DOI: 10.1111/cgf.14971

Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

Fine Back Surfaces Oriented Human Reconstruction for Single RGB-D Images

Xianyong Fang[†], Yu Qian , Jinshen He, Linbo Wang[‡], Zhengyi Liu

School of Computer Science and Technology, Anhui University, China

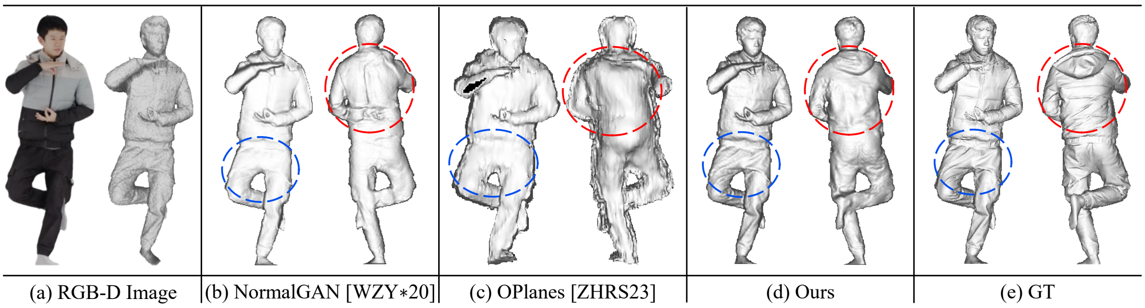


Figure 1: *Human surface reconstruction by different methods for single RGB-D images. The frontal and back surfaces from each method are shown on the left and right of its corresponding subfigure respectively, with the red and blue circles showing example areas. Our method (d) apparently recovers richer surface details than the other two methods, where the back surface is fine without the wrong details of the frontal one (see the red circle on (b)).*

Abstract

Current single RGB-D image based human surface reconstruction methods generally take both the RGB images and the captured frontal depth maps together so that the 3D cues from the frontal surfaces can help infer the full surface geometries. However, we observe that the back surfaces can often be quite different from the frontal surfaces and, therefore, current methods can mess the recovery process by adopting such 3D cues, especially for the unseen back surfaces. We need to do the back surface inference without the frontal depth map. Consequently, a novel human reconstruction framework is proposed, so that human models with fine geometric details, especially for the back surfaces, can be obtained. In this approach, a progressive estimation method is introduced to effectively recover the unseen back depth maps. The coarse back depth maps are recovered by the parametric models of the subjects, with the fine ones further obtained by the normal-maps conditioned GAN. This framework also includes a cross-attention based denoising method for the frontal depth maps. This method adopts the cross attention between the features of the last two layers encoded from the frontal depth maps and thus suppresses the noise for fine depth maps by the attentions of features from the low-noise and globally-structured highest layer. Experimental results show the efficacies of the proposed ideas.

CCS Concepts

• ***Computing methodologies*** → *Parametric curve and surface models; Virtual reality;* ***Reconstruction;***

1. Introduction

Accurate reconstruction of 3D human surfaces can be applied to various areas, such as metaverse and robotics and so on. These days we have witnessed booming studies on it [rlc]. Perhaps the easiest way to do it is just by only one capture of the subject, where the widely adopted methods are based on single RGB or RGB-D images. Single RGB image based methods often have to additionally include depth estimation for better performances [GFM∗19, SLSC23, XYC∗23]. However, estimated depths can be very unstable. On the contrary, single RGB-D image based methods can be

[†] fangxianyong@ahu.edu.cn

[‡] Corresponding author, wanglb@ahu.edu.cn

more robust, thanks to the directly captured depth maps as strong 3D priors. Therefore, we study the single RGB-D image based approach for the human surface estimation.

As far as we know, only very few studies deal with single RGB-D images [WZY*20, ZHRS23] and their performances are limited (Figure 1): The surface details can be lost (see the areas inside the circles on Figure 1), while the frontal details can even appear on the back (see the area inside the red circle of Figure 1 (b)). Checking the principles of those studies, we can see that they take the RGB image and its associated frontal depth map together for surface inference, either as a whole (*e.g.*, OPlanes shown in Figure 1 (c)) or as the fusion of the separately estimated frontal and back surfaces (*e.g.*, NormalGAN shown in Figure 1 (b)).

However, the frontal and back surfaces of humans can often be quite different and, therefore, simply putting the RGB image and the frontal depth map together for direct surface inference can easily mess the results: The frontal geometries may unnecessarily appear on the back due to the 3D cues of the frontal surfaces. This is especially true for RGB-D images: Their frontal depth maps are directly captured, while the back ones are completely unseen. Consequently, the back surfaces can be easily biased towards the known frontal ones. Therefore, we argue that the better way for 3D human surface recovery with RGB-D images can be to infer the unseen back surfaces without the interference from the frontal depth maps.

Effective priors are important for the inference of the unseen back surfaces, while the normal maps adopted in NormalGAN are too general to be strong enough for the human subjects. As for human surface reconstruction, the particular distribution of human geometry should be considered. Especially, the parametric body models, such as SMPL [LMR*15] or SMPL-X [PCG*19], can be effective choices, considering their statistical natures for human shapes [XYC*23, CHW23a, FYR*19, MCL22].

Accordingly, the depth maps of the unseen backs can be accurately estimated by the parametric body models and normal maps of human surfaces. The parametric model can supply a rough initialization on the general shape of the surface, while the normal map can provide surface geometrical details to help refine the initial shape as accurate surface.

In addition, the frontal depth maps are often noisy and should also be effectively denoised before doing the 3D frontal construction. The encoded features of the noisy frontal depth maps contain information of low-noise global structures in high level and high-noise geometrical details in the low level. Therefore, the global structures embedded in the high-level features can guide the denoising of the noisy low-level features for fine frontal feature maps.

Consequently, this article proposes a novel human reconstruction framework to model human surfaces with rich geometrical details, which is especially good at fine back surfaces. For the unseen back surface estimation, it adopts a progressive reconstruction strategy, where the coarse depth map of the back surface is first inferred by the parametric body model and further refined by the normal-maps conditioned Generative Adversarial Networks (GAN) [GPAM*14] for accurate 3D back recovery. For the frontal surface estimation, it introduces a cross attention based denoising method to denoise the frontal depth map, where the cross attentions between the low-noise high-level features and the high-noise low-level features are used to suppress the noise and thus help infer the fine map. The complete 3D human surface is finally obtained by the two 3D surface clouds with their nearest neighbors. Figure 1(d) shows the reconstruction result with geometrical details by our method. It is much better than the other two methods (Figure 1 (b) and (c)) and there are no frontal details on the back by our method (see the area inside the red circle of Figure 1(d)).

Note that the parametric model has already adopted for effective prior cues in various studies, such as those with single images [XYTB22] [XYC*23, ZYW*19], image sequences or videos [CLHG22, CHW23b], and even RGB-D sequences [SXZ*22, YZG*18]. However, there is no study applying it to the reconstruction for the single RGB-D images. The depth image denoising methods have also studied by many researchers as an independent target [JL18, SSC*19, ZW16, GLGT19] or a part of a bigger work [SZB*23, WZY*20, DXD*22]. However, none of them considers the cross-attention based idea on features from the highest two layers.

In summary, our contributions can be summarized as follows.

- A fine back surfaces oriented human reconstruction method for single RGB-D images, which separates the reconstruction into front and back estimations with the fine back surfaces estimated without the guidance of the frontal depth maps.
- A progressive estimation method for the unseen back depth maps, which first takes the parametric models as prior to get the rough depth maps and then refine those maps by the normal-maps conditioned GAN.
- A cross-attention based denoising method for the captured frontal depth maps, which takes the high-level features to attend the low-level features and thus suppress the noise for fine frontal depths.

2. Related Work

This section mainly reviews the development of human surface reconstruction with RGB-D images. Single RGB images and the depth map denoising methods are also briefly reviewed. For more details on the general development of human surface reconstruction, you may refer to [TKB*23, CPZ21].

