Geometry-aware Two-scale PIFu Representation for Human Reconstruction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although PIFu-based 3D human reconstruction methods are popular, the quality of recovered details is still unsatisfactory. In a sparse (*e.g.*, 3 RGBD sensors) capture setting, the depth noise is typically amplified in the PIFu representation, resulting in flat facial surfaces and geometry-fallible bodies. In this paper, we propose a novel geometry-aware two-scale PIFu for 3D human reconstruction from sparse, noisy inputs. Our key idea is to exploit the complementary properties of depth denoising and 3D reconstruction, for learning a two-scale PIFu representation to reconstruct high-frequency facial details and consistent bodies separately. To this end, we first formulate depth denoising and 3D reconstruction as a multi-task learning problem. The depth denoising process enriches the local geometry information of the reconstruction features, while the reconstruction process enhances depth denoising with global topology information. We then propose to learn the two-scale PIFu representation using two MLPs based on the denoised depth and geometry-aware features. Extensive experiments demonstrate the effectiveness of our approach in reconstructing facial details and bodies of different poses and its superiority over state-of-the-art methods.

1 Introduction

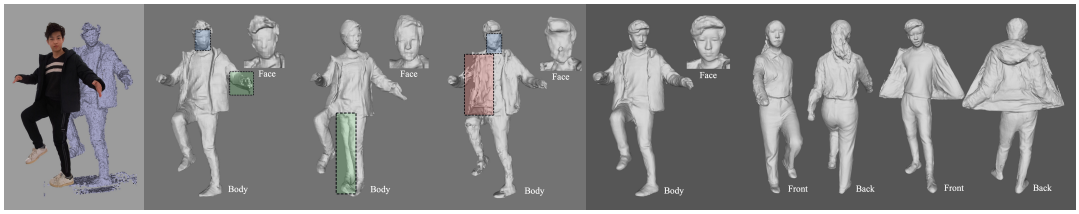
Three-dimensional human reconstruction, which aims to obtain a dense surface geometry from single-view or multi-view human images, is a fundamental topic in computer vision and computer graphics. While reconstructing high-fidelity 3D human models is possible using commercial multi-view/stereo software under the customized studio setting [12, 55, 78, 27, 82], it is highly desirable to lift the studio setting constraint, which may be inaccessible to most users. Low-cost RGBD sensors have recently become popular in 3D human reconstruction, and tracking-based methods are developed to fuse the depth data from RGBD sensors for reconstruction [63]. During fusion, the estimation of non-rigid human body deformation is essential to improve the reconstruction quality [62, 96]. However, it is technically challenging to ensure the stability of the depth fusion algorithm, due to occlusions and severe noise in the depth data.

Recently, learning-based 3D human reconstruction methods have significantly simplified the capture setting. The parametric human model [57, 66, 64] is introduced to reduce the modeling difficulties from complex poses. After training on images-to-model pairs, methods [43, 100] may even reconstruct 3D human shapes from single images. However, these methods typically only obtain naked-like human bodies. As detail requirements increase, the focus of learning-based human reconstruction methods has been shifted to the implicit representation, *e.g.*, pixel-aligned implicit function (PIFu) [70, 71, 37]. PIFu-based methods [70] can reconstruct human bodies with different types of details (*e.g.*, hair and clothing) without utilizing predefined templates. However, they often produce topology

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

errors in the reconstructed human models, especially in the regions that are invisible or where the input depth is highly noisy (*e.g.*, hairs). While PIFuHD [71] partially alleviates this problem by synthesizing the body's normal maps at both front and back sides, the details reconstructed from the synthesized normals may not be consistent with the target. Hence, some methods [37, 95] resort to the multi-view feature fusion scheme to reduce these topology errors.

Although implicit methods develop fast, we observe that the quality of their recovered details is still unsatisfactory under the sparse capture settings (*e.g.*, 3 RGBD sensors). Due to the less overlapping between sparse views, the input depth noise is typically amplified, increasing the difficulty of performing stable reconstruction. In addition, due to the ineffective fusion of RGB and depth features, these methods may not easily reconstruct high-frequency details. As a result, we often observe flat or incorrect facial surfaces, body geometries with topology errors, as shown in Fig. 1(b,c,d).



(a) Raw RGBD (b) PIFu [70] (c) PIFuHD [71] (d) IPNet [3] (e) **Ours** (f) Our Additional Results

Figure 1: Given sparse and noisy inputs, with one of the three RGBD views as shown in (a), existing methods such as Multi-view RGBD-PIFu [70] (b), PIFuHD [71] (c) and IPNet [3] (d) tend to produce over-smoothed facial details (b,d) or topology errors (c,d), due to the amplified noise in sparse views. Our method learns the geometry-aware two-scale PIFu representation, which can produce vivid facial/hair details and accurate bodies under different poses (e,f).

In this paper, we propose a geometry-aware two-scale PIFu representation for reconstructing digital humans from sparse, noisy inputs. Our method is based on two observations. First, depth denoising and 3D reconstruction are complementary to each other. The former task preserves local geometric fidelity, while the latter task provides global topology guidance. Second, while a function of high complexity may not be easy to express, it is much easier to approximate it piecewise (*e.g.*, in two parts). Inspired by these observations, we first formulate depth denoising and 3D reconstruction as a multi-task learning problem. We encode RGB and depth separately, and use fused features to perform depth denoising, enriching deep features with local geometric information. Based on the denoised depth and geometry-aware features, we propose using two MLPs to represent the PIFus for the face (particular region) and the body, respectively. This separate modeling increases the network capacity for handling details at different scales. As shown in Fig. 1(e, f), our method can produce results with high-fidelity body and facial details for different actions. Our main technical contributions are:

1) We propose a novel geometry-aware PIFu method for digital human reconstruction from sparse, noisy RGBD images. Our approach exploits the complementary properties of depth denoising and 3D reconstruction to learn robust geometry information while suppressing the input noise.

2) We propose to learn a two-scale PIFu representation based on geometry-aware features, by using two individual MLPs for the face and the body separately. The two-scale formulation enhances the network capacity in producing high-frequency facial details and smooth body surfaces.

3) Extensive experiments demonstrate that the proposed method can produce high-quality digital human reconstruction results, based on noisy depth maps taken by three Azure Kinect sensors.

