PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction

Zerong Zheng, Tao Yu, Yebin Liu, *Member, IEEE* and Qionghai Dai, *Senior Member, IEEE*

Abstract—Modeling 3D humans accurately and robustly from a single image is very challenging, and the key for such an ill-posed problem is the 3D representation of the human models. To overcome the limitations of regular 3D representations, we propose **Pa**rametric **M**odel-Conditioned **I**mplicit **R**epresentation (PaMIR), which combines the parametric body model with the free-form deep implicit function. In our PaMIR-based reconstruction framework, a novel deep neural network is proposed to regularize the free-form deep implicit function using the semantic features of the parametric model, which improves the generalization ability under the scenarios of challenging poses and various clothing topologies. Moreover, a novel depth-ambiguity-aware training loss is further integrated to resolve depth ambiguities and enable successful surface detail reconstruction with imperfect body reference. Finally, we propose a body reference optimization method to improve the parametric model estimation accuracy and to enhance the consistency between the parametric model and the implicit function. With the PaMIR representation, our framework can be easily extended to multi-image input scenarios without the need of multi-camera calibration and pose synchronization. Experimental results demonstrate that our method achieves state-of-the-art performance for image-based 3D human reconstruction in the cases of challenging poses and clothing types.

Index Terms—Body Pose, Human Reconstruction, Surface Representation, Parametric Body Model, Implicit Surface Function

✦

1 INTRODUCTION

IMAGE-BASED parsing of human bodies is a popular topic in computer vision and computer graphics. Among all the tasks of image-based human parsing, recovering 3D humans from a single RGB image attracts more and more interests given its wide applications in VR/AR content creation, image and video editing, telepresence and virtual dressing. However, as an ill-posed problem, recovering 3D humans from a single RGB image is very challenging due to the lack of depth information and the variations of body shapes, poses, clothing types and lighting conditions.

Benefiting from the huge progress of deep learning techniques, recent studies have tried to address these challenges using learning-based methods [1], [2], [3], [4], [5], [6], [7], [8]. According to their 3D representations, these methods can be roughly classified into two categories: parametric methods and non-parametric methods. Parametric methods like HMD [3] and Tex2Shape [4] utilize a statistical body template model (e.g., SMPL [9]) as the geometrical prior, and learn to deform the template according to the silhouette and the shading information. However, the low dimensional parametric model limits the performance when handling different clothing types such as long dresses and skirts. Non-parametric methods use various free-form representations, including voxel grids [1], silhouettes [2], implicit fields [8] or depth maps [6], [7], to represent 3D human models. Although being able to describe arbitrary clothes, current
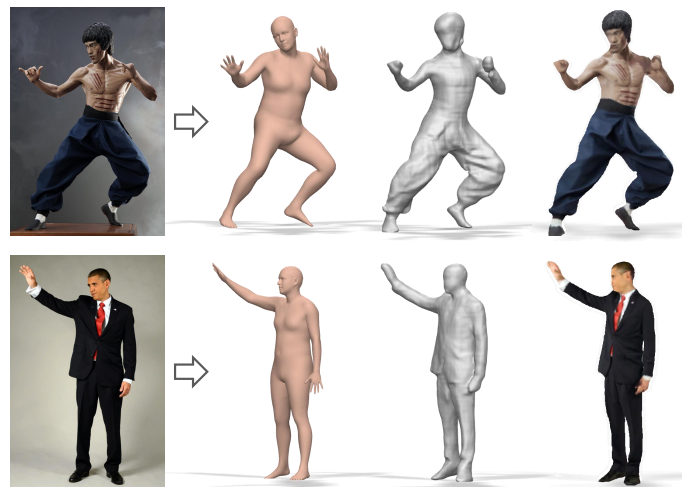

Fig. 1. Example results reconstructed by our method. Our method is able to infer the underlying body shape, the 3D surface geometry and its texture given a single RGB image.

non-parametric methods still suffer from challenging poses and self-occlusions due to the lack of semantic information and depth ambiguities.

Recently, DeepHuman [5] pioneered in conditioning the non-parametric volumetric representation on the parametric SMPL model, and demonstrates robust reconstruction for challenging poses and diverse garments. However, it struggles to recover high-quality geometric details due to the resolution limitation of the regular occupancy volume.

• *Zerong Zheng, Tao Yu, Yebin Liu and Qionghai Dai are with Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, P. R. China.*
• *Corresponding authors: Yebin Liu and Qionghai Dai.*

More importantly, conditioning the non-parametric representation on SMPL models raises a crucial problem: how can we obtain accurate SMPL model estimation given a single image at testing time? In DeepHuman, the SMPL models at testing time are estimated using learning-based methods and/or optimization-based methods. Although the recent years have witness a renaissance in the context of single-image body shape estimation [10], [11], [12], [13], [14], [15], we observe that there is still an *accuracy gap* between the training and testing stages: during training, we can obtain well-aligned SMPL models for the training data by fitting the SMPL template to the high quality scans, while for testing, the SMPL estimations provided by the state-of-the-art single-image methods [12], [13] still cannot align with the keypoints or the silhouettes as perfectly as the SMPL models in the training dataset.

As a result, the neural network trained with the ground-truth image-SMPL pairs cannot generalize well to the testing images which have no ground-truth SMPL annotations. A simple solution is to replace the ground-truth SMPL models with the predicted ones while still using the ground-truth surface scans for training supervision, preventing the network from heavy dependence on accurate SMPL annotations. However, due to depth ambiguities, it is impossible to guarantee well alignments between the predicted SMPL models and the ground-truth scans along the depth axis. Consequently, by doing so we are forcing the networks to make "correct" geometry inference given a "wrong" SMPL reference along the depth axis, which will generate even worse results.

To achieve high-quality geometric detail reconstruction while maintaining robustness to challenging pose and clothing styles, in this paper, we propose *Parametric Model-Conditioned Implicit Representation*, dubbed PaMIR, to incorporate the parametric SMPL model and the free-form implicit surface function into a unified learning and optimization framework. The implicit surface function overcomes the resolution limit of volumetric representation, and enables detailed surface reconstruction capability. To fill the accuracy gap between the ground truth training data and the inaccurate testing input, we further propose a new depth-ambiguity-aware training loss and a body reference optimization module in the PaMIR-based framework, and achieve surface detail reconstruction even under imperfect body reference initialization. Specifically, the three technical contributions are summarized as follow:

- **PaMIR.** Our PaMIR representation has the ability to condition the implicit field on the SMPL prediction, which is realized by a novel network architecture converting an image feature map and the corresponding SMPL feature volume into an implicit surface representation. The SMPL feature volume is directly encoded from the SMPL prediction provided by GraphCMR [13], and serves as a soft constraint for handling challenging poses and/or self-occlusions.
- **Training Losses.** We use the predicted SMPL models at both the training and the testing stages. In order to make the network to be more robust against the inaccurate body reference along the depth direction, instead of using a traditional reconstruction loss, we propose a *depth-ambiguity-aware reconstruction loss* that

adaptively adjusts the reconstruction supervision based on the current SMPL estimation while respecting 2D image observations.

- **Body Reference Optimization.** We also propose a *body reference optimization* method for testing process to further refine the body reference. Our optimization method directly utilizes the network itself as well as its outputs to construct an efficient energy function for body fitting. The optimization further alleviates the error caused by the inaccurate initial SMPL estimation, and consequently closing the accuracy gap of SMPL annotation between training and testing.

Overall, in our framework, the underlying body template and the outer surface are mutually beneficial: on one hand, benefiting from the proposed depth-ambiguity-aware reconstruction loss, even the imperfect body template can be used to provide strong semantic information for implicit surface reconstruction; on the other hand, the deep implicit function of the outer surface is also used to optimize the underlying body template by minimizing the body fitting error directly.

Furthermore, thanks to the usage of the common underlying SMPL model as body reference, our PaMIR representation implicitly builds a correspondence relationship across different models. As a result, our method can be easily extended to multi-image setups like multi-view inputs or video inputs without the requirement for calibration and synchronization. Experiments show that our method is able to reconstruct high-quality human models with various body shapes, poses and clothes, and outperforms state-of-the-art methods in terms of accuracy, robustness and generalization capability.

2 RELATED WORK

Human Reconstruction from multi-view images. Previous studies focused on using multi-view images for human model reconstruction [16], [17], [18]. Shape cues like silhouette, stereo and shading have been integrated to improve the reconstruction performance [17], [18], [19], [20], [21]. State-of-the-art multi-view real-time systems include [22], [23]. Extremely high-quality reconstruction results have also been demonstrated with tens or even hundreds of cameras [24]. To capture detailed motions of multiple interacting characters, more than six hundred cameras have been used to overcome the self-occlusion challenges [25], [26].

In order to reduce the difficulty of system setup, human model reconstruction from sparse camera views has recently been investigated by using CNNs for learning silhouette cues [27] and stereo cues [28]. These systems require about 4 camera views for a coarse-level surface detail capture. Note also that although temporal deformation systems using lightweight camera setups [29], [30], [31] have been developed for dynamic human model reconstruction using skeleton tracking ([29], [32]) or human mesh template deformation [30], these systems require a pre-scanned subject-specific 3D template for deformation optimization.

Parametric Body Estimation from a Single Image. Dense 3D parsing from a single image has attracted substantial interest recently because of the emergence of human

statistical models like SCAPE [33] and SMPL [9]. For example, by fitting the SMPL model to the 2D keypoint detections [34] and other dense shape cues [35], the shape and pose parameters can be automatically obtained from a single image [10], [36]. Instead of optimizing mesh and skeleton parameters, recent approaches proposed to train deep neural networks that directly regress the 3D shape and pose parameters from a single image [11], [12], [13], [14], [15]. However, the parametric body models can only capture the shape and pose of a minimally clothed body, thus lack the ability to represent a 3D human model with more general clothing layers.

