

PIFu for the Real World: A Self-supervised Framework to Reconstruct Dressed Human from Single-view Images

Zhangyang Xiong^{1,2}, Dong Du², Yushuang Wu², Jingqi Dong², Di Kang³,
Linchao Bao³, and Xiaoguang Han^{2,1*}

¹ FNii, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China

² SSE, The Chinese University of Hong Kong-Shenzhen, Shenzhen, China

³ Tencent AI Lab, Shenzhen, China

Abstract. It is very challenging to accurately reconstruct sophisticated human geometry caused by various poses and garments from a single image. Recently, works based on pixel-aligned implicit function (PIFu) have made a big step and achieved state-of-the-art fidelity on image-based 3D human digitization. However, the training of PIFu relies heavily on expensive and limited 3D ground truth data (i.e. synthetic data), thus hindering its generalization to more diverse real world images. In this work, we propose an end-to-end self-supervised network named SelfPIFu to utilize abundant and diverse in-the-wild images, resulting in largely improved reconstructions when tested on unconstrained in-the-wild images. At the core of SelfPIFu is the depth-guided volume-/surface-aware signed distance fields (SDF) learning, which enables self-supervised learning of a PIFu without access to GT mesh. The whole framework consists of a normal estimator, a depth estimator, and a SDF-based PIFu and better utilizes extra depth GT during training. Extensive experiments demonstrate the effectiveness of our self-supervised framework and the superiority of using depth as input. On synthetic data, our Intersection-Over-Union (IoU) achieves to 89.03%, 20% and 28.6% higher compared with PIFuHD and ECON, respectively. For in-the-wild images, our method excels at reconstructing geometric details that are both rich and highly representative of the actual human, as illustrated in Fig. 1 and 11.

Keywords: Depth Estimation · Human Reconstruction · Implicit Function · Self-supervised Learning.

1 Introduction

Image-based human digitization has gained considerable attention in the last decades. It is widely used in games, telepresence, and VR/AR applications. To recover accurate 3D human shapes from sparse 2D observations, various models are proposed, such as parametric models [3, 19, 20, 23, 28, 36, 38, 40], silhouetted models [32], volumetric models [51, 64], and implicit models [8, 43, 44]. Among them, PIFuHD [44] produces high-fidelity 3D reconstruction with impressive geometric details such as wrinkles of clothes, which achieves the state-of-the-art.

Compared with the original version PIFu [43], PIFuHD [44] proposes a two-level pixel-aligned implicit function learning framework for high-resolution 3D reconstructions, where additional normal maps are generated and integrated for fine-grained detail

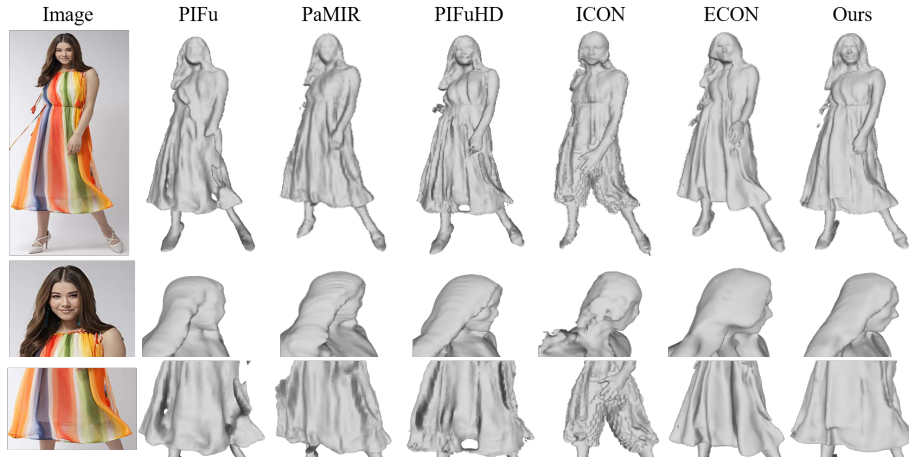


Fig. 1. Comparisons among different methods. PIFu takes a single color image as input (second column). PIFuHD takes both image and predicted-normal maps as input (third column). PaMIR, ICON, and ECON all utilizes SMPL prior (from fourth to sixth line), while ECON also introduces depth maps to optimize the coarse mesh. Ours takes a predicted-depth map as input (seventh column). The predicted-depth maps are shown in right. Ours contains more completed and reasonable geometry detailed.

preservation. However, PIFuHD still suffers from two issues: (i) PIFuHD relies heavily on delicately created 3D ground truth supervision and normal map, while the existing dataset, *e.g.* RenderPeople [1], contains only a few hundred static models covering limited identities, poses, and complex garment geometry, resulting in performance degradation when applied on in-the-wild images; (ii) the image-to-normal-to-shape reconstruction of PIFuHD is constructed with normal maps as the intermediate while the normal estimation is sensitive to texture and shadow in the input image [16, 49, 54].

Considering the limited synthetic normal data cannot cover all varieties of cloth textures and complex shadows in the real world, estimating normal maps from real world images is usually difficult and error-prone. The implicit function may not handle these errors in normal map which lead to artifacts in shape reconstruction quality, as illustrated in Fig. 8 and 11. Thus, the heavy reliance on 3D ground truth and the poor robustness of normal estimation of PIFuHD motivates us to rethink the whole reconstruction procedure. Compared to using normal maps, depth maps, which contain not only rich and detailed shape information but also *directly* define the surface, are able to provide spatial supervision beyond the surface and make it possible to learn a more promising signed distance field (SDF). What’s more, depth estimators generalize well from synthetic data to in-the-wild images (see Fig. 6).

Based on the above observations, we proposed to utilize depth estimation as our auxiliary (intermediate) input to the PIFu network, and further convert depth into effective supervisions that enable our self-supervised PIFu (SelfPIFu). SelfPIFu consists of a normal estimator, a depth estimator and a novel self-supervised SDF-based pixel-aligned implicit function (PIFu) learning module that takes depth map as input. The

self-supervised mechanism considers two kinds of supervisions, *i.e.*, a volume-aware and a surface-aware one, to optimize PIFu using a mixture of synthetic data with 3D human ground truth and in-the-wild images with well-estimated depth maps. Specifically, the volume-aware self-supervision utilizes *volume* supervision from point cloud data, which is transformed from the depth map. The surface-aware self-supervision utilizes a differentiable surface renderer [25] from SDF as supervision. With their help, our SDF-based PIFu effectively learns convincing surface details especially for in-the-wild images. Extensive experiments show that our method outperforms state-of-the-art human digitization approaches on both synthetic and real images.