2 Related Works

Tracking-based Human Reconstruction methods [47, 63, 103, 62, 28, 17, 46, 16, 96, 30, 76, 95] track human motions and infer the non-rigid deformations of the references to reconstruct 3d human mesh in a temporal fusion manner. The references are typically parameterized as the 3D poses/skeletons [22, 93, 94, 87, 96, 30, 31, 52] and/or parametric body models [6, 97, 96, 42] (*e.g.*,SMPL [57, 66, 42]). Some methods combine tracking with segmentation [56] or optical flow estimation [60] to help compute the references for reconstruction. To tackle the occlusion and

large motion problems, some methods develop high-end systems for the dense capture of human performance, consisting of a large number (up to 100) of RGBD sensors [12, 55, 78] or custom color lights [27, 82] (*e.g.*, 1,200 individually controllable light sources in the acquisition setup in [82]).

Learning-based Human Reconstruction methods [102, 74, 1, 21, 84, 61, 101, 80, 88, 95] leverage the neural 3D representation for reconstructing the geometry and/or texture details. Some methods obtain 3D human meshes from a single RGB [74] or RGBD [84] image by incorporating the SMPL template [77], which are ineffective for deformed human bodies. Other methods adopt the image-to-image translation pipeline to regress the 3D mesh via 2D estimations of intermediate textures [1], silhouettes [61], and depths [21], while some approaches jointly exploit 2D and 3D information, *e.g.*, body joints and per-pixel shading information [102], and 2D/3D poses and segmentation map [80]. These methods typically suffer from over-smoothed surfaces in occluded regions.

Recently, the pixel-aligned implicit function (PIFu) [70] has attracted much attention for 3D human reconstruction due to its effective implicit representation, and PIFuHD [71] estimates normal maps to reduce the geometry errors in the occluded regions based on PIFu. Due to its success, many methods promote PIFu with voxel-alignment [99, 37, 34], deformation field [35, 38], real-time approach [48], illumination [2], monocular fusion [51], sparse-view temporal fusion [95], or apply it for point clouds based human reconstruction [11, 3, 59]. However, PIFu-based methods typically lack fine facial details in the sparse capture setting due to its all-in-one implicit 3D representation learning. In this paper, we propose the geometry-aware Two-scale PIFu representation to tackle this problem.

Depth Image Denoising is essential for depth images captured from consumer-level depth sensors (*e.g.*, Azure Kinect). Traditional filtering-based methods [44, 8, 15, 68, 33, 73, 32] enhance depth data from various sensors. Color or infrared images are used to help depth image denoising [14, 91, 53, 65, 54, 20, 90, 58, 32]. These methods typically assume the intensity image to be edge-aligned with the depth image, and they tend to produce artifacts when this assumption is violated. Some methods [26, 72, 63, 13, 36, 9, 62, 29] fuse multiple frames to refine the depth images, but they tend to produce over-smoothed depth images.

Learning-based depth image denoising methods are proposed based on dictionary-learning [45] and deep-learning [25, 50, 49, 41, 88, 75]. Recently, Yan *et al.* [88] propose the DDRNet for deep denoising using fused geometry and color images. Their method tends to produce inconsistent results. Sterzentsenko *et al.* [75] propose a self-supervised method that leverages photometric supervision from a differentiable rendering model to smooth the depth noise. However, the photometric supervision lacks 3D geometry information and their approach tends to produce homogeneous results. In this paper, we propose to formulate depth denoising and 3D reconstruction as a multi-task learning problem. The global topology information facilitates the depth denoising significantly.

Monocular Face Reconstruction methods aim to reconstruct personalized faces from monocular data. The parametric face model, *e.g.*, 3D morphable model (3DMM) [5, 4, 69] or multi-linear blend shapes [7, 81, 23], are typically used in conventional methods. However, these methods may not reconstruct accurate or dense facial geometry. Deep-learning-based methods [39, 92, 19, 24] that incorporate face landmarks or the 3DMM model for face reconstruction also tend to produce coarse reconstructions results. Other methods [79, 40, 19, 67, 98] estimate the facial dense shape instead of the low-dimensional template parameters. All these approaches only reconstruct faces. In this paper, we apply the PIFu representation and extend it to two scales for modeling human bodies and faces. Our formulation enables producing vivid facial expressions with accurate body reconstructions.

3 Proposed Method

We propose the geometry-aware two-scale PIFu method, to address the problem of PIFu-based approaches that reconstruct flat facial and geometry-fallible body surfaces in the sparse capture setting. First, to handle the noise issue of the sparse capture setting, we propose to formulate the depth denoising and the body reconstruction processes in the multi-task learning manner to exploit their complementary properties. Second, although the face only occupies a small proportion of the whole human model, it typically contains more high-frequency (*e.g.*, vivid expressions) than other parts and

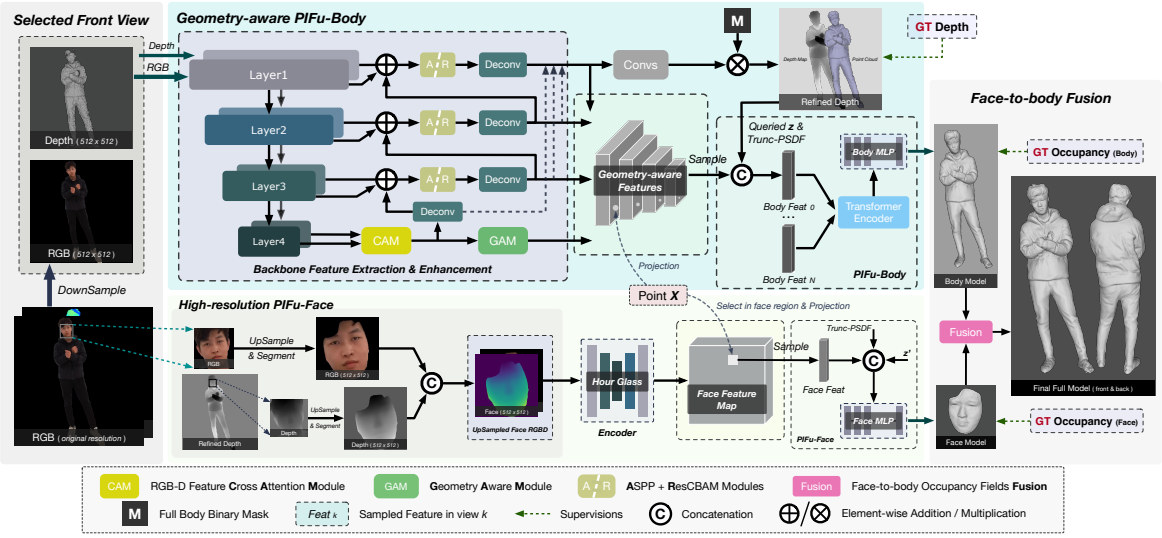


Figure 2: Proposed method overview. Given sparse and noisy RGBDs as inputs, *Geometry-aware PIFu-Body* performs depth denoising and predicts the body occupancy field. *High-resolution PIFu-Face* predicts the face occupancy field with fine-grained details. The body and face occupancy fields are fused to produce final results via the *Face-to-body Fusion* scheme.

plays a vital role in assessing the reconstruction fidelity. To this end, we propose the two-scale PIFu representation to allocate more network capacity for face reconstruction.

