Pretrained Pixel-Aligned Reference Network for 3D Human Reconstruction

Gee-Sern Hsu Yu-Hong Lin Chin-Cheng Chang
National Taiwan University of Science and Technology, Taipei, Taiwan
{jison, m10903430, m11003416}@mail.ntust.edu.tw

Abstract

We propose the Pretrained Pixel-aligned Reference (PPR) network for 3D human reconstruction. The PPR network utilizes a pretrained model embedded with a reference mesh surface and full-view normals to better constrain spatial query processing, leading to improved mesh surface reconstruction. Our network consists of a dual-path encoder and a query network. The dual-path encoder extracts front-back view features from the input image through one path, and full-view reference features from a pretrained model through the other path. These features, along with additional spatial traits, are concatenated and processed by the query network to estimate the desired mesh surface. During training, we consider points on the pretrained model as well as around the ground-truth mesh surfaces, enabling the implicit function to better capture the mesh surface and overall posture. We evaluate the performance of our approach through experiments on the THuman 2.0 and RenderPeople datasets, and compare it with state-of-the-art methods.

1. Introduction

Humans are central subjects in images and videos. Image-based human modeling leads to a myriad of applications from medical imaging to virtual reality. The development of image-based human reconstruction approaches is a major topic in the fields of computer vision. It is a challenging task due to the difficulty in maintaining high fidelity with holistic human representation during the transformation from 2D image to 3D space. The difficulty mostly depends on the 3D representation considered for the reconstruction and hardware limitations. The global constraints have been solved by some approaches, such as the parametric 3D geometric representation [9] and the voxel-based volumetric method [21]. For capturing holistic body meshes, the Skinned Multi-Person Linear Model (SMPL) [9], which is a parametric 3D human body model made by skinning and blend shapes learned from thousands of 3D body scans, which produces human model with facial and body shape details but without clothing in appearance. The voxel-based representation [21] can solve the clothing appearance issue but requires expensive hardware to handle the training of volumetric CNNs.

Parametric models are proposed to generate 3D human meshes with various postures. For example, the SMPLify [4] and HMR [7] estimate the shape and pose coefficients of a SMPL model [9] for an input image. These parametric human body models are made for reconstructing the body posture, rather than the reconstruction of the actual clothed surface. To better reconstruct clothed surface, the deep implicit function approaches [15, 16] make good progress in the reconstruction of 3D clothed human meshes from one single human images. Instead of explicitly parameterizing the output representation, these methods regress a function to determine the surface occupancy for a 3D location, and show its strength in reconstructing high-fidelity 3D geometry without keeping the entire output volume. Recent work [18, 20] combine parametric model with implicit function modeling for better posture reconstruction.

Different from the existing approaches that consider implicit functions for 3D clothed human reconstruction, we leverage a pretrained model built on an implicit function to develop another and better implicit function model. We propose the Pretrained Pixel-Aligned Reference (PPR) Network, which is the first modeling approach that considers an off-the-shelf implicit function model, such as PIFu [15] or PIFuHD [16] as a pretrained model, to reconstruct the 3D clothed human shape for a 2D input image. The RRP net is built on a two-path architecture which captures the front-back view features of the human subject in the image by one path, and the full-view features obtained from the pretrained model by the other path. Both features are concatenated with additional spatial traits and entered to a query network that is built upon a Multilayer Perceptron (MLP) to determine the spacial occupancy of a 3D point, leading to the reconstruction of the desired 3D mesh shape. Our approach is verified by the experiments on benchmark datasets and compared with state-of-the-art methods.

The contributions of this work can be summarized as follows:

- The first approach that explores a trained implicit func-

tion as a pretrained model is proposed and verified through experiments.

- The dual-path architecture with two image encoders that extract different image features is verified effective for reconstructing the 3D mesh surface for a 2D image.

- The proposed pretrained reference sampling scheme is verified more effective than the sampling undertaken by other implicit function modeling approaches.

- The proposed approach demonstrates better performance than state of the art, especially for human subjects with non-standing postures which are generally considered as challenging cases.

The rest of this paper is organized as follows. A brief review on previous work is given in Sec. 2. The proposed approach is presented in Sec. 3, followed by the experiments on benchmark datasets presented in Sec. 4. A conclusion to this study is given in Sec. 5

2. Related Work

2.1. Parametric model

3D human body reconstruction from a single 2D image is challenging because of the lack of the depth information about the whole body and between the body parts. To tackle this challenging task, the parametric 3D human model [4] is proposed to impose body constraints on the estimated model. The estimated parametric model can only capture a standard human body template to supplement the depth information that a single RGB image cannot provide. However, the parametric model cannot capture the various body postures and the clothed surfaces shown on the 2D image. While using multiple parametric models can be more effective representing a human body, it is difficult to use multiple shape representations to handle various geometrical postures and clothed surfaces.