In summary, the contributions of this work are as follows:

- A novel self-supervised framework, including volume-aware and surface-aware SDF learning, is proposed to reconstruct more accurate geometry especially on real-world images.
- We propose to use depth map to better represent the 3D information from the image for more robust learning of implicit function, which can lead to a higher quality of human reconstruction than using normal map.
- Extensive experiments and analysis on both synthetic data and real images support our claim about the advantages of depth-guided PIFu, and validate the effectiveness and superiority of the proposed framework.

The remainder of the paper is organized as follows. In Sec. 2, we briefly provide a survey of related work. Sec. 3 first presents our depth estimator and our implicit function, then introduce the depth-guided self-supervised framework. Sec. 4 conducts extensive evaluations on our SelfPIFu with SOTA. At last, Sec. 5 gives several concluding remarks.

2 Related Work

2.1 Single-view Human Reconstruction

3D human reconstruction from a single RGB image is inherently challenging due to the view occlusion and shape ambiguity. To address this ill-posed problem, pioneer works (*e.g.* SCAPE [3] and SMPL [28]) propose parametric models that are derived from large scanned human datasets to provide strong priors. However, methods based on parametric models are restricted to producing naked human bodies without garment details [3, 6, 19, 20, 23, 28, 36, 38, 40]. Although [65] introduces a hierarchical mesh deformation network to restore detailed shapes, the results are far from the ground truth. Some other methods attempt to define a garment template mesh and deform it to approach the target shape of clothes [4, 5, 17, 50, 66], but they are limited to generating specified garments and coarse details due to the fixed topology and resolution of the template. Besides these template-based methods, silhouette-based [32] and voxel-based models [51, 64] are proposed for human reconstruction with arbitrary topology and geometry of clothes. Yet, they still produce rough geometry using a low-resolution representation due to the heavy calculation consumption.

Recently, implicit models [7, 30, 37, 61] are applied to image-based 3D reconstruction with an arbitrary resolution and fine-grained details, which dominate the field of

3D reconstruction. PIFu [43] introduces a pixel-aligned implicit function for human digitization which extracts pixel-aligned local features to recover detailed geometry. They further propose PIFuHD [44] with a multi-level architecture to generate higher-resolution details in line with the input image, achieving state-of-the-arts. However, PIFuHD requires sophisticated human models for training while the 3D dataset is extremely limited, making PIFuHD fail when applying to in-the-wild images with challenging human poses and diverse garment topology. In addition, PIFuHD relies on the normal generation that is not as accurate as depth estimation for 3D reconstruction. GEO-PIFu [13] uses a structure-aware 3D U-Net to inject global shape topology into a deep implicit function. Their results seem to surpass PIFu, but they still can't compare to PIFuHD as far as high frequency details are concerned. PaMIR [63] brings in a parametric human model to improve the generalization ability of the implicit model, it loses some geometry details. ICON [60] utilizes SMPL-guided clothed-body normal prediction and local-feature based implicit surface reconstruction to achieve impressive results, especially in extreme poses. The author further propose the ECON [59], which introduces intermediate depth maps based on ICON to achieve better results. However, ICON and ECON need to perform an additional optimization process for each image. In the same year, another work Difu [47] also introduces the depth maps as a middle variable to obtain better results.

In this paper, we utilize a robust depth generator to guide PIFu learning, which presents better generalizability than PIFuHD and achieves better performance than PaMIR, ICON & ECON.

2.2 Single-view Depth Estimation

Single-view depth estimation is a fundamental task in computer vision. Traditional methods attempt to figure out depth values from images based on hand-crafted features or geometric priors, such as stereopsis [39], camera focus [9], shading [31], and normals [33]. They tend to produce artifacts when applied to real images with various noise and varying illumination. Recent studies mainly focus on learning-based methods to obtain accurate depth estimation. Most of these works are about scene analysis and reconstruction [11, 22, 24, 41, 42, 53, 58, 62]. [52] presents a large-scale synthetic dataset for human depth estimation. However, the depth is derived from the SMPL model [28] which misses surface details. [49] separates depth estimation into a smooth base shape and a residual detail shape and design a network with two branches to regress them respectively based on really scanned RGBD dataset of clothed humans. The following work [48] extends it to video-based human depth estimation by first calculating an SMPL model at each video frame (*i.e.* the smooth base shape) and then learning the residual detail shape in a self-supervised manner. Even so, the details [48, 49] are not as realistic as the visualization presented in the images. [16] achieves high fidelity results by jointly learning the depth along with the surface normal and warping local geometry among video frames to enforce temporal coherence. In this paper, we also utilize a normal generation to assist our depth estimation, which can provide more accurate geometry priors for 3D human reconstruction.

2.3 Self-supervised 3D Reconstruction

With the limited availability of 3D data, studies are emerging to directly learn 3D shapes from input images without ground truth supervision. These methods can be roughly grouped into three categories based on differentiable rendering [21], view consistency, and frame consistency, respectively. For the single-view reconstruction, rendering-based methods propose a differentiable renderer to formulate the connection between the inferred 3D surface (*e.g.* voxel, mesh, and implicit field) and its 2D renderings (*e.g.* silhouette, depth, normal, and RGB images), which can serve as image-based supervision by comparing the renderings with the input images [14, 18, 25–27, 35]. View-based methods attempt to apply the geometry consistency of different views to self-supervised learning for multi-view reconstruction [10, 45, 46, 56]. Similarly, Video-based methods utilize the consistency and continuity among video frames to restrict the underlying 3D shapes [16, 29, 48, 55]. To improve the generalization ability of 3D human reconstruction with implicit learning, we first learn a robust depth estimation to provide accurate geometry cues and then apply the inferred depth to supervise SDF learning. We not only utilize a differentiable rendering [25] to produce surface-aware supervision but also propose a novel volume-aware supervision based on the inferred depth, both contributing to obtaining high fidelity clothed human shapes without ground truth supervision.

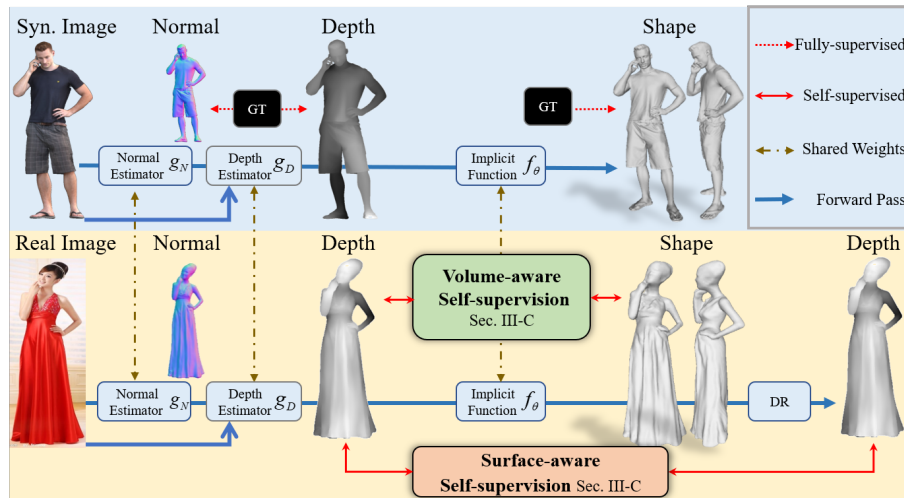


Fig. 2. Overview of our framework. It consists of a normal estimator, a depth estimator, and a depth-guided SDF-based PIFu module that enables self-supervised learning. For synthetic data with 3D ground truth (top row), we use fully-supervised learning method similar to PIFuHD [44]. For in-the-wild images without any labels (bottom row), we first estimate their depth maps, and then optimize the network parameters using a novel depth-guided self-supervised learning method (Sec. 3.3). We visualize the image and point-cloud-like depth map in the middle of frame.