As illustrated in Fig. 2, our network contains three parts: (1) the *Geometry-aware PIFu-Body*, $\mathcal{F}_b$ that predicts the body occupancy field O_b and the refined depth maps from the noisy RGBD images of $\mathcal{N}$ (*i.e.*, three) perspective views; (2) the *High-resolution PIFu-Face*, $\mathcal{F}_f$ which obtains the fine-grained face occupancy field O_f, using refined depth map from $\mathcal{F}_b$ and the high-resolution RGB image of the front view as inputs; (3) the light-weight *Face-to-body Fusion*, $\mathcal{W}$ that reconstructs the full human model by fusing the face and body occupancy fields (*i.e.*, O_b and O_f).

3.1 Geometry-aware PIFu-Body: $\mathcal{F}_b$

PIFu-based methods predict 3D occupancy fields in an implicit manner, enabling image-aligned features to be aware of global topological information, which helps suppress depth noise. For example, in Fig. 3(e), the holes can be filled using only geometric supervision. However, this depth map tends to be over-smoothed, and the reconstructed surface also lacks details (Fig. 3(f)). On the other hand, with only depth supervision, image-to-image depth denoising can fill holes and add details. However, the obtained depth involves incorrect details (*e.g.*, face), as shown in Fig. 3(g), due to the lack of 3D geometric guidance. Based on these observations, we formulate depth denoising and PIFu-based occupancy estimation in a multi-task learning manner. It exploits the global topological information of the 3D occupancy field to guide the denoising process, and the local high-frequency details of refined depth to improve occupancy estimation (Fig. 3(h,i,j)).

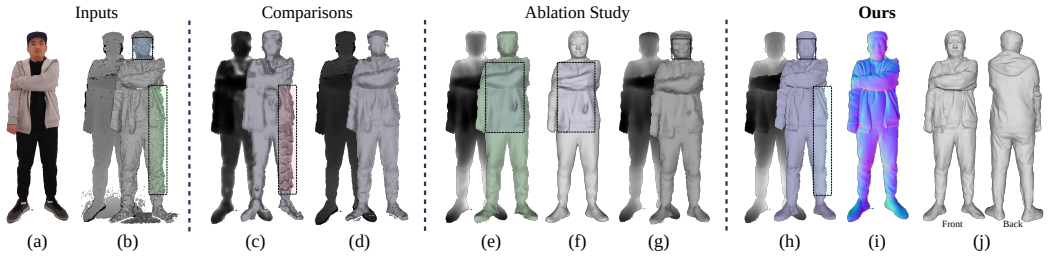


Figure 3: Visualization of depth denoising and the subsequent reconstruction results. Raw RGB and Depth images (a,b). Results of two depth denoising methods, [88] and [75] (c,d). Our results of only using occupancy supervision (e,f). Our refined depth under depth supervision only (g). Our refined depth, its normal map, and full-body mesh (h,i,j).

Formulation of $\mathcal{F}_b$. Given the triplet ($\mathbf{T}$) of RGB ($\mathbf{I}$), depth ($\mathbf{D}$) images, body binary masks ($\mathbf{M}$) from $\mathcal{N}$ perspective views, and the query point $\mathbf{X} \in \mathbb{R}^3$ as inputs, we formulate the $\mathcal{F}_b$ to predict both the body occupancy value $\sigma_b \in [0, 1]$ and the refined depth maps $\mathbf{D}_{rf}$, as:

$$\mathcal{F}_b(\mathbf{X}, \mathbf{T}) = \{\mathcal{M}_b(\mathcal{A}(\{\mathcal{B}(\psi_g(\mathbf{T}_i), \mathbf{x}_i), \mathcal{C}_i(\mathbf{X})\}_{i=1,\dots,\mathcal{N}})), \mathcal{D}(\psi_g(\mathbf{T}_i))\} := \{\sigma_b, \mathbf{D}^i_{rf}\}, \quad (1)$$

where $\psi_g(\cdot)$ is a mapping function that encodes $\mathbf{T}$ into multi-scale Geometry-aware features, $\mathbf{x}_i = \pi_i(\mathbf{X}) \in \mathbb{R}^2$, is the 2D projection of point $\mathbf{X}$ at i-th view, and $\mathbf{z}_i \in \mathbb{R}$ is the depth of $\mathbf{X}$ in the local coordinate system of the i-th view. $\mathcal{C}_i(\mathbf{X}) = [\mathbf{z}_i, p_i(\mathbf{X})]$ where $p_i(\mathbf{X}) = \mathcal{T}(\mathbf{z}_i - \mathcal{B}(\mathbf{D}^i_{rf}, \mathbf{x}_i)) \in [-\delta_p, \delta_p]$, is the truncated PSDF value as in [95]. $\mathcal{B}(\cdot, \mathbf{x}_i)$ is the sampling function to obtain pixel-aligned 2D features $\mathcal{B}(\psi_g(\mathbf{T}_i), \mathbf{x}_i)$ and depth information $\mathcal{C}_i(\mathbf{X})$, which are processed by the multi-view feature aggregation module $\mathcal{A}$ and further fed into the implicit function $\mathcal{M}_b$ for occupancy querying. Meanwhile, the decoder $\mathcal{D}(\cdot)$ processes features $\psi_g(\mathbf{T}_i)$ for depth denoising.

Geometry-aware Features: $\psi_g(\mathbf{T}_i)$. The geometry-aware mapping $\psi_g(\cdot)$ aims to exploit the complementary properties of depth denoising and occupancy field estimation. Hence, it is expected to fuse the multi-modal RGB-D inputs effectively. To handle their modal discrepancy, we use two independent *HRNets* [83] (Fig. 2) to process the RGB ($\mathbf{I}$) and depth ($\mathbf{D}$) inputs respectively, where $\{\mathbf{I}, \mathbf{D}\}$ will be pre-processed to filter out the background via $\{\mathbf{I}', \mathbf{D}'\} = \mathbf{M}(\mathbf{I}, \mathbf{D})$. To fuse the RGB and depth backbone features, we propose a novel Cross Attention Module (CAM) on the highest level of backbone features, and use the CAM output feature to guide the fusion of lower levels in a level-wise manner. We also propose a novel Geometry Aware Module (GAM) to enrich the CAM output features with high-frequency information. The enhanced features form the features $\psi_g(\mathbf{T}_i)$.