Octopus [2] model reconstructs a 3D shape, including the parameters of SMPL plus clothing and hair in a canonical T-pose space. Tex2Shape [3] turns the shape regression into an aligned image-to-image translation problem by using the UV-mapping to unfold the SMPL body surface onto a 2D image. Lazova et al. [8] learn to complete the full 3D texture from partial texture map and generalize to new poses, shapes and viewpoints with SMPL and traditional rendering. BodyNet [17] is an end-to-end trainable network with intermediate network supervision in terms of 2D pose, 2D body part segmentation, and 3D pose. DeepHuman [21] fuses different scales of image features into the 3D space through volumetric feature transformation, which helps to recover accurate surface geometry.

2.2. Deep Implicit Representation

Unlike the parametric models that constrain the solution area or need cubic memory, Chen et al. [5] advocate the use of implicit fields for learning generative models of shapes and introduce an implicit field decoder. Mescheder et al. [10] encodes a description of the 3D output at infinite resolution without excessive memory footprint.Park et al. [12] introduce DeepSDF, a learned continuous Signed Distance Function representation of a class of shapes that enables high quality shape representation, interpolation and completion from partial and noisy 3D input data. Saito et al. [15] propose the implicit function with pixel-aligned for 3D clothed human reconstruction from RGB images. However, features from single RGB image can not provided sufficient information for deep implicit function, and it cause the result often produces artifacts including broken limbs, depth error and geometric noise. To advance their method, [16] input high resolution images and inference front-back normal map to provide more feature to deep implicit function, these features improve the result surface detail and decline the artifacts in previous work. Geo-PIFu [6] is based on a deep implicit function-based representation to learn latent voxel features using a structure-aware 3D U-Net. PaMIR [20] extract features from voxelized SMPL mesh to improve the generalization ability under the scenarios of challenging poses and various clothing topologies. ICON [18] use SMPL fit body to guide the front-back normal map inference and the deep implicit function. These works are robust to varied human poses and decrease the artifacts in the original PIFu [15]. However, the high dependence on SMPL body makes its results severely affected by wrong information form fitting results of the parametric model.

3. Our Approach

As our approach is developed based on PIFu [15] and PIFuHD [16], we first summarize their approaches in Sec. 3.1, and present ours in Sec. 3.2.

3.1. A Brief Review to PIFu and PIFuHD

The core part of PIFu [15] (Pixel-aligned Implicit Function) is an implicit function $f(\mathbf{v}, I)$ defined for a 3D point $\mathbf{v} = (v_x, v_y, v_z) \in \mathcal{R}^3$ and an associated image I as follows.

$$f(\mathbf{v}, I) = \begin{cases} 1, & \text{if } \mathbf{v} \text{ inside the mesh surface.} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

To estimate the implicit function f, PIFu [15] explores the following model,

$$f(\mathbf{v}, I) = g(h(\mathbf{v}_{x,y}, I), z_v) \quad (2)$$

where $\mathbf{v}_{x,y} = (v_x, v_y)$ is the orthogonal projection of $\mathbf{V}$ onto the x-y image plane, $h(\cdot)$ is a feature embedding func-

tion built by a CNN, z_v is the depth along the ray from $\mathbf{v}_{x,y}$, and g is a Multi-Layer Perceptron (MLP) that determines the right z_v^* so that (2) holds.

As the 3D representation of PIFu [15] is bounded by the 512×512 input image resolution, the PIFuHD [16] considers 1024×1024 input image resolution with a double-level approach for the reconstruction of higher resolution. A coarse level, similar to PIFu [15], focuses on integrating the global geometry and producing backbone image features of 128×128 resolution. A fine level adds on more details by using the 1024×1024 input image to produce backbone image features of 512×512 resolution. To predict the back of a subject, PIFuHD [16] explores a front-to-back inference scheme by predicting the normal maps of the frontal and back sides of the subject, and entering the predicted normal maps as part of the features for the 3D reconstruction. For the loss function considered at training, PIFuHD [16] uses the Binary Cross Entropy (BCE) loss instead of the L1 and L2 losses in PIFu [15].

