ANIM: Accurate Neural Implicit Model
for Human Reconstruction from a single RGB-D image

Marco Pesavento[1,3*]   Yuanlu Xu[3]   Nikolaos Sarafianos[3]   Robert Maier[3]   Ziyan Wang[3]

Chun-Han Yao[2]   Marco Volino[1]   Edmond Boyer[3]   Adrian Hilton[1]   Tony Tung[3]

[1]University of Surrey, CVSSP, UK   [2]UC Merced   [3]Meta Reality Labs

Abstract

Recent progress in human shape learning, shows that neural implicit models are effective in generating 3D human surfaces from limited number of views, and even from a single RGB image. However, existing monocular approaches still struggle to recover fine geometric details such as face, hands or cloth wrinkles. They are also easily prone to depth ambiguities that result in distorted geometries along the camera optical axis. In this paper, we explore the benefits of incorporating depth observations in the reconstruction process by introducing ANIM, a novel method that reconstructs arbitrary 3D human shapes from single-view RGB-D images with an unprecedented level of accuracy. Our model learns geometric details from both multi-resolution pixel-aligned and voxel-aligned features to leverage depth information and enable spatial relationships, mitigating depth ambiguities. We further enhance the quality of the reconstructed shape by introducing a depth-supervision strategy, which improves the accuracy of the signed distance field estimation of points that lie on the reconstructed surface. Experiments demonstrate that ANIM outperforms state-of-the-art works that use RGB, surface normals, point cloud or RGB-D data as input. In addition, we introduce ANIM-Real, a new multi-modal dataset comprising high-quality scans paired with consumer-grade RGB-D camera, and our protocol to fine-tune ANIM, enabling high-quality reconstruction from real-world human capture. https://marcopesavento.github.io/ANIM/

1. Introduction

The increasing interest in 3D virtual world creation has led to a substantial demand for easily accessible 3D reconstruction solutions. Consequently, this has emerged as a prominent research domain in computer vision with applications in virtual and augmented reality, gaming, medicine and e-shopping, among others. A recurrent challenge revolves around ensuring the fidelity of the created models

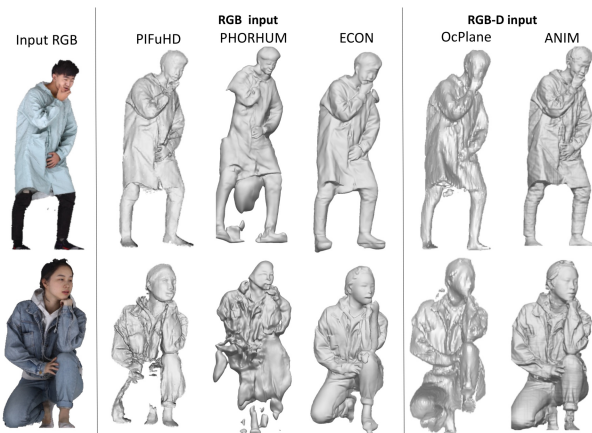*Work performed during an internship at Meta Reality Labs, Sausalito

Figure 1. ANIM enables human shape reconstruction with higher accuracy and without shape distortions compared to the state-of-the-art methods based on monocular RGB-D or RGB input.

and, consequently, the accuracy of the reconstruction methods, especially when reconstructing 3D avatars of real people. To this aim, depth sensors that are nowadays ubiquitous in commercial devices (*e.g.*, LiDAR Depth Camera, AI Stereo Depth Map, Azure Kinect, and Asus) can be leveraged to develop efficient and accurate reconstruction solutions. Our objective is to build high-fidelity models of clothed humans from single RGB-D images by learning an Accurate Neural Implicit Model (ANIM). Monocular approaches based on generative adversarial networks [33, 54] produce realistic front and back depths of the model. However, their fidelity is limited as the prediction depends on the generative ability of the network. While leveraging priors such as parametric body models can produce complete body shapes [37], they often lack details. Several works resort to multiple RGB-D images or monocular videos, combining multiple predictions to reconstruct higher-quality 3D shapes [15, 16, 50]. In contrast to previous works that exclusively process either single RGB images [45–47, 57, 64] or 3D point clouds [8, 12, 52], our proposed approach, which relies on a neural implicit model (a learned Signed Distance Field, or SDF), reconstructs accurate 3D models of clothed humans from a single RGB-D image with significantly higher levels of detail. Related approaches that also