Non-parametric Human Reconstruction from a Single Image. Regarding single-image human model reconstruction without using parametric models, recent studies have adopted techniques based on silhouette estimation [2], template-based deformation ([3], [4]), depth estimation [6], [7], volumetric reconstruction [1], [5] and implicit reconstruction [8]. Although they have achieved promising results, typical limitations still exist when using a single 3D representation: silhouette-based methods like [2] is suffer from lack of details and view inconsistency, template-based deformation methods [3], [4] are unable to handle loose clothes, depth-based methods [6], [7] cannot handle self-occlusions naturally, and volumetric methods [1], [5] cannot recover high-frequency details due to volume resolution limitations. Implicit-function-based methods, like PIFu [8], achieve the state-of-the-art reconstruction, but they only demonstrate reconstruction under very simple poses, and their performance will be severely deteriorated under the scenarios of challenging poses.

Implicit Fields. How to represent 3D surface is a core problem in 3D learning. Explicit representations like point clouds [37], [38], voxel grids [39], [40], triangular meshes [41], [42] have recently been explored to replicate the success of deep learning techniques on 2D images. However, the loss of structure information in point cloud, the memory requirement of regular voxel grid, and the fixed topology of meshes make these explicit representations unsuitable to represent arbitrary high-quality 3D surfaces in neural networks. Implicit surface representations [43], [44], on the other hand, overcome these challenges and demonstrate the best flexibility and expressiveness for topology representation. It defines a surface as the level set of a function of occupancy probability or surface distance. Recent works [?], [45], [46], [47], [48], [49] have shown promising results on generative models for 3D shape inference based on implicit representations. Drawing inspiration from PIFu [8], we also use deep implicit fields in our PaMIR representation in order to recover more geometrical details.

3 SURFACE REPRESENTATION

3.1 Deep Implicit Function

A deep implicit function defines a surface as the level set of an occupancy probability function F, e.g. $F(p) = 0.5$, where $p \in \mathbb{R}^3$ denotes a 3D point and $F : \mathbb{R}^3 \mapsto [0, 1]$ is represented by a deep neural network. In practice, in order to represent the surface of a specific object using a general

neural network, the function F also takes a condition variable (e.g., image feature of the object) as input and thus can be written as:

$$F(p, c) : \mathbb{R}^3 \times \mathcal{X} \mapsto [0, 1] \tag{1}$$

where $c \in \mathcal{X}$ is the condition variable that encodes the overall shape of a specific surface and can be custom designed in accordance of applications [45], [46], [47]. Intuitively, F predicts the continuous inside/outside probability field of a 3D model, in which iso-surface can be easily extracted

In PIFu [8], the authors combined the condition variable with the point coordinate and formulate a pixel-aligned implicit function as:

$$F(C(p)) : \mathbb{R}^3 \mapsto [0, 1] \tag{2}$$

$$C(p) = (S(\mathbf{F}_I, \pi(p)), Z(p))^{\mathrm{T}} \tag{3}$$

where $\mathbf{F}_I = E_I(\mathbf{I})$ represents the image feature map from the deep image encoder $E_I(\cdot)$, $\pi(p)$ the 2D projection of p on the feature map $\mathbf{F}_I$, $S(\cdot, \cdot)$ is the sampling function used to sample the value of $\mathbf{F}_I$ at pixel $\pi(p)$ using bilinear interpolation, and $Z(p)$ is the depth value of p in the camera coordinate space. A weak perspective camera is assumed in both PIFu and this paper. Intuitively, given a pixel on the image, PIFu casts a ray through that pixel along the direction of z-axis and estimates the values of occupancy probability on that ray based on the local image feature of the given pixel. Thanks to the usage of pixel-level features as condition variable, PIFu can reconstruct fine-scale detailed surfaces that are well aligned to the input images.

PIFu [8] demonstrates high-quality human digitization for fashion poses, like standing or walking. However, we argue that purely relying on 2D image feature is insufficient to handle severe occlusions and large pose variations for single-image 3D human recovery, especially when ideal high quality dataset (covering sufficient poses, shapes and cloth types) is inaccessible. Specifically, the reasons are two-folds. On one hand, under the scenarios of self-occlusions, there may be multiple peaks of occupancy probability along the z-rays, but it is hard to infer these types of changes consistently and accurately only based on the local pixel-level 2D features. On the other hand, without the awareness of the underlying body shape and pose, the reconstruction results are prone to seriously incomplete or asymmetric bodies, like breaking arms and unequal legs, in the cases of challenging poses.

3.2 Parametric body Model

Inspired by DeepHuman [5], we integrate the parametric body model, SMPL [9], to regularize the human reconstruction. SMPL is a function $M(\cdot)$ that maps pose θ and shape β to a mesh of n_S vertices:

$$M(\beta, \theta) = W(T(\beta, \theta), J(\beta), \mathbf{W}))$$
$$T(\beta, \theta) = \mathbf{T} + B_s(\beta) + B_p(\theta) \tag{4}$$

where linear blend-skinning $W(\cdot)$ with skinning weights $\mathbf{W}$ poses the T-pose template $\mathbf{T} + B_s(\beta) + B_p(\theta)$ based on its skeleton joints $J(\cdot)$. As shown in DeepHuman, a SMPL model fitted to the input image can serve as a strong geometry prior, and guarantee pose and shape aware
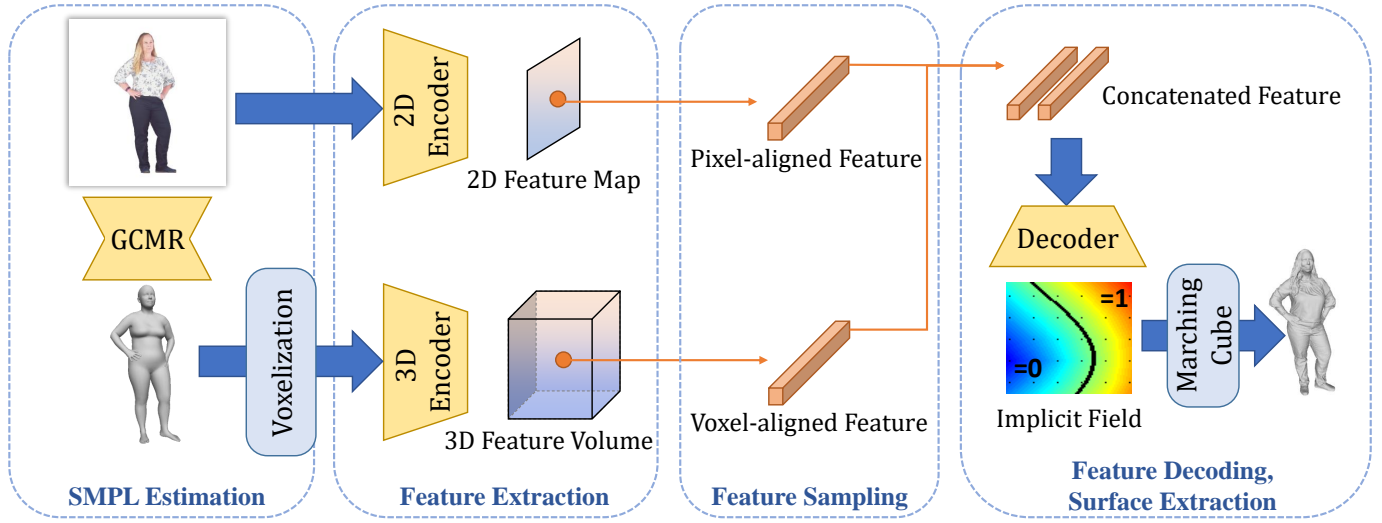
Fig. 2. Overview of the PaMIR representation and the corresponding network architecture. Given an input image, we first estimate a corresponding SMPL model in the SMPL estimation step. In the following step, the image and the SMPL model are converted into a feature map and a feature volume respectively. Conditioned on the feature vectors sampled in the feature sampling step, the implicit function value for each 3D point can be obtained in the final step.

human model reconstructions. However, estimating SMPL model given a single image is a fundamentally ill-posed problem. Recent studies have explored to learn from large-scale datasets to address this challenge [11], [12], [13], [14], [15]. In this work we use the state-of-the-art network in GCMR [13] as the backbone for body shape inference, but it can be replaced by other equivalent networks.

3.3 PaMIR: Parametric Model-Conditioned Implicit Representation

To combine the strengths of parametric body models and non-parametric implicit field, we introduce Parametric Model-Conditioned Implicit Representation (PaMIR). Specifically, in PaMIR, we define the definition of $C(p)$ in Eqn.(3) as:

$$C(p) = \left(S\left(\mathbf{F}_I, \pi(p)\right), S\left(\mathbf{F}_V, p\right)\right)^{\mathrm{T}} \tag{5}$$

where $S\left(\mathbf{F}_I, \pi(p)\right)$ has the same definition as Eqn.(3), while $\mathbf{F}_V$ is the feature volume and $S\left(\mathbf{F}_V, p\right)$ represents the voxel-aligned volumetric feature at p sampled from $\mathbf{F}_V$. The feature volume $\mathbf{F}_V$ is obtained by firstly converting the SMPL mesh into an occupancy volume $\mathbf{V}_O$ through mesh voxelization and then encoding it through a 3D encoder E_V, i.e., $\mathbf{F}_V = E_V(\mathbf{V}_O)$. Given the per-pixel feature vector $S\left(\mathbf{F}_I, \pi(p)\right)$ of p as well as its per-voxel feature vector $S\left(\mathbf{F}_V, p\right)$, we learn an implicit function $F(C(p))$ in Eqn.(2) that can classify whether p is inside or outside the surface. Note that although the introduction of 3D dense feature volume may increase memory consumption, in practice, we found that a relatively small feature volume is enough for soft semantic regularization. Moreover, we omit $Z(p)$ in $C(p)$ since $S\left(\mathbf{F}_V, p\right)$ already contains 3D coordinate information. The illustration of the proposed PaMIR representation and the corresponding network is shown in Fig.2.