3 Method

An overview of our framework is shown in Fig. 2. Given a single-view image, we first employ a normal estimator and a depth estimator to generate the corresponding depth and then feed the depth into a pixel-aligned implicit function (PIFu) module to predict the 3D human shape. For synthetic data with 3D labels, we use a fully supervised mechanism like PIFu [43] and PIFuHD [44] (top row).

Due to limited image-3D shape pairs as supervision, we propose a novel self-supervised training mechanism, termed as SelfPIFu, to improve the generalization ability of the existing PIFu based methods. Specifically, the inferred depth is further used to provide *volume-aware* and *surface-aware* supervision through depth-guided space sampling and differentiable implicit field rendering. With their help, we can effectively optimize PIFu with in-the-wild images that do not have 3D geometric GT during training, resulting in largely improved generality (bottom row).

3.1 Normal and Depth Estimation

Depth map not only generates 3D cues as direct input for implicit learning but also indirectly provides two self-supervised mechanisms which are volume-aware SDF learning and surface-aware SDF learning for implicit training without 3D ground truth. To obtain accurate depth estimation, we adopt the advanced HDNet [16] that jointly learns an auxiliary normal. *i.e.*:

$$g_N(x; I) = N_{pred}, \quad (1)$$

$$g_D(x; I, N) = D_{pred}, \quad (2)$$

where I represents the input image, where $x \in \mathbb{R}^2$ is the xy-location in the image, N_{pred} and D_{pred} represent the normal map and depth map at the corresponding location respectively. As we know, surface normal is responsive to the local texture, winkle, and shade [49, 54]. Although the geometric information provided by the normal map may be incorrect, it is still useful for subsequent depth map learning. Like the results in [49] and [16].

We train the normal and depth estimator fully supervised on RenderPeople [1] data with ground truth and minimize the following overall loss:

$$L_N = \lambda_{cos} \cos^{-1} \left(\left(\frac{N_{pred}}{\|N_{pred}\|} \right) \cdot \left(\frac{N_{gt}}{\|N_{gt}\|} \right) \right) + \lambda_n \|N_{gt} - N_{pred}\|_1 \quad (3)$$

$$L_D = \|D_{gt} - D_{pred}\|_1, \quad (4)$$

where λ_{cos} and λ_n relative weights between losses, N_{pred} and D_{pred} are predicted normal map and depth map, N_{gt} and D_{gt} represent the ground truth. The depth estimator achieves strong generalization on in-the-wild images, just like Fig. ??, more results will be shown in the next Section.

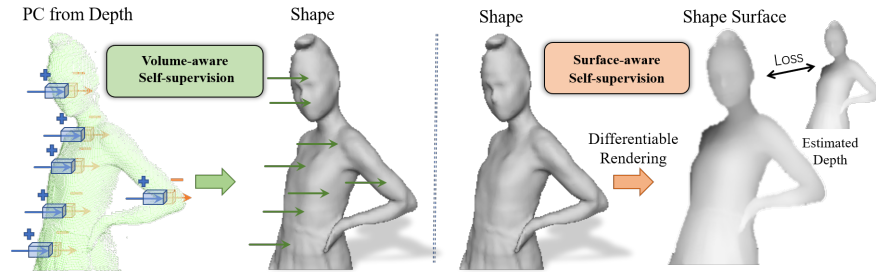


Fig. 3. Depth-guided self-supervised learning. *Volume-aware* self-supervised SDF learning (left) utilizes on-/near-surface points converted from the estimated depth map to provide pseudo *volume* supervision. *Surface-aware* self-supervised SDF learning (right) imposes a *surface-wise* self-supervision by comparing DIST [25] rendered and estimated depth maps. More details are described in Sec. 3.3.

3.2 SDF-based Pixel-aligned Implicit Function from Depth

We use a signed distance function (SDF) instead of occupancy to represent 3D geometry, because SDF field is continuous in the 3D space and can interpolation reasonable details when using marching cube to get a mesh. After we obtain a robust depth estimator to infer geometry information from the input image, the goal of 3D human digitization is to model a function f which can calculate the SDF value s of an arbitrary query point $p \in \mathbb{R}^3$ under the observation of an inferred depth map, *i.e.*:

$$f_{\theta}(g(x, D_{pred}), z(p)) = s, s \in \mathbb{R}, \quad (5)$$

where $x = \pi(p)$ is the 2D projection of query point p , $g(x, D)$ is depth feature at x , $z(p)$ is the depth value of p in the camera coordinate space, and θ is the trainable parameters of our PIFu module.

Our implicit function module f_{θ} consists of a depth encoder (instead of an image encoder) and an SDF decoder. The depth encoder is a fully convolutional network using an hourglass architecture [34], while the SDF decoder is made up of multi-layer perceptrons (MLPs). Implementation details can be found in Sec. 4.

3.3 Depth-guided Self-supervised Learning

The training of PIFu requires plenty of 3D clothed human shapes spanning diverse identities, poses, and complex garment geometry. However, 3D data is limited and expensive to acquire. On the other hand, sufficiently diverse social media images are available online and the state-of-the-art depth estimator usually gives accurate enough depth estimates with convincing geometry details. So we propose two self-supervised mechanisms in Fig. 3 to make best use of the well-estimated depth maps (e.g. Fig. 6) to improve the accuracy, robustness, and generalization ability of the SDF-based PIFu.

Volume-aware self-supervised SDF learning. Inspired by recently neural implicit shape modeling works such as IGR [12], we propose to utilize *volume supervision* provided by the point cloud, which is transformed from the estimated depth map, as pseudo

labels to better optimize/regularize the implicit field. Specifically, we back-project the depth map into the camera coordinate space using orthogonal camera model to obtain 3D points. Because these points should be on the underlying human surface, their SDF values are set to 0. In addition, the variation of SDF value near the surface is supposed to be stable due to the continuity of the SDF field. We can randomly sample N points along the camera view direction with a threshold σ to generate pseudo labels for training. Specifically, we assign points far from the viewpoint with negative values and points close to the viewpoint with positive values. The absolute values are equal to the distances from the sampled points to the surface.