2.1. RGB-D image based human surface reconstruction

Reconstruction of human surfaces by RGB-D images has been studied for a long time [KLL*13, DF14]. The most popular way is to take multiple views or videos. For example, Dou and Fuchs *et al.* [DF14] assumed a pre-scanned version of the static environment as a prior and gradually reconstructed the dynamic humans by non-rigid registration. Another famous example is DynamicFusion [NFS15], aiming at real-time non-rigid volumetric fusion. Various follow-ups [IZN*16, SBI18, GXY*17] have proposed to improve the performances with various constraints and priors. Among them, SMPL [LMR*15] is adopted as the parametric body prior [YZG*18]. However, those studies [ZSG*18] often take the traditional optimization based methods such as the non-rigid ICP, where multiple steps with manually defined energy functions are adopted to do the fusion.

Along with the booming of deep learning, new methods [YZG*21, DXD*22, CFF*22, BNT21] for RGB-D sequences appeared. For example, Function4D [YZG*21] combines temporal volumetric fusion and deep implicit functions, where dynamic sliding fusion is applied to do effective neighbor depth fusion and detail-preserving deep implicit functions are further used to infer geometrical details and generate textures; Dong *et al.* [DXD*22] proposed a novel geometry-aware PIFu method which exploits the complementary properties of depth denoising and 3D reconstruction and learns a two-scale PIFu representation for face and body separated reconstruction; Cai *et al.* [CFF*22] proposed a template-free method to recover high-fidelity dynamic humans based on the neural SDF and neural radiance fields (NeRF) [MST*21]. Recently, Zheng *et al.* [ZLWY23] incorporated the occupancy field and albedo field with an additional visibility field and reconstructed the model by a novel TransferLoss to implicitly enforce the alignment between the visibility field and occupancy field.

There are few studies [WZY*20, ZHRS23] on single RGB-D images. NormalGAN [WZY*20] takes a fast adversarial learning based approach, where the back-view depth is inferred from the frontal-view depth under the help of the normal-maps conditioned GAN. However, this frontal surface based back inference may introduce unnecessary frontal details. Zhao *et al.* [ZHRS23] formulated single-view RGB-D human reconstruction as an occupancy-plane-prediction task, where the occupancy planes indicate the occupancy at every pixel location for the corresponding 3D point and are more flexible than those of classical voxel-grid representations. However, this method cannot robustly recover the surface details without prior constraints. In addition, the parametric model has not been considered by these single RGB-D based methods.

2.2. Single RGB image based human surface reconstruction

The parametric models are often adopted by single RGB images [JZH*20, LIPM19, SLSC23, XYTB22, XYC*23]. For example, ICON [XYTB22] first estimates detailed clothed-human normals (front/back) and then regresses the implicit surface with the guidance of SMPL estimation. ECON [XYC*23] further extends ICON by estimating the depth maps of both the frontal and back surfaces. We, however, adopt this type of ideas into the RGB-D image based work.

Depths [GFM*19, SLSC23, XYC*23] and 3D displacements [JZH*20], normal maps [XYTB22, XYC*23], albedos [AZS22], silhouettes [NSH*19], implicit functions [SHN*19, SSSJ20, XYTB22, XYC*23, LZX*23], GAN [JJW*23] and NeRF [HHP*23] are the popular ideas except parametric models adopted in the existing methods. Recently, SHERF [HHP*23] adopts a hierarchical feature bank to supply enough information for high fidelity NeRF reconstruction; Liao *et al.* [LZX*23] combined deep learning and traditional optimization to obtain the general shape and refined surface details respectively.

2.3. Depth image denoising

Traditional methods often take filters to denoise the depth image [CBTT08, DBPT10, RSD*12] and use the RGB images as support [YYDN07, LKH07, PKT*11]. Kwon *et al.* [KTL15] proposed a dictionary-learning based data-driven method. However, traditional methods rely heavily on the manually designed noise features.

Deep learning based methods are now popular [ZW16, FLJW22, YWW*18, JL18, SSC*19, GLGT19, LHAY19]. For example, DDRNet [YWW*18] adopts a cascaded CNN structure to denoise and further refines the depth maps for high qualities of both low and high frequencies. Sterzentsenko *et al.* [SSC*19] proposed a fully convolutional deep auto-encoder that learns to denoise depth maps. However, it requires multiple views. SGN [GLGT19] adopts a top-down self-guidance architecture to effectively incorporate multiscale information from the shuffling operation. Recently, Yan *et al.* [YLZ*20] adopted a group based nuclear norm and learning graph model for depth image denoising.

Recent studies often take the denoisng as a part of bigger work [SZB*23, WZY*20, DXD*22]. For example, NormalGAN [WZY*20] takes normal map constrained GAN to denoise the frontal depths for the full surface reconstruction. Dong *et al.* [DXD*22] included the denoising process in the geometry-aware PIFu-Body module, which the global topological information of the 3D occupancy field guides the denoising process. Similarly, we consider the depth map denoising as a component of the bigger human estimation work but take the cross-attention based idea.

3. Our Proposed Method

Existing single RGB-D image based human reconstruction methods take the RGB images and their frontal depth maps together to estimate the 3D human surfaces. It highly depends on the frontal depth map for 3D cues of the whole surface. However, there are often wide differences between the frontal and back surfaces of humans. Therefore, current methods may mess the recovery process with low quality estimation and even the frontal details on the final back surfaces.

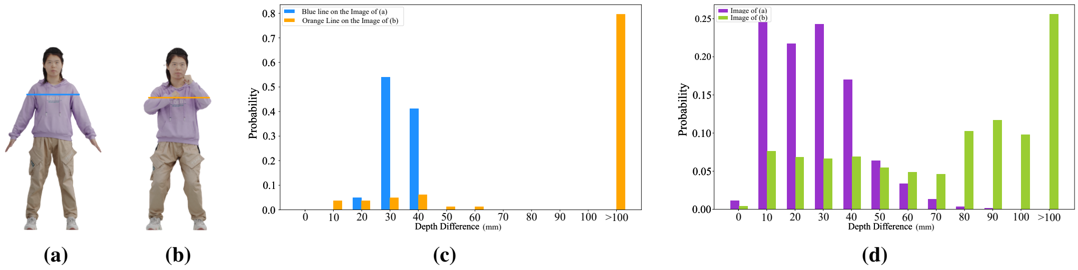
Let's first discuss our observation on the depth differences.

3.1. Differences between frontal and back surfaces

We observe that the front and back of humans are different and sometimes this difference is very significant. Figure 2 shows such an experiment with two typical RGB-D images from THuman2.0 [YZG*21], where one subject in two different poses is captured by two images (Figure 2a and Figure 2b). The imaged surfaces of the subject in these two images are geometrically different.

Statistically local and global comparisons are applied to explore the depth differences between the frontal and back surfaces for these two images. Here, for fairness, the depth of each point on a surface is updated first by its relative depth on that surface, *i.e.*, it is subtracted by the max depth of its surface. The relative depths equally capture the local shape of the object and thus are adopted here for our purpose of frontal and back comparison.

Locally, the statistical depth differences between the 3D frontal and back surfaces according to each example line specified from almost the same positions are compared (Figure 2c). The histogram



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

Figure 2: *Statistical comparison in histogram on the depth differences between the frontal and back surfaces of RGB-D images. (a) and (b) represent two typical RGB-D images with the rows under the blue and orange lines specified for local comparison; (c) shows the local comparison by two histograms for depth differences under the two lines in (a) and (b); (d) shows the global comparison by two histograms for the two images shown in (a) and (b). Note: 1) The depth of each point is first subtracted by the max depth of its surface as relative depth before doing the comparisons; 2) each number labeling the difference, dep, on the two horizontal axes, except '0' and '>100' means the range* $(dep-10, dep]$.

along the orange line on the image of Figure 2b have nearly three-thirds of depth differences bigger than 100mm (79.52%), while there is even no depth difference bigger than 40mm for that along the blue line of Figure 2a. Most depth variations along the blue line approximately range from 20mm to 40mm (95.09%). The biggest depth differences along the orange line is 162.0mm, while the biggest one on the blue line is only 40mm. It can be seen from these two histograms that the depth variations between the frontal and back surfaces along these lines for the two images are quite different.

Globally, we compare the depth differences between the full frontal and back surfaces for these two poses (Figure 2d). Again sharp contrast can be found for the depth differences for the two images on the whole image space. More than one-fourth of depth differences for (b) is bigger than 100mm (25.56%), while there is completely no depth differences bigger than 90mm for (a) whose depth differences are mostly between 0mm and 40mm (88.61%).