In PaMIR, the free-form implict representation is regularized by the semantic features of the parametric model.

As we can see in the following sections, introducing SMPL model as a body reference has several advantages:

1) Pose Generalization. For surface geometry reconstruction, the SMPL input can be regarded as an initial guess for the network, which helps to factor out pose changes of the subject and eliminate depth ambiguities, thus making the network mainly focusing on surface detail reconstruction. As a result, our method has more robust performance compared with PIFu especially under challenging poses.
2) Multi-modal Prediction. Unlike PIFu [8] or Occupancy Network [47] that only condition on the input images, our PaMIR representation also conditions on the underlying bodies. Thus, in contrast to reconstruct one model for each input image, our method can reconstruct various plausible models given different but plausible body poses (Fig.12).
3) Easy Extension to Multi-image Setups. With the common underlying SMPL model as a body reference, our PaMIR representation also implicitly builds a correspondence relationship across different models and images. As a result, our method can be easily extended to multi-image settings like multi-view input and video input without explicit calibration and synchronization.

4 METHOD

4.1 Overview

Our PaMIR-based 3D human reconstruction is implemented as a neural network. An overview of the network is illustrated in Fig.2. Given a single image $\mathbf{I}$ as input, our method first feeds it into the GCMR network to estimate an initial SMPL model, which is then converted into an occupancy volume through voxelization. In the feature extraction step, the input image is encoded into a feature map $\mathbf{F}_I$ by a 2D convolution network, while the occupancy volume are encoded into a feature volume $\mathbf{F}_V$ by a 3D convolution
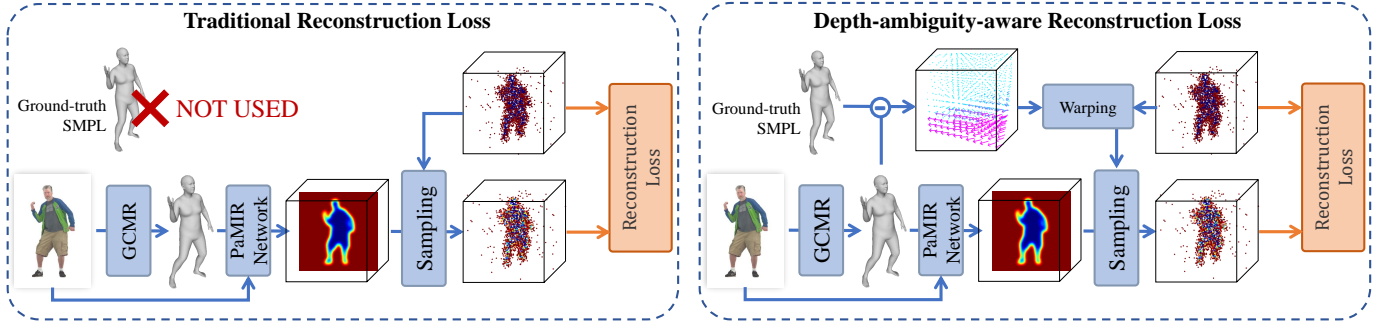
Fig. 3. Comparison of the traditional reconstruction loss and our proposed depth-ambiguity-aware reconstruction loss. Unlike the traditional reconstruction loss, we shift the point samples according to the depth difference between the ground-truth SMPL and the predicted one. In this way the network is able to learn to infer an implicit field registered with the predicted 3D pose.

network. For each point in the 3D space, its pixel-aligned image feature $S(\mathbf{F}_I, \pi(p))$ and voxel-aligned volume feature $S(\mathbf{F}_V, p)$ are sampled in the feature map and the feature volume, respectively. The two feature vectors are then concatenated and translated to an occupancy probability value by a feature-to-occupancy decoder as formulated in Eqn.(2) and (5).

To alleviate the reliance on accurate SMPL annotations, we use the training images, the SMPL models estimated by GCMR and the corresponding ground-truth meshes to train the other parts of the network. To deal with the depth inconsistency between the predicted SMPL models and the ground-truth scans, we carefully design a depth-ambiguity-aware reconstruction loss as elaborated in Sec.4.2. Moreover, for inference, we propose body reference optimization, and optimize the human body template as well as the implicit function in an iterative manner in Sec.4.3. To obtain the final reconstruction results, we densely sample the occupancy probability field over the 3D space and extract the iso-surface of the probability field at threshold 0.5 using the Marching Cube algorithm. Texture inference can be performed in a similar way to geometry inference (Sec.4.4). Our framework can be easily extended to multi-image setups, which is described in Sec.5.

4.2 Depth-ambiguity-aware Reconstruction Loss

To train our network, we sample 3D points in 3D space around the human model, infer their occupancy probabilities and construct a per-point reconstruction loss. The traditional reconstruction loss is defined as the mean square error between the predicted occupancy probability of the point samples and the ground-truth ones. However, we argue that with a single-image setup, depth ambiguity is inevitable for many body poses and hence it is unpractical to force the network output to be perfectly identical with the ground-truth. Furthermore, if we utilize the predicted SMPL models as reference of 3D shape and pose, using traditional reconstruction loss will lead to negative impact on the reconstruction accuracy because the predicted models may not be well aligned with the ground-truth along the z-axis. To deal with this issue, we propose a novel depth-ambiguity-aware reconstruction loss defined as:

$$\mathcal{L}_R = \frac{1}{n_p} \sum_{i=1}^{n_p} |F(C(p_i + \Delta p_i)) - F^*(p_i)|^2, \qquad (6)$$

where n_p is the number of point samples, p_i a 3D point sample indexed by i, $F^*(p_i)$ the ground-truth occupancy value of p_i, and $\Delta p_i = (0, 0, \Delta z_i)^{\mathrm{T}}$ is the compensating translation along z-axis. Δz_i is calculated on-the-fly during network training using the explicit correspondences between the predicted SMPL model and the ground-truth SMPL model according to

$$\Delta z_i = \sum_{j \in \mathcal{N}(i)} \frac{\omega_{j \to i}}{\omega_i} \left(Z(v_j) - Z(v_j^*) \right), \qquad (7)$$

where $\mathcal{N}(i)$ is the nearest SMPL vertex set of p_i and $|\mathcal{N}(i)| = 4$, $\omega_{j \to i}$ the corresponding blending weight, ω_i the weight normalizer, v_j and v_j^* are the j-th vertex of the predicted SMPL model and the corresponding ground-truth SMPL model, respectively. The blending weight is defined according to the distance between the point sample p_i and its neighboring SMPL vertex v_j:

$$\omega_{j \to i} = \exp\left(-\frac{\|p_i - v_j\|}{2\sigma^2} \right),$$
$$\omega_i = \sum_{j \in \mathcal{N}(i)} \omega_{j \to i}. \qquad (8)$$

Intuitively, with the depth-ambiguity-aware loss, we are guiding the network to output plausible surface corresponding to the predicted SMPL model but not the exact ground-truth occupancy volume. In this way the network is able to learn to infer an implicit field registered with the predicted 3D pose. The illustration of depth-ambiguity-aware reconstruction loss is shown in Fig.3.

4.3 Body Reference Optimization

Although our training scheme already prevents our network from being heavily dependent on the accuracy of SMPL estimation, the inconsistency between the image observation and the SMPL estimation may still lead to reconstruction artifacts. Fortunately, the predictions of our method can be further refined at inference time. Specifically, we can improve the accuracy of SMPL estimation by minimizing:

$$(\beta^*, \theta^*) = \arg\min_{\beta, \theta} \mathcal{L}_B + \lambda_R \mathcal{L}_{REG} \qquad (9)$$

where $\mathcal{L}_B$ is the body fitting loss used to encourage the alignment of the predicted implicit function and SMPL model and $\mathcal{L}_{REG}$ is a regularization term penalizing the
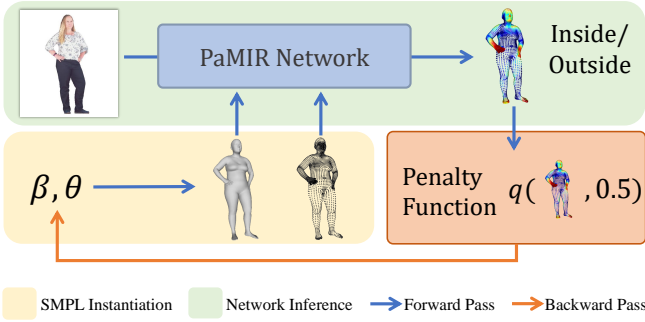
Fig. 4. Illustration of body reference optimization. The estimated SMPL parameters are firstly decoded into a SMPL mesh and then the occupancy values of its vertices are infered by our PaMIR network. Through minimizing the body fitting loss, the SMPL parameters are optimized iteratively.

difference between (β^*, θ^*) and the initial prediction. The body fitting loss is defined as following:

$$\mathcal{L}_B = \frac{1}{n_S} \sum_{j=1}^{n_S} q\left(F\left(c(v_j)\right) - 0.5\right) \tag{10}$$

where n_S is the number of SMPL vertices, v_j is the j-th predicted SMPL vertex and $q(\cdot)$ is the penalty loss defined as:

$$q(x) = \begin{cases} |x| & x \geq 0, \\ \frac{1}{\eta}|x| & x < 0, \end{cases} \tag{11}$$

where $\eta = 5$. In other words, $p(\cdot)$ applies more penalization on the vertices that fall outside of the predicted surface while allowing the body to shrink into the surface in case of loose clothes like skirts and dresses. The regularization term is defined as

$$\mathcal{L}_{REG} = |\beta - \beta_{init}|_2^2 + |\theta - \theta_{init}|_2^2, \tag{12}$$

where $(\beta_{init}, \theta_{init})$ is the initial SMPL parameters estimated by GCMR. Note that other constraints like 2D keypoint detection results can also be added to further improve the fitting performance.