estimate an implicit representation of the 3D shape from a single RGB-D image either lack pixel-alignment with the input, as in OPlanes [62], or estimate depth from the RGB input, limiting therefore the fidelity, and rely on 3D parametric models, as DiFU [49]. Our representation is a pixel-voxel-aligned implicit model of the reconstructed surface learned using a combination of multi-resolution 2D feature extractor and a specific SparseConvNet U-Net (the Volume Feature Extractor or VFE) to process multi-resolution 2D and volumetric features. A depth-supervision strategy is also introduced to further enhance the SDF estimation. We demonstrate the advantages of using RGB-D images over alternative methods that suffer from depth ambiguity or that reproduce low-fidelity details, as illustrated in Figure 1. Our extensive experiments show that ANIM outperforms existing methods that reconstruct 3D human shapes from single RGB images, surface normals, point cloud or RGB-D data. In practice, consumer-grade RGB-D cameras produce noisier data compared to high-end 3D scanners. This directly impacts the 3D reconstruction quality since the surface estimation builds on features learned from the input. In order to reduce this impact and to achieve high-fidelity reconstruction with consumer-grade camera, we propose to learn a model trained with high-quality 3D ground-truth data paired with real noisy RGB-D data as input. Since datasets with these characteristics are currently unavailable, we introduce the multi-modal dataset ANIM-Real which includes 3D scans reconstructed from a high-resolution multi-view camera system aligned with RGB-D data captured by a consumer-grade camera. Fine-tuning ANIM with ANIM-Real enables to better handle sensor noise and to obtain high-quality 3D shape models from real-world capture.

In summary, our contributions are:

- A novel network architecture ANIM that includes a pixel-voxel-aligned implicit representation obtained from the 3D Volume Feature Extractor and 2D multi-resolution features to reconstruct accurate and high-fidelity 3D human shape from a single-view RGB-D image.

- A novel depth-supervision strategy that refines the SDF learning of the 3D points lying on the reconstructed surface by leveraging the input point cloud.

- The multi-modal dataset ANIM-Real comprising synchronous captures from a high-quality 3D human scanner aligned with a consumer-grade RGB-D camera, and a protocol to fine-tune ANIM real-world human capture.

- Unprecedented quantitative and qualitative results for human shape reconstruction from single RGB-D images.

2. Related Work

Reconstruction from single-view images. Single-view 3D human reconstruction has been approached using a wide range of methods and representations. Representations used in this domain include voxels [28, 53, 63], two-way depth maps [17, 48], visual hull [42], parametric models [2–4, 10, 29, 30, 35, 43, 59]. These methods cannot reproduce high-quality 3D human shapes, with only minimally clothed reconstruction. In contrast, implicit function representations have shown great promise for the task of human digitization from a single image [22, 23, 25, 32, 34, 45–47]. One of the first approaches to adopt this representation was PIFu [46], which exploits pixel-aligned image features rather than global features to preserve local details of the input image as the occupancy of any 3D point is predicted. SuRS [45] demonstrates fine-scale detail can be recovered even from low-resolution input images using a super-resolution learning framework. PaMIR [64] concatenates 3D features extracted from an estimated SMPL model with 2D features. In PIFuHD [47], the quality of the reconstruction is improved by using surface normals and a coarse-to-fine implicit function framework. Alldieck *et al*. [5] improved upon PIFuHD by estimating the 3D geometry, the surface albedo, and shading, from a single image in a joint manner. ICON [57] improves the estimation of front and back normals used for the reconstruction by guiding it with a parametric model whilst ECON [58] addresses the problem of clothing reconstruction of the SMPL body by feeding the estimated normals into a d-BiNI optimizer.