Cross Attention Module (CAM). The CAM aims to fuse the RGB and depth features by computing their non-local correlations. Since depth is typically noisier than RGB information, we first compute the self-attention map based on the RGB feature and then use it to reweight both RGB and depth features before they are fused. The architecture of CAM is shown in Fig. 4, of which the implementation is based on the non-local model [85] but we extend it to handle the RGB-D fusion. Specifically, the fused feature Y can be written as:

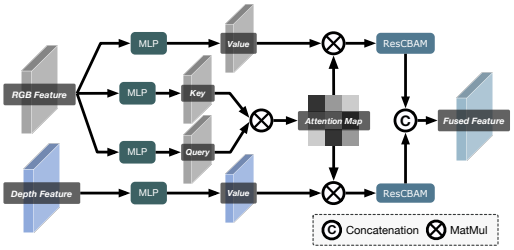


Figure 4: Cross Attention Module (CAM).

$$Y = \mathcal{R}_r(g_r(F_r) \otimes \kappa(F_r)) \copyright \mathcal{R}_d(g_d(F_d) \otimes \kappa(F_r)), \quad (2)$$

where $\mathcal{R}_r$ and $\mathcal{R}_d$ represent two ResCBAMs [86] used for calculating the local attentions for fusion, © is the channel-wise concatenation and $\otimes$ denotes the matmul operation. F_r and F_d indicate the RGB and depth backbone features output by Layer4 (Fig. 2). $\kappa(F_r)$ computes the non-local feature affinity map as $\kappa(F_r) = softmax(\theta(F_r)^T \otimes \phi(F_r))$, where θ and ϕ are learnable linear embedding functions. g_r, g_d are two functions to compute the value features of F_r and F_d.

Geometry Aware Module (GAM). The GAM aims to enrich the fused features of CAM with high-frequency information for reconstructing geometric details. To this end, we propose to calculate the depth contrasted feature between a local region and its surroundings to capture high-frequency depth variations, as shown in Fig. 5. Specifically, we split out the RGB and depth features: F'_r, F'_d, according to the channels number, and calculate the contrasted feature for F'_d as:

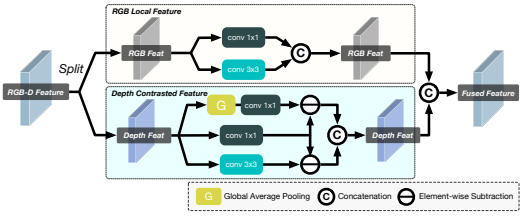
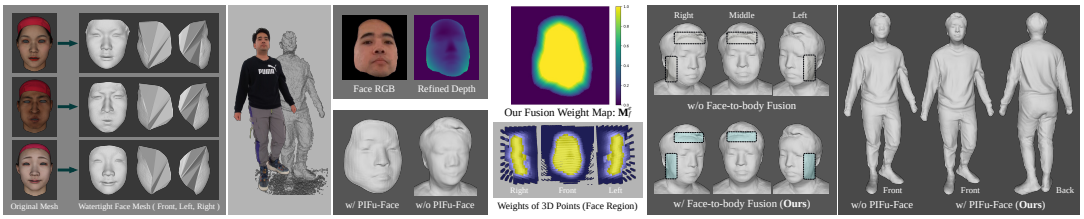


Figure 5: Geometry Aware Module (GAM).

$$C_d = (f_l(F'_d) - f_g(F'_d)) \copyright (f_l(F'_d) - f_l(\mathcal{G}(F'_d))), \quad (3)$$

where f_l denotes the local convolution with a 1x1 kernel and f_g denotes the context convolution with a 3x3 kernel(dilate rate=x). $\mathcal{G}$ is a global average pooling operation. © is the concatenation operation. The second item after © represents the local depth feature of the pixels relative to the global feature F'_d. As a result, C_d amplifies high-frequency signals at depth transitions, benefiting the prediction of these details (*e.g.*, hairs in Fig. 1). For F'_r, we maintain its information through local convolutions.

3.2 High-resolution PIFu-Face: $\mathcal{F}_f$



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

Figure 6: Facial mesh closure on *FaceScape* [89] dataset for training (a). Raw RGBD input images (b). Our reconstructed face model (w/ and w/o PIFu-Face) (c). The weight map $\mathbf{M}_f^e$ and 3D weights for *Face-to-body Fusion* (d). Comparison results on *Face-to-body Fusion* and PIFu-Face (e,f).

Face regions typically have more high-frequency components than the body (*e.g.*, mouth *v.s.*soft clothing) (Fig. 6(a,b,f)). Even enhanced with depth denoising, the all-in-one PIFu ($\mathcal{F}$) still struggles to represent high-frequency facial details (*e.g.*, nose & mouth in Fig. 6(c,f)). Hence, we propose to express the implicit function $\mathcal{F}$ in a piece-wise manner (*i.e.*, $\mathcal{F}_b$ and $\mathcal{F}_f$) to reduce the complexity of joint occupancy estimation while producing vivid facial and body details. Specifically, we propose to learn the high-frequency facial PIFu representation ($\sigma_f \in [0, 1]$) conditioned on the high-resolution face image ($\mathbf{I}_f$ cropped from the frontal view) and the corresponding denoised facial depth map $\mathbf{D}_{rf}^f$.

Formulation of $\mathcal{F}_f$. Given the above inputs (denoted as $\mathbf{T}_f = \{\mathbf{I}_f, \mathbf{D}_{rf}^f, \mathbf{M}_f\}$), we formulate the $\mathcal{F}_f$ along the query point $\mathbf{X}_f \in \mathbb{R}^3$ within the face regions as:

$$\mathcal{F}_f(\mathbf{X}_f, \mathbf{T}_f) = \mathcal{M}_f(\mathcal{B}(\mathcal{H}_f(U^{\uparrow}(\mathbf{T}_f)), \mathbf{x}_f), \mathcal{C}_f(\mathbf{X}_f)) := \sigma_f, \tag{4}$$

where $\mathcal{H}_f$ denotes the feature extractor for facial images $\mathbf{T}_f$. The function $U^{\uparrow}(\cdot)$ up-samples $\mathbf{T}_f$ to the same resolution as $\mathbf{T}$. $\mathcal{M}_f$ is the implicit function for querying O_f. $\mathcal{C}_f(\mathbf{X}_f)$ is defined the same as $\mathcal{C}_i(\mathbf{X})$, but we use the up-sampled masked facial depth $\mathbf{M}_f(U^{\uparrow}(\mathbf{D}_{rf}^f))$ to compute $p_f(\mathbf{X})$.