The insight behind the formulation of body fitting loss is that the predicted SMPL may not be perfectly aligned with the image observation and consequently, the output implicit function is a compromise between these two information. Thus, by minimizing the body fitting loss, we can eliminate the inconsistency between the SMPL prediction and the image observation. As a result, more consistent SMPL prediction as well as more accurate surface inference can be obtained. The idea of our body reference optimization is illustrated in Fig.4.

We visualize the optimization process in Fig.5. In this figure, the initial pose of the right arm is incorrect, which leads to reconstruction artifacts. However, as the optimization is carried out, the right arm gradually moves towards the right position. In the meantime, the reconstructed mesh becomes more and more plausible. To clarify, the reconstruction results in this experiment do not contradict the results in Fig.14: in this experiment, the right arm is almost invisible and consequently it is difficult for the network to infer the correct geometry with an inaccurate arm pose.
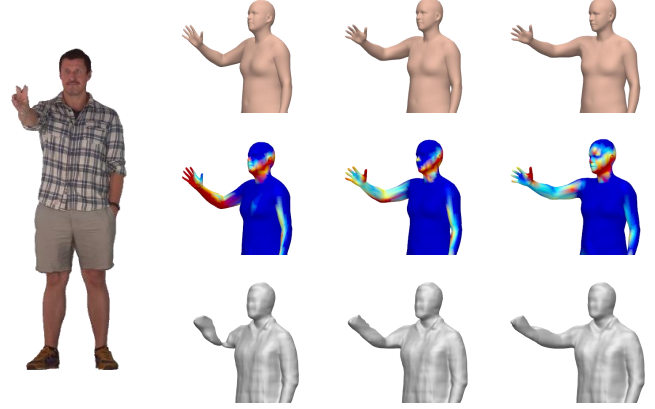

Fig. 5. Visualization of the body optimization process. The leftmost column: the input image. The 2nd to 4th columns: the reconstruction results before reference body optimization, the intermediate status of optimization and the results after optimization. We present the SMPL models in the top row and the reconstructed models in the bottom row. The median row shows the occupancy probability of SMPL vertices: blue color represents $F(C(v)) = 1.0$, while red means $F(C(v)) = 0.0$.

4.4 Texture Inference

Following the practice of Texture Field [50] and PIFu [8], we also regard the surface texture as a vector function defined in the space near the surface, which can support texturing of shapes with arbitrary topology and self-occlusion. To perform texture inference, we make some simple modification to our network. Specifically, we define the output of the decoder in Fig.2 as an RGBα vector field instead of a scalar field. The RGB value is the network prediction of the color of a specific point on the mesh surface, while the alpha channel is used to blend the predicted value with the observed one:

$$C(p_i) = \alpha * S\left(\mathbf{I}, \pi(p)\right) + (1 - \alpha) * C'(p_i) \tag{13}$$

where $C'(p_i)$ is the color prediction provided by the network, i.e., the first three channels of the decoder output, α is the last channel of the decoder output and $C(p_i)$ is the final color prediction. Thus, the reconstruction loss in Eqn.6 is modified accordingly to:

$$\mathcal{L}_R = \frac{1}{n_p} \sum_{i=1}^{n_p} |C(p_i + \Delta p_i, c(p_i + \Delta p_i)) - C^*(p_i)| + \\ |C'(p_i + \Delta p_i, c(p_i + \Delta p_i)) - C^*(p_i)|, \tag{14}$$

where $C^*(p_i)$ is the ground-truth vertex texture and L1 loss is used to avoid color over smoothing. Intuitively, the network learns to infer the color of the whole surface and also determine which part of the surface is visible so that we can directly sample color observation on the image. We show the effect of the alpha channel in Fig.6.

Note that we do not decompose shading component from the original color observation to obtain the "real" surface texture. To obtain shading-free texture, one can use state-of-the-art albedo estimation method like [51] to firstly remove shading from the input images and then feed the processed images to our texture inference network.
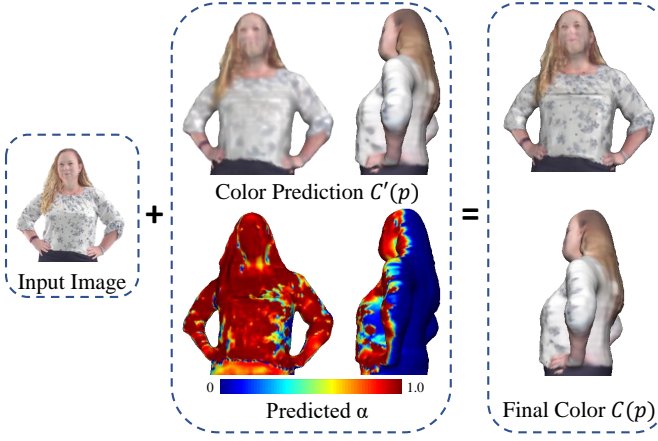
Fig. 6. Illustration of our texture inference process. The texture network first recovers the color on the whole surface as well as an alpha channel which is used to blend the predicted color with the observation. In this way, the input image is maximally utilized and more texture details are recovered (Please zoom in to compare the texture on the shirt).
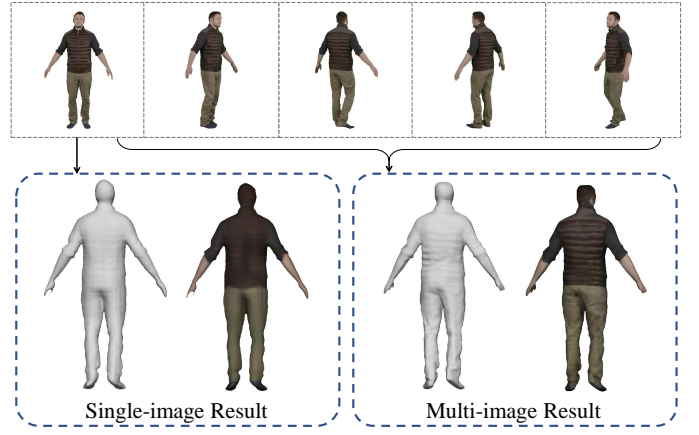

Fig. 7. Comparison between the single-image result and the multi-image result. By adding four more frames (without calibration and synchronization infomation) as input, our method is able to recover the surface details on the back. In contrast, the back area is over-smoothed in the single-image setting. Zoom in for better view.

5 EXTENSION TO MULTI-IMAGE SETUP

With the common underlying SMPL model as body reference, our PaMIR representation also implicitly builds a correspondence relationship across different models. As a result, our method can be easily extended to multi-image settings like multi-view input or video input. Specifically, given N_I images of the same person $\mathbf{I}_t, t = 0, 1, ..., N_I - 1$, we first use the optimization method described in Sec.4.3 to obtain the SMPL models aligned to each input images, $M(\theta_t, \beta_t)$. Note that the optimization step is performed on each individual image independently; for future work more constraints like shape consistency and pose consistency constraints can be incorporated to the energy function. With these SMPL template parameters we can establish correspondences across two images. For example, for a point p_i in the model space of the reference image $\mathbf{I}_0$, its corresponding point $p_i^{(t)}$ in the model space of $\mathbf{I}_t$ can be calculated as

$$p_i^{(t)} = \mathbf{M}_i^{0 \to t} p_i,$$
$$\mathbf{M}_i^{0 \to t} = \sum_{j \in \mathcal{N}(p_i)} \frac{\omega_{j \to i}}{\omega_i} \mathbf{M}_j^{(t)}(\theta_t) \left(\mathbf{M}_j^{(0)}(\theta_0) \right)^{-1} \quad (15)$$

where $\mathbf{M}_j^{(0)}(\theta_0)$ and $\mathbf{M}_j^{(t)}(\theta_t)$ are the linear blending skinning (LBS) matrices of the j-th SMPL vertex in the reference frame and the t-th frame, respectively, $\mathcal{N}(p_i)$, $\omega_{j \to i}$ and ω_i share the same definition as Eqn.(7).

To perform geometry inference using multiple image frames, we follow the practice of PIFu [8] and decompose the feature-to-occupancy network $F(\cdot)$ in Eqn.2 into two components:

$$F = F_2 \circ F_1 \quad (16)$$

where F_1 is a feature embedding network while F_2 is an occupancy reasoning network. For a 3D point p_i in the reference frame $\mathbf{I}_0$, we first sample its feature vector $C^{(0)}(p_i)$ as described in Sec.3.3. Then we calculate the corresponding point $p_i^{(t)}$ in the t-th frame and also sample its corresponding feature vector $C^{(t)}(p_i^{(t)})$, where $t = 1, 2, ..., N_I - 1$. The feature embedding network F_1 takes these feature vectors

and encodes them into latent feature embedding vectors $\Phi_i^{(k)}, k = 0, 1, ..., N_I - 1$. The latent embedding vectors across different frames are aggregated into one embedding $\bar{\Phi}_i$ using mean pooling, which is then mapped to the target implicit field $F(p_i)$ by the occupancy reasoning network F_2. The multi-image network is fine-tuned from the single-image network using multi-view rendering of 3D human models.

Note that PIFu also demonstrates results given mutli-view inputs. However, the multi-view input should be well calibrated and synchronized. In contrast, neither calibration or synchronization is necessary in our method because we can utilize the SMPL estimation to build the correspondence across different views. Moreover, our method can handle the cases where body poses are not identical across images, e.g., video input. This feature enables video-based personalized avatar creation. A similar application has been demonstrated in [52], but our method can support more challenging cloth topologies and recover more surface details. Overall, in terms of multi-image setup, our method is more general and practical than state-of-the-art methods. A comparison between the single-image result and the multi-image result is presented in Fig.7. As shown in the figure, by adding four more video frames as input, our method is able to recover the surface details on the back that is invisible in the first frame.

