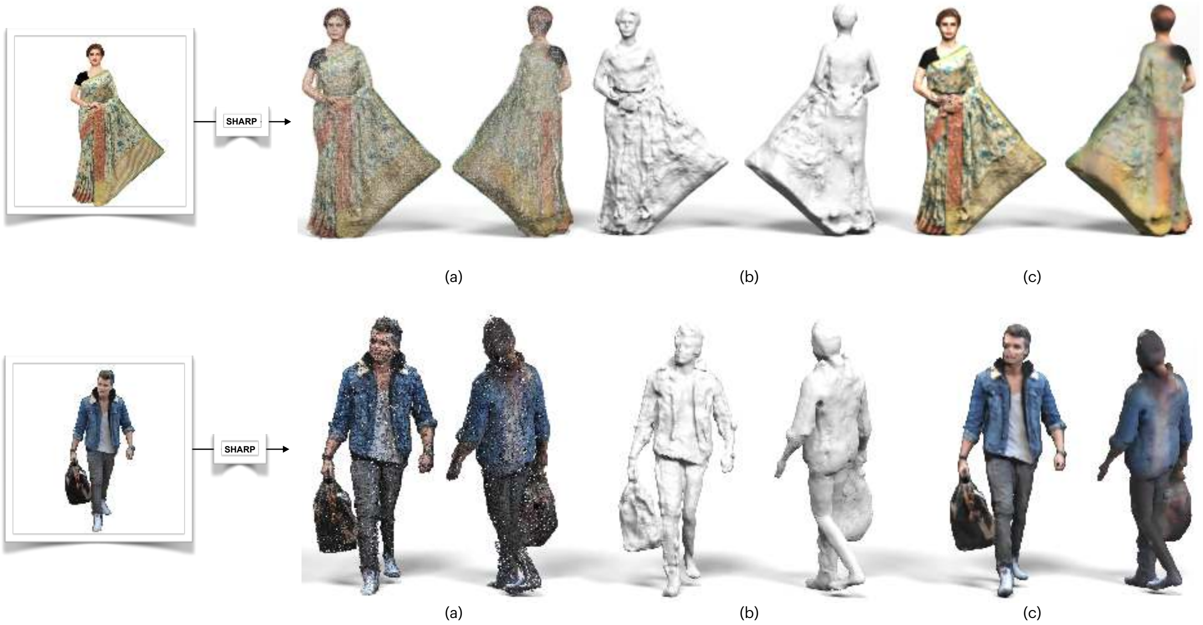

Fig. 1: Results of our method on in-the-wild images. Point cloud, uncolored and colored mesh is shown in (a), (b) & (c), respectively.

model like SMPL (Loper et al, 2015) to a monocular input image by learning from image features (Kanazawa et al, 2018; Güler et al, 2018; Omran et al, 2018; Lin et al, 2021; Kolotouros et al, 2021). SMPL prediction is improved when Multi-view input images are provided as shown in (Liang et al, 2019). However, such parametric SMPL mesh does not capture geometrical details owing to person-specific appearance and clothing. The other class of non-parametric reconstruction techniques pose no such body prior constraints (Saito et al, 2019, 2020; Natsume et al, 2019; Varol et al, 2018; Bhatnagar et al, 2020; Venkat et al, 2018) and hence can potentially handle loose clothing scenarios.

In particular, the recent implicit function learning models, like PIFu (Saito et al, 2019) and PIFuHD (Saito et al, 2020), estimate voxel occupancy by utilizing pixel-aligned RGB image features computed by projecting 3D points onto the input image. However, the pixel-aligned features suffer from depth ambiguity as multiple 3D points are projected to the same pixel. Another interesting work, Geo-PIFu (He et al, 2020) attempted to refine implicit function estimation by combining volumetric features and pixel-aligned features together to resolve local feature ambiguity. As an alternate representation for 3D objects/scenes, some of the recent works model scenes as multiple (depth) plane images (MPIs) (Tucker and Snavely, 2020) in camera frustum. 3D human body reconstruction has also been attempted in the same vein by predicting front and back depth maps in (Gabeur et al, 2019; Smith et al, 2019). However, the front-back representation fails to handle self-occlusions caused by body parts.

In our recent work peeledhuman (Jinka et al, 2020), we introduced *PeeledHuman*; a novel non-parametric shape representation of the human body to address the self-occlusion problem. PeeledHuman representation encodes the 3D human body shape as a set of depth and RGB peel maps. Depth (and RGB) peeling is performed by ray-tracing on the 3D body mesh and extending each ray beyond its first intersection to obtain the peel maps. This provides an elegant, sparse 2D encoding of body shape, which inherently addresses the self-occlusion problem. However, the non-parametric approaches do not explicitly seek to impose global body shape consistency and hence, produces implausible body shape and pose.

The aforementioned problems can be addressed by introducing a body shape prior while reconstructing humans in loose clothing. The volume-to-volume translation network proposed in DeepHuman (Zheng et al, 2019) attempts to combine image features with the SMPL prior in a volumetric representation. ARCH (Huang et al, 2020; He et al, 2021) proposed to induce a human body prior by sampling points around a template SMPL mesh before evaluating occupancy labels for each point. However, sampling around the canonical SMPL surface is insufficient to reconstruct humans with articulated poses in loose clothing. Similarly, PaMIR (Zheng et al, 2021) proposes to voxelize SMPL body and feed it as an input to the network, which conditions the implicit function around the SMPL feature volume. However, volumetric feature estimation is still computationally expensive and is limited by the resolution. Moreover, in PaMIR, texture and geometry cannot be inferred in an end-to-end fashion and require two separate networks. Additionally, all these existing SMPL prior-based methods do not effectively exploit the rich surface representation as they either voxelize or sample points around the SMPL surface.

The continuous surface representation provided by SMPL prior is valuable as it models the natural curvature of body parts which cannot be easily recovered with non-parametric methods. Some of the existing methods have been successfully shown to deform SMPL surfaces locally to accommodate relatively tight clothing scenarios (Alldieck et al, 2019a; Bhatnagar et al, 2019; Patel et al, 2020; Alldieck et al, 2019b; Lahner et al, 2018; Zhu et al, 2019). Nevertheless, they fail to handle loose clothing scenarios, as the surface of garments can also have complex geometrical structures that are only partially dependent on the underlying body shape and pose, where non-parametric methods have mainly been successful. Interestingly, we can retain the best of these two approaches by deforming SMPL surface locally while reconstructing the remaining surface details (loose clothing) with no body prior constraints. More specifically, one can decouple the reconstruction of 3D clothed body surface into two complementary partial reconstruction tasks: (a) to recover the person-specific body surface details by locally deforming the SMPL prior, (b) to recover the remaining surface details of the loose clothing that cannot be recovered by just deforming the SMPL prior.

In regard to the representation of the 3D surface, while implementing the above two tasks, PeeledHuman representation seems to be a good choice owing to its sparse encoding of a 3D surface into 2D maps. More importantly, such representation also enables a seamless fusion of the two partial reconstructions due to the spatially aligned nature of these maps. 3D geometry can be extracted from PeeledHuman representation by simply back-projecting the peel maps to generate point cloud. Recent works (Ma et al, 2021a,b) have shown that point clouds are a good way to model clothing deformations arising from articulated pose.

Thus, this work proposes SHARP, a novel 3D body reconstruction method that can successfully handle significantly loose clothing, self-occlusions and arbitrary viewpoints. SHARP takes SMPL body encoded in PeeledHuman representation (Jinka et al, 2020), aligned to the input image as a prior to the reconstruction framework. The *SMPL prior peel maps*, along with the monocular RGB image, is fed as the input to our framework, which initially predicts *residual peel maps*, *auxiliary peel maps*, along with *RGB peel maps*. Here, the residual peel maps represent the pixel-wise depth offsets from SMPL prior peel maps in the view direction. On the other hand, auxiliary peel maps model the complementary geometrical details of the surface, which are not handled by residual peel maps. Subsequently, predicted residual and auxiliary peel maps are fused to obtain *fused peel maps*, capturing the geometry of the unified clothed body. The final fused peel maps, along with predicted RGB peel maps are back-projected to obtain the colored point cloud. We finally recover the mesh after minimal post-processing of the corresponding point cloud followed by meshification using Poisson Surface Reconstruction(Kazhdan et al, 2006). The fused peel maps can model arbitrarily loose clothing and can handle accessories (e.g., bags) as well, as shown in Figure 1. Unlike other existing methods that use adversarial loss and 3D Chamfer loss, the proposed problem formulation enables our network to learn only with L_1 losses on 2D maps, which reduces the training time. Since, the clothed human body can be recovered in the form of point cloud directly from the back-projection of final fused peel maps, the inference time is also significantly reduced. We have described layerwise back-projection in greater detail in **section 2** of the supplementary draft.

Additionally, many state-of-the-art methods for reconstructing 3D human bodies (Saito et al, 2019, 2020; Zheng et al, 2021; Natsume et al, 2019; Huang et al, 2020) train their models on expensive commercial datasets which are not publicly available. These datasets have 3D human body scans which resemble real humans. This data helps the learning-based models to generalize well on unseen real-world scenarios. Unfortunately, the majority of existing datasets available in the public domain (Bhatnagar et al, 2019; Zheng et al, 2019; Bertiche et al, 2020; Tiwari et al, 2020) either consist of 3D body models in relatively tighter clothing, lack high-frequency geometrical & texture details, or are synthetic in nature. Recently, THUman2.0 (Yu et al, 2021) dataset released in the public domain has high-quality 3D body scans captured using a dense DSLR rig. Although they provide human scans with relatively loose clothing styles, their data lacks significantly loose garment types which occlude the lower body completely, e.g., long-skirt/tunic/saree. Moreover, the dataset is reconstructed with the multi-camera setup which has its known limitations. To bridge these gaps, we collected ***3DHumans***, a dataset of 3D human body scans with a wide variety of clothing styles and varied poses using a commercial structured-light sensor (accurate up to 0.5mm). We are able to retain high-frequency geometrical and textural details, as shown in Figure 6. We also benchmark some of the SOTA methods on this dataset and report superior performance of our method. To summarize, our contributions are:

1. We propose SHARP, a novel approach to fuse parametric and non-parametric shape representations for reconstructing 3D body model in loose clothing from an input monocular (RGB) image.
2. Our proposed end-to-end learnable encoder-decoder framework infers color and geometrical details of body shape in a single forward pass at a lower inference time as compare to SOTA methods.
3. We collected ***3DHumans***, a dataset of 3D human body scans that has a wide variety of clothing and body poses with rich textural and geometrical details. The dataset will be released in the public domain to further accelerate the research.