The above experiment shows that the front and back of a human can be quite different, which often appears along with pose changes. Therefore, direct inference of the back surfaces with only the RGB images and their frontal depth maps, such as the method adopted in NormalGAN [WZY*20], does not have enough knowledge on the back distributions and makes frontal details easily appear on the back surfaces (Figure 1b and 1c). This type of work messes the front and back without considering their differences and, therefore, failures with smooth frontal and back surfaces are often unavoidable.

This investigation leads us to think of a novel method to obtain detail-rich human models for single RGB-D images, where the fine back surfaces can be inferred without the guidance of the frontal depth maps. We will briefly introduce this method in the following.

3.2. Outline of our method

In this method, for the main target of fine back surfaces, the parametric body model is considered as the additional prior to help estimate the unseen back besides the detail-rich normal maps. Certainly, effectively denoising the frontal maps is also important for accurate front reconstruction. Consequently, a framework with two branches for the front and back estimation can be structured (Figure 3).

Assume the input RGB-D image S consisting of one RGB image I and one frontal depth map D_F, $S = \{I, D_F\}$. In the frontal estimation, D_F is noisy and thus should be denoised before constructing the 3D frontal surface. Here, a cross-attention based method is used for this purpose, where the low-noise high-level features and the high-noise low-level features are cross attended for denoised frontal depth map D'_F. For the back estimation, a progressive idea is used for the unseen back. The parametric model P acts as the strong prior to get the rough back depth map $D_B^{(R)}$ and the fine refined depth map D'_B is finally obtained with GAN constrained by detail-rich normal priors. The frontal and back depth maps are reprojected back as 3D frontal and back point clouds C_F and C_B respectively which are finally fused together as the whole 3D human model M_S. Note, the expressive SMPL-X model [PCG*19] is the parametric body model used.

Orthogonal projection is assumed as the way of capturing the input images for the framework, which is easier to apply and can better match the front and back than perspective projection. In practice, the depth maps by a consumer depth camera like Microsoft Kinect capture close subjects and, therefore, the appearance differences got by perspective and orthogonal projections are small and can be omitted. Therefore, our method directly takes the depth map as input without considering the orthogonalization pre-process adopted by NormalGAN [WZY*20].

Now, let's discuss the details of the frontal and back estimation processes in the two separate branches.

3.3. Cross-attended frontal surface construction

The frontal depth map directly captures the geometry of the frontal surface with significant noise due to the camera sensors. Therefore, it must be denoised for accurate 3D construction of the surface. Generally, the high-level features containing global structure information are less noisy than the low-level features which is rich with

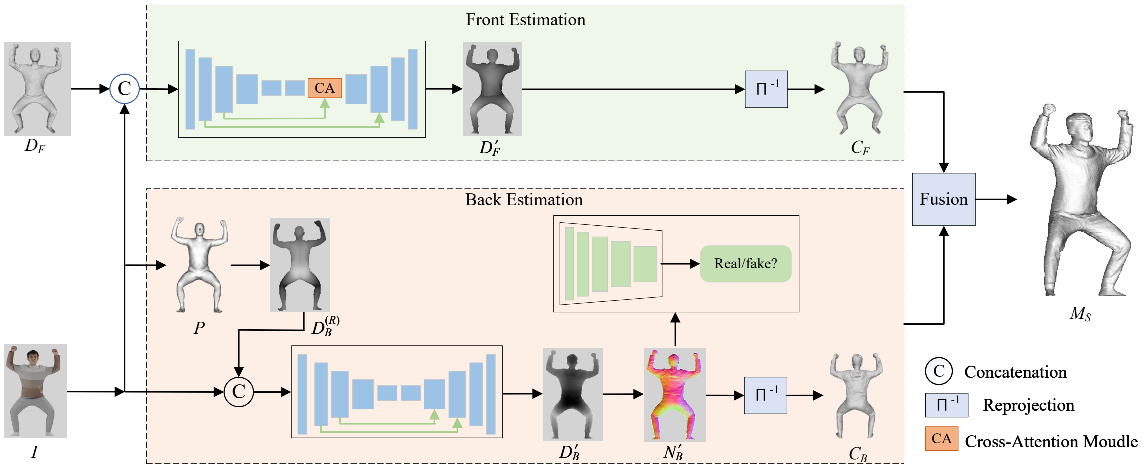


Figure 3: *The pipeline of the proposed method. There are two branches for estimating the frontal surface cloud C_F and back surface cloud C_B respectively from the input RGB-D image consisting of one RGB image I and its depth map D_F. A cross attention based method is adopted in the front estimation to denoise the frontal depth map D_F to be D'_F. A progressive back depth estimation method is applied in the back estimation to obtain the back depth map D'_B by the parametric model P and normal-maps conditioned GAN. The final surface M_S is obtained by the fusion of C_F and C_B.*

geometrical details. Therefore, we can use the low-noise global structures of the high-level features to attend the high-noise low-level features to recover the details while suppressing the noise. Therefore, a cross-attention based method is proposed.

It adopts a U-Net style of design (see the front estimation part in Figure 3), consisting of four layers for the encoder and the decoder respectively. The concatenation of the RGB image I and depth map D_F is the input while the denoised frontal depth map D'_F is the output. A cross-attention module CA is applied on the refined highest bottleneck features with the lower-level features. The low-noise high-level features with rich structural information attend the high-noise low-level features and thus help obtain noise-free features with fine details for D'_F.

Let's discuss the structure of the cross-attention module CA (Figure 4). The higher level features are the bottleneck features $\boldsymbol{f}_b$ while the lower level ones are those from the previous level $\boldsymbol{f}_3$. Even lower features are not considered in this module because bigger dimension differences will make the information lost significantly when doing upsampling based fusion.

In particular, $\boldsymbol{f}_b$ are self-attended to obtain the attention maps $\boldsymbol{f}_m$ for fusion with the lower features. The self-attention is defined with the transformations from two parallel MLP $\mathcal{M}_1$ and $\mathcal{M}_2$, so that the output features, query and key, are refined into lower dimensions. The attention maps $\boldsymbol{f}_m$ are computed as

$$\boldsymbol{f}_m = \sigma(\mathcal{M}_1(\boldsymbol{f}_b) \times \mathcal{M}_2(\boldsymbol{f}_b)^T), \tag{1}$$

where σ represents the *sigmoid* function and × denotes matrix multiplication.

$\boldsymbol{f}_m$ are then applied to the lower features $\boldsymbol{f}_3$ for doing the weighted summation of $\boldsymbol{f}_b$ and $\boldsymbol{f}_3$ as the output features,

$$\boldsymbol{f}'_b = (\boldsymbol{f}_m \times \mathcal{M}_3(\boldsymbol{f}_3))||\boldsymbol{f}_b, \tag{2}$$

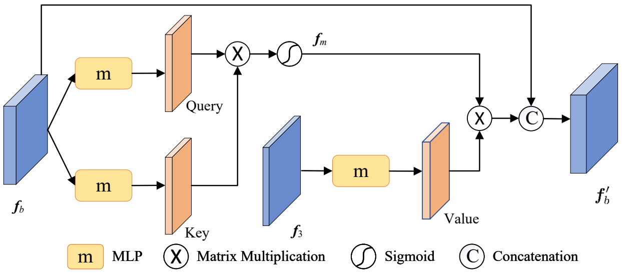


Figure 4: *The principle of the cross-attention module (CA). The self-attended bottleneck features $\boldsymbol{f}_b$ and the features from the previous layer, $\boldsymbol{f}_3$, are cross attended and combined to obtain the final refined features $\boldsymbol{f}'_b$ for further decoding.*

where || represents concatenation. Here, like the self-attention process of $\boldsymbol{f}_b$, $\boldsymbol{f}_3$ are transformed by another MLP $\mathcal{M}_3$ as lower but refined dimensional data, value, before doing the cross-attention computation.

Further decoding $\boldsymbol{f}'_b$ by the denoising network will finally get the denoised frontal depth map D'_F. Then, the frontal point cloud C_F can be obtained by re-projecting (Π^{-1}) D'_F to 3D space according to the camera matrix A,

$$C_F = \Pi^{-1}(D'_F, A). \tag{3}$$

Figure 5 shows an example denoising result by this cross-attention based method. The CA module can effectively suppress the noise for a fine frontal depth map so that its 3D frontal surface (Figure 5b and 5b) has rich details but without apparent noise.