Since we consider only pose changes across views and neglect surface detail deformations (e.g., wrinkle movements) across frames, challenging pose deviation and inconsistent geometry between frames will cause inaccurate feature fusion and reconstruction artifacts. Besides, the multi-image network is trained using only multi-view rendering of human models without pose deviations. For future work, we can introduce 4D data to resolve these issues.

6 EXPERIMENTS

In this section we evaluate our approach. Details about the implementation and the evaluation dataset are given in Sec.6.1 and Sec.6.2 respectively. In Sec.6.3 we demonstrate

Fig. 8. Our results on natural images. From left to right: the 1st is the input images, the 2nd and 3rd column show the SMPL models estimated by our method (network inference + optimization), the 4th to 6th demonstrates our geometry reconstruction results, and the last three column demonstrates the texture inference results. The results demonstrate the ability of our method to reconstruct high-quality models and its robust performance to tackle various human poses and clothing styles.

that our method is able to reconstruct human models with challenging poses. We then compare our method against state-of-the-art methods in Sec.6.4. We also evaluate our contributions both qualitatively and quantitatively in Sec.6.5. The quantitative evaluation results are given in Tab.3. We mainly use geometry reconstruction performance for evaluation, which is our main focus.

6.1 Implementation Details

Network Architecture. For image feature extraction, we adapt the same 2D image encoders in PIFu [8] (i.e., Hourglass Stack [54] for geometry and CycleGAN [55] for texture), which take as input an image of 512×512 resolution and outputs a 256-channel feature map with a resolution of 128×128. For volumetric feature extraction, we use a 3D convolution network which consists of two convolution layers and three residual blocks. Its input resolution is $128 \times 128 \times 128$ and its output is a 32-channel feature volume with a resolution of $32 \times 32 \times 32$. We replace batch normalization with group normalization to improve the

training stability. The feature decoder is implemented as a multi-layer perceptron, where the number of neurons is $(288, 1024, 512, 256, 128, ch)$, where $ch = 1$ for the geometry network while $ch = 4$ for the texture network.

Training Data. To achieve state-of-the-art reconstruction quality, we collect 1000 high-quality textured human scans with various clothing, shapes, and poses from Twindom[1]. We randomly split the Twindom dataset into a training set of 900 scans and a testing of 100 scans. To augment pose variety, we also randomly sample 600 models from DeepHuman [5] dataset. We render the training models from multiple view points using Lambertian diffuse shading and spherical harmonics lighting [56]. We render the images with a weak perspective camera and image resolution of 512×512. To obtain the ground-truth SMPL annotations for the training data, we apply MuVS [57] to the multi-view images for pose computation and then solve for the shape parameters to further register the SMPL to the scans. For point sampling during training, we use the same scheme proposed in PIFu
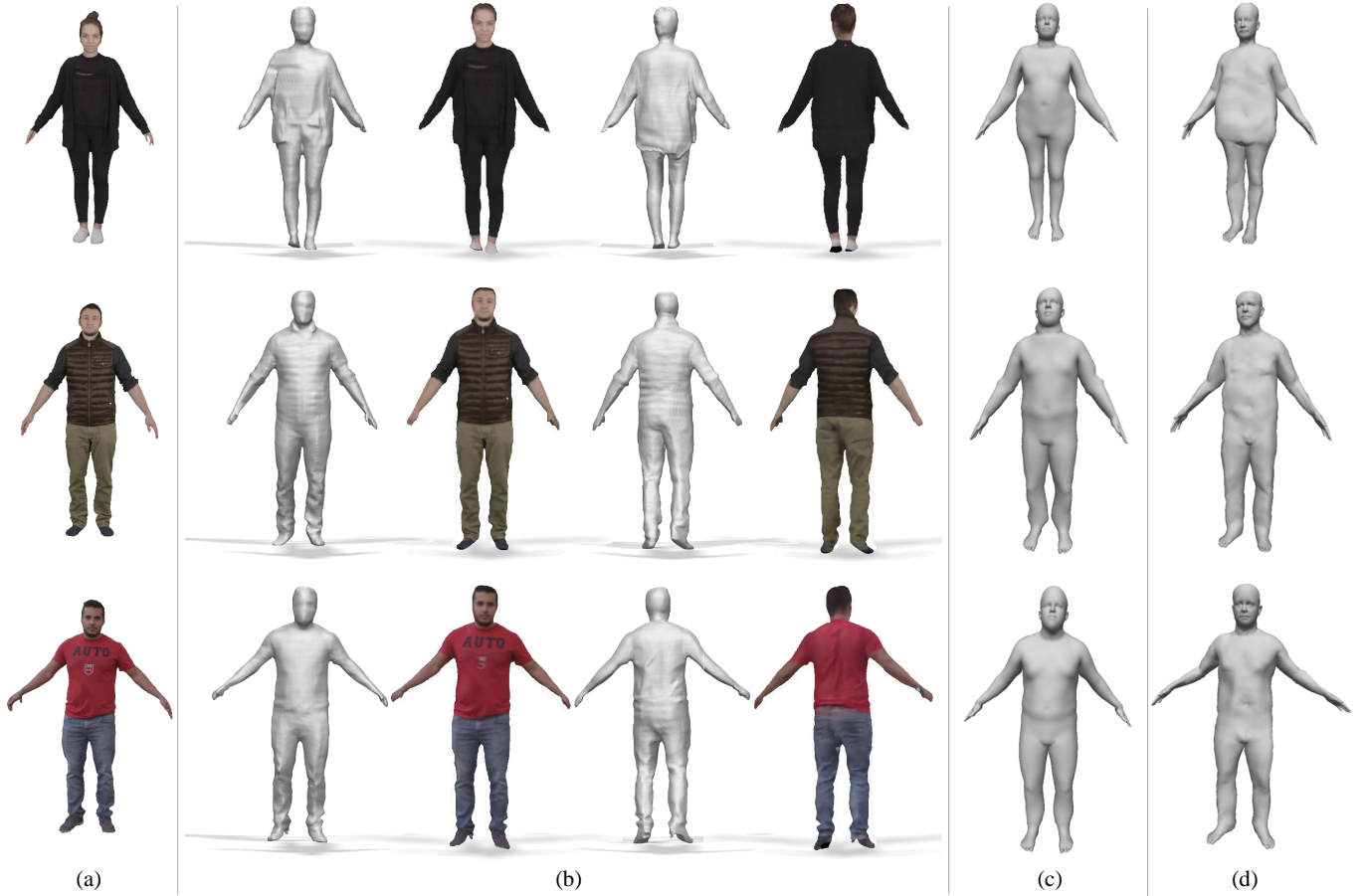
1. https://web.twindom.com/

Fig. 9. Multi-image human model reconstruction on VideoAvatar dataset: (a) the first column is the first images of the input sequences, (b) the reconstructed models by our method, (c) results by [52], (d) Results by [53]. Our method can reconstruct full-body models with high-resolution details, proving the capability of our PaMIR-based reconstruction method.

[8], in which the authors combine uniform sampling and adaptive sampling based on the surface geometry and use embree algorithm [58] for occupancy querying.

Network training. We use Adam optimizer for network training with the learning rate of 2×10^{-4}, the batch size of 3, the number of epochs of 9, and the number of sampled points of 5000 per subject. The learning rate is decayed by the factor of 0.1 at every 10000-th iteration. We also combine predicted and ground-truth SMPL models during network training. Specifically, in each batch, we randomly select one image to replace the predict SMPL model with the ground-truth one when constructing the depth-ambiguity-aware reconstruction loss. In this way we can guarantee the best performance is obtained once the underlining SMPL becomes more accurate. The multi-image network is fine-tuned from the models trained for single-view network with three random views of the same subject using a learning rate of 2×10^{-5} and a batch size of 1. Before training the PaMIR network, we first fine-tune the pre-trained GCMR network on our training set.

Network testing. When testing, our network only requires an RGB image as input and outputs both the parametric model and the reconstructed surface with texture. To maximize the performance, We run the body reference optimization step for all results unless otherwise stated. Fifty iterations are needed for the optimization and take about 40 seconds.

Network complexity. We report the module complexity in Tab.1. To reconstruct the 3D human model given an RGB image, we use the hierarchical SDF querying method proposed in OccNet [47] to reduce network querying times. Overall, taking the body reference optimization step into account, it takes about 50s to reconstruct the geometry of the 3D human model and 1s to recover its surface color.

TABLE 1
The number of parameters and execution time of each network module

Module	#Parameters	Execution Time
GCMR	46,874,690	0.15s*
Geometry Network	27,225,105	0.25s†
Texture Network	13,088,268	0.29s†

* Measured using one RGB image.
† Measured using one RGB image and 10k query points.