2 Related Work

Parametric Body Fitting. Estimating the 3D parametric human body models, like SMPL (Loper et al, 2015), SMPL-X (Pavlakos et al, 2019), SCAPE (Anguelov et al, 2005) etc., from a monocular image using deep learning methods (Bogo et al, 2016; Kanazawa et al, 2018) has achieved a great success with robust performance. In particular, HMR (Kanazawa et al, 2018) proposes to regress SMPL parameters while minimizing re-projection loss with the known 2D joints. Different priors have been used to refine the parametric estimates as in (Varol et al, 2017; Omran et al, 2018; Kolotouros et al, 2019a; Kanazawa et al, 2019; Kolotouros et al, 2021; Lin et al, 2021). Despite these approaches being computationally efficient, they lack realistic human appearance and clothing details. Methods for modelling details like hair/cloth/skin by estimating offsets from SMPL vertex have been proposed, but they work on very tight clothing and can not model the loose clothing deformation arising from pose . (Bhatnagar et al, 2019; Venkat et al, 2019; Kolotouros et al, 2019b).

Non-parametric Body Reconstruction: Recovering 3D human body from multi-camera setup requires traditional techniques like voxel carving, triangulation, multi-view stereo, shape-from-X (Azevedo et al, 2009; Dou et al, 2016; Bogo et al, 2017; Mulayim et al, 2003). Stereo cameras and consumer RGBD sensors are highly susceptible to noise. In the domain of deep learning, initially, voxel methods gained popularity as 3D voxels are a natural extension to 2D pixels (Venkat et al, 2018; Varol et al, 2018; Zheng et al, 2019). SiCloPe (Natsume et al, 2019) estimates human body silhouettes in novel views to recover underlying 3D shape from 2D contours. Recently, implicit function learning methods for human body reconstruction became popular, which use pixel-aligned features to learn neural implicit function over a discrete occupancy grid (Saito et al, 2019, 2020). However, these methods suffer from sampling redundancy as they have to sample points in a grid to infer the surface, majority of which do not lie on the actual surface. They also suffer from depth ambiguity as multiple 3D points map to the same pixel-aligned feature. Animating clothed humans with template garments is proposed in (Corona et al, 2020). However, this method cannot produce the textural details of the garments. In our recent work peeledhuman (Jinka et al, 2020), we proposed a sparse 2D representation of 3D surface by estimating and storing the intersection of the surface with ray.(Mildenhall et al, 2020) where it samples points along the camera ray to evaluate RGBσ on these samples.

Prior-based Non-Parametric Body Reconstruction: ARCH (Huang et al, 2020; He et al, 2021) learns a deep implicit function by sampling points around the 3D clothed body in the canonical space. But, the transformation of the clothed mesh from canonical space to arbitrary space is done by learning SMPL-based skinning weights which can not handle the deformation of the loose clothing. These methods rely on large scale dataset of 3D human scans to train the model, and suffer from reconstruction errors and weak generalization capability although demonstrating good results. (Bhatnagar et al, 2020) also proposes to combine strengths of parametric and non-parametric models. However, it takes input as sparse point cloud which is difficult to obtain in-the-wild settings. Geo-PIFu (He et al, 2020) utilizes structure-aware latent voxel features, along with pixel-aligned features to learn a neural implicit function. PaMIR (Zheng et al, 2021) learns a deep implicit function conditioned on the features which are a combination of 2D features obtained from image and 3D features obtained from the SMPL body volume. However, voxel features are computationally expensive and of low resolution. DeepHuman (Zheng et al, 2019) leverages dense semantic representations from SMPL as an additional input. Nevertheless, similar to Geo-PIFu, DeepHuman is also a volumetric-regression based approach and hence, incurs a high computational cost. Moreover, similar to PIFu, these deep implicit methods require separate networks for learning geometry and texture.

3D Human Body Datasets: Deep-Learning based 3D human body reconstruction solutions rely on the data available at hand. Not only the shear amount of samples, but the quality of geometry and texture is also important in order to drive the learning. Many 3D human body datasets have been proposed, some of which only contain body-shape information, while some also include clothing details on top of it. TOSCA (Bronstein et al, 2008) dataset contains synthetic meshes of fixed topology with artist-defined deformations. SHREC (Li et al, 2012) and FAUST (Bogo et al, 2014) provide meshes and deformation models created by an artist that cannot reproduce what we find in the real world. BUFF (Zhang et al, 2017) contains 3D scans with relatively richer geometry details, but the number of subjects, poses and clothing style is very limited and not sufficient to generalize deep learning models. Another synthetic dataset CLOTH3D (Bertiche et al, 2020) incorporates loose clothing by draping 3D modeled garments on SMPL in Blender. It has a wide variety of clothing styles, but due to the nature of SMPL body model, details like hair and skin are absent. THUman1.0 (Zheng et al, 2019) dataset provides a large number of human meshes with varied poses and subjects. However, the texture quality is low and cannot mimic real-world subjects. SIZER (Tiwari et al, 2020) dataset provides real scans of 100 subjects, wearing garments in 4 different sizes of 10 fixed garments classes. But all the scans are in A-pose which is insufficient for a deep learning model to generalize to different poses. THUman2.0 (Yu et al, 2021) dataset provides a large number of high-quality textured meshes of different subjects in various poses. It also incorporates varied clothing styles and high-frequency geometrical details like hair and wrinkles etc. However, loose wrapped clothing styles, which completely occlude the full body, are still absent.

3 Method

In this section, we first outline PeeledHuman representation for encoding 3D shapes and discuss briefly about SMPL, followed by the details of our proposed framework.

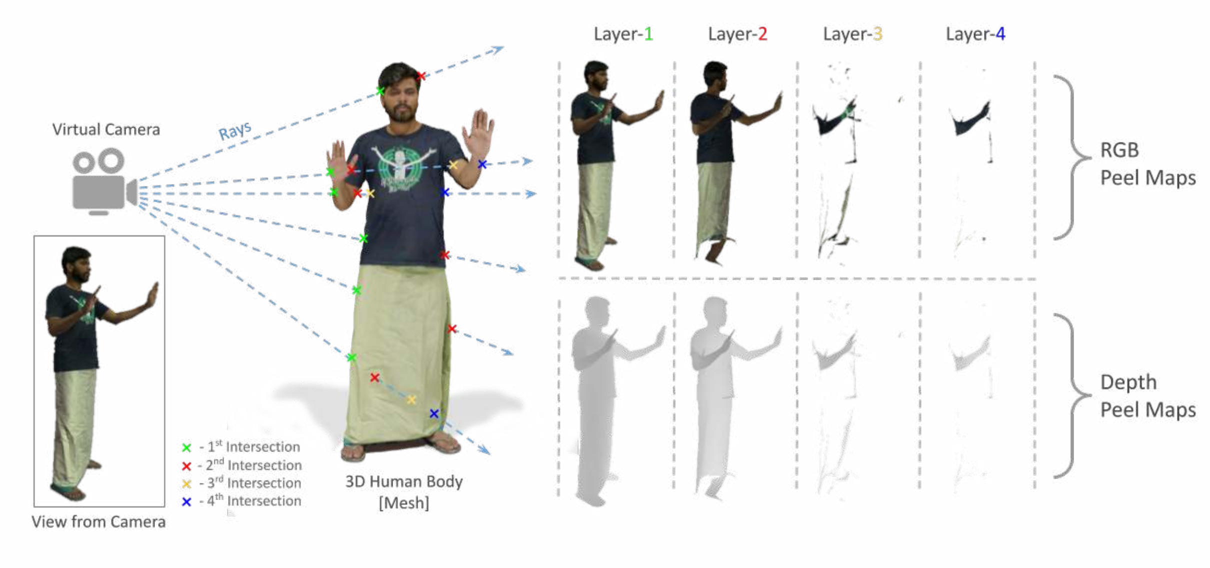

Fig. 2: PeeledHuman representation to encode 3D human body into 2D maps.

3.1 Background

3.1.1 PeeledHuman Representation

Our PeeledHuman representation is a sparse, non-parametric, multi-layered encoding of 3D shapes (Jinka et al, 2020). The human body mesh is placed in a virtual scene and a set of rays are emanated from the camera center through each pixel towards the mesh. The first set of ray-intersections with the mesh are recorded as the first layer depth peel map d_1 and RGB peel map r_1, capturing visible surface details nearest to the camera. This is similar to RGBD images captured from sensors, like Kinect. Subsequently, the rays are extended beyond the first intersection point (piercing through the intersecting surface) to hit the surface behind it. The corresponding depth and RGB values are recorded in the next layer peel maps, represented by d_i and r_i respectively, as shown in Figure 2. We use total $i = 4$ layers of peeled representation in this work. This representation is efficient, as it only stores ray-surface intersection in the form of sparse 2D maps, unlike voxels and implicit function representations, which are redundant in their representation.

3.1.2 SMPL Parametric Body Model

Skinned Multi-Person Linear (SMPL) (Loper et al, 2015) is a parametric 3D model of the human body that is based on vertex-based skinning and blend shapes and is learned from thousands of 3D body scans. SMPL factors the full human body mesh into the pose ($\theta \in \mathbb{R}^{72}$) and shape ($\beta \in \mathbb{R}^{10}$) parameters. θ for each joint is defined as the axis angle rotation relative to its parent in the kinematic tree, while β represents the shape PCA coefficients learned from various body scans.

SMPL starts with an artist-created mean template mesh $\mathcal{T} \in \mathbb{R}^{6890\times3}$ and blend skinning weights $\mathcal{W} \in \mathbb{R}^{6890\times24}$. Template mesh, based on its skeleton joints, $\mathcal{J}(\cdot)$ is deformed through two blend functions $\mathcal{B}_s(\beta)$ and $\mathcal{B}_p(\theta)$. Shape blend-shape function $\mathcal{B}_s(\beta)$ performs the per-vertex displacements, sculpting the person's identity, whereas pose-dependent blend-shape function $\mathcal{B}_p(\theta)$ takes a vector of pose parameters θ as input and maps them to another set of additive per-vertex displacements. Pose-dependent blend-shape function accounts for dynamic soft tissue deformation caused by the pose deviation from the rest-pose. Finally, the deformed template mesh $\mathcal{T}(\theta + \mathcal{B}_s(\beta) + \mathcal{B}_p(\theta))$ is transferred to the final mesh $\mathcal{M}(\theta, \beta)$ through a linear blend skinning (LBS) function $W(\cdot)$ as:

$$\mathcal{M}(\theta, \beta) = W(\mathcal{T}(\theta + \mathcal{B}_s(\beta) + \mathcal{B}_p(\theta)), \mathcal{J}(\beta), \mathcal{W}) \quad (1)$$