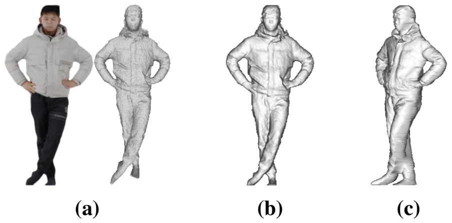


Figure 5: *Demonstration of the cross-attention based denoising method for the frontal depth maps. (a): Source RGB-D image; (b) reconstructed frontal surface after denoising; and (c) side view of the reconstruction.*

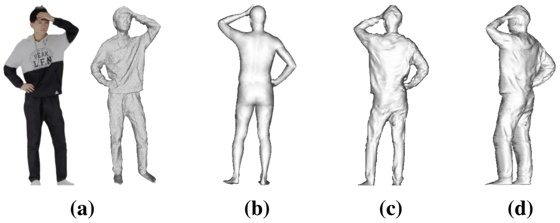


Figure 6: *Demonstration of the progressive estimation method for the back depth map. (a): Source RGB-D image; (b) coarse SMPL-X based back surface; (c): finally reconstructed back surface after the progressive estimation; and (d) side view of the reconstruction.*

3.4. Progressive modeling of the unseen back surface

The estimation of the unseen back surface (see the back estimation part in Figure 3) is mainly fulfilled by two progressive steps for accurate back depth maps: coarse estimation by the parametric body model and refinement by normal-maps conditioned GAN. The refined depth maps are then re-projected as 3D back cloud for further fusion.

For the coarse estimation, the full-body parametric model P can be estimated from the input image by some existing studies, such as PIXIE [FCB*21] and PyMAF-X [ZTZ*23]. Here, the coarse back depth map $D_B^{(R)}$ can be computed by projecting the model to the view plane,

$$D_B^{(R)} = \Pi(P, A). \tag{4}$$

Note that this re-projection can also be obtained by an existing renderer, such as PyTorch3D or OpenGL.

The refinement is based on the normal-maps constrained GAN because the normal maps are geometrically richer than the depth maps and thus can be adopted for better depth map refinement (See our ablation in Section 4.5 for more information). Here the generator $\mathcal{G}$ takes a U-Net based eight-layer structure. It accepts $D_B^{(R)}$ as input and outputs the refined back depth map with surface details,

$$D'_B = \mathcal{G}(D_B^{(R)}). \tag{5}$$

This refined depth is adversarially trained by a normal-maps based discriminator $\mathcal{D}$ based on the Markovian discriminator [IZZE17], penalizing structures by local image patches and thus ensuring efficient generation.

The normal maps used in $\mathcal{D}$ are obtained as follows. The normal of the i-th point $\boldsymbol{p}_i$, $\boldsymbol{l}_i$, can be approximated with its neighbors indexed as the set C,

$$\boldsymbol{l}_i = \overline{\sum_{j,k \in C} (\boldsymbol{p}_j - \boldsymbol{p}_i) \times (\boldsymbol{p}_k - \boldsymbol{p}_i)}, \tag{6}$$

where $\bar{\ }$ denotes the normalization operation.

Figure 6 shows an example from the above progressive estimation. The coarse parametric model based back (Figure 6b) provides important cues to do further refinement, even though it's not accurate. The final back (Figure 6c and 6d) is fine with rich details, thanks to the normal maps empowered GAN.

After above two progressive steps, the 3D back cloud C_B can be obtained through re-projection in the same way as computing the frontal cloud (Equation 3).

3.5. Nearest-neighbor based fusion

Now comes to the final fusion, where the frontal and back clouds, C_F and C_B, are combined with the nearest neighbors between them. The triangulation is applied to those neighboring points with interpolation to avoid abrupt changes due to apparent gaps. The details on how to do this fusion can be found in NormalGAN [WZY*20].

To further improve the results, the Poisson mesh editing can also be applied to smooth the cloud for better output. However, we don't use the Poisson method in our experiment for fair comparison with existing methods.

3.6. Loss

Two types of loss, frontal loss and back loss, are considered according to the training targets on either frontal or back surfaces.

3.6.1. Frontal losses

The frontal loss should first include the difference between the estimated frontal depth maps D'_F and their ground-truth depth maps $\hat{D}_F$,

$$L_D(D'_F, \hat{D}_F) = \|D'_F - \hat{D}_F\|_1. \tag{7}$$

However, the depths themselves are not enough because of the lack of 3D supervision and thus may lead to unstable results. Therefore, normal differences between the estimated (N'_F) and the ground-truth ones ($\hat{N}_F$)) are also adopted as supervision,

$$L_N(N'_F, \hat{N}_F) = \|N'_F - \hat{N}_F\|_1. \tag{8}$$

In addition, the features of the normal map estimated by VGG19 [SZ14] is also adopted to tune the performances for better geometrical details. The corresponding loss is measured by the weighted sum of differences from features of i-th layer, $\mathcal{V}_i$, between the estimated and ground-truth norm maps.

$$L_V(N'_F, \hat{N}_F) = \sum_i \lambda_{v_i} \|\mathcal{V}_i(N'_F) - \mathcal{V}_i(\hat{N}_F))\|_1, \tag{9}$$

where λ_{v_i} are the weights.

Consequently, the total frontal loss L_{front} can be formulated as

$$L_{front} = \gamma_D L_D(D'_F, \hat{D}_F) + \gamma_N L_N(N'_F, \hat{N}_F) + \gamma_V L_V(N'_F, \hat{N}_F). \quad (10)$$

where $\gamma_{\{\cdot\}}$ are the coefficients.

3.6.2. Back losses

For the back, the depth difference $L_D(D'_B, \hat{D}_B)$ (Equation 7) between the predicted and the ground-truth back depth maps, D'_B and $\hat{D}_B$, should be included for training the generator $\mathcal{G}$. The typical GAN loss [GPAM*14] is also adopted based on the estimated normal N'_B and its ground truth $\hat{N}_B$,

$$L_G(N'_B, \hat{N}_B) = \boldsymbol{E}_{\hat{N}_B}[log\,\mathcal{D}(\hat{N}_B)] + \boldsymbol{E}_{N'_B}[log\,(1 - \mathcal{D}(\mathcal{G}(N'_B)))]. \quad (11)$$

However, GAN may not converge due to the possible drastic changes of the normals. Therefore an additional feature matching loss [WLZ*18] is included to constrain the discriminator. It compares the feature differences between predicted normal maps and their ground truths from different layers of $\mathcal{D}$,

$$L_M(N'_B, \hat{N}_B) = \sum_{k=2}^{T-1} \|\mathcal{D}_k(N'_B) - \mathcal{D}_k(\hat{N}_B)\|_1, \quad (12)$$

where $\mathcal{D}_k$ represents the k-th layer of $\mathcal{D}$ with total T layers.

In addition, $D_B^{(R)}$ by the parametric model is important for back depth refinement and further 3D estimation. Therefore, an additional silhouette loss measuring the difference of binary masks between the estimated parametric model and imaged subject, P_L and H_L, are also taken,

$$L_P(P_L, H_L) = \|P_L - H_L\|_1. \quad (13)$$

Finally, the VGG loss L_V (Equation 9) is also considered. Consequently, the total generator loss, $L_{back}^{(\mathcal{G})}$, can be formulated as

$$L_{back}^{(\mathcal{G})} = \beta_D L_D(D'_B, \hat{D}_B) + \beta_G L_G(N'_B, \hat{N}_B) + \beta_M L_M(N'_B, \hat{N}_B) + \beta_P L_P(P_L, H_L) + \beta_V L_V(N'_B, \hat{N}_B). \quad (14)$$

where $\beta_{\{\cdot\}}$ are the coefficients.

For the disriminator $\mathcal{D}$, only the normal differences is considered and, therefore, its loss is

$$L_{back}^{(\mathcal{D})} = L_G(N'_B, \hat{N}_B). \quad (15)$$

4. Experimental Results

4.1. Implementation settings

THuman 2.0 [YZG*21] is adopted as the experimental dataset where 500 models including their SMPL-X models are chosen to train and test the proposed framework. The training and testing sub-sets are divided by the ratio 4 : 1. All images are prepared by Blender [Fou] as 424 × 512 according to the depth map size in Kinect V2. Two data augmentation methods are used: 14 RGB-D images of each model are obtained by randomly rotating it horizontally between -30° and 30°, with each depth map then randomly perturbed by multiple Gaussian noises as NormalGAN [WZY*20].