3.2. Pretrained Pixel-Aligned Reference Model

The differences between our approach and PIFu/PIFuHD are threefold. The first is the 512×512 input image resolution, same as for PIFu, but the expressiveness and details of our reconstructed 3D surface can be competitive to those of PIFuHD, substantially reducing the hardware requirements for handling high resolution reconstruction. The second is the two-path architecture which extracts the front-back features from the input image I by one path, and extracts the full-view normal map features from a pretrained model by the other. The third is the pretrained reference sampling scheme that samples the 3D points along the pretrained model mesh surface and imposes large variances to the points with large errors from the ground truth at training.

3.2.1 Two-Path Architecture with Pretrained Model

Fig. 1 shows the configuration of our model. It consists of a front-back view image encoder, a full-view image encoder, and a query network made of a multi-layer perceptron (MLP). The front-back image encoder, which is made of a Stacked Hourglass (SHG) with 4 sub-modules and each sub-module composed of 8 convolution blocks, and functions similarly to the coarse-level network in PIFuHD, extracts the feature map $\Phi \in \mathcal{R}^{256^3}$ from the input image $I \in \mathcal{R}^{512^2\times 3}$ and the frontal and back normal maps $N_f \in \mathcal{R}^{512^2\times 3}$, $N_b \in \mathcal{R}^{512^2\times 3}$. The full-view image encoder, which is made of a U-Net with 5 downscaling double convolution layers and 5 upscaling double convolution layers, extracts the feature vector $\Psi \in \mathcal{R}^{512^2\times 128}$ from the full-view normal maps $N_r^p \in \mathcal{R}^{512^2\times 3}$, $N_l^p \in \mathcal{R}^{512^2\times 3}$ obtained from the pretrained model. Considering a 3D point $\mathbf{v} = (v_x, v_y, v_z)$ aligned to a pixel at $\mathbf{v}_{x,y} = (v_x, v_y)$ on the input I, the MLP takes as input of the aligned features $\phi_{x,y} \in \mathcal{R}^{256}$, $\psi_{y,z} \in \mathcal{R}^{128}$ (from the two feature maps Φ and Ψ, respectively) concatenated with the coordinates v_x, v_z^p and the depth difference $d_z = v_z - v_z^p$, where v_z^p is the depth at the pixel-aligned mesh of the pretrained model, and generates a scalar output $u \in [0, 1]$. $\mathbf{v}$ is considered outside of the mesh surface if $u > \tau_u$, where τ_u is a threshold determined in the experiment, or inside of the mesh surface if otherwise.

The details of our approach can be organized into the following steps:

1. Enter the input image I to a pretrained model M_p to obtain the pretrained 3D mesh $M_p(I)$ and the pretrained normal maps N_f^p, N_b^p, N_l^p and N_r^p for the frontal, backside, left-side and right-side views, respectively. Additionally, we also consider the frontal and backside normals, N_f, N_b, obtained by the image-to-image translation as in [16] for comparison purpose.

2. The two-path structure is composed of two component networks h_1 and h_2 for extracting respectively the front-back feature map Φ and full-view feature map Ψ, which can be written as follows:

$$\Phi(I) = h_1(N_f, N_b, I) \tag{3}$$

$$\Psi(I, M_p(I)) = h_2(N_f^p, N_b^p, N_l^p, N_r^p, I) \tag{4}$$

3. For a 3D point $\mathbf{v} = (v_x, v_y, v_z)$, we first compute the depth difference $d_z = v_z - v_z^p$, and sample the pixel-aligned image feature vectors from $\Phi(I)$ and $\Psi(I)$, namely $\phi_{x,y} = h_1(v_{x,y}, N_f, N_b, I)$ and $\psi_{y,z} = h_2(v_{x,y}, N_f^p, N_b^p, N_l^p, N_r^p, I)$. The concatenated vector $[\phi_{x,y}, \psi_{y,z}, v_x, v_z^p, d_z]$ enters the MLP $g(\cdot)$ to determine the occupancy of $\mathbf{v}$. The same marching cube as in [15] is performed to extract the meshes from the 3D occupancy inferred by $g(\cdot)$.