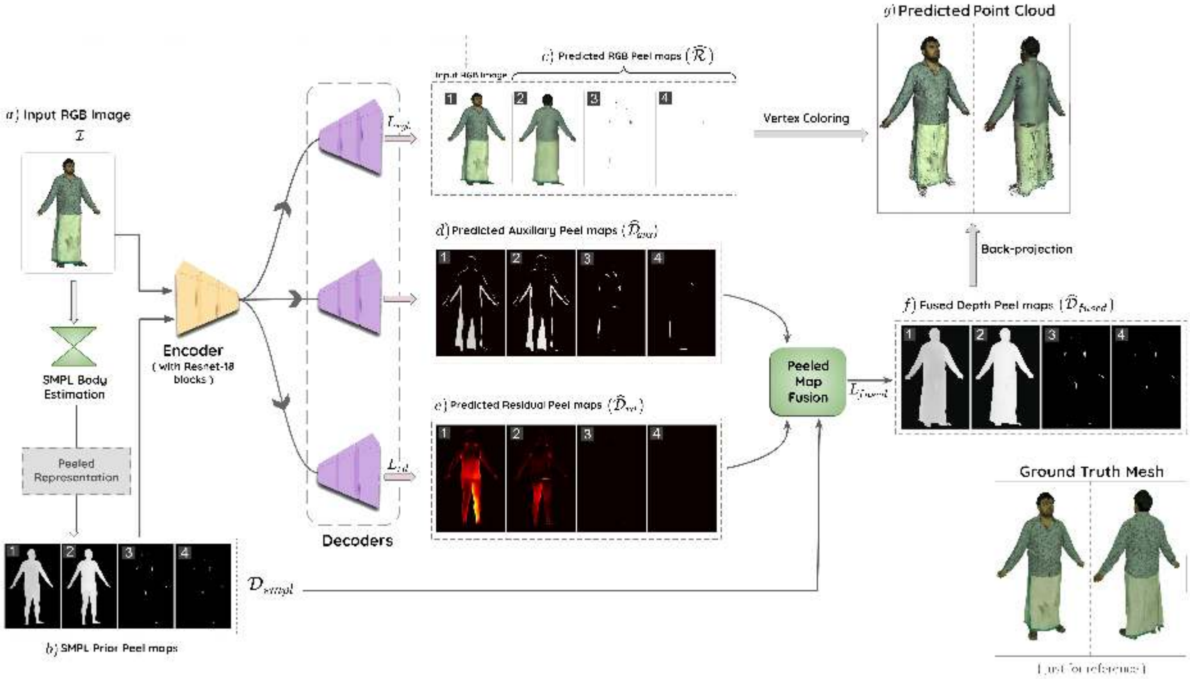

Fig. 3: **Pipeline:** We use an off-the-shelf method to estimate SMPL prior from the input image $\mathcal{I}$, and encode it into peeled representation ($\mathcal{D}_{smpl}$). This, along with image $\mathcal{I}$, is fed to an encoder. Subsequently, three separate decoders branches predict RGB peel maps ($\widehat{\mathcal{R}}$), auxiliary peel maps ($\widehat{\mathcal{D}}_{aux}$) and residual peel maps ($\widehat{\mathcal{D}}_{rd}$), respectively. Finally, a layer-wise fusion of $\widehat{\mathcal{D}}_{aux}$, $\widehat{\mathcal{D}}_{rd}$ and $\mathcal{D}_{smpl}$ is performed to obtain fused peel maps $\widehat{\mathcal{D}}_{fused}$, which is then back-projected along with $\widehat{\mathcal{R}}$ to obtain a vertex colored point-cloud. (The ground truth mesh is shown for comparison only.)

3.2 Overview

We aim to reconstruct a 3D textured human body model of a person in arbitrary pose and clothing from a given monocular input image $\mathcal{I}$, as shown in Figure 3.

Here, we discuss the steps involved in our proposed method.

1. SMPL shape and pose parameters (i.e., $\beta \in \mathbb{R}^{10}$, $\theta \in \mathbb{R}^{72}$) along with parameters of weak perspective camera (s, t_x, t_y) are estimated from ProHMR (Kolotouros et al, 2021). We convert the estimated SMPL to depth peel maps which acts as a shape prior $\mathcal{D}_{smpl}$ (Figure 3) as outlined in subsubsection 3.3.1.
2. Later, input image $\mathcal{I}$ (with background removed) is concatenated with $\mathcal{D}_{smpl}$ and is fed as an input to the shared encoder in our network.
3. Subsequently, three decoders predict different outputs through separate branches, namely, RGB peel maps $\widehat{\mathcal{R}}$, auxiliary peel maps $\widehat{\mathcal{D}}_{aux}$ and residual peel maps $\widehat{\mathcal{D}}_{rd}$, as shown in Figure 3 (c)-(e). The topmost decoder branch predicts only three RGB peel maps as the input $\mathcal{I}$ naturally acts as the first RGB peel map.
4. The SMPL prior peel maps $\mathcal{D}_{smpl}$, residual peel maps $\widehat{\mathcal{D}}_{rd}$ and auxiliary peel maps $\widehat{\mathcal{D}}_{aux}$ are further combined using SMPL mask Γ_i (estimated using Equation 3) to obtain the final fused peel maps $\widehat{\mathcal{D}}_{fused}$.
5. Finally, a colored point-cloud is obtained by back-projecting $\widehat{\mathcal{D}}_{fused}$ and $\widehat{\mathcal{R}}$ to camera coordinate system, as shown in Figure 3 (g). This point-cloud is further post-processed, and then meshified using Poisson Surface Reconstruction (PSR) (Kazhdan et al, 2006).

To illustrate the importance of a shape prior in the prediction of peel maps, we compare SHARP with peeledhuman (Jinka et al, 2020). The peeledhuman network predicts inconsistent body parts as shown in Figure 5 (a). This is because of the fact that there are no geometrical constraints imposed on the structure of predicted body parts. The introduction of prior enables SHARP to reconstruct the human body with plausible body parts and accurate pose as shown in Figure 5 (b).

3.3 Pipeline Details

Here, we discuss in detail about our pipeline, which involves peeled shape prior, residual & auxiliary peel maps and finally, peel map fusion. We also explain in detail the loss functions used to train SHARP.

3.3.1 Peeled Shape (SMPL) Prior

We initially use (Kolotouros et al, 2021) to estimate the SMPL pose and shape parameters (β, θ), along with weak-perspective camera parameters (s, t_x, t_y). The SMPL mesh is brought into the camera coordinate system using (s, t_x, t_y), and then encoded into depth peel maps by passing camera rays through each pixel, as explained in subsubsection 3.1.1, i.e., for every pixel p in layer i, depth value of the point intersected by the camera ray is stored:

$$\mathcal{D}_{smpl} = \{(d_p^i) : \forall p \in \mathcal{I}, i \in \{1,2,3,4\}, d \in \mathbb{R}\} \tag{2}$$

We initialize a layer-wise binary SMPL mask Γ_i by applying thresholding on SMPL prior peel maps. Additionally, we condition this mask on a pre-estimated binary foreground mask $\mathcal{F}$. The foreground mask $\mathcal{F}$ covers only the clothed human in the input image and can be obtained using off-the-shelf background segmentation methods e.g. PGN (Gong et al, 2018) . We use $\mathcal{D}_{smpl}^i$ and $\mathcal{F}$ to estimate the per-layer SMPL mask Γ_i as :

$$\Gamma_i = \begin{cases} 1, & \text{if } \mathcal{D}_{smpl}^i \odot \mathcal{F} > 0 \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In essence, Γ_i for each layer is estimated by retaining only the overlapping regions in corresponding SMPL prior peel map and the foreground mask. This helps refine the initial SMPL mask Γ_i by eliminating parts of the SMPL prior peel maps that falls outside the human body & clothing silhouette, thereby enabling our method to partially overcome the misalignment of SMPL prior with the input image. Note that, $\mathcal{F}$ is common across all the layers. The refined SMPL mask Γ_i is subsequently used for peel map fusion in Equation 3.3.3.

3.3.2 Residual and Auxiliary Peel Maps

To estimate view specific deformations from the SMPL prior input, we propose to predict residual peel maps $\widehat{\mathcal{D}}_{rd}$ by computing additive pixel-wise offsets from the input SMPL depth peel maps $\mathcal{D}_{smpl}$. For every pixel p in layer i of peeled SMPL prior, we predict offset along z-axis [1]:

$$\widehat{\mathcal{D}}_{rd} = \{(\widehat{\delta}_p^i) : \forall p \in \mathcal{I}, i \in \{1,2,3,4\}, \widehat{\delta} \in \mathbb{R}\} \tag{4}$$

For pixels, which depict the projection of bare body parts, network predicts minimal offsets ($\widehat{\mathcal{D}}_{rd}$), thereby capturing the person-specific appearance features like hairline and facial details while preserving overall structure of the body parts.

Thus, each layer of the residual peel maps provides pixel-wise displacements of the corresponding layer of peeled SMPL prior maps along the view-direction (z-axis). These residual deformations only cover the pixels in the input image for which SMPL prior is present. For the remaining pixels of clothed body, we propose to learn their depth values using a separate branch in the form of auxiliary peel maps.

$$\widehat{\mathcal{D}}_{aux} = \{(\widehat{d}_{aux}^i) : \forall p \in \mathcal{I}, i \in \{1,2,3,4\}, \widehat{d} \in \mathbb{R}\} \tag{5}$$

Figure 4 provide 3D visualization of ($\mathcal{D}_{smpl}$ + $\widehat{\mathcal{D}}_{rd}$) and $\widehat{\mathcal{D}}_{aux}$ by back-projecting respective partial depth peel maps. We can observe that these two capture the complementary geometrical details of 3D body and clothing.

3.3.3 Peel Map Fusion

The predicted residual and auxiliary peel maps independently capture complimentary surface details and are subsequently fused to obtain the geometry of unified clothed body. We propose to obtain final fused peel depth maps by layer-wise fusion of $\mathcal{D}_{smpl}$, $\widehat{\mathcal{D}}_{rd}$ and $\widehat{\mathcal{D}}_{aux}$ expressed as:

[1] The camera is placed at (0, 0, 10), Y axis is up and -Z axis is forward, while meshes are placed at origin.

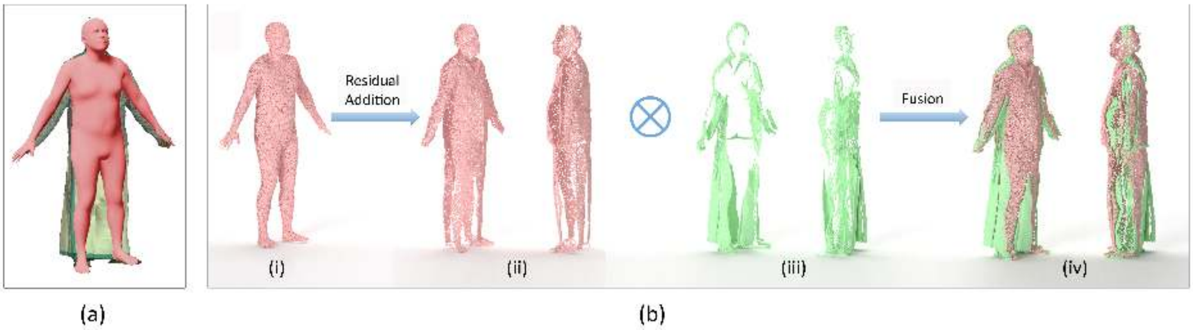

Fig. 4: **(a) SMPL prior overlayed on the input image**: The residual peel maps recover depth along the pixels over which SMPL prior is present across all the layers. For the remaining pixels, auxiliary peel maps are used to recover depth. **(b) 3D representation of fusion:** (i) Point cloud obtained $\mathcal{D}_{smpl}$ is shown in red. (ii) Point cloud obtained from $(\mathcal{D}_{smpl} + \widehat{\mathcal{D}}_{rd})$ is shown from two views in red. (iii) Point cloud obtained from $\widehat{\mathcal{D}}_{aux}$ is shown from two views in green. (iv) Final point cloud obtained from $\widehat{\mathcal{D}}_{fused}$.