Reconstruction from RGB-D and point clouds. Works that use RGB-D images as input for the task of clothed human reconstruction fall into two categories taking a single RGB-D image [54, 62] or multiple sequences of RGB-D images [11, 16, 33, 40] as input. Methods that consider RGB-D sequences have to fuse multiple partial and noisy observations into a coherent model. In contrast, single view RGB-D methods have to tackle the problem of shape completion [31, 61] due to partial observations leading to incomplete reconstructions. Approaches that estimate an implicit representation from a single RGB-D image cannot achieve the same level of quality and accuracy as the proposed method [49, 62]. DiFU [49] estimates the implicit representation by using a SMPL [35] voxel encoder, an U-Net depth estimator, and a scale regressor. The back and front depth maps are estimated from the RGB image, introducing noise and limiting the quality of the reconstructed shapes. OcPlanes [62] adopts a plane-aligned occupancy function to align the feature extracted from the input image to the input depth. Replacing the local pixel-alignment with a global alignment reduces the quality of the reconstruction. Point clouds are an alternative representation explored for the task of 3D human reconstruction[9, 12, 38, 39]. IF-Net [12] exploits partial point clouds and learns implicit functions using latent voxel features. IP-Net [9] further develops this idea by incorporating SMPL[35] into the pipeline to enable animatable reconstructions. In comparison to prior work, ANIM reconstructs 3D human shapes from single-view RGB-D images at an unprecedented level
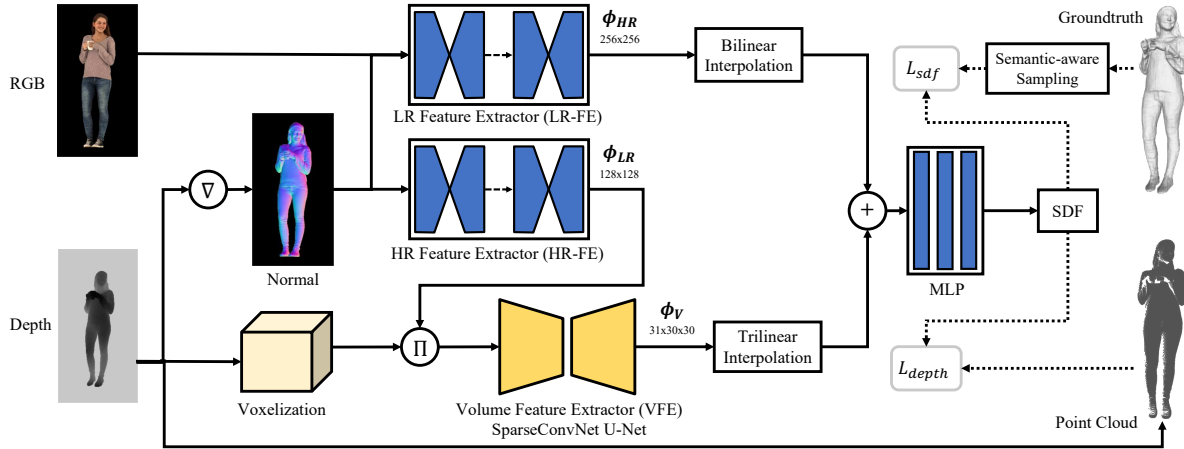
Figure 2. ANIM architecture. Our proposed framework has three major components: i) a multi-resolution appearance feature extractor for color and normal inputs (LR-FE and HR-FE), ii) a novel SparseConvNet U-Net (Volume Feature Extractor or VFE) that efficiently extracts geometry features from 3D voxels and low-resolution image features, iii) an MLP that estimate the implicit surface representation of full-body humans. + denotes concatenation, Π means fetching pixel-aligned 2D LR features and concatenating with 3D voxels, and ∇ indicates gradient operation applied to retrieve normals from depth map (using neighboring pixel cross-product).

of detail and reconstruction accuracy via the proposed novel architecture and depth-supervision strategy.

Single-view RGB-D datasets for human reconstruction. Available datasets containing RGB-D data from consumer-grade cameras are primarily designed for different tasks such as people re-identification [6, 41] or human activity recognition [13, 14, 18, 55, 56]. 3D ground-truth shapes are not provided in these datasets since the body skeleton is sufficient to achieve the intended tasks. Human3.6m [27] provides 3D human shape as ground truth along with depth maps. However, the quality of the shapes is limited, lacking clothing and details of the human body. Training neural implicit models with this dataset restricts the ability to learn high-fidelity clothed 3D human shapes. Recently, SynWild [20] used RGB and IR cameras to create the 3D ground truth but the semi-synthetic dataset is created by rendering the monocular video with a virtual camera, which is not affected by real-world noise. We propose the novel multi-modal dataset ANIM-Real that includes high-quality 3D human shapes reconstructed from a multi-view camera system, aligned and synchronized with real-world RGB-D data acquired with a consumer-grade camera.