To summarize the above, the implicit function derived from our approach can therefore be expressed as follows,

$$F(\mathbf{v}, I, M_p(I)) = g(\phi_{x,y}, \psi_{x,y}, v_x, v_z^p, d_z) \tag{5}$$

where $\phi_{x,y} = h_1(v_{x,y}, N_f, N_b, I)$, $\psi_{x,y} = h_2(v_{x,y}, N_f^p, N_b^p, N_l^p, N_r^p, I)$

3.2.2 Sampling with Pretrained Model Reference and Loss Function

Based on the spatial sampling conducted in [15] and [16], we propose an improved spatial sampling scheme. The sampling in [15] and [16] combines uniform sampling and adaptive sampling based on the surface geometry. In the training phase, they randomly sample points on the mesh

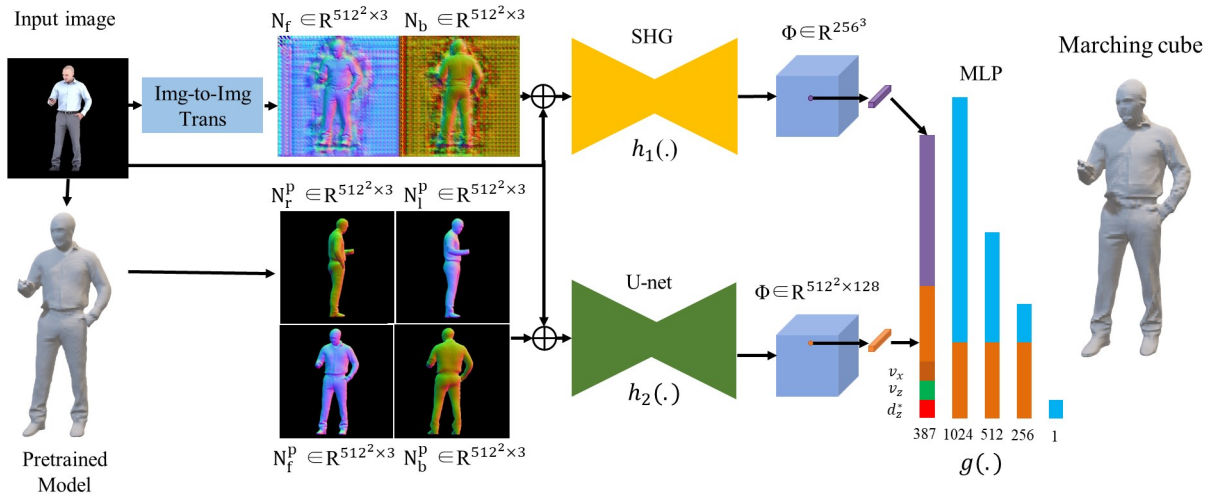

Figure 1. The proposed Pretrained Pixel-Aligned Reference Network is composed of two feature encoders $h_1(.)$, $h_2(.)$ and MLP, each layer of the MLP makes a skip connection with the full view feature from $h_2(.)$.

surface as the ground truth model is available, and add in offsets with Gaussian distribution with a constant variance to perturb their positions around the mesh surface. These sample points are blended with another set of uniformly sampled points within the bounding box that encloses the human subject. We found that this sampling scheme can be difficult handling partial occlusion cases, and propose the following improved sampling scheme.

1. As the pretrained model is available in our framework, we first sample points from the pretrained mesh surface.

2. Calculate the amount of difference between the sampled point and the ground truth (gt) mesh surface.

3. Offset the surface sample points using the calculated difference as the variance of the normal distribution.

The above scheme offers the varying variance which is relatively large at the points that differ significantly from the ground truth, and is small at the points that differ insignificantly from the ground truth. When inferencing, the points with large/small variance can move for a large/small distance, resulting in a better spatial occupancy accuracy.

Considering a set of sampling points $\mathbf{V}_n = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n]$, we use the following binary cross entropy loss L to train our model,

$$L = \sum_{i=1}^{n} (F_{gt}(\mathbf{v}_i) \log F(\mathbf{v}_i, I, M_p(I)) + (1 - F_{gt}(\mathbf{v}_i)) \log(1 - F(\mathbf{v}_i, I, M_p(I))) \quad (6)$$

$F(\mathbf{v}_i, I, M_p(I))$ is the point occupancy prediction made by $g(\cdot)$ for $\mathbf{v}_i$, $F_{gt}(\mathbf{v}_i)$ is ground-truth occupancy value which is 1 for inside point and 0 for outside point.

4. Experiments

We first introduce the databases, the experimental settings and the implementation details in Sec. 4.1, then the evaluation metrics in Sec. 4.2, and then the comparison with the state of the art in Sec. 4.3.

4.1. Dataset and Experimental Settings

We trained our method and retrained other methods on the same 450 scans from THuman dataset [19] for a fair comparison, and used the SMPL-X meshes provided by Thuman dataset for the methods that requires 3D body prior. The 3D mesh of each subject is rendered from every other degree along the yaw direction with an elevation fixed with $0°$. We evaluated the performance on 245 scan meshes selected from Renderpeople [1].