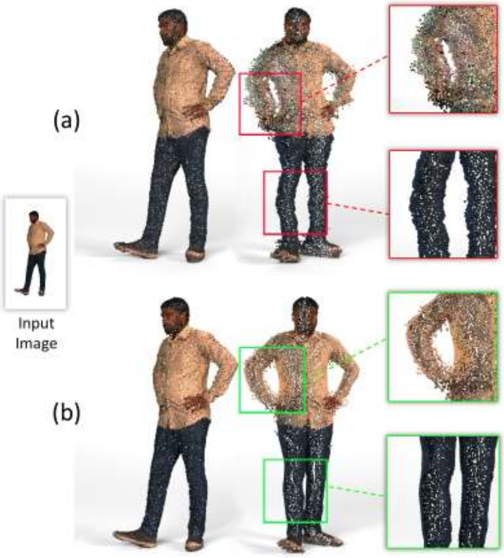

Fig. 5: (a) Distorted body parts in the prediction from peeledhuman (Jinka et al, 2020). (b) Reconstruction obtained from SHARP.

$$\widehat{\mathcal{D}}_{fused} = (\mathcal{D}_{smpl} + \widehat{\mathcal{D}}_{rd}) \otimes \widehat{\mathcal{D}}_{aux} \quad (6)$$

where $\otimes$ is the proposed layer-wise fusion operator as explained below. Here,

$$\widehat{\mathcal{D}}^i_{fused} = \Gamma_i \odot (\widehat{\mathcal{D}}^i_{rd} + \mathcal{D}^i_{smpl}) + (1 - \Gamma_i) \odot \widehat{\mathcal{D}}^i_{aux} \quad (7)$$

here, $\odot$ is element-wise multiplication and for each i^{th} layer $\widehat{\mathcal{D}}^i_{aux} \in \widehat{\mathcal{D}}_{aux}$, $\widehat{\mathcal{D}}^i_{rd} \in \widehat{\mathcal{D}}_{rd}$ and $\mathcal{D}^i_{smpl} \in \mathcal{D}_{smpl}$.

In summary, we have decoupled the task of recovering the clothed 3D human body surface into predicting residual and auxiliary peel maps. We later fused these partial reconstructions into a single unified 3D surface. Our approach ensures geometrically consistent body parts as the residual peel maps predict minimal offsets on the pixels belonging to the bare body where there is no clothing, thereby retaining body-specific geometry.

3.4 Loss Functions

We use encoder-decoder architecture for our predictions in SHARP. We train our network with losses on 2D peel map predictions. Our final learning objective is defined as:

$$L = L_{fuse} + \lambda_{rd} L_{rd} + \lambda_{rgb} L_{rgb} + \lambda_{sm} L_{sm} \quad (8)$$

where λ_{rd}, λ_{rgb} and λ_{sm} are regularization parameters for L_{rd}, L_{rgb}, L_{sm}, respectively. We provide the formulation for the individual loss terms below.

$$L_{fuse} = \sum_{i=1}^{4} \left\| \widehat{\mathcal{D}}^i_{fused} - \mathcal{D}^i_{fused} \right\|_1 \quad (9)$$

L_{fuse} is the sum of L_1 norm between ground truth depth peel maps $\mathcal{D}_{fused}$ and predicted fused peel maps $\widehat{\mathcal{D}}_{fused}$ for each i^{th} layer.

$$L_{rd} = \sum_{i=1}^{4} \left\| \widehat{\mathcal{D}}^i_{rd} - \mathcal{D}^i_{rd} \right\|_1 \quad (10)$$

L_{rd} constraints the residual peel map predictions to that of ground truth offsets. Note that we are training auxiliary peel maps branch without any explicit loss on $\widehat{\mathcal{D}}_{aux}$. The gradients through auxiliary peel map branch back-propagates using L_{rd} and L_{fuse}.

We also enforce per layer first order gradient smoothness of the predicted $(\widehat{\mathcal{D}}^i_{rd} + \mathcal{D}^i_{smpl})$ and ground truth $(\mathcal{D}^i_{rd} + \mathcal{D}^i_{smpl})$ as well as between ground truth and predicted $\widehat{\mathcal{D}}_{fused}$ maps. L^{fuse}_{sm} ensures smoothness between the two predicted surfaces.

$$L_{sm} = L^{rd}_{sm} + L^{fuse}_{sm} \tag{11}$$

where,

$$\begin{aligned} L^{rd}_{sm} &= \sum_{i=1}^{4} \left\| \triangledown(\mathcal{D}^i_{rd} + \mathcal{D}^i_{smpl}) - \triangledown(\widehat{\mathcal{D}}^i_{rd} + \mathcal{D}^i_{smpl}) \right\|_1 \\ L^{fuse}_{sm} &= \sum_{i=1}^{4} \left\| \triangledown\mathcal{D}^i_{fused} - \triangledown\widehat{\mathcal{D}}^i_{fused} \right\|_1 \end{aligned} \tag{12}$$

Additionally, We also train our network with L_1 loss between predicted and ground truth RGB peel maps (L_{rgb}).

4 3DHumans Dataset

As mentioned in section 1, one of the key bottlenecks that hinder progress in the field of 3D human body reconstruction is the lack of publically available real-world datasets that contain high-frequency texture and geometrical details.

To this end, we present 3DHumans, a dataset of around 250 scans containing people in diverse body shapes in various garments styles and sizes. We cover a wide variety of clothing styles ranging from loose robed clothing like saree (a typical South-Asian dress) to relatively tight-fitting shirt and trousers, as shown in Figure 6. The dataset consists of around 150 male and 50 unique female subjects. Total male scans are about 180 and female scans are around 70. In terms of regional diversity, for the first time, we capture body shape, appearance and clothing styles for the South-Asian population. We will release this data in the public domain for academic use.[2]

The 3DHumans dataset is created using the Artec3D Eva hand-held structured light scanner. The scanner has a 3D point accuracy of up to 0.1mm and 3D resolution is 0.5mm. For each 3D human scan, we also provide the SMPL body aligned to it, using (Zheng et al, 2021; Pavlakos et al, 2019).

5 Experiments & Results

In this section, we present the experimental details, datasets and training protocol for SHARP. We also show qualitative and quantitative comparisons with current state-of-the-art methods.

5.1 Implementation Details

We employ a multi-branch encoder-decoder network for SHARP, which is trained in an end-to-end fashion. The network takes the input image concatenated with SMPL peel maps in 512×512 resolution. The shared encoder is consist of a convolutional layer and 2 downsampling layers which have $64, 128, 256$ kernels of size $7{\times}7$, $3{\times}3$ and $3{\times}3$, respectively. This is followed by ResNet blocks which take downsampled feature maps of size $128{\times}128{\times}256$. The decoders for predicting $\widehat{\mathcal{D}}_{fused}$, $\widehat{\mathcal{D}}_{rd}$ and $\widehat{\mathcal{R}}$, consist of two upsampling layers followed by a convolutional layer, having same kernel sizes as of the shared encoder. Sigmoid activation is used in $\widehat{\mathcal{D}}_{fused}$ and $\widehat{\mathcal{D}}_{rd}$ decoder branches, while a tanh activation is used for the $\widehat{\mathcal{R}}$ decoder branch. The $\widehat{\mathcal{D}}_{rd}$ output values are scaled to a $[-1, 0.5]$ range which is found empirically.

We use the Adam optimizer with an exponentially decaying learning rate starting from 5×10^{-4}. Our network takes around 18 hrs to train for 20 epochs on 4 Nvidia GTX 1080Ti GPUs with a batch size of 8 and λ_{rd}, λ_{fuse}, λ_{rgb} and λ_{sm} are set to $1, 1, 0.1$ and 0.001, respectively, found empirically. We use trimesh (Dawson-Haggerty et al., 2019) library for rendering the peel maps.

5.2 Other Datasets

In addition to our 3DHumans dataset (section 4), we perform both qualitative and quantitative evaluations on the following publicly available datasets.

CLOTH3D (Bertiche et al, 2020) is a collection

[2] http://cvit.iiit.ac.in/research/projects/cvit-projects/sharp-3dhumans-a-rich-3d-dataset-of-scanned-humans

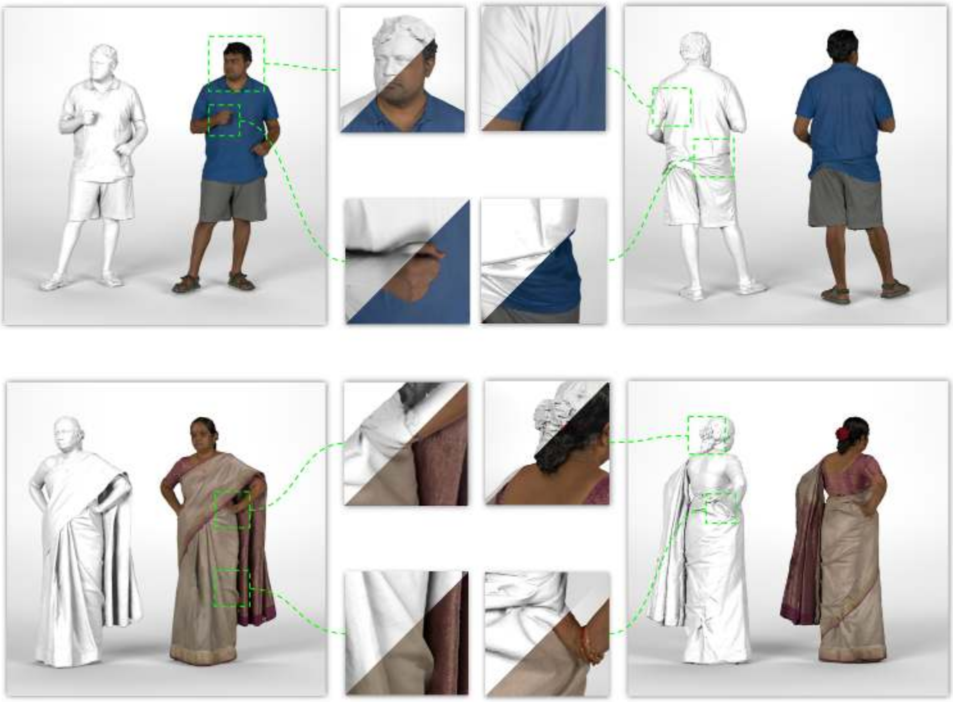

Fig. 6: High-frequency geometrical and textural details present in our 3DHumans dataset.

of 6500 synthetic sequences of SMPL meshes with garments draped onto them, simulated with MoCap data. Each frame of a sequence contains garment and corresponding SMPL body. The garment styles range from skirts to very loose robes. We augment this data by capturing SMPL texture maps with minimal clothing to simulate realistic body textures using (Alldieck et al, 2019a). For each sequence, five frames are randomly sampled. Please refer to the supplementary draft **(section 1)** for understanding the data preparation step and results of SHARP on CLOTH3D.