3. Methodology

ANIM learns an implicit function f to reconstruct accurate and high-fidelity human shapes from a single RGB-D image. We present an end-to-end framework that takes an RGB-D image as input and estimates the SDF of the person. Specifically, as illustrated in Fig. 2, ANIM extracts a high-resolution (HR) 2D feature to encode high-frequency details and a low-resolution (LR) feature to maintain holistic reasoning from a concatenation of the input colour and normal, considering their shared image-space properties. The low-resolution features serve as a prior for a novel SparseConvNet [19] U-Net, which extracts geometric features

by processing 3D voxels created from the depth map and concatenated with its low-resolution image-space features. Given appearance and geometry features, an MLP predicts the SDF of the reconstructed subject. We train the framework end-to-end with a novel depth-supervision strategy that refines the estimation of the SDF of the 3D points close to the reconstructed surface by leveraging the input point cloud. Compared with related methods, our approach fuses information across multiple modes and is thus more robust to depth ambiguity and challenging poses.

3.1. Accurate Implicit Surface Estimation

Assuming the 3D clothed human to be reconstructed as a one-layer watertight mesh, we represent it with an implicit surface function f. The value $f(x)$ of a point $x \in \mathbb{R}^3$ denotes the distance of this point to its closest surface. To obtain a surface, we can simply threshold f to obtain an isosurface $f(x) = \tau$. The surface to be reconstructed is then defined as the zero level-set of f:

$$f' = \{x : f(x) = 0, \ x \in \mathbb{R}^3\}. \tag{1}$$

Fine surface details are stored in high frequency and need to be represented on the final shape, which has to be robust to depth ambiguity. Recent approaches show the effect of representing the shape with an implicit function aligned with the input data. PIFu [46] introduced the concept of pixel-alignment to increase the quality of 3D human shapes by projecting a 3D point $x \in \mathbb{R}^3$ in the image feature $\phi(I)$ of an input RGB image I. PaMIR [64] proposes a voxel-aligned implicit function to leverage spatial information from a parametric model to avoid depth ambiguity. We propose a novel architecture that learns a high-fidelity implicit surface representation s_{HF} that is both pixel-aligned with the input image I and surface normal S_N and voxel-aligned with the voxel created from the input depth map D:

$$\hat{s}_{HF} = f_x\Big(\phi_{HR}(\pi(I, S_N)), z(x), \gamma(D)\Big), \hat{s}_{HF} \in \mathbb{R}, \quad (2)$$

where $\phi_{HR}(I, S_N)$ are the HR features extracted from the concatenation between the input image I and the surface normal S_N, π is the orthographic projection, $\gamma(D) = \phi_V(\phi_{LR}(I, S_N), D)$ is the feature extracted from the depth D linked with LR feature ϕ_{LR} retrieved from I, and S_N. $z(x)$ is the depth value of x. To estimate the implicit representation $\hat{s}_{HF}$, ANIM comprises the following modules.

2D Feature Extractor. 2D pixel-aligned features are obtained from the input image I and surface normals S_N. LR features $\phi_{LR}(I, S_N)$ are extracted with a stacked hourglass network (LR-FE) to guarantee a large receptive field, which is required to maintain holistic reasoning [46]. LR features are also used by the 3D feature extractor since the SparseConvNet requires that each input voxel is linked to an embedding. Instead of using random embedding like previous works [44], performance improves when features extracted from the input image are used (see Sec. 5.2). High-quality details cannot be reconstructed if only LR features are used. To embed local details of I and S_N, we introduce a second stacked hourglass architecture (HR-FE) to retrieve HR features $\phi_{HR}(I, S_N)$. These HR features are pixel-aligned with the input data via orthographic projection π.