For the two-path encoders, $h_1(\cdot)$ is made of an Hourglass network [11] and $h_2(\cdot)$ is made of a U-Net [14], as described in Sec. 3.2. The configurations of both networks are shown in Fig. 2 and 3. The input image I is 512^2 pixels in resolution. The front and back surface normal images that enter $h_1(\cdot)$ are made by the surface normal inference network provided by the PIFuHD [16]. The feature map $\Phi(I)$ from $h_1(\cdot)$ is 256^3 in dimension, and the feature map $\Psi(I, M_p(I))$ from $h_2(\cdot)$ is $512^2 \times 128$. The MLP query network $g(\cdot)$ has an input layer of 387 dimension, three hidden layers with 1024, 512, 256 neurons, and one scalar output. During training, we sampled 8,000 points along the mesh surface of the pretrained model perturbed with zero-

mean Gaussian and minimum variance 0.1 and maximum 0.3. Our network was trained with a batch size of 2, we use RMSProp with learning rate to 0.003 and weight decay by a factor of 0.1 on a single Nvidia RTX 3090 GPU.

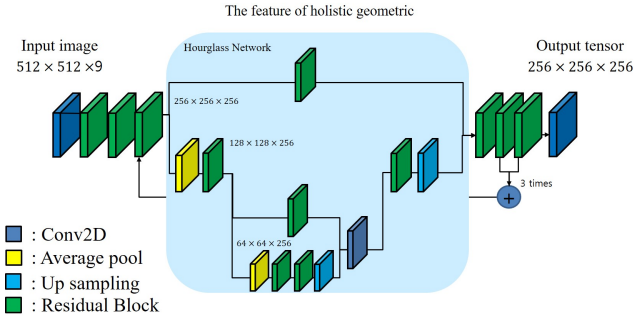

Figure 2. Hourglass network configuration for making the image encoder $h_1(\cdot)$

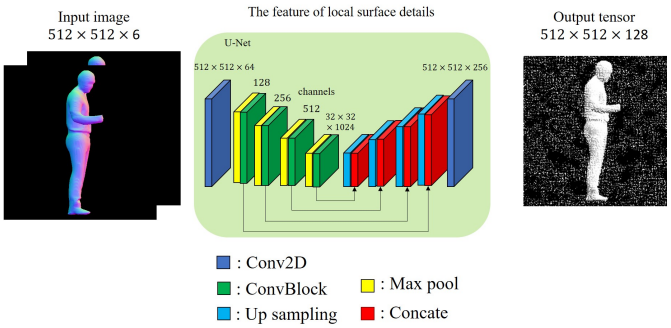

Figure 3. U-Net configuration for making the image encoder $h_2(\cdot)$

4.2. Evaluation Metrics

Different data format requires different evaluation metrics. Parametric methods [4, 7, 13] compute the standard mean 3D body joint error and the vertex offset between the predicted model and the ground truth. However, for the spatial occupancy estimation methods [15, 21], the estimated 3D models can be directly compared with the ground truth to examine the similarity and global quality by using the Chamfer Distance (CD) and the Points To Surface Distance (P2S).

The CD calculates the distance between the closest points across two meshes. It is calculated by summing the distance between each point in one mesh and the closest point in the other, and dividing this sum by the number of points in the mesh to obtain an average distance. The smaller the CD, the smaller the difference between the reconstructed mesh and the ground-truth mesh, and the better the reconstruction quality. The CD is defined as follows,

$$D_{CD}(P,G) = \frac{0.5}{|P|}\sum_{p\in P} d_G(p) + \frac{0.5}{|G|}\sum_{g\in G} d_P(g) \quad (7)$$

where P and G represent the predicted and ground truth vertices, respectively; $d_A(B)$ is a minimum distance function based on Set-A to evaluate the distance from Set-B.

The P2S metric utilizes unidirectional distance to evaluate the offset between the predicted vertices and the closest surface of the ground truth mesh. It can be calculated as follows:

$$D_{p2s}(P, M_g) = \frac{1}{|P|}\sum_{p\in P} d_{M_g}(p) \quad (8)$$

where M_g is ground truth model, P is the predicted vertices, and d_{M_g} is minimum distance from P to the closest surface on the M_g.

4.3. Comparison with Other Approaches

We first use PIFu as our pretrained model and compare the results with other state-of-the-art methods, including PIFu [15], PIFuHD [16], PaMIR [20], and ICON [18]. The comparison with different pretrained models is reported in the next section. To ensure a fair comparison, we compare the pretrained weights from the official Github repository with our retrained weights obtained by using the aforementioned methods. Table 1 presents the quantitative comparisons. Our approach outperforms other methods. Samples of the qualitative results are demonstrated in Figure 4. Our approach yields better reconstructed mesh surfaces and better posture.