After obtaining the pseudo labels, we utilize them to supervise our PIFu learning with a L_1 loss. Note that SDF is good at preserving geometric details but its learning is a classic regression task that is harder to train than a classification one, e.g., the occupancy learning. Therefore, We increase the penalty for results with opposite signs. Our volume-aware loss function L_{vol} can be defined as

$$L_{vol} = \sum_i^N \|s_i - s_i^{gt}\|_1 + \lambda_m \left(\sum_i^N \|s_i - s_i^{gt}\|_1 \times M_i \right) \quad (6)$$

$$M_i = \begin{cases} 1, & \text{if } s_i \times s_i^{gt} < 0 \\ 0, & \text{if } s_i \times s_i^{gt} \geq 0 \end{cases} \quad (7)$$

where s_i is the SDF value of the sample point p_i , M_i is the opposite signs penalty mask, N is the number of sample points, and the superscript “ gt ” means the ground truth, *i.e.* the pseudo labels.

Surface-aware self-supervised SDF learning. In addition to the *volume* supervision from discrete samples, we also propose continuous *surface* supervision based on an differentiable renderer DIST [25]. DIST uses a differentiable sphere tracing algorithm to render the underlying surface of an SDF field to 2D observations including a depth map. We modify DIST, which originally uses a perspective projection, to be compatible with our SelfPIFu, which uses orthogonal projection⁴ We measure the difference between the DIST rendered depth D_r and the estimated depth D_{pred} (from the depth estimator g_D) with a L_2 loss function:

$$L_{surf} = \|D_r - D_{pred}\|_2. \quad (8)$$

In our experiment, we train the PIFu module with a weighted loss L , *i.e.*

$$L = L_{vol} + \lambda L_{surf}, \quad (9)$$

where λ is a weight of the loss L_{surf} . Both volume-aware and surface-aware self-supervised SDF learning mechanisms contribute to the self-supervision of single-view 3D human reconstruction (see Table 4.2 and Fig. 9).

Difference with PIFuHD. In contrast to PIFuHD, our PIFu module presents three main differences: (i) Different input. It takes the inferred depth (not normal) as input since image-based depth estimation can recover more correct geometry information than the

⁴ We use parallel ray tracing instead of sphere tracing.

normal generation. (ii) Different representation. It adopts SDF to represent a 3D shape since SDF is consecutive and can capture more detailed geometry than an occupancy field. (iii) Different architecture. Due to the contribution of depth and SDF, it uses only one-level PIFu and a shallow network to achieve comparable details with PIFuHD.

4 Experiments

We conduct extensive experiments on both synthetic and real data to compare SelfPIFu with other state-of-the-art (SOTA) methods. Quantitative and qualitative results have justified the effectiveness of SelfPIFu and its individual components.

4.1 Datasets, Metrics, and Implementation Details

Datasets. We use two types of data in our experiments: 1) synthetic clothed human data from RenderPeople with rendered images and their corresponding 3D shape, and 2) real-world images or videos without 3D ground truths. On the synthetic dataset, we use 278 subjects of RenderPeople data for training and 22 subjects for evaluation. Similar to the process in PIFuHD [44], for each training subject, we render an image, a depth map, and a normal map every azimuth degree with varying lighting, producing 360 training triplets in total. The depth maps, which are not utilized in PIFu/PIFuHD, are rendered using OpenGL [2]. On the realistic dataset, we use a total of 15,320 real images, of which 1,400 are collected from the TikTok dataset [16], and 13,920 are crawled from the Internet by ourselves to include more diverse identities, poses, and appearances/clothes, among which 320 images are randomly chosen for evaluation.

Evaluation Metrics. To evaluate the quality of the generated 3D human shape, we adopt commonly used Chamfer distance (CD), point-to-surface distance (P2S), and intersection-over-union (IoU) between the generated shape and the ground truth. For real images without 3D ground truth, we visualize the generated results.

Implementation Details. The design of the normal and depth estimator in SelfPIFu is shown in Fig. 4. In the implicit function f_θ in Fig. 2, the encoder adopts hourglass [34] and uses group normalization [57] instead of batch normalization [15] as in PIFuHD [44]. The feature maps output by the encoder is $128 \times 128 \times 256$, where 256 is the channel dimension, and is fed into an MLPs with $\{257, 512, 256, 128, 1\}$ neuron(s) each layer with skip connections at the second and third layers.

For the training on the synthetic data, we train the network for 100 epochs using an Adam optimizer with a learning rate (LR) of 0.0001 and batch size of 8. After the whole framework is pretrained, we freeze the normal and depth estimator, and finetune the implicit function f_θ using the proposed self-supervised training mechanism on both the synthetic data and real world images. In this stage, the LR is reduced to one-tenth and the batch size is set to 2. All images used for training are resized to 512×512 as in PIFuHD. In the implicit function training, we sample 8000 points with a mixture of uniform sampling and importance sampling around the surface with standard deviations of 3cm. In the self-supervised training, λ is set to 0.618. λ_{cos} and λ_n are set to 1.25 and 1 in (3). The threshold σ in 3.3 is set to 1.5cm.

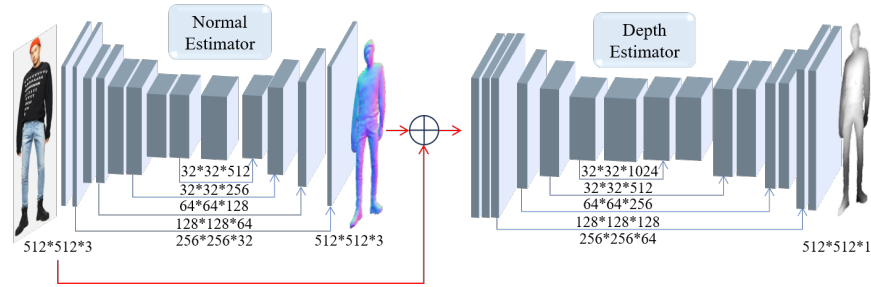


Fig. 4. The structures of normal and depth estimator. All kernel size of convolutions set to $3*3$.

We follow the [16] to train the depth estimator. It is worth noting that during the training process of the depth estimator, incorporating a normal map in the network’s input and adding constraints from the surface normal map to the depth map can effectively improve the accuracy of depth map estimation.

4.2 Evaluations

Depth v.s. Normal as Input. We first experimentally validate our claim that the depth map is a better choice as input for human reconstruction.

Comparison on Synthetic data: Two groups of quantitative experiments are performed on the RenderPeople dataset to show the superiority of using depth as input over using normal as input. Notice that in order to avoid the influence of other factors, such as normal estimator and depth estimator, in *Depth v.s. Normal as Input* part, all inputs are ground-truth data rendered from textured meshes.