6.2 Evaluation Dataset

For qualitative evaluation on single-image reconstruction, we utilize real-world full-body images collected from the
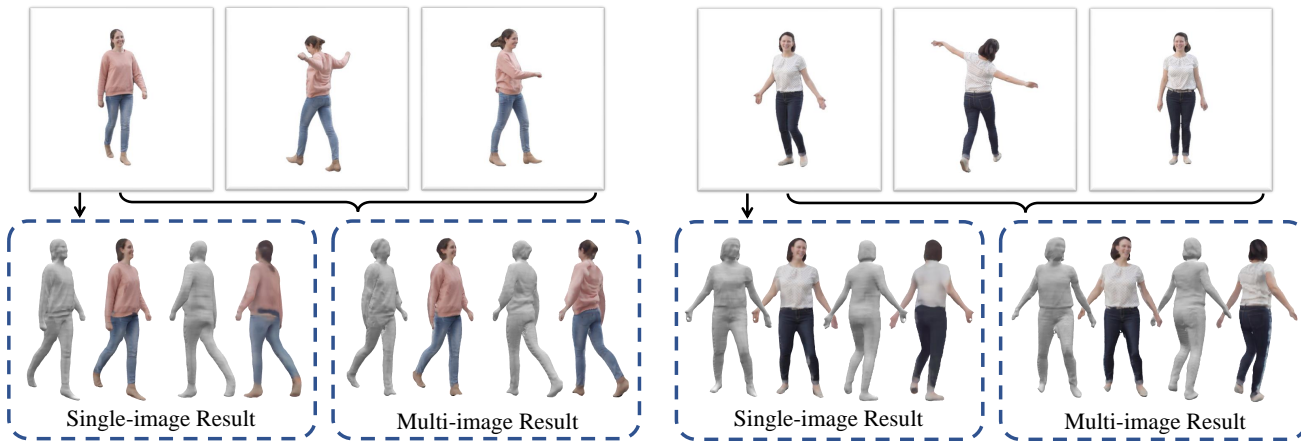
Fig. 10. Comparison between the single-image reconstruction results and the multi-image results from input images with moderate pose deviations. In these cases, our method is still able to integrate information from different frames.

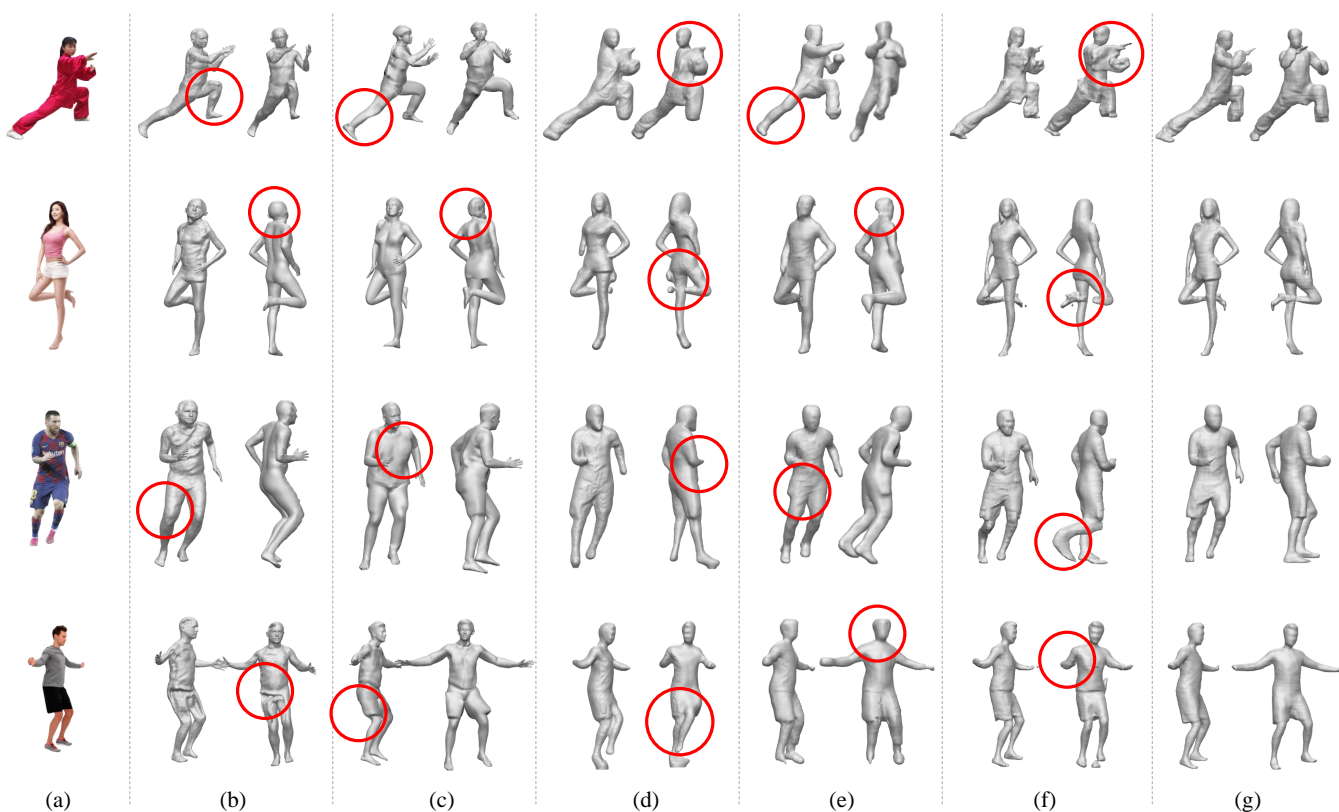(a)   (b)   (c)   (d)   (e)   (f)   (g)

Fig. 11. Qualitative comparison against state-of-the-art methods for single-image human model reconstruction: (a) input images, (b) results by HMD [3], (c) Tex2Shape [4], (d) Moulding Humans [6], (e) DeepHuman [5], (f) PIFu [8], (g) ours.

DeepFashion dataset [59] and from the Internet. We remove the background of these real-world image using neural semantic segmentation [60] followed by Grabcut refinement [61]. For qualitative evaluation on multi-image reconstruction, we use the open-source dataset from VideoAvatar [52] which contains 24 RGB sequences of different subjects turning 360 degree with a rough A-pose in front of a camera. The testing split of the Twindom dataset is used for quantitative comparison. In addition, for quantitative comparison and evaluation we also use the BUFF dataset [62], which provides 26 4D human sequences with various clothes and over 9000 scans in total. To reduce computation burden, We

randomly sample 300 scans from the buff dataset. We render both the Twindom testing data and the BUFF data from 12 views spanning every 30 degrees in yaw axis using the same method in Sec.6.1.

6.3 Results

We demonstrate our approach for single-image 3D human reconstruction in Fig.1 and Fig.8. The input images in Fig.1 and Fig.8 covers various body poses (dancing, Kungfu, sitting and running), and also covers different clothes (loose pants, skirts, sports suits and casual clothes). The results

(a)   (b)   (c)
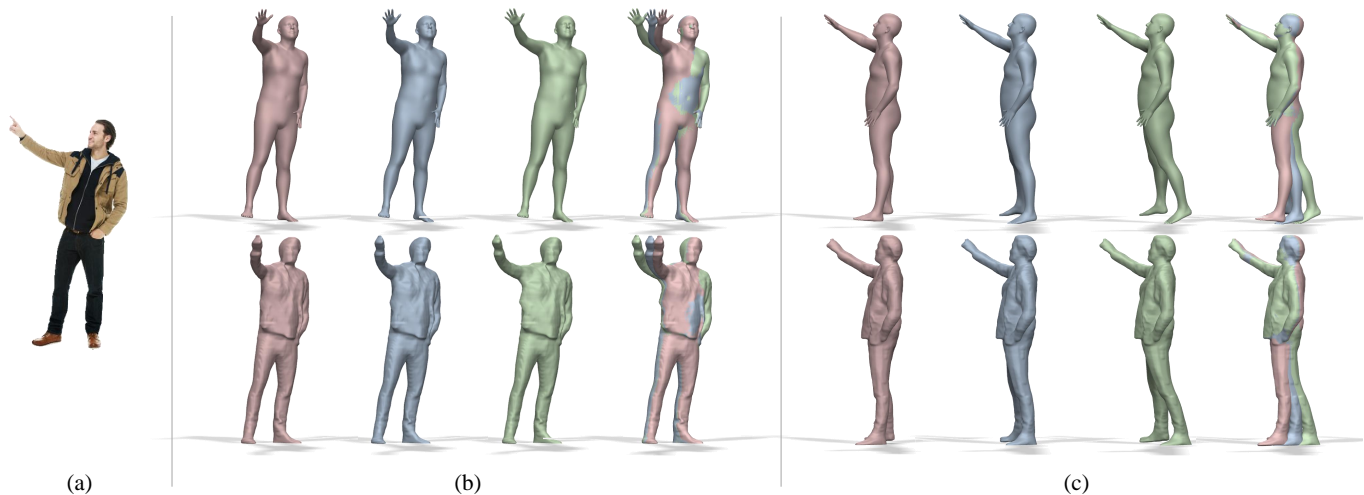
Fig. 12. Two examples of multi-modal output. Given an input image (a), our method can output different possible reconstruction results (top row of (b)(c)) corresponding to different body pose hypothesis (bottom row of (b)(c)). From the overlapped figure (the last columns of (b)(c)), we can see that the right arm and the left leg of reconstructed mesh adjusts accordingly with the right arm and the left leg of the body hypothesis, while other parts of the mesh keep consistent. Better view in color.

TABLE 2
Numerical comparison results.

Dataset	BUFF		TWINDOM	
Method	P2S (cm)	Chamfer (cm)	P2S (cm)	Chamfer (cm)
HMD [5]	2.48	3.92	2.50	4.01
Moulding Humans [6]	2.25	2.68	2.84	3.35
DeepHuman [5]	2.15	2.80	2.35	2.97
PIFu [8]	1.93	2.22	2.34	2.65
Ours	1.52	1.92	1.80	2.12
Ours using ground-truth SMPL	0.709	0.936	0.744	1.00

demonstrate the ability of our method to reconstruct high-quality 3D human models and its robust performance to tackle various human poses and clothing styles.

We also test our performance for multi-image human model reconstruction using the VideoAvatar dataset in Fig.9. We uniformly sample 5 frames for each sequence for surface reconstruction. Note that for each subject, the poses in different frames are different due to the body movements and no camera extrinsic parameter is provided, so PIFu is not suitable to reconstruct human models in this case. In contrast, our method can reconstruct full-body models with high-resolution details, proving the generalization capability of our PaMIR representation. We also present two example results of applying multi-image feature fusion on images with moderate pose deviation in Fig.10. As the figure shows, our method is still able to reconstruct the overall shapes of the subjects in these cases.