THUman1.0 (Zheng et al, 2019) consists of 6800 human meshes registered with SMPL body in varying poses and garments. The dataset was obtained using consumer RGBD sensors. Although the dataset has diverse poses and shapes, it has relatively tight clothing examples with low-quality textures. Please refer supplementary for results on this dataset. Note that the dataset is originally called the **THUman** dataset, we refer it to as **THUman1.0** to avoid the confusion.

THUman2.0 (Yu et al, 2021) is a collection of 500 high quality 3D scans captured using dense DSLR rig. The ataset offers wide variety of poses. However, very loose clothing styles like robed skirts are still lacking. Each mesh in the provided dataset is in different scale. We have brought all the meshes in the same scale by registering SMPL to the scans and performed our experiments.

5.3 Evaluation Metrics

To quantitatively evaluate performance of SHARP, we use the following evaluation metrics:

Point-to-Surface (P2S) Distance: Given a set of points and a surface, P2S measures the average L2 distance between each point and the nearest point to it on the given surface. We use P2S to measure the deviation of the point cloud (back-projected from predicted fused peel maps) from the ground truth mesh.

Chamfer Distance (CD): Given two sets of points S_1 and S_2, Chamfer distance measures the discrepancy between them as follows:

	Our Dataset			THUman2.0 Dataset		
Method	CD ($\times 10^{-5}$) ↓	P2S ↓	Normal ↓	CD ($\times 10^{-5}$) ↓	P2S ↓	Normal ↓
PIFu	20.79	0.00826	0.054	23.72	0.0091	0.036
Geo-PIFu	15.73	0.0092	0.058	17.01	0.0092	0.041
PaMIR	12.54	0.00714	0.054	6.05	**0.0049**	0.038
PeeledHuman	20.88	0.0094	0.061	23.34	0.0094	0.054
Ours	**7.718**	**0.0051**	**0.045**	**6.044**	0.00529	**0.034**

Table 1: Quantitative comparison on 3DHumans and THUman2.0 datasets.

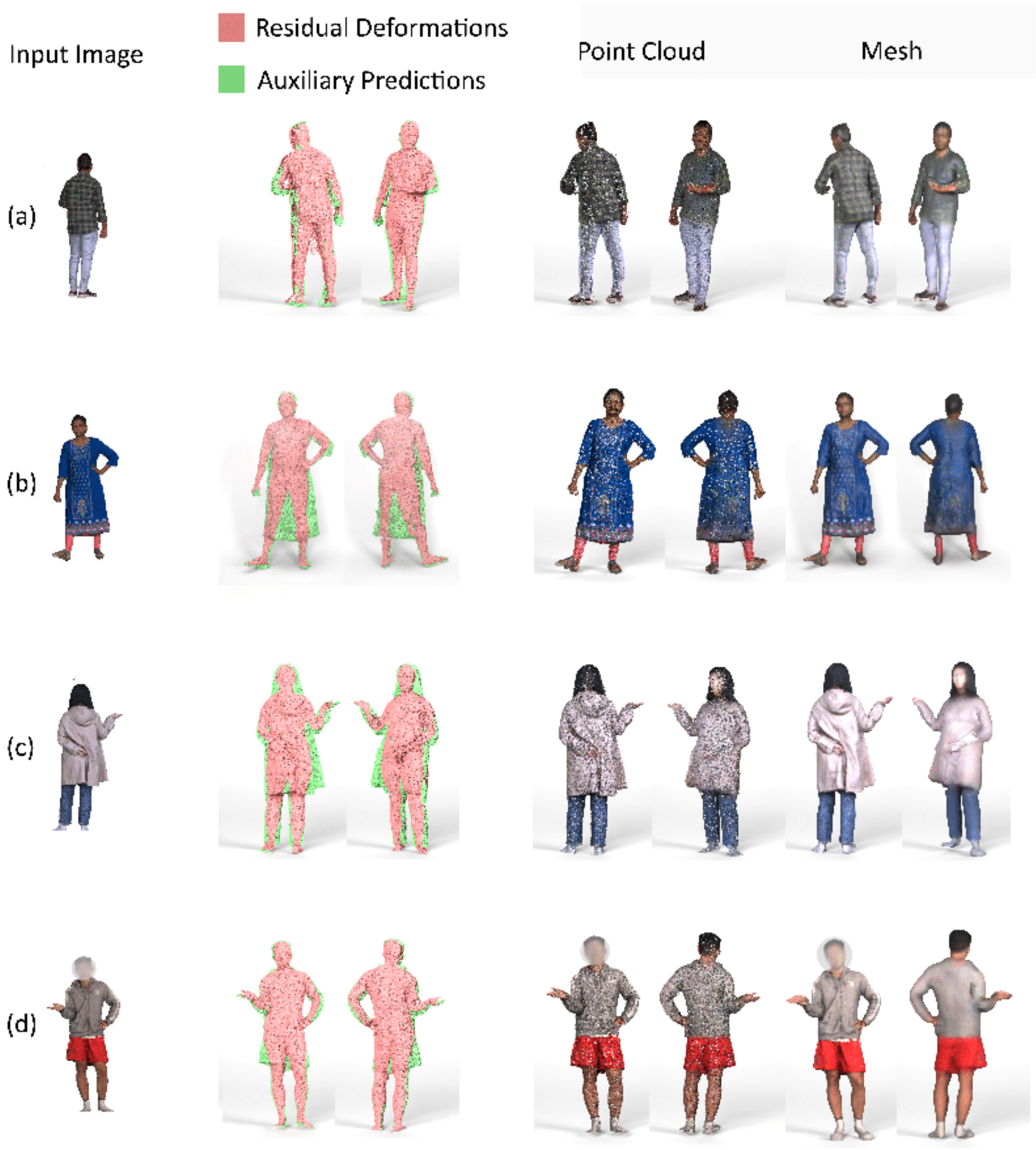

Fig. 7: Results on 3DHumans (a, b) and THUman2.0 datasets (c, d).

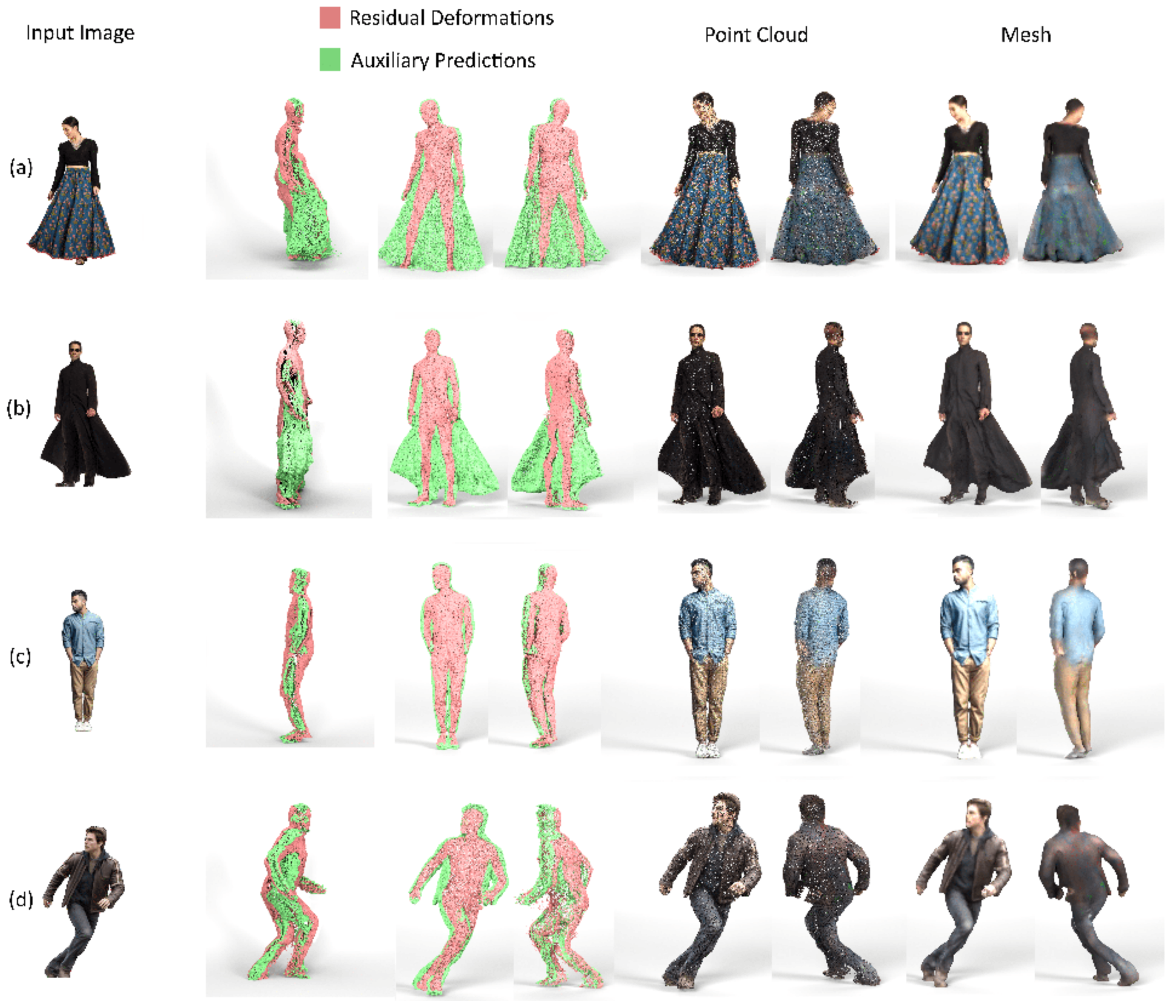

Fig. 8: Results on in-the-wild images.

Method	CD ↓	P2S ↓
JumpSuit	0.00031	0.00872
Dress	0.0012	0.021
Top+Trousers	0.00057	0.0118

Table 2: Performance of our method on clothing styles of CLOTH3D dataset.

$$\begin{aligned} d_{CD}(S_1,S_2) = & \sum_{x\epsilon S_1} min_{y\epsilon S_2}\|x-y\|_2^2 \\ & + \sum_{y\epsilon S_2} min_{x\epsilon S_1}\|x-y\|_2^2 \end{aligned} \tag{13}$$

Normal Re-projection Loss: To evaluate the fineness of reconstructed quality, we compute normal reprojection loss introduced in (Saito et al, 2019). We render the predicted and ground truth normal maps in the image space from the input viewpoint. We then calculate the L2 error between these two normal maps.