Experiments with different settings of the proposed PPR network are reported in the next section.

Method	Chamfer	P2S	sec/item
PIFu [15]	2.26	2.14	**15.76**
PIFu	2.6	2.47	
PIFuhd [16]	2.09	2.02	18.11
PaMIR [20]	2.07	2.03	17.01
PaMIR	2.23	2.16	
ICON [18]	1.96	1.84	65.34
ICON*	2.05	1.93	
Ours	**1.81**	**1.84**	86.14

Table 1. Quantitative comparison on RenderPeople for single-view reconstruction. Ours method is PPR using Pifu as pretrained reference. We retrained other SOTA method in the same dataset and label with star. Units for point-to surface and Chamfer distance are in cm

4.4. Ablation Study

We conducted an ablation study to examine different settings of the proposed approach, including different setups

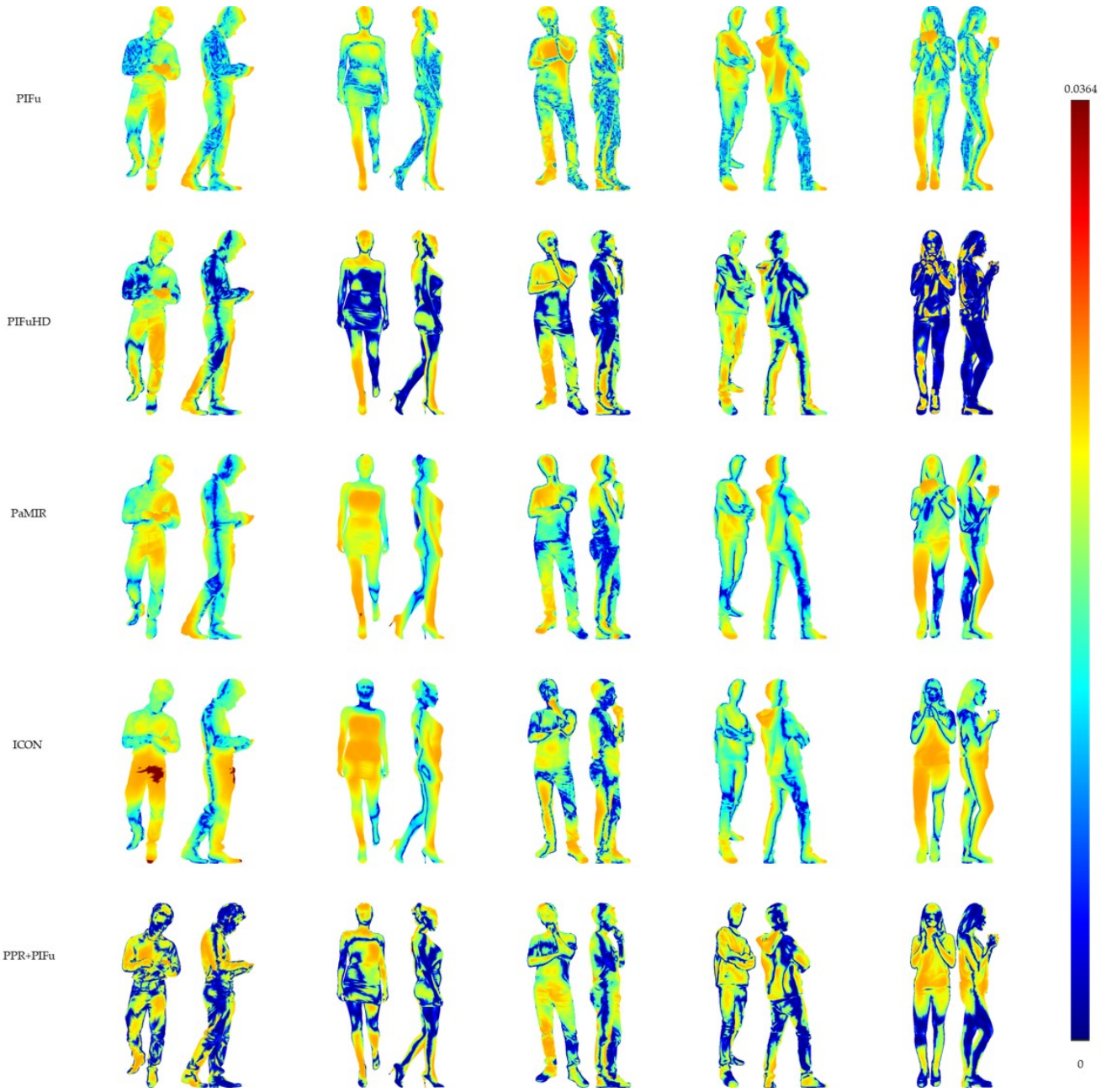

Figure 4. Errors in the reconstructed meshes to the ground truth, scaled to the color bar on the right.

of the two-path encoders, different sampling schemes and embedding with different pretrained models.