The whole idea is implemented by Python and trained on two NVIDIA® V100 graphics cards. Separate trainings are conducted for the frontal and back branches. The batch size of the front branch is set to 16, while that of the back one is set to 32. Adam optimizer with learning rate 0.001 is adopted, where 45 and 25 epochs are applied for the front and back estimation respectively.

Other hyper parameters are set as follows. γ_D, γ_N and γ_V in Equation 10 are set to 1.0, 20.0 and 20.0 respectively, while β_D, β_G, β_M, β_P and β_V in Equation 14 are set to 10.0, 5.0, 20.0, 20.0 and 1.0 respectively. There are five layers in VGG and, therefore, correspondingly five weights in λ_{v_i} of Equation 9 are $1/32$, $1/16$, $1/8$, $1/4$ and 1 from $i = 1$ to $i = 5$.

4.2. Performance comparison settings

Five methods are adopted for performance comparison, including NormalGAN [WZY*20], the occupancy planes based method or OPlanes [ZHRS23], PIFuHD [SSSJ20], ICON [XYTB22] and ECON [XYC*23]. Two of them, NormalGAN and OPlanes, are for single RGB-D images with the left three for single RGB images. For fair reasons, 1) the orthogonalization step of NormalGAN is removed; and 2) all those methods are directly tested with their open codes. Note that the optimal $ECON_{EX}$ version of ECON is adopted for comparison due to its superior performances especially in recovering geometrical details.

Three metrics are adopted to show and compare the performances of different methods: Chamfer distances (CD), point to surface distances (P2S) and normal errors (L2). CD (cm) is one-directional point-to-surface distance; P2S (cm) is a bi-directional point-to-surface distance; L2 is the average difference between normal images separately rendered from reconstructed and ground-truth surfaces by four angles $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$. Generally, their average values are used to evaluate the performances of different methods when doing comparison.

Different methods may obtain models with different sizes and positions. Therefore, for fair comparison, all reconstructed models from different methods are normalized by re-scaling them to exactly fit a $1 \times 1 \times 1$ bounding box with their origins overlapped with the box center. Normal maps are computed by orthogonal projection as 512×512 with PyOpenGL for quantitative comparison.

4.3. Qualitative results

The performance comparison of different methods with images relative easy to reconstruct is shown in Figure 7. All other methods cannot recover as many details as ours for the frontal surfaces. They also return over-smooth results (*e.g.*, ICON, ECON and PIFUHD) or incorrect details with frontal ones (NormalGAN) for the back surfaces. The results of OPlanes seem to be the worst with rough and blurry results, partly due to the direct complete surface estimation from both RGB images and depth maps and partly due to the lack of priors. In general, our method can obtain best results not only on the fine unseen back surfaces but also on the fine frontal surfaces among all methods.

Experiments with more challenge RGB-D images are also taken (Figure 8), where the depths of the fronts change significantly. In this case, the frontal noise is generally higher than that shown in

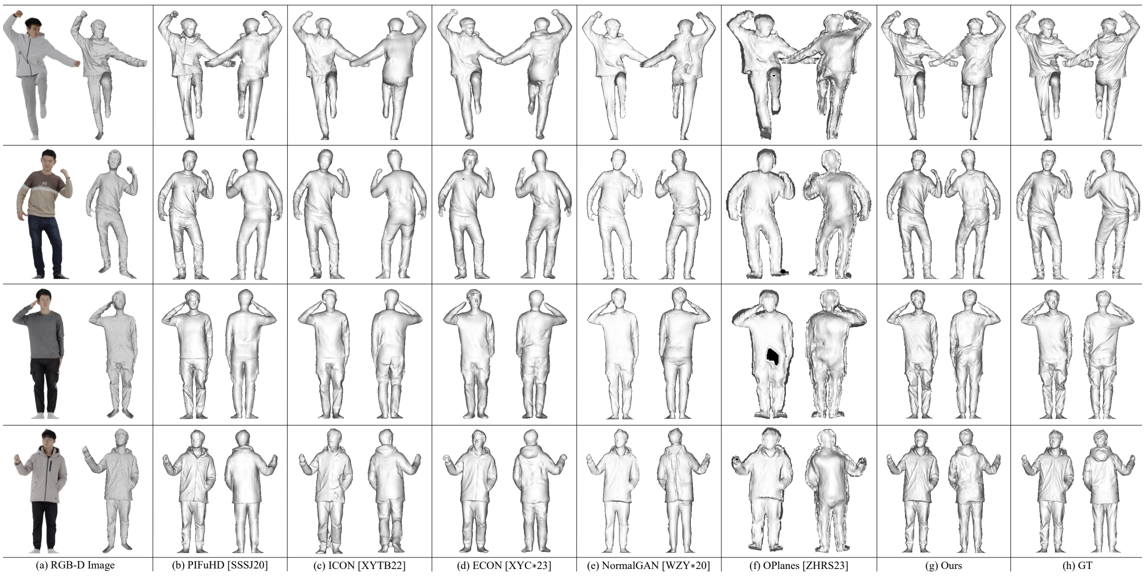

(a) RGB-D Image (b) PIFuHD [SSSJ20] (c) ICON [XYTB22] (d) ECON [XYC*23] (e) NormalGAN [WZY*20] (f) OPlanes [ZHRS23] (g) Ours (h) GT

Figure 7: *Performance comparison of different methods on easy images where their subjects are well posed and thus imaged without dramatic depth variations. The frontal and back surfaces from each method are shown on the left and right of its corresponding subfigure respectively.*

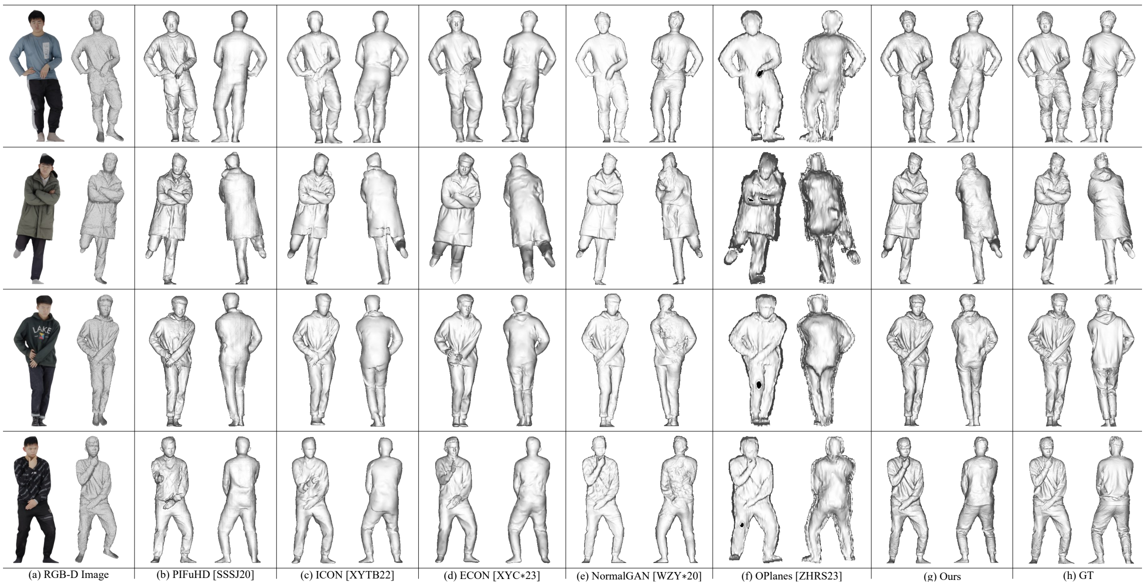

(a) RGB-D Image (b) PIFuHD [SSSJ20] (c) ICON [XYTB22] (d) ECON [XYC*23] (e) NormalGAN [WZY*20] (f) OPlanes [ZHRS23] (g) Ours (h) GT

Figure 8: *Performance comparison of different methods on hard images. Note that depth variations are significant due to the challenging poses. The frontal and back surfaces from each method are shown on the left and right of its corresponding subfigure respectively.*

Figure 7 and the unseen back can be easily mis-estimated without proper handling. The single RGB image based methods (PIFuHD, ICON and ECON) cannot obtain results with fine details. Existing signal RGB-D image based methods still obtain the over-smooth results as before, with too many frontal details due to the direct frontal-to-back inference (NormalGAN). Our method can obtain the best results with lots of correct details in both frontal and back surfaces, thanks to the proposed cross-attention based denoising method for frontal depth maps and the parametric model incorporated progressive estimation method for the unseen back depth maps.