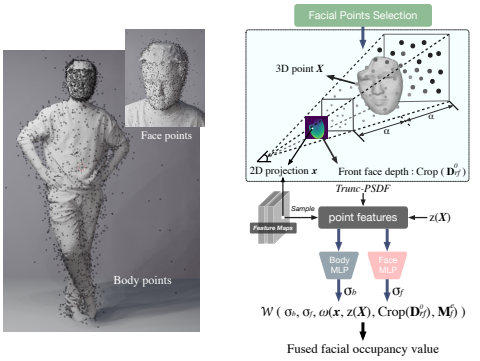
Facial Points Selection. To determine the facial points for inferring σ_f, we select the points among all the querying points of which the projection $\mathbf{x}_f \in \mathbb{R}^2$ falls inside the bounding-box $\mathbf{R}_f$ of $\mathbf{D}_{rf}^f$ and absolute PSDF value is less than α as $\mathbf{X}_f$ (Fig. 7(b)). We set a flag $v_f(\cdot)$ to mark the facial points as:

$$v_f(\mathbf{X}) = \begin{cases} 1 & \mathbf{x}_f \in \mathbf{R}_f \ \& \ \mathrm{abs}(\mathbf{z}_f - \mathcal{B}(\mathbf{D}_{rf}^f, \mathbf{x}_f)) < \alpha \\ 0 & else \end{cases}. \tag{5}$$

3.3 Face-to-body Occupancy Fields Fusion: $\mathcal{W}$

Simply merging the reconstructed face and body (*i.e.*, replacing σ_b with σ_f for the facial points) would result in the discontinuity artifacts at the stitching (Fig. 6(e), 1st row). To address this issue, we propose to fuse O_b and O_f via adaptive weights calculated in 3D space. As shown in Fig. 6(d), we compute a 2D fusion weight map ($\mathbf{M}_f^e$) in the x-y plane by eroding edges of the facial mask $\mathbf{M}_f$. Along the z axis, we compute the weights through a Gaussian distribution model of the PSDF values. Then, we formulate the joint probability distribution of the final fusion weight ω as:

$$\omega = \mathcal{B}(\mathbf{M}_f^e, \mathbf{x}_f) \cdot \exp(-\beta \cdot (\mathbb{P}(\mathbf{D}_{rf}^f, \mathbf{X}_f))^2), \tag{6}$$



(a) Points Sampling (b) Occupancy Fields Fusion

Figure 7: Equal facial and body points sampling (a). Face-to-body fusion (b).

where $\mathbb{P}(\cdot, \cdot)$ denotes the function to calculate the PSDF value, *i.e.*, $\mathbb{P}(\mathbf{D}_{rf}^f, \mathbf{X}_f) = \mathbf{z}_f - \mathcal{B}(\mathbf{D}_{rf}^f, \mathbf{x}_f)$ and the parameter β is set to $1e^3$ in default. In Eq. 6, the first term yields smaller values from the center of the face region to its boundaries, which improves the smoothness around the stitching. The second term emphasizes the occupancy values computed by $\mathcal{F}_f$ before and after the face surface. Hence, we can leverage ω to fuse the σ_b and σ_f as: $\mathcal{W}(\sigma_b, \sigma_f, \omega) = \omega \cdot \sigma_f + (1 - \omega) \cdot \sigma_b$.

3.4 Loss Function

We adopt the extended Binary Cross Entropy (BCE) loss [71] to supervise the predicted occupancy values σ_b and σ_f on the sampled body and facial points $\tilde{\mathbf{X}} = [\mathbf{X}_b, \mathbf{X}_f]$, which can be written as:

$$L_\sigma = \mu_0 \cdot \sum_{\mathbf{X}_b \in \mathcal{S}_0} \mathcal{L}_B(\sigma_b, \sigma_b^*) + \mu_1 \cdot \sum_{\mathbf{X}_f \in \mathcal{S}_1} \mathcal{L}_B(\sigma_f, \sigma_f^*), \tag{7}$$

where σ_b^*, σ_f^* are the ground-truth occupancy values for $\mathbf{X}_b$ and $\mathbf{X}_f$. $\mathcal{S}_0$ and $\mathcal{S}_1$ denote the sampled sets. $\mathcal{L}_B$ represents the BCE loss, μ_0 and μ_1 are the weights to balance PIFu-Body and PIFu-Face.

Regularization Term. We propose a *Regularization Loss* (L_{reg}) to reduce the artifacts in depth-jumping regions during multi-views aggregation, as:

$$L_{reg} = \sum_{\mathbf{X}_b \in \mathcal{S}_j} \mathcal{L}_2(\mathbf{n}(\mathbf{X}_b, \mathbf{T}), \mathbf{n}(\mathbf{X}_b + \epsilon, \mathbf{T})), \tag{8}$$

where $\mathcal{S}_j \in \mathcal{S}_0$ is the set of points projected on the depth-jumping regions (refer to the supplemental for details of obtaining these regions). The parameter ϵ is a small random uniform 3D perturbation. $\mathbf{n}(\mathbf{X}_b, \mathbf{T}) \in \mathbb{R}^3$ is the normal vector at $\mathbf{X}_b$, defined as $\nabla_{\mathbf{X}_b}\mathcal{F}_b(\mathbf{X}_b, \mathbf{T}) / \|\nabla_{\mathbf{X}_b}\mathcal{F}_b(\mathbf{X}_b, \mathbf{T})\|_2$. Eq. 8 encourages the normals of $\mathbf{X}_b$ to be consistent with those of the points sampled in its neighborhood, hence enhancing the smoothness along the stitching boundaries.

Depth Denoising Term. We penalize the per-pixel difference between $\mathbf{D}_{rf}$ and the rendered ground-truth depth map $\mathbf{D}_{gt}$. We also penalize the error of calculated normal maps to prevent $\mathbf{D}_{rf}$ from becoming blurred. The loss for depth denoising is formulated as:

$$L_D = \rho_D \cdot \sum_{s}^{S} \lambda_s \mathcal{L}_1(\mathbf{d}_{rf}^s(p), \mathbf{d}_{gt}^s(p)) + \rho_N \cdot \mathcal{L}_2(\mathbf{n}_{rf}(p), \mathbf{n}_{gt}(p)), \tag{9}$$

where $\mathbf{d}_{rf}^s(p)$ denotes the s-scale (S=4, the scale of $\psi_g(\mathbf{T})$) predicted depth value of $\mathbf{D}_{rf}$ in pixel p. $\mathbf{n}_{rf} \in \mathbb{R}^3$ is the normal vector of $\mathbf{N}_{rf}$ in pixel p, where $\mathbf{N}_{rf}$ is the normal map computed from $\mathbf{D}_{rf}$. $\mathbf{d}_{gt}^s(p)$ and $\mathbf{n}_{gt}$ are the corresponding ground-truth value and vector. $\mathcal{L}_1$ and $\mathcal{L}_2$ are the smooth $L1$ loss and $L2$ loss. The weights ρ_D, ρ_N and λ_s are used for balancing different loss terms.