Table 2 shows the performance with different settings of the PPR network, including using only the encoder $h_1(\cdot)$ only, both encoders using the same designs ($h_1(\cdot) + h_1(\cdot)$ for both being $h_1(\cdot)$ or $h_2(\cdot)$), and for the case that $h_1(\cdot)$ and $h_2(\cdot)$ exchanges designs. Fig. 5 shows samples for a qualitative comparison. We also show the performance with different traits added to the image feature, including without d_z^* (the distance from the query point to the pretrained model), and with 2-dimensional coordinates (v_x, v_z) and 3-dimensional coordinates (v_x, v_y, v_z). It is verified that the addition with d_z^* and 2-dimensional (v_x, v_z) yields the best performance.

We also compared the Gaussian perturbation with constant variance on the spatial sampling points. As shown in the middle part of Table 2, large constant variance leads to large discrepancy, and the most appropriate setting is the

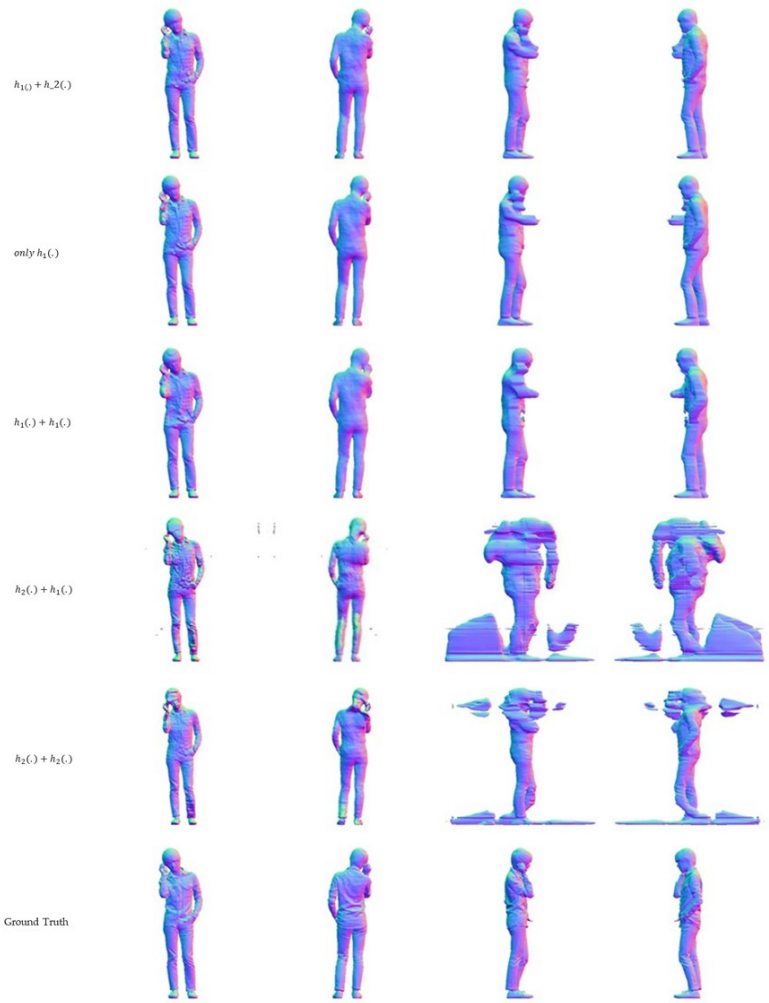

Figure 5. Qualitative comparison of different setups of the two-path image encoders.

proposed sampling scheme described in Sec. **??**. The bottom part of Table 2 shows the comparison with different pretrained models, namely the PIFu, PIFuHD and SMPL-X. Fig. 6 shows samples to demonstrate the 4-view normals of using different pretrained models for a qualitative comparison. It is experimentally verified that the proposed settings yields the best performance.

Our method is capable of using any pretrained results for training. We conducted experiments using pretrained results from PIFu, PIFuHD, and SMPL-X to evaluate our approach. We found that PIFu and PIFuHD suffer from a lack of depth features, resulting in similar performance when used as PPR models. However, using SMPL-X as the PPR model yielded better pose fit for some reconstructions that required deeper information. Nonetheless, poor fitting results in certain standing poses also impacted the reconstruction performance. To further improve our method, we trained it on the Thuman dataset.