6.4 Comparison

We qualitatively compare our method with several state-of-the-art methods including HMD [3], Tex2Shape [4], Moulding Humans [6], DeepHuman [5] and PIFu [8]. Among them, HMD [3] and Tex2Shape [4] are parametric methods based on SMPL [9] model deformation, PIFu [8] uses a deep implicit function as geometry representation, Moulding Humans [6] uses the combination of a front depth map and a back depth map for representation, and DeepHuman [5]

combines volumetric representation with the SMPL model. Note that Tex2Shape only computes shapes, so we use the pose parameters obtained by our method to skin the its results. For simplicity, we omit comparisons with the works that have already been compared like BodyNet [1] and SiCloPe [2]. In our experiments, DeepHuman, PIFu and Moulding Humans are all retrained on the our dataset, while parametric methods like HMD and Tex2shape are not because our dataset contains loose garments like dresses which will deteriorate the performance of those methods.

We conduct qualitative comparison in Fig.11. As shown in the figure, HMD and Tex2Shape have difficulties dealing with loose clothes and cannot reconstruct the surface geometry accurately; Moulding Human [6] fails to handle challenging poses (due to the lack of semantic constraints), and produce broken body parts when self-occlusions occur, which is the essential limitation of its double depth representation; DeepHuman [5] cannot recover high-frequency geometrical details although it succeeds to reconstruct the rough shapes from the images; PIFu [8] struggles to reconstruct the models in challenging poses and also suffer from self-occlusions. In contrast, our method is able to reconstruct plausible 3D human models under challenging body poses and various clothing styles. In terms of surface quality and pose generalization capacity, our method is superior to other state-of-the-art methods.

We quantitatively compare our method with the state-of-the-art methods using both Twindom testing dataset and

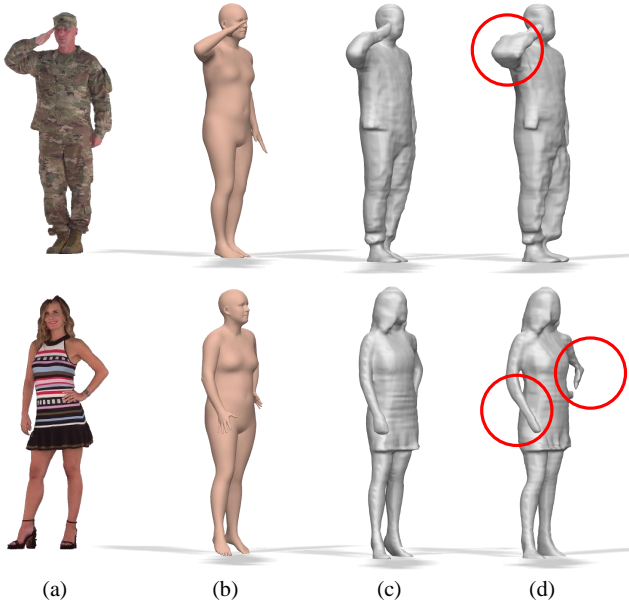(a)   (b)   (c)   (d)

Fig. 13. Qualitative evaluation on depth-ambiguity-aware training loss: (a) the input images, (b) the corresponding SMPL models estimated by GCMR, (c) our reconstruction results, (d) the reconstruction results of the baseline network.

BUFF rendering dataset to evaluate the geometry reconstruction accuracy. Similar to the experiments in PIFu [8], we use point-to-surface error as well as the Chamfer distance as error metric. The numerical results are presented in Tab.2. The quantitative comparison shows that our method outperforms the state-of-the-art methods in terms of surface reconstruction accuracy. We also provide the errors when ground-truth SMPL annotations are available to present the upper limit of our reconstruction accuracy if the SMPL estimation is perfect. Overall, our method is more general, more robust and more accurate than HMD [3], Moulding Humans [6], DeepHuman [5] and PIFu [8].

For multi-image setups, we also conduct qualitative comparison against state-of-the-art methods [52] and [53]. The method proposed in [52] is an optimization-based method which deforms the SMPL template according to the silhouettes. The optimization is performed on the whole video sequence. In contrast, the method in [53] is a learning-based method that deforms the SMPL template based on only 8 views of images. The comparison results are shown in Fig.9. From the results we can see that although all methods are able to reconstruct the overall shapes correctly, our method is able to recover more surface details than the other two methods. This is because our non-parametric representation allows more flexible surface reconstruction. We do not perform quantitative comparison because there is no such benchmark available.

6.5 Evaluation

6.5.1 PaMIR Representation

The superiority of our PaMIR representation is already proved through the comparison experiments in Sec.6.3 and 6.4. In this evaluation, we demonstrate the advantage of our PaMIR representation from another aspect: it can support

TABLE 3
Numerical Ablation Study.

Method*	P2S (cm)
GT SMPL	1.71
Pred SMPL wo/ DAAL	1.81
Pred SMPL w/ DAAL	1.60
Full (Pred SMPL w/ DAAL + Optimization)	1.52

* GT = ground-truth, Pred = predicted, DAAL = depth-ambiguity-aware loss

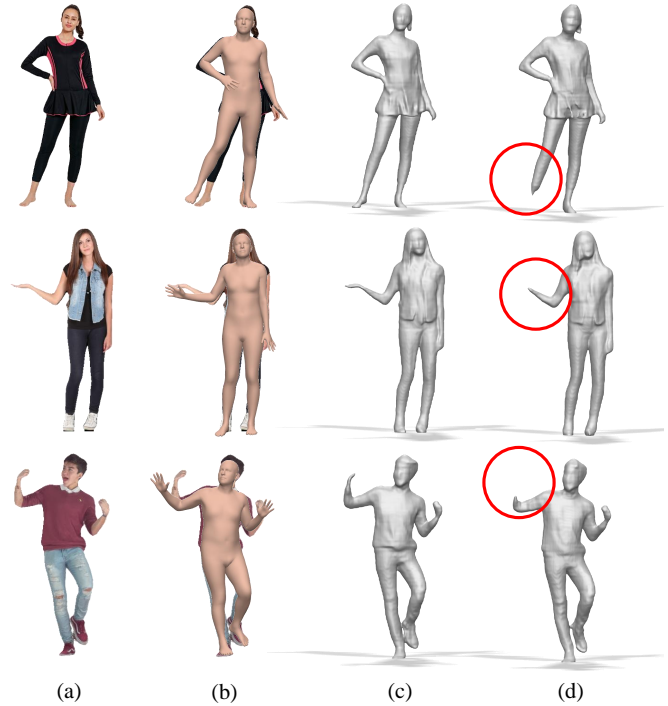(a)   (b)   (c)   (d)

Fig. 14. Qualitative evaluation on our training scheme: (a) the input images, (b) the estimated SMPL models rendered on top of the input images, (c) our reconstruction results, (d) the reconstruction results of the baseline network.

multi-modal outputs. To be more specific, our method can output multiple possible human models corresponding to different body pose hypotheses. Two examples are shown in Fig.12. In both experiments we manually adjust one part of the body in order to generate multiple pose hypotheses. As a result, the reconstruction results change in accordance with the input SMPL models. In contrast, non-parametric method like PIFu [8] or Moulding Humans [6] can only generate a specific mesh for each image, showing their limited and overfit capability.

6.5.2 Depth-ambiguity-aware Loss

To conduct ablation study on our depth-ambiguity-aware loss, we implement a baseline network that is trained using the traditional reconstruction loss and compare it against our network. After the same number of training epochs, we test both network on real-world input images. Two example results are shown in Fig.13. As we can see in the figure, with our depth-ambiguity-aware reconstruction loss, the network is able to learn to reconstruct 3D human models that have consistent poses with the estimated SMPL

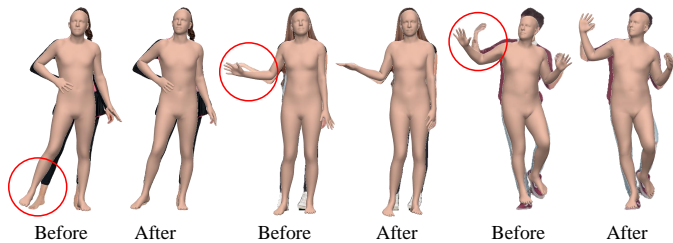Before   After   Before   After   Before   After

Fig. 15. Evaluation of reference body optimization. Note that after body reference optimization, the SMPL models are more accurately aligned with the image observations.

TABLE 4
Mean Per Joint Position Error (MPJPE, unit: cm) Before/After Body Reference Optimization.

Dataset	BUFF	TWINDOM
Before Optimization	2.65	2.85
After Optimization	2.49	2.74

models. On the contrary, the models reconstructed by the baseline network without our depth-ambiguity-aware loss are less consistent with the SMPL models and have more artifacts. The quantitative study is presented in Tab.3. From the numeric comparison between the 3rd and 4th rows of Tab.3, we can conclude that our depth-ambiguity-aware loss avoids the negative impact of the depth inconsistency between predicted SMPL models and ground-truth scans, and improves the accuracy of the reconstruction results.

6.5.3 Training Scheme

We also implement another baseline network that is trained using the ground-truth SMPL annotations to evaluate our training scheme. We compare this baseline network with our network on different input images. In Fig.14, we present some cases in which the predicted SMPL models are not well aligned with image keypoints and silhouettes. As we can see in Fig.14, under the scenarios of inaccurate SMPL estimation, the baseline network fails to reconstruct complete full-body models, while our network is able to provide plausible results in accordance with image observations. Note that this feature is of vital importance. For instance, in the last example of Fig.14, the right forearm of the baseline model is completely missing. In this case, the reference body optimization step could not work because no matter how the right forearm moves, the occupancy probability value of its vertices are always closed to zero. The quantitative study is presented in Tab.3. From the numerical comparison between the 2nd and 4th rows of Tab.3, we can conclude that our training scheme improves the accuracy of our reconstruction results at inference time when no accurate SMPL annotation is available.