Experiment on the real RGB-D images captured by Kinect V2 is also conducted. The parametric models for all images are obtained by PIXIE [FCB*21] and consequently the proposed reconstruction model is retrained. These images are rich with noise that is difficult to suppress by all methods. However, our method can still obtain better results, especially on the back surfaces when comparing to existing methods (Figure 10).

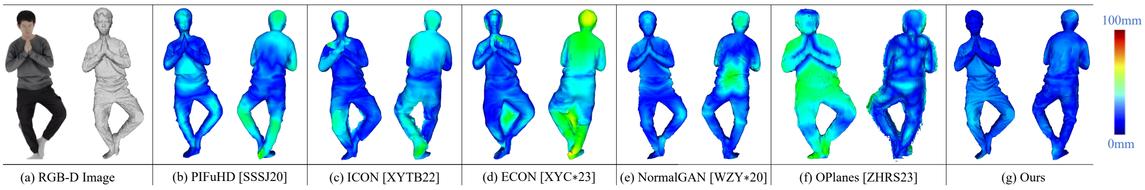


Figure 9: *Performance comparison of different methods by error maps. The frontal and back surfaces from each method are shown on the left and right of its corresponding subfigure respectively.*

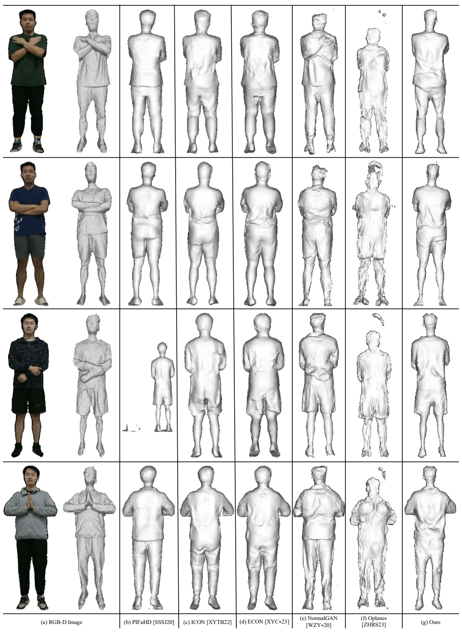


Figure 10: *Performance comparison of different methods for real RGB-D images. Reconstruction results for back surfaces are shown.*

4.4. Quantitative results

The statistical results of different methods on the accuracy of full human surface are collected (Table 1). Our method achieves the best results among all results for all metrics.

Table 1: *Statistical results of different methods. The best results are shown in bold.*

Method	CD $\times 10^{-2}$ ↓	P2S $\times 10^{-2}$ ↓	L2 ↓
PIFuHD [SSSJ20]	0.1658	0.1680	0.1197
ICON [XYTB22]	0.1916	0.1965	0.1339
ECON [XYC*23]	0.2121	0.2206	0.1516
NormalGAN [WZY*20]	0.1590	0.1549	0.1053
OPlanes [ZHRS23]	0.2105	0.2317	0.1941
Ours	**0.1227**	**0.1254**	**0.0897**

Further details from different methods on either frontal or back surfaces are also collected (Table 2). Here only the normal based L2 is used because it can better depict the geometrical accuracy than CD and P2S for such unclosed surfaces. Our method still obtains the best results among all methods for either the frontal or the back surfaces.

Table 2: *Statistical results of different methods for the frontal and back surfaces only by L2. Average means the average L2 error from both frontal and back surfaces. The best results are shown in bold.*

Methods	Front ↓	Back ↓	Average ↓
PIFUHD [SSSJ20]	0.0602	0.0632	0.0617
ICON [XYTB22]	0.0688	0.0711	0.0700
ECON [XYC*23]	0.0818	0.0820	0.0819
NormalGAN [WZY*20]	0.0558	0.0670	0.0614
OPlanes [ZHRS23]	0.1896	0.2067	0.1982
Ours	**0.0509**	**0.0612**	**0.0560**

The comparison among different methods can also be visualized by error maps (Figure 9). Our method can best capture the ground-truth model than all other methods, which again demonstrates the superiority of ours.

4.5. Ablation

Two ablation studies are conducted according to the two methods proposed in the front and back branches of our framework respectively. Our method reconstructs the frontal and back surfaces independently and, therefore, the L2 metric for either the frontal or back normal maps alone is also additionally used in each ablation study.

For the methods of front estimation, three configurations of different denoising ideas for frontal depth maps are tested, including U-Net, U-Net with the normal-maps conditioned GAN (U-Net+nGAN) which is also used by NormalGAN [WZY*20] and our proposed U-Net and CA combined cross-attention based methods (U-Net+CA). Figure 11 gives the example visual comparison results. Our method (Attention) achieves fine results and is better than other methods which obtain incorrect geometries. Table 3 gives the statistic results on the reconstructed frontal surfaces, which again shows that our proposed method is best among all configurations.

For the back estimation, three configurations of different back depth estimation methods are taken, including the normal-maps conditioned GAN (nGAN), SMPL-X initialized back depth maps plus depth-maps conditioned GAN (SMPL+dGAN) and our

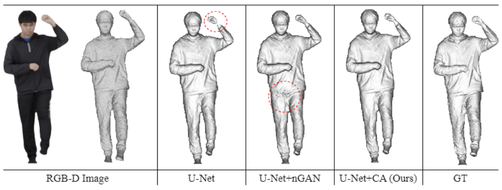


Figure 11: *Example reconstructions for different front estimation ideas in the ablation study. The red circles show the incorrect parts on the surfaces.*

Table 3: *Statistical results for different denoising ideas for frontal depth maps. The best result is shown in bold. Note: Front L2 represents the average L2 error for the frontal surfaces.*

Methods	Front L2 ↓	CD $\times 10^{-2}$ ↓	P2S $\times 10^{-2}$ ↓	L2 ↓
U-Net	0.0538	0.1344	0.1386	0.0966
U-Net+nGAN	0.0530	0.1330	0.1392	0.0904
U-Net+CA (Ours)	**0.0509**	**0.1227**	**0.1254**	**0.0897**

method, *i.e.*, the SMPL-X initialized back depth maps plus normal-maps conditioned GAN (SMPL+nGAN). Figure 12 gives the example visual comparison results. Our method (SMPL+nGAN) achieves better back results than other methods which can exhibit apparent errors. Table 4 gives the statistic results on the reconstructed back surfaces. Again, this table shows that our progressive depth map estimation method for the unseen back is the most effective among all configurations.

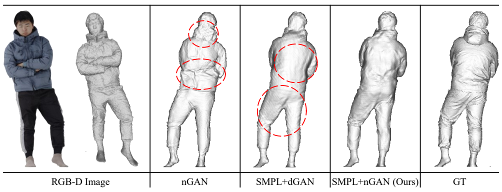


Figure 12: *Example reconstructions for different back estimation ideas in the ablation study. The red circles show the incorrect parts on the surfaces.*

Table 4: *Statistical results for different estimation ideas for back depth maps. The best result is shown in bold. Note: Back L2 represents the average L2 error for the back surfaces.*

Methods	Back L2 ↓	CD$\times 10^{-2}$ ↓	P2S$\times 10^{-2}$ ↓	L2 ↓
nGAN	0.0670	0.1574	0.1551	0.1002
SMPL+dGAN	0.0631	0.1628	0.1521	0.0954
SMPL+nGAN (Ours)	**0.0611**	**0.1227**	**0.1254**	**0.0897**

4.6. Additional experiment on light condition

Light condition may also affect the reconstruction performances. Intuitively, it might be more and more difficult to rebuild the geometries when the light turns lower and lower with the subject appearances more and more ambiguous. Here an additional experiment is taken where 20 rendering light powers from a direct point light equally decreased from 2000W and 100W according to Blender [Fou] are used. Figure 13 shows that the geometrical details of an example model almost lose nothing during the diminishing process. This is because the depth images and parametric models are not affected by such light changes and, therefore, still provide important cues to help recover surface details even under weak light conditions (*e.g.*, 100W).