5. Conclusion

Our Pretrained Pixel-Aligned Reference (PPR) network takes a different approach to 3D human reconstruction than previous methods that estimate implicit functions. Instead, our network uses a two-path architecture that incorporates a pretrained model based on an implicit function. This allows us to leverage the advantages of an off-the-shelf pretrained model and use four-view normals as powerful references. The two-path architecture uses different encoders along each processing path to extract different features. These encoders form another set of implicit functions that enable us to better reconstruct the 3D mesh surface from a single 2D image. Our experiments on benchmark datasets show that the proposed approach outperforms other methods.

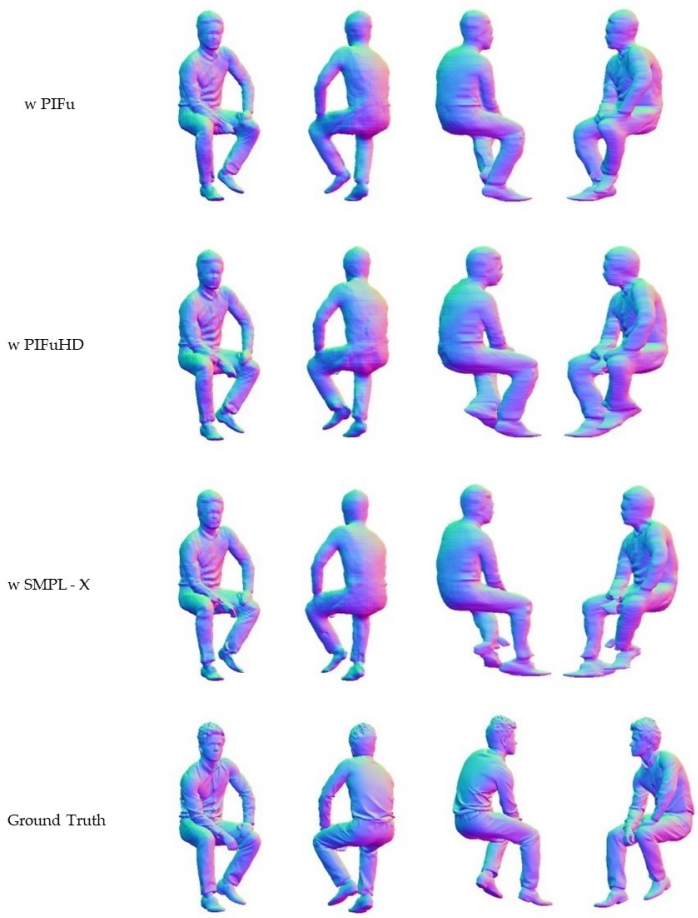

Figure 6. The 4-view normal maps for different pretrained models.

Method	Chamfer	P2S
Only $h_1(\cdot)$	2.02	2.01
$h_1(\cdot) + h_1(\cdot)$	1.94	2.00
$h_2(\cdot) + h_1(\cdot)$	11.54	17.23
$h_2(\cdot) + h_2(\cdot)$	5.45	7.82
wo d_z^*	1.91	1.88
(v_x, v_z)	**1.81**	**1.84**
(v_x, v_y, v_z)	1.84	1.86
$\sigma = 5$	1.89	1.88
$\sigma = 10$	1.87	1.88
$\sigma = 15$	1.92	1.94
PPR-PIFu	**1.81**	**1.84**
PPR-PIFuHD	1.82	1.83
PPR-SMPL-X	1.90	1.91

Table 2. Ablation Study results. 1) Different network settings and spatial features components. 2) constants standard deviation settings for sampling. 3) Using different model as the PPR result.

Moreover, as the PPR network only considers the four-view normals of the pretrained model, other characteristics of the model could also be useful. We are currently exploring the use of the mesh surface of the pretrained model to further improve performance, and we will share our results as soon as they become available.

References

[1] Renderpeople. https://renderpeople.com/, 2018. 4

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2293–2303, 2019. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 1, 2, 5

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[6] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072*, 2020. 2

[7] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 5

[8] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 2

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 1

[10] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 4

[12] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[13] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 5

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[15] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 3, 5

[16] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1, 2, 3, 4, 5

[17] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 2

[18] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13296–13306, 2022. 1, 2, 5

[19] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 4

[20] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021. 1, 2, 5

[21] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 1, 2, 5