The whole loss function can be defined as $L = L_\sigma + \lambda_{reg} \cdot L_{reg} + \lambda_D \cdot L_D$, where the weights λ_{reg}, λ_D are the balance terms. Refer to the supplemental for detailed implementation information.

4 Experiments

Datasets and Evaluation Metrics. We use the *THuman2.0* [95] dataset which contains 500 high-quality 3D human scans to train and validate our network. We split the dataset into training and test sets with a ratio of 4:1. We rotate each scan along the yaw axis and render RGBD body portraits at every 6-degree rotation with CUDA acceleration. For PIFu-Face, we pretrain $\mathcal{F}_f$ using the *FaceScape* [89] dataset, which contains 3D head models of different people and expressions. We crop the face regions and make the cropped meshes watertight (Fig. 6(a)). For the input raw depth maps, following [18], we synthesize the sensor noise on $\mathbf{D}_{gt}$ to obtain $\mathbf{D}$. For σ_b^* and σ_f^*, we follow the sampling strategy of PIFuHD [71] to sample body and facial points (Fig. 7(a)), and compute their occupancy values as the ground-truth labels.

We adopt the Point-to-Surface (P2S) distance(cm) and Chamfer distance(cm) for mesh, *L2* ($1e^{-1}$) and Cosine distance ($1e^{-3}$) for normals as the metrics to measure the errors between the reconstructed and the ground-truth surfaces. Lower metric values indicate better performance.

4.1 Comparisons with the State-of-the-arts

We compare our method with six state-of-the-arts human reconstruction methods, including the Multi-view RGBD-PIFu [70], PIFuHD [71], StereoPIFu [37], IPNet [3], DoubleFusion [96] and Function4D [95]. Among them, the DoubleFusion [96] and Function4D [95] are tracking-based, Multi-view RGBD-PIFu [70], PIFuHD (single RGB image) [71], StereoPIFu (stereo RGB images) [37] are PIFu-based, and the IPNet [3] is implicit function based.

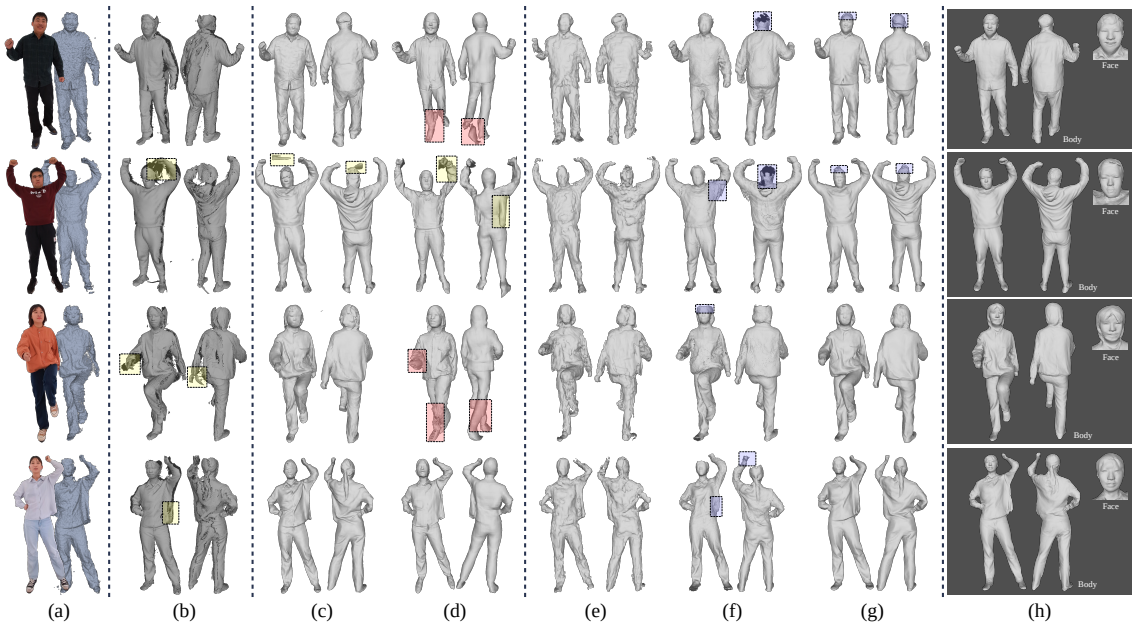


Figure 8: Qualitative comparisons on our captured real data, between the proposed method and five latest state-of-the-art methods. RGBD images of the front view (a). Our refined depths (fused point clouds) (b). Multi-view RGBD-PIFu [70] (c). PIFuHD [71] (d). IPNet [3] (e). DoubleFusion [96] (f). Function4D [95] (g). Ours (h). Zoom in to see the details.

Quantitative Comparisons. Tab. 1 reports the quantitative comparisons on our test set between the proposed method and four PIFu (or implicit function)-based approaches. Our method ranks in the first place under all metrics, exceeding the second-best results with a 16.72% reduction in Chamfer and P2S distances, a 9.50% reduction in *L*2 and Cosine distances. The PIFuHD [71] reconstructs 3D human bodies from a single RGB image, which cannot handle the topology errors, resulting in the low reconstruction accuracy. The IPNet [3] receives the fused point clouds as inputs, but it still cannot handle the significant noise issues in depths. It also has to fit the SMPL, which may not handle the complex poses. The StereoPIFu [37] can produce reasonable results from the front view but cannot handle the topological errors hidden in other perspectives. Obtaining stereo pairs from multiple views may be a solution, but it significantly increases the computational cost introduced by its 3D voxel features. For RGBD-PIFu [70], we use multi-view RGBDs as inputs to handle the topology errors. However, it still fails to produce reliable results (*e.g.*, artifacts in hairs, over-smoothed faces) when depth maps contain larger noise. In contrast, our method achieves state-of-the-art performance by learning the geometry-aware two-scale PIFu representation.