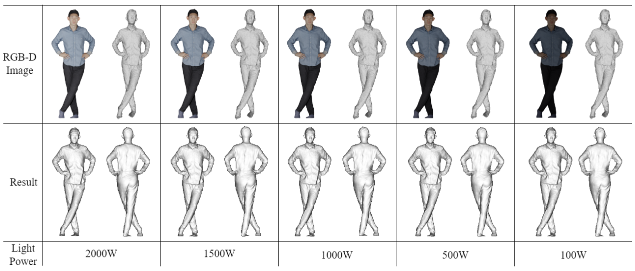


Figure 13: *Example reconstruction results of both the frontal and back surfaces of a model under five different light powers.*

Quantitative results are also collected (Figure 14). All measurements change little during the diminishing process, even though those by L2 slightly go higher with the light turning lower. These results again justify the visual performances shown in Figure 13. We can therefore conclude that this method is generally not affected by the light variation.

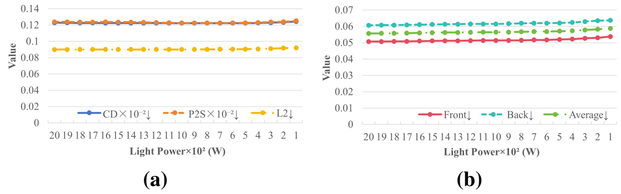


Figure 14: *Statistical results of our method under different light powers. (a): Measurements by CD, P2S and L2; (b): L2 errors and their averages for frontal and back surfaces.*

4.7. Limitations

Our method may fail when there are self-occlusions caused by the arms far away from the main body (Figure 15). In this case, the information from the occluded body parts cannot be successfully inferred from the frontal depth maps because the distances between the arms and body are too big. Consequently, the occluded arm parts and the body will be directly connected due to the triangulation based reconstruction.

In addition, the parametric body model is important for the success of the proposed method. However this model sometimes cannot be successfully estimated by existing methods, especially when the subjects are in some challenging poses, such as in hunkering down or flying kick. In this case, our method will also fail. More

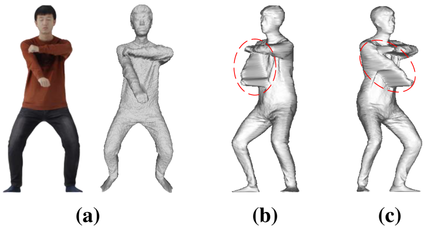


Figure 15: *Demonstration of the failure from the distant-to-body part caused self-occlusion. (a): RGB-D Image; (b): one side view of the reconstruction; and (c): another side view of the reconstruction*

robust pose and shape estimation method will make our method apply to more challenging RGB-D images.

5. Conclusions

Existing single RGB-D image based methods for the human surface reconstruction lack surface details with even frontal details often appearing in the unseen back. Noticing that the unseen back is better to be predicted without the interference of the frontal depth map due to the possibly significant difference between them, we introduce a novel framework for effective reconstruction with rich geometrical details, especially fine back surfaces. This framework takes the parametric human model as the strong prior and thus proposes a progressive depth map estimation method for the unseen back. This method combines the parametric body model and normal-maps conditioned GAN to subsequently obtain the coarse depth maps and further refined ones. This framework also includes a cross-attention based denoising method to improve the quality of the captured frontal depth maps. Experimental results show the advantages of the proposed approach over existing methods.

Acknowledgement

This work is supported by Natural Science Foundation of Anhui Province (2108085MF210) and Key Natural Science Fund of Department of Education of Anhui Province (KJ2021A0042).

References

[AZS22] Alldieck T., Zanfir M., Sminchisescu C.: Photorealistic monocular 3D reconstruction of humans wearing clothing. In *CVPR* (2022), pp. 1506–1515. 3

[BNT21] Burov A., Niessner M., Thies J.: Dynamic surface function networks for clothed human bodies. In *ICCV* (2021), pp. 10754–10764. 3

[CBTT08] Chan D., Buisman H., Theobalt C., Thrun S.: A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-Camera and Multi-Modal Sensor Fusion Algorithms and Applications-M2SFA2 2008* (2008). 3

[CFF*22] Cai H., Feng W., Feng X., Wang Y., Zhang J.: Neural surface reconstruction of dynamic scenes with monocular RGB-D camera. In *NIPS* (2022). 3

[CHW23a] Cao Y., Han K., Wong K.-Y. K.: SeSDF: Self-evolved signed distance field for implicit 3D clothed human reconstruction. In *CVPR* (2023), pp. 4647–4657. 2

[CHW23b] Cao Y., Han K., Wong K.-Y. K.: SeSDF: Self-evolved signed distance field for implicit 3D clothed human reconstruction. In *CVPR* (2023), pp. 4647–4657. 2

[CLHG22] Chen L., Li J., Huang H., Guo Y.: CrossHuman: Learning cross-guidance from multi-frame images for human reconstruction. In *ACM Multimedia* (2022), pp. 2483–2494. 2

[CPZ21] Chen L., Peng S., Zhou X.: Towards efficient and photorealistic 3D human reconstruction: A brief survey. *Visual Informatics 5*, 4 (2021), 11–19. 2

[DBPT10] Dolson J., Baek J., Plagemann C., Thrun S.: Upsampling range data in dynamic environments. In *CVPR* (2010), pp. 1141–1148. 3

[DF14] Dou M., Fuchs H.: Temporally enhanced 3D capture of room-sized dynamic scenes with commodity depth cameras. In *IEEE VR* (2014), pp. 39–44. 2

[DXD*22] Dong Z., Xu K., Duan Z., Bao H., Xu W., Lau R.: Geometry-aware two-scale PIFu representation for human reconstruction. *NIPS 35* (2022), 31130–31144. 2, 3

[FCB*21] Feng Y., Choutas V., Bolkart T., Tzionas D., Black M. J.: Collaborative regression of expressive bodies using moderation. In *3DV* (2021), pp. 792–804. 6, 8

[FLJW22] Fan L., Li Y., Jiang C., Wu Y.: Unsupervised depth completion and denoising for RGB-D sensors. In *ICRA* (2022), pp. 8734–8740. 3

[Fou] Foundation B.: Blender. https://www.blender.org/. [Online; accessed 28-May-2023]. 7, 10

[FYR*19] Fang X., Yang J., Rao J., Wang L., Deng Z.: Single RGB-D fitting: Total human modeling with an RGB-D shot. In *VRST* (2019), pp. 1–11. 2

[GFM*19] Gabeur V., Franco J.-S., Martin X., Schmid C., Rogez G.: Moulding humans: Non-parametric 3D human shape estimation from single images. In *ICCV* (2019), pp. 2232–2241. 1, 3

[GLGT19] Gu S., Li Y., Gool L. V., Timofte R.: Self-guided network for fast image denoising. In *ICCV* (2019), pp. 2511–2520. 2, 3

[GPAM*14] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial nets. In *NIPS* (2014), vol. 27, p. 2672–2680. 2, 7

[GXY*17] Guo K., Xu F., Yu T., Liu X., Dai Q., Liu Y.: Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM TOG 36*, 4 (2017), 1. 2

[HHP*23] Hu S., Hong F., Pan L., Mei H., Yang L., Liu Z.: SHERF: Generalizable human NeRF from a single image. *arXiv preprint arXiv:2303.12791* (2023). 3

[IZN*16] Innmann M., Zollhöfer M., Niessner M., Theobalt C., Stamminger M.: VolumeDeform: Real-time volumetric non-rigid reconstruction. In *ECCV* (2016), Springer, pp. 362–379. 2

[IZZE17] Isola P., Zhu J.-Y., Zhou T., Efros A. A.: Image-to-image translation with conditional adversarial networks. In *CVPR* (2017), pp. 1125–1134. 6

[JJW*23] Jiang S., Jiang H., Wang Z., Luo H., Chen W., Xu L.: HumanGen: Generating human radiance fields with explicit priors. In *CVPR* (2023), pp. 12543–12554. 3

[JL18] Jeon J., Lee S.: Reconstruction-based pairwise depth dataset for depth image enhancement using CNN. In *ECCV* (2018), pp. 422–438. 2, 3