6.5.4 Reference Body Optimization

To evaluate the effectiveness of our reference body optimization step, we compare the body fitting results before and after optimization using the evaluation images in Sec.6.5.3. The results are presented in Fig.15. As shown in the figure,

TABLE 5
Normal Reprojection Error with Inputs Images From Various Numbers of View Points.

View Num	1	2	3	4
Normal Reprojection Error	0.161	0.147	0.139	0.135

the optimization step can further register the SMPL model to the image observation, resulting into more accurate body pose estimation. This is also proven in the quantitiative evaluation in Tab.4. From the numerical results in the last two rows of Tab.3, we can also see that the mesh reconstruction is also improved after reference body optimization.

6.5.5 Multi-image Setup

To evaluate the detail changes using more or less images, we conduct a quantitative evaluation in Tab.5. Specifically, we perform feature fusion with 1, 2, 3 and 4 views of the same model and extract the meshes. Then we measure the normal reprojection error [8] from 8 uniform viewpoints to evaluate the detail improvements. The numerical results presented in Tab.5 suggest that

7 DISCUSSION

Conclusion. Modeling 3D humans accurately and robustly from a single RGB image is an extremely ill-posed problem due to the varieties of body poses, clothing types, view points and other environment factors. Our key idea to overcome these challenges is factoring out pose estimation from surface reconstruction. To this end, we have contributed a deep learning-based framework to combine the parametric SMPL model and the non-parametric deep implicit function for 3D human model reconstruction from a single RGB image. Benefiting from the proposed PaMIR representation, the depth-ambiguity-aware reconstruction loss and the reference body optimization algorithm, our method outperforms state-of-the-art methods in terms of both robustness and surface details. As shown in the supplementary video, we can obtain temporally consistent reconstruction results by applying our method to video frames individually. We believe that our method will enable many applications and further stimulate the research in this area.

From a more general perspective of view, we have made a step forward towards integrating semantic information into the free-form implicit fields which have attracted more and more attention from the research community for its flexibility, representation power and compact nature. For example, similar topics include the reconstruction of hand and face using such PaMIR representation. We also believe our semantic implicit fields would become one important future topic and inspire many other studies in 3D vision. For example, the multi-modality of our method (Fig.12) can inspire research on image-based geometry editing; the practice of integrating geometry template into implicit functions can be used to improve the robustness of other 3D recovery tasks. Moreover, the semantic implicit functions can also be adopted in research on 3D perception, parsing and understanding.
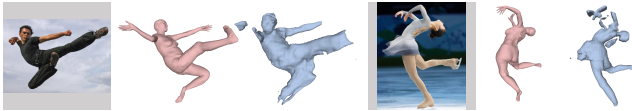
Fig. 16. Failure cases. For extremely challenging poses, our method fails to generate plausible reconstruction results.

Limitation and Future Work. Our method needs high-quality human scans for training. However, it is highly costly and time consuming to obtain a large-scale dataset of high-quality human scans. Moreover, the currently available scanners for human bodies require the subject to keep static poses in a sophisticated capturing environment, which makes them incapable to capture real-world human motions in the wild and consequently our training data is biased towards simple static poses like standing. Therefore, although the proposed method already makes a step forward in terms of generalization capability, it still fails in the cases of extremely challenging poses, see Fig.16. One important future direction is to alleviate the reliance on ground-truth by exploring large-scale image and video dataset for unsupervised training.

REFERENCES

[1] 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.

[2] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," *CoRR*, vol. abs/1901.00049, 2019. [Online]. Available: http://arxiv.org/abs/1901.00049

[3] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] 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.

[5] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[6] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, "Moulding humans: Non-parametric 3d human shape estimation from single images," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero, "Facsimile: Fast and accurate scans from an image in less than a second," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[8] S. Saito, , Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[9] 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)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[10] 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 *ECCV*, 2016, pp. 561–578.

[11] 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 *3DV*, 2018.

[12] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE CVPR*, 2018.

[13] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, 2019.

[14] J. Liang and M. C. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[15] Y. Xu, S.-C. Zhu, and T. Tung, "Denserac: Joint 3d pose and shape estimation by dense render-and-compare," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[16] T. Kanade, P. Rander, and P. J. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," *IEEE MultiMedia*, vol. 4, no. 1, pp. 34–47, 1997.

[17] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 21–31, 2007.

[18] Y. Liu, Q. Dai, and W. Xu, "A point-cloud-based multiview stereo algorithm for free-viewpoint video," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 407–418, 2010.

[19] C. Wu, K. Varanasi, Y. Liu, H. Seidel, and C. Theobalt, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *IEEE ICCV*, 2011, pp. 1108–1115.

[20] M. Waschbüsch, S. Würmlin, D. Cotting, F. Sadlo, and M. H. Gross, "Scalable 3d video of dynamic scenes," *The Visual Computer*, vol. 21, no. 8-10, pp. 629–638, 2005.

[21] D. Vlasic, P. Peers, I. Baran, P. E. Debevec, J. Popovic, S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 174:1–174:11, 2009.

[22] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor *et al.*, "Fusion4d: real-time performance capture of challenging scenes," *ACM Trans. Graph*, vol. 35, no. 4, p. 114, 2016.

[23] M. Dou, P. L. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi, "Motion2fusion: real-time volumetric performance capture," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 246:1–246:16, 2017.

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

[25] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *IEEE ICCV*, 2015, pp. 3334–3342.

[26] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *IEEE CVPR*, 2017.

[27] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton, "Volumetric performance capture from minimal camera viewpoints," in *ECCV*, 2018, pp. 591–607.

[28] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li, "Deep volumetric video from very sparse multi-view performance capture," in *ECCV*, 2018, pp. 351–369.

[29] D. Vlasic, I. Baran, W. Matusik, and J. Popovic, "Articulated mesh animation from multi-view silhouettes," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 97:1–97:9, 2008.

[30] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 98:1–98:10, 2008.

[31] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *IEEE CVPR*, 2009, pp. 1746–1753.

[32] Y. Liu, J. Gall, C. Stoll, Q. Dai, H. Seidel, and C. Theobalt, "Markerless motion capture of multiple characters using multiview image segmentation," *IEEE T-PAMI*, vol. 35, no. 11, pp. 2720–2735, 2013.

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

[34] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE CVPR*, 2017, pp. 1302–1310.

[35] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE CVPR*, 2018.

[36] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *IEEE CVPR*, 2017, pp. 4704–4713.

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *arXiv preprint arXiv:1612.00593*, 2016.

[38] C.-H. Lin, C. Kong, and S. Lucey, "Learning efficient point cloud generation for dense 3d object reconstruction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[39] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 922–928.

[40] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.

[41] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "Atlasnet: A papier-m\^ ach\'e approach to learning 3d surface generation," *arXiv preprint arXiv:1802.05384*, 2018.

[42] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *ECCV*, 2018.

[43] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans, "Reconstruction and representation of 3d objects with radial basis functions," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 67–76.

[44] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 303–312. [Online]. Available: https://doi.org/10.1145/237170.237269

[45] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5939–5948.

[47] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[48] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3d shape reconstruction and completion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[49] C. M. Jiang, A. Sud, A. Makadia, J. Huang, M. Niessner, and T. Funkhouser, "Local implicit grid representations for 3d scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[50] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger, "Texture fields: Learning texture representations in function space," in *Proceedings IEEE International Conf. on Computer Vision (ICCV)*, 2019.

[51] Y. Kanamori and Y. Endo, "Relighting humans: occlusion-aware inverse rendering for full-body human images," in *SIGGRAPH Asia 2018 Technical Papers*. ACM, 2018, p. 270.

[52] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *IEEE CVPR*, 2018.

[53] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019.

[54] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499.

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[56] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *IEEE CVPR*, 2017, pp. 4627–4635.

[57] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black, "Towards accurate markerless human shape and pose estimation over time," in *International Conference on 3D Vision (3DV)*, 2017.

[58] I. Wald, S. Woop, C. Benthin, G. S. Johnson, and M. Ernst, "Embree: A kernel framework for efficient cpu ray tracing," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 143:1–143:8, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2601097.2601199

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

[60] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE T-PAMI*, 2018.

[61] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut -interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (SIGGRAPH)*, August 2004.

[62] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zerong Zheng received the B.S. degree in Department of Automation, Tsinghua University in July 2018. He is currently pursuing Ph.D. degree in Department of Automation, Tsinghua University, advised by Prof. Yebin Liu. . His current research interests include computer vision and computer graphics.

Tao Yu Tao Yu receieved the B.S. degree in Measurement & Control Technology and Instrumentation from Hefei University of Technology, China, in 2012. He receieved his Ph.D. degree in Precision Instrument and Machinery from Beihang University, China, in 2019. He then joined Tsinghua University as a postdoc. His research lies at the intersection between computer vision, computer graphics and machine learning.

Yebin Liu Yebin Liu received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 2002 and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He is currently an associate professor with Tsinghua University. He was a research fellow in the Computer Graphics Group of the Max Planck Institute for Informatik, Germany, in 2010. His research areas include computer vision, computer graphics, and computational photography.

Qionghai Dai received the M.S. and Ph.D. degrees in computer science and automation from Northeastern University, Shenyang, China, in1994 and 1996, respectively. He is currently a professor in the Department of Automation and the director of the Broadband Networks and Digital Media Laboratory, Tsinghua University,Beijing. He is also an academician of Chinese Academy of Engineering. He has authored or co-authored more than 200 conference and journal papers and two books. His research interests include computational photography and microscopy, computer vision and graphics, and intelligent signal processing. He is associate editor of the Journal of Visual Communication and Image Representation, the IEEE Transactions on Neural Networks and Learning Systems, and the IEEE Transactions on Image Processing.