Methods	Mesh		Normal	
	P2S$\times 10^{-2}\downarrow$	Chamfer$\times 10^{-2}\downarrow$	$L2\times 10^{-1}\downarrow$	Cosine$\times 10^{-3}\downarrow$
RGBD-PIFu [70]	0.3335	0.3188	0.207	0.824
PIFuHD [71]	1.7268	1.7423	0.512	1.576
StereoPIFu [37]	0.5832	0.5425	0.328	1.193
IPNet [3]	0.8563	0.7247	0.196	0.751
Ours	**0.2652**	**0.2775**	**0.176**	**0.685**

Table 1: Quantitative comparisons to state-of-the-art methods on our test dataset. The best results are marked in **bold**.

Qualitative Comparisons. Fig. 8 shows the reconstruction results of five existing methods and our method on our captured real data. Although refined depths $\mathbf{D}^i_{rf}$ can be fused (*e.g.*, TSDF-Fusion in [63]) to obtain 3D human models (Fig. 8(b)), the reconstructed results contain large holes and low-quality regions due to sparse inputs and multi-view inconsistencies. The multi-view RGBD-PIFu and the PIFuHD methods often suffer from floating geometry (Fig. 8(c)) and topology errors (Fig. 8(d)) due to non-negligible depth noise and the lack of other view information. For IPNet, even if we take the fused point clouds (from 3 frames) as inputs, the topological errors are still obvious (Fig. 8(e)), and facial details are missing. We also test the DoubleFusion [96] on our captured data. As shown in Fig. 8(f), due to the low quality of the original depths and the sparse embed-graph of the whole body, this volumetric fusion-based approach tends to smooth out regions where we expect to see the high-frequency information (*e.g.*, faces). In addition, when a pose differs significantly from the initial

pose, the reconstructed mesh tends to be over-stretched (blue boxes). The Function4D [95] tracks the former and latter frames for the current frame and fuses the multi-frame point clouds to produce new depth maps. When the motion changes are small, their method is not easy to reconstruct reasonable high-frequency details (flat faces, hairs in Fig. 8(g)). Our approach leverages the depth denoising to produce robust depth (*e.g.*, hairs in Fig. 8(h)) for the PIFu-based reconstruction process. Moreover, our two-scale PIFu represents the face and body separately to produce vivid details.

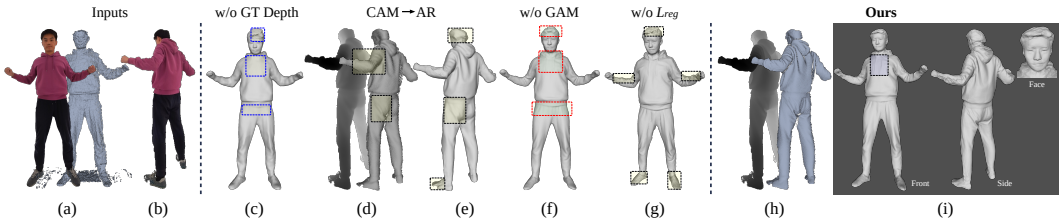


Figure 9: Ablation study. RGBD images of two views (a,b). Reconstruction results and refined depth of ablated versions (c-g). Our results (h,i). Zoom in to see details.

4.2 Ablation Study

Depth & Occupancy Supervision. Tab. 2(1st row) shows that by removing the depth denoising task, the performance drops by 14.84% and 18.98% in terms of Mesh and Normal metrics. The decoder $\mathcal{D}(\cdot)$ produces over-smoothed depths and the reconstructed results also lack details, as shown in Fig. 3(e,f) and Fig. 9(c). On the other hand, by removing the 3D supervision, the denoised depth may contain incorrect details; See Fig. 3(g). This experiment verifies the effectiveness of our multi-task formulation.

Model	Mesh		Normal	
	P2S$\times 10^{-2}\downarrow$	Chamfer$\times 10^{-2}\downarrow$	$L2\times 10^{-1}\downarrow$	Cosine$\times 10^{-3}\downarrow$
w/o GT Depth	0.3143	0.3085	0.213	0.801
CAM → AR	0.2826	0.2939	0.194	0.759
w/o GAM	0.2695	0.2810	0.178	0.694
w/o GAM & CAM → AR	0.2901	0.2877	0.202	0.773
w/o $\mathcal{L}_{reg}$	0.2760	**0.2691**	0.185	0.719
Ours	**0.2652**	0.2775	**0.176**	**0.685**

Table 2: Ablation study. AR indicates the combination of ASPP [10] and ResCBAM [86]. CAM → AR indicates replacing CAM with AR.

CAM & GAM. The 2nd-to-4th rows of Tab. 2 evaluate our CAM and GAM modules for RGBD fusion and geometry-aware enrichment, respectively. By replacing the CAM with the common combination of ASPP [10] and ResCBAM [86] (*i.e.*, AR in Fig. 2) or removing the GAM, both the performances drop. We also observe that using AR without GAM yields worse performance on average than removing the GAM. Fig. 9(d,e,f) illustrates the differences, where we can see that these two modules help reconstruct the high-frequency details (*e.g.*, clothes folds of Fig. 9(h,j)).

PIFu-Face: $\mathcal{F}_f$. The visual comparisons in Fig. 6(c,f) show that the PIFu-Face, *i.e.*, $\mathcal{F}_f$ significantly improves the ability of our method to reconstruct high-frequency facial details.

Face-to-body Fusion: $\mathcal{W}$. The visual comparison in Fig. 6(e) shows that our weighted face-to-body fusion can eliminate the artifacts generated by simply merging the reconstructed face and body.

Regularization loss: L_{reg}. Fig. 9(g) shows that without the proposed L_{reg}, our method tends to produce jagged noise on stitched boundaries where depth may not be consistent.

5 Conclusion

In this paper, we propose the geometry-aware two-scale PIFu representation to reconstruct the digital human body with fine-grained facial details. The first novelty is that we formulate the depth denoising and implicit occupancy estimation in a multi-task learning manner, to exploit their complementary properties. The second novelty lies in that we formulate the two-scale PIFu via two MLPs to represent the face and body separately, to reduce the complexity of modeling high-frequency facial details. Finally, a lightweight face-to-body fusion scheme fuses the face and body occupancy fields to generate reliable reconstruction results of high fidelity. However, it remains challenging for our method to generate high-quality hand details. Compared to face regions, hands have smaller sizes but larger variations of geometric shapes and spatial positions, making them extremely challenging to be reconstructed from the sparse noisy inputs. We are interested in exploring this topic in the future.

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019.

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 2022.

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *ECCV*, 2020.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *conference on Computer graphics and interactive techniques*, 1999.

[5] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE TPAMI*, 2003.

[6] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *ICCV*, 2015.

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE TVCG*, 2013.