In the first group, we test the usefulness of various input combinations, including RGB image (denoted as “I”), normal map (denoted as “N”), depth map (denoted as “D”), and their combinations (e.g. “ID” for using image and depth as input). We follow the standard fully-supervised training process of PIFu [43] (i.e. w/o the proposed self-supervised training branches in Sec. 3.3) in these experiments. The reconstructed shape of each variant is evaluated by the IoU with the ground truth. During the training process, we randomly sample 10k near-surface points from the GT mesh of the synthetic data, and computed the IoU on the training and testing sets, respectively.

The reconstruction results measured with IoU are shown in Fig. 5. In training set, using only image (i.e. “I”) as input results in 91.8% IoU. Including normal as input (“IN” or “N”), the accuracy increases to around 94.0%. Including depth as input (“D”, “ND”, “IND”), the accuracy increases to around 95.8%. Clearly, including depth as input significantly improve the reconstruction IoU. One possible reason is that the gap between depth and 3D geometry is smaller than the gap between normal/image and 3D geometry, which means depth is more informative than normal and normal is more informative than image for this shape reconstruction. When an more informative exists in the input, the extra less informative input modality can only provide minimal gain (e.g. 95.8% for “D” vs 95.8% for “ID”, 95.9% for “ND” vs 95.9% for “IND”)

Compare with training set, we care more about the results on testing set. Similar trend has been observed on the test set while the gap between depth and other modality

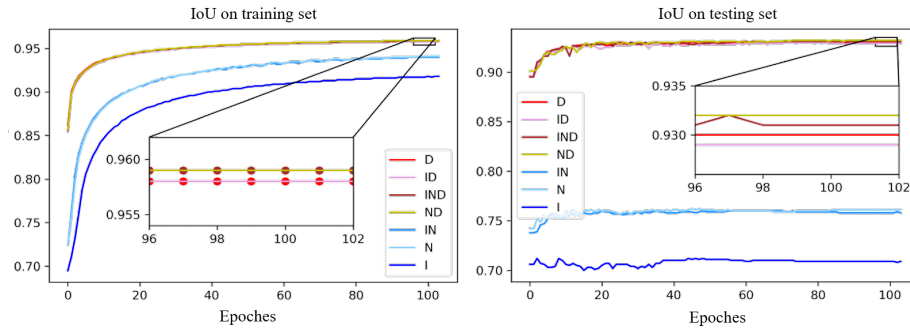


Fig. 5. Comparison of using different input information as input, including image (denoted as “I”), normal map (“N”), depth map (“D”), and their combinations (e.g. “ID” for image and depth). Training (top) and test (bottom) loss curves during training are plotted.

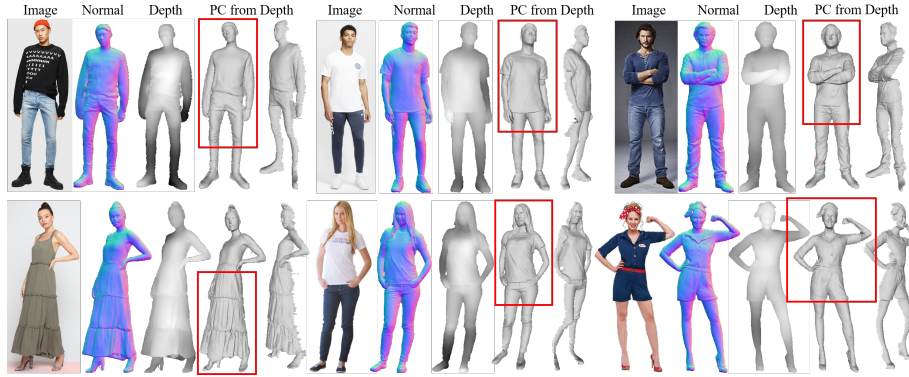


Fig. 6. Visual comparison of the estimated normal maps and depth maps on real world images. Even when tested on in-the-wild images, the estimated depth maps and point clouds (PC) contain high-fidelity details, which are effectively transferred into the final 3D geometry by the proposed depth-guided self-supervised learning module. See Sec. 3.3 for details and Fig. 8 for comparisons.

becomes even bigger, demonstrating that including depth as input could substantially improve the generalization ability. For example, the IoU is above 93.0% when including depth as input. While the IoU drastically drops under 76% when removing depth as input. For example, when the input is a single image, the reconstruction IoU is only around 70.8%. With a normal map as or among the input, the IoU rises to around 75.8%.

Visual comparison of different kinds of input combinations on synthetic data is shown in Fig. 7. Using the depth map as input, the shape is the most complete and visually closer to GT. In particular, we observe that in the edge region, the shape becomes incomplete when the input contains a normal or an image. This may be because the depth provides sufficient information about the frontal space geometry. Additional image and/or normal input instead introduce additional uncertainties that affect the prediction of the geometry by the implicit field function. Considering the visualization results of the reconstruction, we use the depth map as the only input to our SlePIFu.

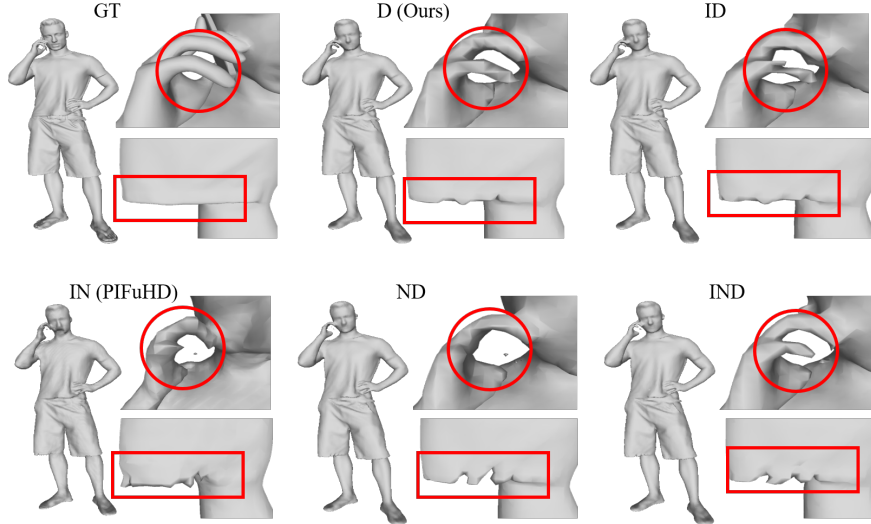


Fig. 7. Comparison of using different input information as input on synthetic data with GT mesh, including depth map (denoted as “D”) and the combinations of image, depth map, and normal. (e.g. “ID” for image and depth)

In the second group, we follow the pipeline of SelfPIFu, where the proposed depth-guided self-supervision is included to make better use of the depth information for reconstruction. This module further increases IoU up to 0.5% (e.g. 93.0% to 93.5% for “D”) The reconstruction IoUs of using different types of intermediate are listed in the last column Table 4.2. These results show the great potential of using depth for high-quality reconstruction, not only in a supervised training scenario, but also when using self-supervision methods for further improvements.