5.4 Quantitative Evaluation

We evaluate the aforementioned metrics on 3DHumans & THUman2.0 datasets and comapred the results with PIFu (Saito et al, 2019), PaMIR (Zheng et al, 2021), Geo-PIFu (He et al, 2020) and PeeledHuman (Jinka et al, 2020). We trained

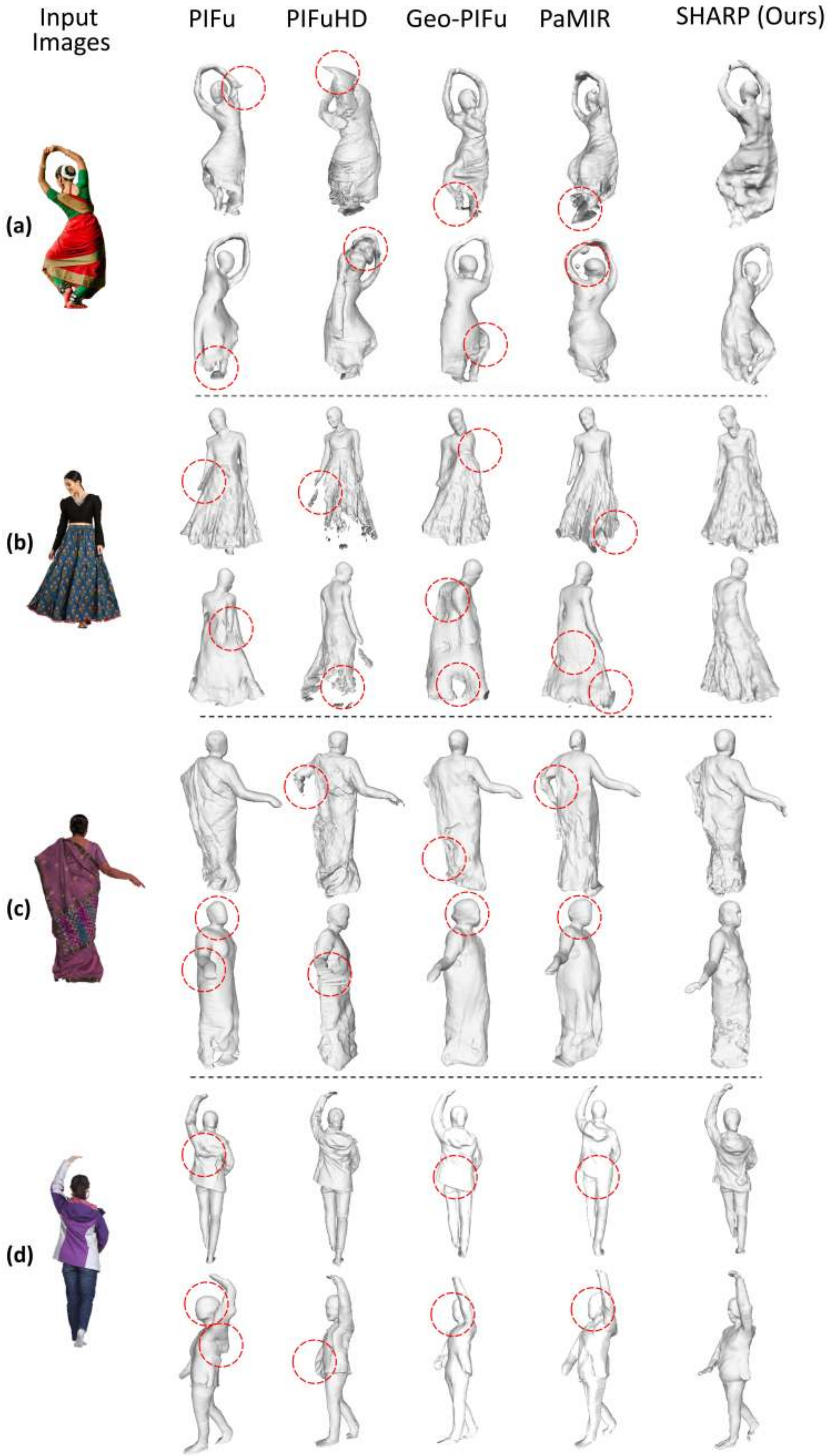

Fig. 9: Qualitative comparison of SOTA methods. (a) and (b) are in-the-wild images, (c) and (d) are from 3DHumans and THUman2.0 datasets respectively, shown in two different views.

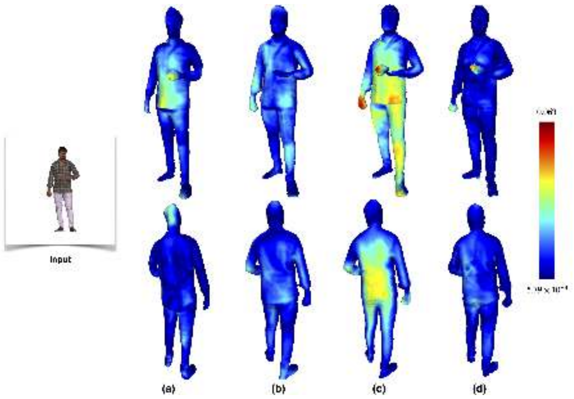

Fig. 10: **P2S Plot:** Point-to-surface plots on the reconstructed outputs from (a) PaMIR, (b) Geo-PIFu, (c) PIFu and (d) SHARP.

all the models from scratch on these datasets under the same train/test split. We transform all the predicted models from different methods to the canonical coordinates of the ground truth mesh and report metrices in Table 1. The quantitative comparison concludes that our method outperforms the SOTA methods.

Unlike peeledhuman (Jinka et al, 2020) that uses a generative network, we use a simple encoder-decoder architecture. We trained PaMIR with approximately thrice the amount of data (SHARP is trained on 70 views per mesh, while for PaMIR, 200 views per mesh are used). Geo-PIFu needs to be trained for coarse and query networks separately and complete training takes three days to train on 3DHumans in our setup.

Additionally, Table 2 summarizes quantitative analysis on the CLOTH3D dataset where we evaluate CD and P2S metrics on different styles of clothing to indicate the generalization of our method across various clothing styles. We also provide comparisons with THUman1.0 and CLOTH3D datasets in the supplementary.

5.5 Qualitative Evaluation

We show the reconstructions obtained by our method using THUman2.0 and 3DHumans datasets (Figure 7), where we also show point clouds obtained by back-projecting residual and auxiliary peel maps. Figure 7 (a) and (b) are samples from our dataset, (c) and (d) are from THUman2.0 dataset. Please refer supplementary for additional results on CLOTH3D dataset. One can observe that our model can handle various styles of clothing (including *tunic*) covering the lower body parts and with a wide variety of poses. Residual and auxiliary peel maps captured the complimentary surface details as visualized in red and green in Figure 7. In order to test the generalizability of our method on unseen in-the-wild images, we show results on random internet images using our method in Figure 8. Similar to PIFu (Saito et al, 2019), we use an off-the-shelf method to remove the background from these images before passing them to our network. It can be noted that our method is able to reconstruct the human body with self-occlusions and tackle a wide variety of clothing styles, ranging from tight to loose clothing with diverse poses. Notably, our method is able to generalize well on unseen, very loose clothing styles present in Figure 8 (a) & (b).

In Figure 9, we show qualitative comparison with SOTA methods. PIFu and PIFuHD do not use body prior, which leads to missing and distorted body parts. Geo-PIFu predicts a volumetric prior before performing implicit reconstruction. On the other hand, PaMIR uses SMPL prior as input. Hence, both methods tends to produce smoother geometry as they use voxelized representation, which is known to smooth out the geometrical details. It can be noted that our method retains high-frequency surface details as shown in Figure 9. Additionally, we also show comparison with our previous work peeledhuman (Jinka et al, 2020) in Figure 11. We observed that our formulation yielded superior results over peeledhuman which also uses the PeeledHuman representation sans SMPL prior.

All the aforementioned methods have been trained on our 3DHumans dataset except for PIFuHD. Since the training code for PIFuHD is not yet available, we use the model provided by the authors. In order to fairly compare with other methods, we selected a body with tight clothing and generated plots of P2S error of all methods trained on our dataset as visualized in Figure 10. One can infer from these plots that our approach yields superior performance in terms of distribution of P2S error over the reconstructed surface.

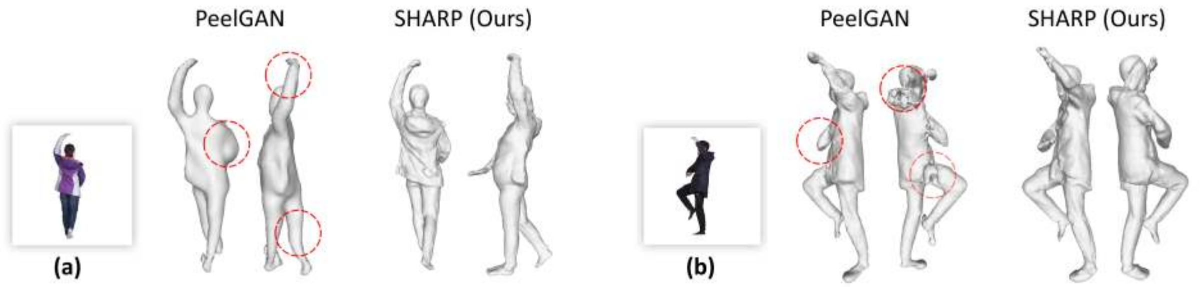

Fig. 11: Qualitative comparison of peeledhuman and SHARP.

Method	No. of parameters	Execution Time
PaMIR(Geo+Tex)	40M(27M+13M)	4.03s(3.9s+0.13s)
Geo-PIFu(coarse+fine)	30.6M (14.9M+15.7M)	16.32s(0.32s+16s)
Ours	**22M**	**0.09s**

Table 3: Comparison of complexity analysis.

Method	CD ↓	P2S ↓
Ours w.o. L_{sm}	8.3652	0.0053
Ours w.o. fusion	9.98	0.0054
Ours	**7.718**	**0.0051**

Table 4: **Ablation Study:** Effect of loss functions.

Network	CD ↓	P2S ↓
U-Net	8.417	0.0052
Hourglass	15.6	0.0068
ResNet(ours)	**7.71**	**0.0051**

Table 5: **Ablation Study:** Effect of different architectures.

5.6 Network Complexity

We report a detailed analysis of the execution time of SOTA methods in Table 3. All the numbers are computed on a single NVIDIA GTX 1080Ti GPU with a single input image. PaMIR needs feed-forward of two networks to infer shape and geometry. On the other hand, Geo-PIFu needs to infer coarse volumetric shape followed by fine shape. We calculate the feed-forward execution time for the complete forward pass of Geo-PIFu and PaMIR as these methods need multiple forward passes while inferring. Note that ours is an end-to-end inference model which predicts both shape and color in a single forward pass efficiently with 0.09 seconds, which is significantly faster when compared to the aforementioned methods. Additionally, our network is lightweight, consisting of 22 million parameters, while PaMIR and Geo-PIFu has 40 and 30.6 million parameters, respectively.

Blocks	parameters	CD ↓	P2S ↓
6	8.26M	22.81	0.0073
9	12.17M	8.9	0.0053
18	22M	**7.71**	**0.0051**

Table 6: **Ablation Study:** Effect of ResNet blocks.

6 Discussion

In this section, We perform ablative studies on various components of the network. We run all experiments on our proposed 3DHumans dataset. We also discuss in detail about the post-processing steps, along with the limitations and failure cases of our method.