[8] Derek Chan, Hylke Buisman, Christian Theobalt, and Sebastian Thrun. A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[9] Jiawen Chen, Dennis Bautembach, and Shahram Izadi. Scalable real-time volumetric surface reconstruction. *ACM ToG*, 2013.

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *CVPR*, 2020.

[12] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM TOG*, 2015.

[13] Yan Cui, Sebastian Schuon, Sebastian Thrun, Didier Stricker, and Christian Theobalt. Algorithms for 3d shape scanning with a depth camera. *IEEE TPAMI*, 2012.

[14] James Diebel and Sebastian Thrun. An application of markov random fields to range sensing. 2005.

[15] Jennifer Dolson, Jongmin Baek, Christian Plagemann, and Sebastian Thrun. Upsampling range data in dynamic environments. In *CVPR*, 2010.

[16] Mingsong Dou, Philip L. Davidson, S. Fanello, S. Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: real-time volumetric performance capture. *ACM ToG*, 2017.

[17] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM ToG*, 2016.

[18] Péter Fankhauser, Michael Bloesch, Diego Rodriguez, Ralf Kaestner, Marco Hutter, and Roland Siegwart. Kinect v2 for mobile robot navigation: Evaluation and modeling. In *International Conference on Advanced Robotics*, 2015.

[19] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[20] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *ICCV*, 2013.

[21] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *ICCV*, 2019.

[22] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009.

[23] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM TOG*, 2016.

[24] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE TPAMI*, 2021.

[25] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *CVPR*, 2017.

[26] Sigurjon Arni Gudmundsson, Henrik Aanaes, and Rasmus Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3d estimation. *International Journal of Intelligent Systems Technologies and Applications*, 2008.

[27] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM TOG*, 2019.

[28] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *ICCV*, 2015.

[29] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM ToG*, 2017.

[30] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM TOG*, 2019.

[31] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020.

[32] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust image filtering using joint static and dynamic guidance. In *CVPR*, 2015.

[33] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE TPAMI*, 2012.

[34] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *NeurIPS*, 2020.

[35] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021.

[36] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 2012.

[37] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *CVPR*, 2021.

[38] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020.

[39] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[40] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic facial geometry inference using deep neural networks. In *CVPR*, 2018.

[41] Junho Jeon and Seungyong Lee. Reconstruction-based pairwise depth dataset for depth image enhancement using cnn. In *ECCV*, 2018.

[42] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018.

[43] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018.

[44] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM ToG*, 2007.

[45] HyeokHyen Kwon, Yu-Wing Tai, and Stephen Lin. Data-driven depth map refinement via multi-scale sparse representation. In *CVPR*, 2015.

[46] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-view dynamic shape refinement using local temporal integration. In *ICCV*, 2017.

[47] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM ToG*, 2009.

[48] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *ECCV*, 2020.

[49] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *ECCV*, 2016.

[50] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE TPAMI*, 2019.

[51] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *CVPR*, 2020.

[52] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *CVPR*, 2021.

[53] Marvin Lindner, Andreas Kolb, and Klaus Hartmann. Data-fusion of pmd-based distance-information and high-resolution rgb-images. In *International Symposium on Signals, Circuits and Systems*, 2007.

[54] Ming-Yu Liu, Oncel Tuzel, and Yuichi Taguchi. Joint geodesic upsampling of depth images. In *CVPR*, 2013.

[55] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE TVCG*, 2009.

[56] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011.

[57] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015.

[58] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *CVPR*, 2014.

[59] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *ICCV*, 2021.

[60] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *ICCV*, 2015.

[61] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *CVPR*, 2019.

[62] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *CVPR*, 2015.

[63] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR*, 2011.

[64] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *ECCV*, 2020.

[65] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. High quality depth map upsampling for 3d-tof cameras. In *ICCV*, 2011.

[66] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[67] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, 2017.

[68] Christian Richardt, Carsten Stoll, Neil A Dodgson, Hans-Peter Seidel, and Christian Theobalt. Coherent spatiotemporal filtering, upsampling and rendering of rgbz videos. In *Computer Graphics Forum*, 2012.

[69] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *CVPR*, 2016.

[70] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019.

[71] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020.

[72] Sebastian Schuon, Christian Theobalt, James Davis, and Sebastian Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. In *CVPR*, 2009.

[73] Xiaoyong Shen, Chao Zhou, Li Xu, and Jiaya Jia. Mutual-structure for joint filtering. In *ICCV*, 2015.

[74] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *ICCV*, 2019.

[75] Vladimiros Sterzentsenko, Leonidas Saroglou, Anargyros Chatzitofis, Spyridon Thermos, Nikolaos Zioulis, Alexandros Doumanoglou, Dimitrios Zarpalas, and Petros Daras. Self-supervised deep depth denoising. In *ICCV*, 2019.

[76] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, 2020.

[77] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021.

[78] Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. Scanning 3d full human bodies using kinects. *IEEE TVCG*, 2012.

[79] Anh Tun Trn, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, 2018.

[80] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 2018.

[81] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. In *ACM SIGGRAPH 2006 Courses*. 2006.

[82] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia*. 2009.

[83] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 2021.

[84] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *ECCV*, 2020.

[85] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[86] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[87] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM ToG*, 2018.

[88] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *ECCV*, 2018.

[89] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, 2020.

[90] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. Color-guided depth recovery from rgb-d data using an adaptive autoregressive model. *IEEE TIP*, 2014.

[91] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér. Spatial-depth super resolution for range images. In *CVPR*, 2007.

[92] Yu Yanga, Xiao-Jun Wu, and Josef Kittler. Landmark weighting for 3dmm shape fitting. *arXiv preprint arXiv:1808.05399*, 2018.

[93] Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. Performance capture of interacting characters with handheld kinects. In *ECCV*, 2012.

[94] Mao Ye and Ruigang Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *CVPR*, 2014.

[95] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021.

[96] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, 2018.

[97] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *CVPR*, 2019.

[98] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *ICCV*, 2019.

[99] Yebin Liu Zerong Zheng, Tao Yu and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, 2021.

[100] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.

[101] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, 2019.

[102] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019.

[103] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM TOG*, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, we discuss this in Sec. 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Yes, we discuss this in the supplemental material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We explain the implementation details in the supplemental material.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All four metrics are computed as the average of 5 runs with different degrees of input depth noise, as stated in the supplemental material.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This information is provided in the supplemental material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The THuman2.0 and FaceScape datasets that we used are encrypted to prevent unauthorized access. We have submitted a request form and a request mail to obtain the datasets.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Our captured data contains personal information but with permission, as stated in the supplemental material.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]