Table 1. The IoUs (%) between reconstructed shapes and the ground truths on the RenderPeople dataset with respect to different input combinations.

Image	Normal	Depth		IoU (%)	IoU _{self} (%)
✓			I	70.8	
	✓		N	76.1	
		✓	D	93.0	93.5 (+0.5)
✓	✓		IN	75.8	
✓		✓	ID	92.9	93.5 (+0.5)
	✓	✓	ND	93.1	93.5 (+0.4)
✓	✓	✓	IND	93.2	93.5 (+0.3)

We compare using different input information, including image (denoted as “I”), normal map (“N”), depth map (“D”), and their combinations (e.g. “ID” for image and depth). “IoU_{self}” means adding our proposed depth-guided self-supervised module during training.

Depth Map Contains Details. To better demonstrate the details contained in the estimated depth map, we convert depth mesh to 3D point cloud and render it from different views. Specifically, we first project depth to 3D (with orthogonal camera model) to get the 3D point cloud, and then generate a mesh from it for rendering. Fig. 6 shows the estimated normal maps and depth maps from in-the-wild images, and rendered point cloud image from depth.

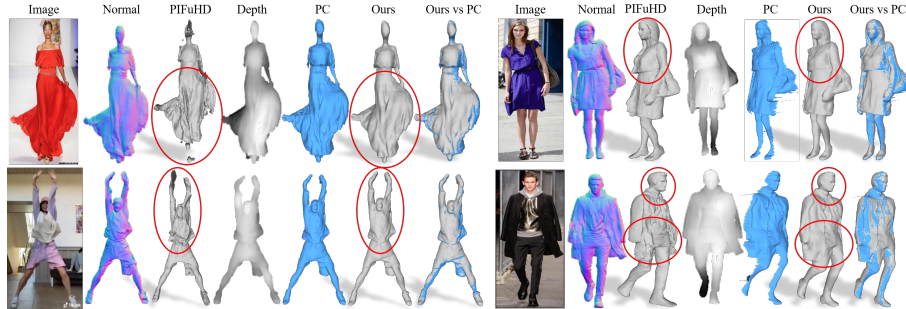


Fig. 8. Visualization of reconstruction results on PIFuHD and Ours. PIFuHD sometimes generates false disconnected regions when applied on real images (first row). In comparison, Ours can recover a complete human shape with almost no small shards. Furthermore, PIFuHD usually reconstructs false details, *e.g.* unrealistic face features and clothes wrinkles (second row).

Table 2. User Study on different variants of our SelfPIFu.

	N	D	S	V	Score
M1	✓				0.07
M2		✓			0.08
M3		✓	✓		0.23
M4		✓	✓	✓	0.63

“N”, “D”, “S”, “V” denote normal, depth, surface-aware, and volume-aware self-supervision, respectively.

Comparison with PIFuHD on Real World Images: We visualize the comparison results of PIFuHD (normal-based) and Ours (depth-based) in Fig. 8. PIFuHD sometimes generates false disconnected regions when applied on real images (first row in Fig. 8). In comparison, Ours can recover a complete human shape with almost no small shards. Furthermore, PIFuHD usually reconstructs false details, *e.g.* unrealistic face features and clothes wrinkles (second row in Fig. 8). In comparison, PIFu-D can better handle such cases with complete shapes and more plausible details.

Ablation Study on the Self-supervision Mechanism. The proposed self-supervision mechanism contains two parts, *i.e.* surface-aware and volume-aware SDF learning. Ablative experiments are conducted on synthetic data to demonstrate the effectiveness of

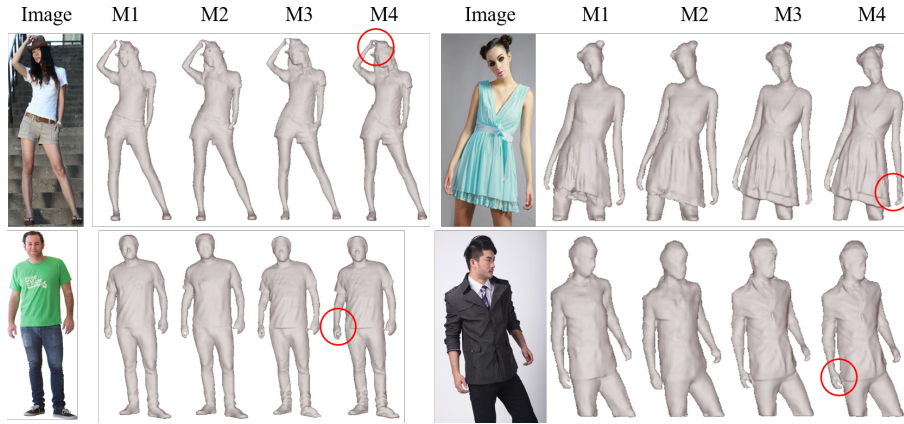


Fig. 9. Ablative study on the self-supervised mechanism. “M1-M4” represent using normal map as input (M1), using depth map as input (M2), using depth map as input plus surface-aware self-supervision (M3), and using depth map as input plus both surface-aware and volume-aware self-supervision (M4).

either of them, respectively. We visualize the reconstructed meshes of several samples that are randomly chosen from the collected real images in Fig. 9, where four results are obtained by taking as input normal map only (denoted as M1), depth map only (M2), depth map plus surface-aware self-supervision (M3), and depth map plus both surface-aware and volume-aware self-supervision (M4), respectively. It is shown that both two self-supervision mechanisms can improve the reconstruction quality when using the depth map as the intermediate input, especially in the details of reconstruction. We also conduct a user study for human evaluation on the reconstruction results. 10 samples are randomly chosen and evaluated by more than 50 persons for evaluation. The participants are asked to choose the best one among four reconstruction results (shuffled differently for each set) given the original single-view human image. The ratio of votes for each version is computed as its final evaluation score. The results of this user study are listed in Table 2, which show the improvements brought by the proposed self-supervision mechanism.

4.3 Comparison with the State-of-the-art

Reconstruction of Synthetic Images. We implement five SOTA methods, PIFu [43], PIFuHD [44], PaMIR [63], ICON [60], ECON [59], for comparison with the proposed SelfPIFu. On the RenderPeople dataset, the evaluation results on three metrics are presented in Table 3. As seen, our SelfPIFu outperforms all listed methods consistently on all three metrics, achieving a new SOTA performance on RenderPeople with an average of 0.7934cm for Chamfer distance, 1.5377cm for P2S distance, and 89.03% for IoU. It surpasses PIFuHD by around 0.4cm, 0.54cm, and 20.08% on the Chamfer distance, P2S distance, and IoU, respectively.