Network	CD ↓	P2S ↓
Addition	8.24	0.0052
Average	8.82	0.0058
Concat	**7.57**	**0.0049**
Ours*	7.71	0.0051

Table 7: **Ablation Study:** Comparison of various Fusion Strategies. Ours is only end-to-end trainable mechanism as opposed to Addition, Average and Concat fusion.

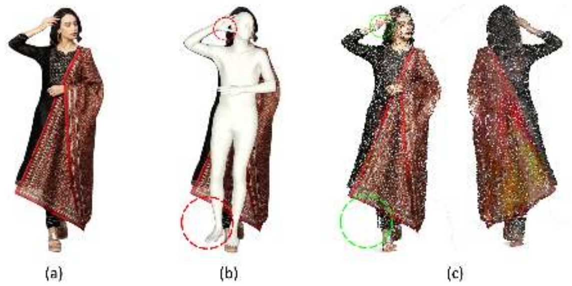

Fig. 12: **Handling noisy shape prior:** (a) Input image, (b) SMPL prior misaligned with the input image, (c) Point cloud output from SHARP.

6.1 Ablation Study: Architectural Choices

Impact of loss functions: In Table 4, we demonstrate the impact of various loss functions on the output point cloud. First, we evaluate SHARP without smoothness loss (L_{sm}) and observe that it leads to an increase in Chamfer distance and P2S error, which is caused by the noise in prediction of fused peel maps.

Secondly, to evaluate the importance of the peel map fusion, we train our network without fusion. In this setting, we used only two decoder branches, one for predicting RGB peel maps and the other for predicting depth peel maps. This lead to smooth, predictions, which misses out body-specific geometrical details, further increasing CD and P2S values.

Impact of various backbone networks: We evaluate the performance of SHARP on various backbone network architectures. In particular, we used U-Net (Ronneberger et al, 2015) and stacked hourglass network (Newell et al, 2016) as backbone networks along with residual networks. All the backbone networks are trained with same loss functions as described in subsection 3.4. We report the performance of these networks in Table 5. Residual network outperforms both Unet and hourglass networks in this multi-branch prediction task. We also observed that hourglass network was not able to predict four layer RGB peel maps.

Impact of number of ResNet blocks: We also evaluate the performance of SHARP by varying the number of ResNet blocks as shown in Table 6. We train our network on 3DHumans with 6, 9 and 18 blocks. Using only 6 ResNet blocks, which is almost one-third of the original network, SHARP is able to achieve similar performance as PIFu (please refer Table 1). Using 9 ResNet blocks, we are able to achieve closer numbers to majority of existing SOTA methods. We observed that the further increase in the number of ResNet blocks did not yield any significant improvement.

Fusion Strategies We analyse the performance of SHARP with various fusion strategies of peel maps. In this experiment, we perform feature level fusion instead of auxiliary and residual depth peel map fusion. We train this fusion network in a coarse-to-fine strategy where initially we replace the auxiliary peel map branch with predicting complete depth peel maps $\widehat{D}_{peel}$. We train this network with losses L_{rd}, L_{sm} and L_1 loss on predicted and ground truth peel maps. We then, take this network as initialization to train fusion module where we take intermediate features of $\widehat{D}_{peel}$ and $\widehat{D}_{rd}$ branches respectively. Refer supplementary (Figure 4) for the architecture diagram. We fuse them using three strategies (a) addition, (b) average and (c) concatenation. These fused features are then passed to upsampling and convolutional layers to predict final fused depth peel maps. Here, we freeze the weights of the network except the layers after the feature fusion. We call it as *Late Fusion* as it requires pre-trained network.

We report the performance in Table 7 and learn that late fusion with concatenation results in better performance. However, we note the training for late fusion is not end-to-end as described and we adopt end-to-end trainable network with fusion proposed in Equation 6 as our final choice.

6.2 Handling Noisy Shape Prior

The shape prior based reconstruction methods are susceptible to noisy initialization from incorrect prior. Generally, this leads to incorrect pose conditioning, which further deteriorates the final reconstruction. Our method can partially handle such noisy prior as we use refined per-layer SMPL mask Γ_i (introduced in subsubsection 3.3.1, Equation 3) to mask out the regions of the SMPL prior peel maps which fall outside the clothed human silhouette in input image. Thus, residual deformation $\widehat{\mathcal{D}}_{rd}$ predicted for the misaligned regions of the SMPL prior is not considered during fusion, and we are able to avoid the errors in reconstruction due to such misalignments. Figure 12 shows a case of noisy SMPL prior for an in-the-wild image and the final reconstruction output of SHARP, where it is able to recover from incorrect prior in the leg and hand region.

6.3 Peeled Representation Layers

In this work, we used four layers of peeled representation for human body recovery. However, our formulation is generalizable to arbitrary number of layers. In Figure 13, we show the performance of SHARP on real images (not included in training data distribution) where six layers are needed to capture the geometry. Note that SHARP is able to recover geometry under the case of severe self-occlusion and with skewed viewpoints. To train this network, we initially train network with four layers and then using this model to initialize weights for model with six layers.

6.4 Post-processing

The output of our network is prone to slight noise in the predicted peel maps, resulting in sparse outliers in the back-projected point cloud, as shown in Figure 15 (a). These outliers are removed by density-based filtering, where we fit spheres with 16 neighbours on each point. The points, which are inside the spheres having a radius greater than the threshold (0.01 in our case), are removed to obtain a clean point cloud, as shown in Figure 15 (b). Finally, the filtered point cloud might have some small holes which are subsequently filled by meshification using Poisson Surface Reconstruction (PSR) Figure 15 (c).

6.5 Limitations

Ambiguity due to textural edges: 3D reconstruction from a monocular RGB image, being an ill-posed problem, is susceptible to interpreting the textural edges as geometrical details. In Figure 14, we show reconstructions from our method and PaMIR, where both the methods incorrectly interpret textural details of a flat clothing surface as geometrical details and hallucinate geometrical structures, which are non-existent.

Failure cases: One of the key challenges faced by majority of existing prior-based methods is self-intersection of body parts in the prior, mainly due to challenging poses. In Figure 16, a failure case of our approach is shown where the network reconstructs the occluded regions accurately, but fails to recover from interpenetrating body parts, present in the input SMPL prior (hands penetrating the legs).

7 Conclusion

Reconstructing 3D clothed human body from a monocular RGB image is an extremely ill-posed problem due to skewed viewpoints, depth ambiguities, complex poses and arbitrary clothing styles. Although many solutions exist which can recover clothed human body in relatively tighter clothing, they fail to generalize when it comes to in-the-wild loose clothing scenarios. To this end, we have contributed a novel end-to-end trainable deep learning framework, SHARP, which uses a sparse and efficient fusion of parametric body prior with non-parametric PeeledHuman representation, and is able to reconstruct human body in arbitrarily loose clothing.

In more general perspective, we built on our sparse non-parametric 2D shape representation and proposed an efficient strategy to fuse it with parametric shape prior. We train a compact, encode-decoder based network using a set of L1 losses on 2D maps, while reconstructing complex 3D geometry. The proposed formulation is

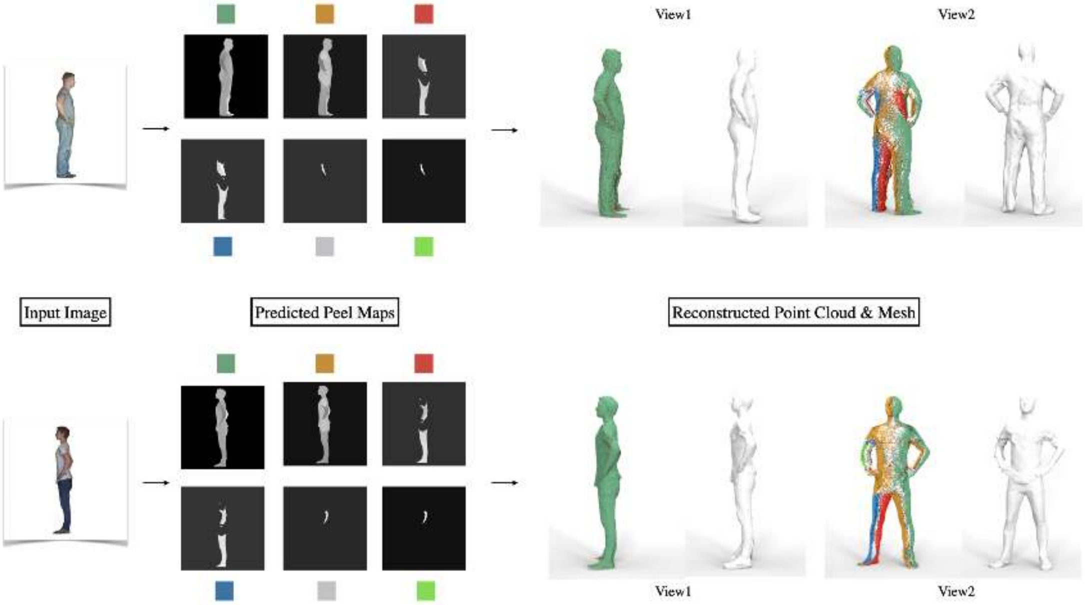

Fig. 13: Performance of our method on six peel layer representation. We show the predicted final fused depth peel maps (with corresponding color coding) along with backprojected point cloud (points from a layer is color coded with the same color as indicated in depth peel maps) and reconstructed mesh respectively.

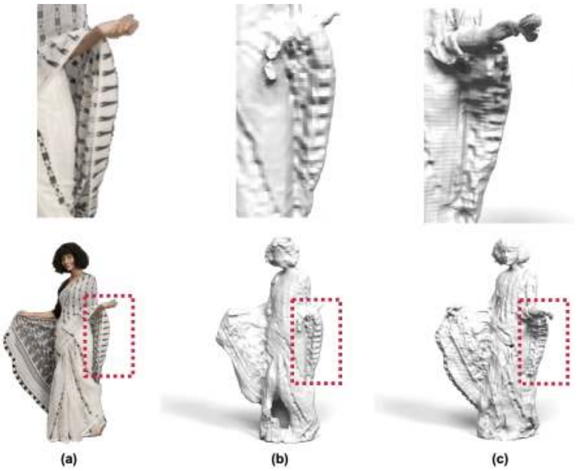

Fig. 14: **Texture-Geometry Ambiguity:** High-frequency textural details can be interpreted as geometrical details by monocular deep reconstruction techniques. (a) Input image, (b) PaMIR and (c) SHARP.

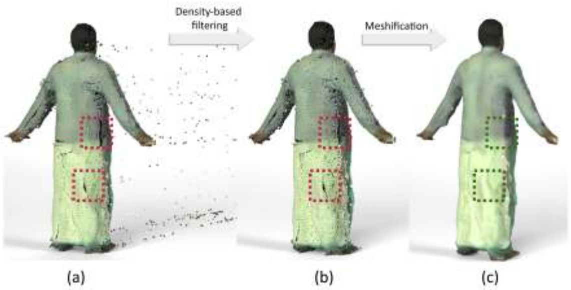

Fig. 15: Effect of post-processing.

sparse in terms of representation, resulting in low inference time of the network.