[JZH*20] Jiang B., Zhang J., Hong Y., Luo J., Liu L., Bao H.: BCNet: Learning body and cloth shape from a single image. In *ECCV* (2020), Springer, pp. 18–35. 3

[KLL*13] Keller M., Lefloch D., Lambers M., Izadi S., Weyrich T., Kolb A.: Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *3DV* (2013), pp. 1–8. 2

[KTL15] Kwon H., Tai Y.-W., Lin S.: Data-driven depth map refinement via multi-scale sparse representation. In *CVPR* (2015), pp. 159–167. 3

[LHAY19] Li Y., Huang J.-B., Ahuja N., Yang M.-H.: Joint image filtering with deep convolutional networks. *IEEE T-PAMI 41*, 8 (2019), 1909–1923. 3

[LIPM19] Lazova V., Insafutdinov E., Pons-Moll G.: 360-degree textures of people in clothing from a single image. In *3DV* (2019), IEEE, pp. 643–653. 3

[LKH07] Lindner M., Kolb A., Hartmann K.: Data-fusion of PMD-based distance-information and high-resolution RGB-images. In *2007 International Symposium on Signals, Circuits and Systems* (2007), vol. 1, pp. 1–4. 3

[LMR*15] Loper M., Mahmood N., Romero J., Pons-Moll G., Black M. J.: SMPL: A skinned multi-person linear model. *ACM TOG 34*, 6 (2015), 1–16. 2

[LZX*23] Liao T., Zhang X., Xiu Y., Yi H., Liu X., Qi G.-J., Zhang Y., Wang X., Zhu X., Lei Z.: High-fidelity clothed avatar reconstruction from a single image. In *CVPR* (2023), pp. 8662–8672. 3

[MCL22] Moon G., Choi H., Lee K. M.: Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPR* (2022), pp. 2308–2317. 2

[MST*21] Mildenhall B., Srinivasan P. P., Tancik M., Barron J. T., Ramamoorthi R., Ng R.: NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM 65*, 1 (2021), 99–106. 3

[NFS15] Newcombe R. A., Fox D., Seitz S. M.: DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR* (2015), pp. 343–352. 2

[NSH*19] Natsume R., Saito S., Huang Z., Chen W., Ma C., Li H., Morishima S.: SiCloPe: Silhouette-based clothed people. In *CVPR* (2019), pp. 4480–4490. 3

[PCG*19] Pavlakos G., Choutas V., Ghorbani N., Bolkart T., Osman A. A. A., Tzionas D., Black M. J.: Expressive body capture: 3D hands, face, and body from a single image. In *CVPR* (2019), pp. 10975–10985. 2, 4

[PKT*11] Park J., Kim H., Tai Y.-W., Brown M. S., Kweon I.: High quality depth map upsampling for 3D-TOF cameras. In *ICCV* (2011), pp. 1623–1630. 3

[rlc] rlczddl: Awesome 3D human reconstruction. https://github.com/rlczddl/awesome-3d-human-reconstruction. [Online; accessed 28-May-2023]. 1

[RSD*12] Richardt C., Stoll C., Dodgson N. A., Seidel H.-P., Theobalt C.: Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. *Computer Graphics Forum 31*, 2pt1 (2012), 247–256. 3

[SBI18] Slavcheva M., Baust M., Ilic S.: SobolevFusion: 3D reconstruction of scenes undergoing free non-rigid motion. In *CVPR* (2018), pp. 2646–2655. 2

[SHN*19] Saito S., Huang Z., Natsume R., Morishima S., Kanazawa A., Li H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV* (2019), pp. 2304–2314. 3

[SLSC23] Song D.-Y., Lee H., Seo J., Cho D.: DIFu: Depth-guided implicit function for clothed human reconstruction. In *CVPR* (2023), pp. 8738–8747. 1, 3

[SSC*19] Sterzentsenko V., Saroglou L., Chatzitofis A., Thermos S., Zioulis N., Doumanoglou A., Zarpalas D., Daras P.: Self-supervised deep depth denoising. In *ICCV* (2019), pp. 1242–1251. 2, 3

[SSSJ20] Saito S., Simon T., Saragih J., Joo H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR* (2020), pp. 84–93. 3, 7, 9

[SXZ*22] Su Z., Xu L., Zhong D., Li Z., Deng F., Quan S., Fang L.: RobustFusion: Robust volumetric performance reconstruction under human-object interactions from monocular RGBD stream. *IEEE T-PAMI* (2022). 2

[SZ14] Simonyan K., Zisserman A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 6

[SZB*23] Sun M., Zheng Y., Bao T., Chen J., Jin G., Wu L., Zhao R., Jiang X.: Uni6Dv2: Noise elimination for 6D pose estimation. In *26th International Conference on Artificial Intelligence and Statistics* (2023), vol. 206, pp. 1832–1844. 2, 3

[TKB*23] Tretschk E., Kairanda N., BR M., Dabral R., Kortylewski A., Egger B., Habermann M., Fua P., Theobalt C., Golyanik V.: State of the art in dense monocular non-rigid 3D reconstruction. *Computer Graphics Forum 42*, 2 (2023), 485–520. 2

[WLZ*18] Wang T.-C., Liu M.-Y., Zhu J.-Y., Tao A., Kautz J., Catanzaro B.: High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR* (2018), pp. 8798–8807. 7

[WZY*20] Wang L., Zhao X., Yu T., Wang S., Liu Y.: Normal-GAN: Learning detailed 3D human from a single RGB-D image. In *ECCV* (2020), Springer, pp. 430–446. 2, 3, 4, 6, 7, 9

[XYC*23] Xiu Y., Yang J., Cao X., Tzionas D., Black M. J.: ECON: Explicit clothed humans obtained from normals. In *CVPR* (2023), pp. 512–523. 1, 2, 3, 7, 9

[XYTB22] Xiu Y., Yang J., Tzionas D., Black M. J.: ICON: Implicit clothed humans obtained from normals. In *CVPR* (2022), IEEE, pp. 13286–13296. 2, 3, 7, 9

[YLZ*20] Yan C., Li Z., Zhang Y., Liu Y., Ji X., Zhang Y.: Depth image denoising using nuclear norm and learning graph model. *ACM Transactions on Multimedia Computing, Communications, and Applications 16*, 4 (2020), 1–17. 3

[YWW*18] Yan S., Wu C., Wang L., Xu F., An L., Guo K., Liu Y.: DDRNet: Depth map denoising and refinement for consumer depth cameras using cascaded CNNs. In *ECCV* (2018), pp. 151–167. 3

[YYDN07] Yang Q., Yang R., Davis J., Nistér D.: Spatial-depth super resolution for range images. In *CVPR* (2007), pp. 1–8. 3

[YZG*18] Yu T., Zheng Z., Guo K., Zhao J., Dai Q., Li H., Pons-Moll G., Liu Y.: DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR* (2018), pp. 7287–7296. 2

[YZG*21] Yu T., Zheng Z., Guo K., Liu P., Dai Q., Liu Y.: Function4D: Real-time human volumetric capture from very sparse consumer RGBD sensors. In *CVPR* (2021), pp. 5746–5756. 3, 7

[ZHRS23] Zhao X., Hu Y.-T., Ren Z., Schwing A. G.: Occupancy planes for single-view RGB-D human reconstruction. In *AAAI* (2023). 2, 3, 7, 9

[ZLWY23] Zheng R., Li P., Wang H., Yu T.: Learning visibility field for detailed 3D human reconstruction and relighting. In *CVPR* (2023), pp. 216–226. 3

[ZSG*18] Zollhöfer M., Stotko P., Görlitz A., Theobalt C., Niessner M., Klein R., Kolb A.: State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum 37*, 2 (2018), 625–652. 2

[ZTZ*23] Zhang H., Tian Y., Zhang Y., Li M., An L., Sun Z., Liu Y.: PyMAF-X: Towards well-aligned full-body model regression from monocular images. *IEEE T-PAMI* (2023). 6

[ZW16] Zhang X., Wu R.: Fast depth image denoising and enhancement using a deep convolutional network. In *ICASSP* (2016), pp. 2499–2503. 2, 3

[ZYW*19] Zheng Z., Yu T., Wei Y., Dai Q., Liu Y.: DeepHuman: 3D human reconstruction from a single image. In *ICCV* (2019), pp. 7739–7749. 2