In addition to objective comparisons, we also made visual comparisons in Fig. 10. In frontal view, the differences between the results of the different methods are small,

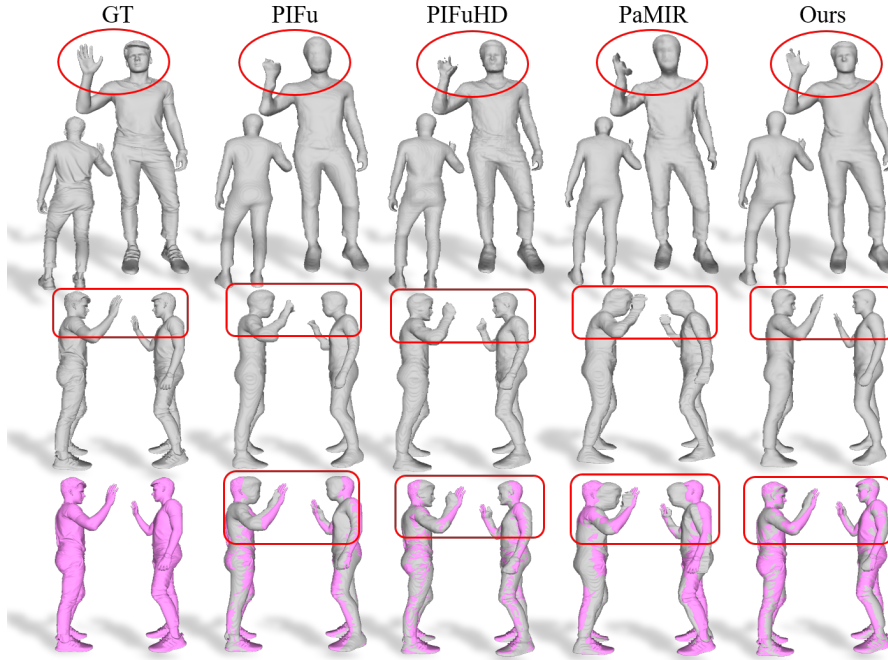


Fig. 10. Visual comparison on synthetic data. We compare reconstructed meshes from PIFu [43], PIFuHD [44], PaMIR [63], ICON [60], ECON [59], and ours. The first row shows front- and back-view results. The second row shows two side-view results and overlaid on GT to demonstrate the reconstruction errors.

which seems to be different from the results of the objective comparison. However, their differences are very obvious from the side-view. The results of PaMIR, ICON and ECON show significant offset, like head, hand and foot due to depth ambiguity, which is effectively addressed by our SelfPIFu since it uses depth map as input and, more importantly, is trained on more diverse in-the-wild images enabled by the proposed depth-guided self-supervised learning.

Reconstruction of Real Images. To compare with SOTA methods on the human reconstruction of real images, we provide qualitative results and also conduct user studies for human evaluation.

Qualitative Comparison: we visualize the reconstruction result of SelfPIFu and other SOTA approaches. As shown in Fig. 11, the reconstruction products from SelfPIFu are of higher quality than others. Compared with PIFuHD, we get similar observations as in Sec. 4.2, can generate high-quality shapes with less artifacts and more accurate estimated location (i.e. depth/global translation) than a normal-based one. Compared with PIFu, PaMIR and ICON, our SelfPIFu can recover significantly more details, and also more complete and reasonable shapes. ECON also uses the depth map as the intermediate variable, and combined with SMPL to optimize a good geometric details, but the pose priority will introduces errors into the final results, especially the palm area, that cannot be ignored.

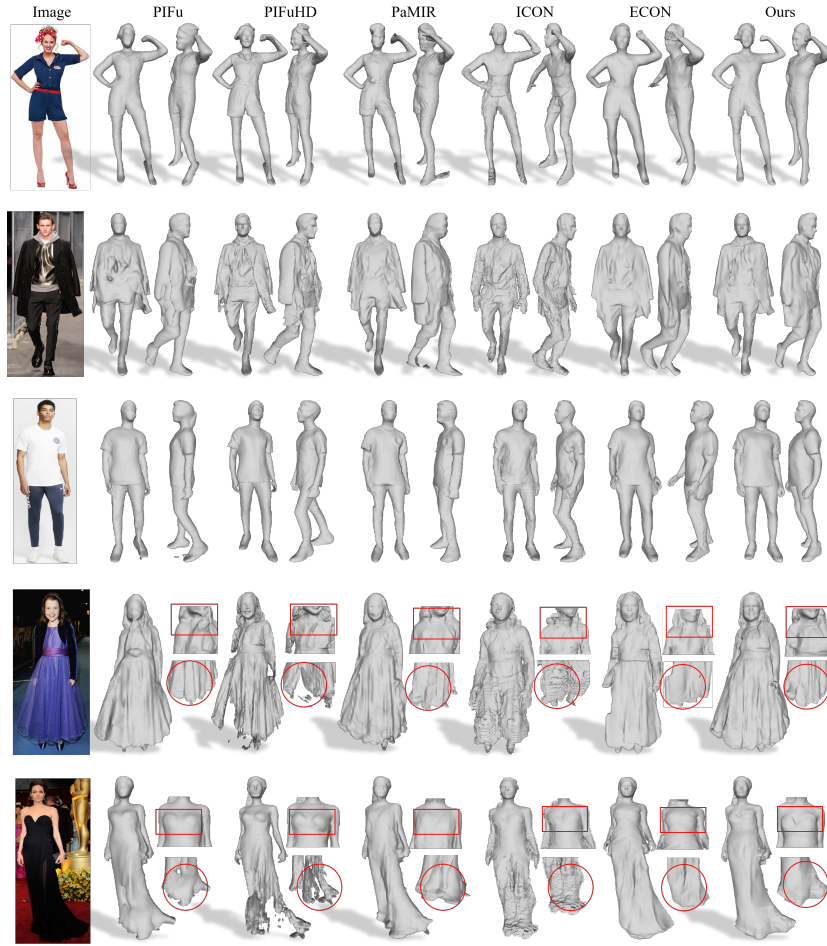


Fig. 11. Visual results on real world images. Reconstructions from PIFu [43], PIFuHD [44], PaMIR [63], ICON [60], ECON [59], and ours are shown. Zoom in to see more details.

5 Conclusion

In this paper, we propose a novel self-supervised framework, named SelfPIFu, which is able to utilize abundant and diverse in-the-wild images lacking 3D GT during training, resulting in largely improved reconstructions when tested on unconstrained in-the-wild image. Firstly, we empirically find that the estimated depth usually contains plausible details and is more robust than the estimated normal, and propose to use the depth map as the intermediate input for single-view 3D human reconstruction. Then we further propose to exploit the inferred depth map as supervision to guide the learning of the implicit function by designing a novel depth-guided self-supervised SDF-based PIFu learning module. This module contains two components, i.e. volume-aware and surface-aware supervision, to better utilize the depth information. As a result, our Self-

Table 3. Quantitative evaluation on the Renderpeople dataset.