We evaluated our framework on various publicly available datasets and reported superior qualitative and quantitative performance as compared to state-of-the-art methods. Since, data is a key bottleneck in the field of deep learning based 3D human body reconstruction, we contributed 3DHumans dataset and intend to release it in the public domain to further accelerate the research. Our dataset contains 3D human body scans of high-frequency textural and geometrical details, with a wide variety of the body shapes in various clothing styles.

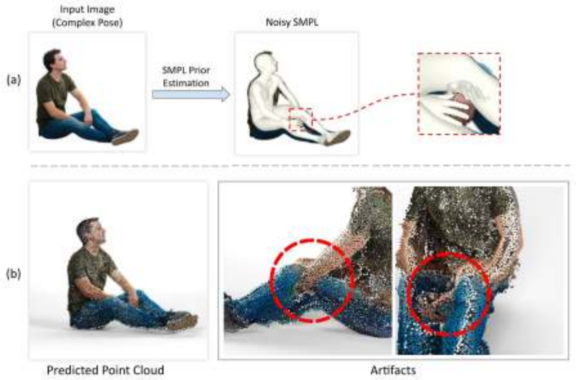

Fig. 16: **Failure Case** : **(a)** Noisy SMPL estimation (hands are intersecting with the legs) due to highly complex pose. **(b)** Artifacts in the predicted point cloud.

Although per-frame reconstruction of SHARP yields reasonable intra-frame consistency without any explicit temporal conditioning (as shown in the supplementary video), it will be interesting to explore extension of our method to learn over video sequences where it is difficult to get high quality ground-truth data. Another interesting direction is to incorporate learning from multi-view images for better reconstruction results. Additionally, performance of our method can be further improved by addressing the texture-geometry ambiguity and recovering from challenging scenarios such as self-intersecting body parts.

References

Alldieck T, Magnor M, Bhatnagar BL, et al (2019a) Learning to reconstruct people in clothing from a single RGB camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2019.00127

Alldieck T, Pons-Moll G, Theobalt C, et al (2019b) Tex2Shape: Detailed full human body geometry from a single image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2019.00238

Anguelov D, Srinivasan P, Koller D, et al (2005) SCAPE: Shape completion and animation of people. ACM Transactions on Graphics (TOG) 24(3):408–416. https://doi.org/10.1145/1186822.1073207

Azevedo TC, Tavares JMR, Vaz MA (2009) 3D Object Reconstruction from Uncalibrated Images Using an Off-the-Shelf Camera, pp 117–136. https://doi.org/10.1007/978-1-4020-9086-8_7

Baak A, Müller M, Bharaj G, et al (2011) A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2011.6126356

Bertiche H, Madadi M, Escalera S (2020) CLOTH3D: Clothed 3d humans. In: Proceedings of the European Conference on Computer Vision (ECCV), https://doi.org/10.1007/978-3-030-58565-5_21

Bhatnagar BL, Tiwari G, Theobalt C, et al (2019) Multi-Garment Net: Learning to dress 3D people from images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2019.00552

Bhatnagar BL, Sminchisescu C, Theobalt C, et al (2020) LoopReg: Self-supervised learning of implicit surface correspondences, pose and shape for 3D human mesh registration.

In: Advances in Neural Information Processing Systems (NeurIPS)

Bhatnagar, B. L., Sminchisescu, C., Theobalt, C., Pons-Moll, G. . Combining implicit function learning and parametric models for 3d human reconstruction. In European Conference on Computer Vision (pp. 311-329). Springer, Cham.

Bogo F, Romero J, Loper M, et al (2014) Faust: Dataset and evaluation for 3d mesh registration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 3794–3801, https://doi.org/10.1109/CVPR.2014.491

Bogo F, Kanazawa A, Lassner C, et al (2016) Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV), https://doi.org/10.1007/978-3-319-46454-1_34

Bogo F, Romero J, Pons-Moll G, et al (2017) Dynamic FAUST: Registering human bodies in motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2017.591

Bronstein AM, Bronstein MM, Kimmel R (2008) Numerical geometry of non-rigid shapes. Springer Science & Business Media, https://doi.org/10.1007/978-0-387-73301-2

Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., and Moreno-Noguer, F. (2021). SMPLicit: Topology-aware generative model for clothed people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Dawson-Haggerty et al. (2019) Trimesh library. URL https://trimsh.org/

Dou M, Khamis S, Degtyarev Y, et al (2016) Fusion4D: Real-time performance capture of challenging scenes. ACM Transactions on Graphics (TOG) 35(4):1–13. https://doi.org/10.1145/2897824.2925969

Gabeur V, Franco JS, Martin X, et al (2019) Moulding humans: Non-parametric 3d human shape estimation from single images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2019.00232

Gall J, Stoll C, De Aguiar E, et al (2009) Motion capture using joint skeleton tracking and surface estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp 1746–1753, https://doi.org/10.1109/CVPR.2009.5206755

Gong K, Liang X, Li Y, et al (2018) Instance-level human parsing via part grouping network. https://doi.org/10.1007/978-3-030-01225-0_47, 1808.00157

Güler RA, Neverova N, Kokkinos I (2018) DensePose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2018.00762

Habermann M, Xu W, Zollhofer M, et al (2020) DeepCap: Monocular human performance capture using weak supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR42600.2020.00510

He T, Collomosse J, Jin H, et al (2020) GeoPIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS)

He T, Xu Y, Saito S, Soatto S, and Tung, T. "ARCH++: Animation-ready clothed human reconstruction revisited". In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021.

Huang Z, Xu Y, Lassner C, et al (2020) ARCH: Animatable reconstruction of clothed humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/cvpr42600.2020.00316

Jinka SS, Chacko R, Sharma A, et al (2020) PeeledHuman: Robust shape representation for textured 3D human body reconstruction. In: Proceedings of the IEEE Conference on 3D Vision (3DV), https://doi.org/10.1109/3DV50981.2020.00098

Kanazawa A, Black MJ, Jacobs DW, et al (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2018.00744

Kanazawa A, Zhang JY, Felsen P, et al (2019) Learning 3D human dynamics from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2019.00576

Kazhdan M, Bolitho M, Hoppe H (2006) Poisson surface reconstruction. In: Proceedings of Eurographics symposium on Geometry processing (SGP), https://doi.org/10.2312/SGP/SGP06/061-070

Kolotouros N, Pavlakos G, Black MJ, et al (2019a) Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2019.00234

Kolotouros N, Pavlakos G, Daniilidis K (2019b) Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2019.00463

Kolotouros N, Pavlakos G, Jayaraman D, et al (2021) Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 11,605–11,614

Lahner Z, Cremers D, Tung T (2018) Deepwrinkles: Accurate and realistic clothing modeling. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 667–684, https://doi.org/10.1007/978-3-030-01225-0_41

Li B, Godil A, Aono M, et al (2012) Shrec'12 track: Generic 3d shape retrieval. In: Proceedings of Eurographics Workshop on 3D Object Retrieval (3DOR), pp 119–126, https://doi.org/10.2312/3DOR/3DOR12/119-126

Lin K, Wang L, Liu Z (2021) Mesh graphormer. arXiv preprint arXiv:210400272

Liang, Junbang, and Ming C. Lin. ”Shape-aware human pose and shape reconstruction using multi-view images.” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

Loper M, Mahmood N, Romero J, et al (2015) SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (Proc SIGGRAPH Asia) 34(6):248:1–248:16. https://doi.org/10.1145/2816795.2818013

Ma Q, Saito S, Yang J, et al (2021a) Scale: Modeling clothed humans with a surface codec of articulated local elements. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)

Ma Q, Yang J, Tang S, et al (2021b) The power of points for modeling humans in clothing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)

Mildenhall B, Srinivasan PP, Tancik M, et al (2020) Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV), https://doi.org/10.1007/978-3-030-58452-8_24

Mulayim AY, Yilmaz U, Atalay V (2003) Silhouette-based 3-D model reconstruction from multiple images. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 33(4):582–591. https://doi.org/10.1109/TSMCB.2003.814303

Natsume R, Saito S, Huang Z, et al (2019) Siclope: Silhouette-based clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2019.00461

Newcombe RA, Fox D, Seitz SM (2015) DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2015.7298631

Omran M, Lassner C, Pons-Moll G, et al (2018) Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In: Proceedings of the IEEE Conference on 3D Vision (3DV), https://doi.org/10.1109/3DV.2018.00062

Patel C, Liao Z, Pons-Moll G (2020) TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR42600.2020.00739

Pavlakos G, Choutas V, Ghorbani N, et al (2019) 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), https://doi.org/10.1109/CVPR.2019.01123

Saito S, Huang Z, Natsume R, et al (2019) PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2019.00239

Saito S, Simon T, Saragih J, et al (2020) PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/cvpr42600.2020.00016

Shotton J, Fitzgibbon A, Cook M, et al (2011) Real-time human pose recognition in parts from single depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1145/2398356.2398381

Smith D, Loper M, Hu X, et al (2019) Facsimile: Fast and accurate scans from an image in less than a second. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5330–5339, https://doi.org/10.1109/ICCV.2019.00543

Tiwari G, Bhatnagar BL, Tung T, et al (2020) Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 1–18, https://doi.org/10.1007/978-3-030-58580-8_1

Tucker R, Snavely N (2020) Single-view view synthesis with multiplane images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/cvpr42600.2020.00063

Varol G, Romero J, Martin X, et al (2017) Learning from synthetic humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2017.492

Varol G, Ceylan D, Russell B, et al (2018) BodyNet: Volumetric inference of 3D human body shapes. In: Proceedings of the European Conference on Computer Vision (ECCV), https://doi.org/10.1007/978-3-030-01234-2_2

Venkat A, Jinka SS, Sharma A (2018) Deep textured 3D reconstruction of human bodies. In: Proceedings of British Machine Vision Conference (BMVC)

Venkat A, Patel C, Agrawal Y, et al (2019) HumanMeshNet: Polygonal mesh recovery of humans. In: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), https://doi.org/10.1109/ICCVW.2019.00273

Wei X, Zhang P, Chai J (2012) Accurate real-time full-body motion capture using a single depth camera. ACM Transactions on Graphics (TOG) 31(6):1–12. https://doi.org/10.1145/2366145.2366207

Yu T, Zheng Z, Guo K, et al (2021) Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

(CVPR), pp 5746–5756, https://doi.org/10.1109/CVPR46437.2021.00569

Zhang C, Pujades S, Black MJ, et al (2017) Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 4191–4200, https://doi.org/10.1109/CVPR.2017.582

Zheng Z, Yu T, Wei Y, et al (2019) DeepHuman: 3D human reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), https://doi.org/10.1109/ICCV.2019.00783

Zheng Z, Yu T, Liu Y, et al (2021) Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence https://doi.org/10.1109/TPAMI.2021.3050505

Zhu H, Zuo X, Wang S, et al (2019) Detailed human shape estimation from a single image by hierarchical mesh deformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4491–4500, https://doi.org/10.1109/CVPR.2019.00462