Method	CD (cm)	P2S (cm)	IoU (%)
PIFu [43]	1.2473	2.4219	72.24
PIFuHD [44]	1.1976	2.3435	68.23
PaMIR [63]	2.2563	4.4823	65.09
ICON [60]	1.7857	3.4812	56.77
ECON [59]	2.0607	4.4395	60.42
SelfPIFu	0.7934	1.5377	89.03

PIFu improves the reconstruction quality compared with the normal-based PifuHD and generalize better on real world images. We believe that the SelfPIFu can lead to more accurate and robust reconstructions, particularly in real-world settings.

Limitation. We utilize depth as an intermediate input in the overall SelfPIFu framework. Naturally, the accuracy of depth map affects the final results. Although the depth estimator predicts plausible and detailed results in standard and easy poses (e.g. standing-like poses), the robustness of the depth estimator it suffers from extreme poses.

Future work. The most straightforward approach to address extreme poses is to include such poses in the training data. However, given the challenges associated with acquiring extensive training data, an alternative approach is to explore the incorporation of SMPL priors into the depth estimator in future research.

References

1. <https://renderpeople.com/>
2. <https://www.opengl.org/>
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics* **24**(3), 408–416 (2005)
4. Bertiche, H., Madadi, M., Escalera, S.: CLOTH3D: clothed 3d humans. In: *European Conference on Computer Vision*. pp. 344–359. Springer (2020)
5. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5420–5430 (2019)
6. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In: *European conference on computer vision*. pp. 561–578. Springer (2016)
7. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5939–5948 (2019)
8. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6970–6981 (2020)
9. Darrell, T., Worn, K.: Pyramid based depth from focus. In: *Proceedings CVPR’88: The Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 504–505. IEEE Computer Society (1988)
10. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: *2019 IEEE/CVF Confer-*

- ence on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 285–295. IEEE Computer Society (2019)
11. Fei, X., Wong, A., Soatto, S.: Geo-supervised visual depth prediction. *IEEE Robotics and Automation Letters* **4**(2), 1661–1668 (2019)
 12. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. *CoRR* **abs/2002.10099** (2020), <https://arxiv.org/abs/2002.10099>
 13. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In: *NeurIPS* (2020)
 14. Hu, T., Wang, L., Xu, X., Liu, S., Jia, J.: Self-supervised 3d mesh reconstruction from single images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6002–6011 (2021)
 15. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3d human body reconstruction from a single image via volumetric regression. In: *ECCV Workshops*. pp. 0–0 (2018)
 16. Jafarian, Y., Park, H.S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12753–12762 (2021)
 17. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. In: *European Conference on Computer Vision*. pp. 18–35. Springer (2020)
 18. Jiang, Y., Ji, D., Han, Z., Zwicker, M.: Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1251–1261 (2020)
 19. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8320–8329 (2018)
 20. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7122–7131 (2018)
 21. Kato, H., Beker, D., Morariu, M., Ando, T., Matsuoka, T., Kehl, W., Gaidon, A.: Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057* (2020)
 22. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: *2016 Fourth international conference on 3D vision (3DV)*. pp. 239–248. IEEE (2016)
 23. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 6050–6059 (2017)
 24. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5162–5170 (2015)
 25. Liu, S., Zhang, Y., Peng, S., Shi, B., Pollefeys, M., Cui, Z.: Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2019–2028 (2020)
 26. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7708–7717 (2019)
 27. Liu, S., Saito, S., Chen, W., Li, H.: Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems* **32** (2019)
 28. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* **34**(6), 248:1–248:16 (2015)

29. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5667–5675 (2018)
30. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
31. Miller, G.: Efficient algorithms for local and global accessibility shading. In: Proceedings of the 21st annual conference on Computer graphics and interactive techniques. pp. 319–326 (1994)
32. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4480–4490 (2019)
33. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. *ACM transactions on graphics (TOG)* **24**(3), 536–543 (2005)
34. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
35. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3504–3515 (2020)
36. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 International Conference on 3D Vision (3DV). pp. 484–494. IEEE (2018)
37. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019)
38. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
39. Poggio, G.F., Poggio, T.: The analysis of stereopsis. *Annual review of neuroscience* **7**(1), 379–412 (1984)
40. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)* **34**(4), 120 (2015)
41. Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 283–291 (2018)
42. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3313–3322 (2019)
43. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172* (2019)
44. Saito, S., Simon, T., Saragih, J., Joo, H.: 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. pp. 84–93 (2020)
45. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7763–7772 (2019)

46. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: European Conference on Computer Vision. pp. 53–70. Springer (2020)
47. Song, D.Y., Lee, H., Seo, J., Cho, D.: Difu: Depth-guided implicit function for clothed human reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8738–8747 (2023)
48. Tan, F., Zhu, H., Cui, Z., Zhu, S., Pollefeys, M., Tan, P.: Self-supervised human depth estimation from monocular videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 650–659 (2020)
49. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7750–7759 (2019)
50. Tiwari, G., Bhatnagar, B.L., Tung, T., Pons-Moll, G.: SIZER: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In: European Conference on Computer Vision (ECCV). Springer (August 2020)
51. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 20–36 (2018)
52. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 109–117 (2017)
53. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)
54. Wang, L., Zhao, X., Yu, T., Wang, S., Liu, Y.: Normalgan: Learning detailed 3d human from a single rgb-d image. In: European Conference on Computer Vision. pp. 430–446. Springer (2020)
55. Wen, Y., Liu, W., Raj, B., Singh, R.: Self-supervised 3d face reconstruction via conditional estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13289–13298 (2021)
56. Wu, F., Bao, L., Chen, Y., Ling, Y., Song, Y., Li, S., Ngan, K.N., Liu, W.: Mvf-net: Multi-view 3d face morphable model regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 959–968 (2019)
57. Wu, Y., He, K.: Group normalization. In: ECCV. pp. 3–19 (2018)
58. Xian, K., Zhang, J., Wang, O., Mai, L., Lin, Z., Cao, Z.: Structure-guided ranking loss for single image depth prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 611–620 (2020)
59. Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J.: ECON: Explicit Clothed humans Optimized via Normal integration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
60. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: Implicit Clothed humans Obtained from Normals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13296–13306 (June 2022)
61. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In: Advances in Neural Information Processing Systems. pp. 490–500 (2019)
62. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 204–213 (2021)

63. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence* (2021)
64. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 7739–7749 (2019)
65. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4491–4500 (2019)
66. Zhu, H., Cao, Y., Jin, H., Chen, W., Du, D., Wang, Z., Cui, S., Han, X.: Deep Fashion3D: A dataset and benchmark for 3d garment reconstruction from single images. In: *European Conference on Computer Vision*. pp. 512–530. Springer (2020)