3D Feature Extractor. Learning spatial relationships in 3D space is fundamental to solving the problems of depth ambiguity derived by the single-view input. We process a voxel retrieved from the input depth D with a novel SparseConvNet U-Net style architecture: Voxel Feature Extractor (VFE). Due to the requirements of the VFE, the LR feature $\phi_{LR}(I, S_N)$ are linked to the voxels to provide additional information from the 2D RGB input before extracting the 3D features. Instead of using a 3D convolutional neural network as in [22, 64], we extract voxel-aligned features $\gamma(D)$ with 3D sparse convolution layers, which have been proven to be efficient when the input is sparse such as in the point cloud created from a single view. This also ensures a performance gain at training and testing times with faster speed in the order of magnitude compared to the 3D ConvNets. Voxel-alignment is obtained by trilinear interpolation of 3D points x with ϕ_V.

Multi-layer Perceptron (MLP). The 2D pixel-aligned features are concatenated with the 3D voxel-aligned feature and processed by a multi-layer perceptron that models the implicit function f_x and estimates the final SDF $\hat{s}_{HF}$.

3.2. Depth-Supervision Strategy

To improve the learning of the SDF of the 3D reconstructed surface, we propose to leverage the depth channel of the RGB-D input and estimate an implicit representation of the input sparse point cloud ζ by extracting pixel-aligned feature $\phi_{HR}(I, S_N)$ from HR-FE and voxel-aligned features $\gamma(\zeta) = \phi_V(\phi_{LR}(I, S_N), \zeta)$ from the VFE. An implicit function f_ζ representing ζ is learned with an MLP that

shares the weight with the one previously applied:

$$\hat{s}_\zeta = f_\zeta\Big(\phi_{HR}(\pi(I, S_N)), z(x_\zeta), \gamma(\zeta)\Big), \quad \hat{s}_\zeta \in \mathbb{R}, \quad (3)$$

where x_ζ are the points of the input point cloud, which are projected into $\phi_{HR}(I, S_N)$ for pixel alignment. The SDF of x_ζ should be 0 since x_ζ lies on the surface of the reconstructed shape. The network significantly improves its ability to estimate which points lie on the surface (Sec. 5.2).

The network is trained end-to-end with two Huber losses, one to train the implicit function f_x of 3D points x sampled on the 3D ground-truth shape:

$$L_{sdf} = \begin{cases} 0.5\,(\hat{s}_{HF} - s_{HF})^2, & \text{if } \|\hat{s}_{HF} - s_{HF}\|_2 < \delta, \\ \delta\,(|\hat{s}_{HF} - s_{HF}| - 0.5\delta), & \text{otherwise}, \end{cases} \quad (4)$$

and the other for the depth-supervision strategy:

$$L_{depth} = \begin{cases} 0.5\,(\hat{s}_\zeta - s_\zeta)^2, & \text{if } \|\hat{s}_\zeta - s_\zeta\|_2 < \delta, \\ \delta\,(|\hat{s}_\zeta - s_\zeta| - 0.5\delta), & \text{otherwise}, \end{cases} \quad (5)$$

where s_ζ is the ground-truth label for the N_ζ points of the depth map ζ and δ is a threshold for estimation correctness. The learning is thus supervised with both points sampled in the 3D space of the full body to learn its full representation, and with points from the depth to improve the SDF estimation of points that lie on the visible surface, resulting in a more accurate representation of high-quality details.

Inference. Instead of using M^3 random 3D points distributed in a 3D grid in space as related works, we consider the input point cloud to create a 3D grid to sample the SDF. A bounding box is created around the sparse point cloud augmented with Gaussian sampling. The resolution of the novel 3D grid is computed as $(m \times H, m \times W, m \times D)$ where $m = \sqrt[3]{M^3/L}$, $M = 256$, $L = H \times W \times D$ and (H, W, D) is the dimension of the bounding box. The points are not randomly distributed in a squared grid but are concentrated in the region where the shape is generated. The 2D and 3D features are extracted from the RGB-D input and aligned with the grid of points. The final SDF is estimated with the MLP and the shape is obtained by extracting iso-surface $f_x = 0$ of the probability field $\hat{s}_{HF}$ via Marching Cubes [36].

Extension to consumer-grade RGB-D camera. RGB-D data acquired in real-world scenarios are affected by the noise propagated in the capture systems. Applying ANIM directly on data acquired with a consumer-grade camera (*e.g.* Azure Kinect) reveals severe reconstruction artifacts due to the sensor noise (Fig. 3a). ANIM should learn the noise added to captured data to perform effectively in real-world scenarios. A solution is to fine-tune ANIM with real noisy data. However, datasets that combine RGB-D data with high-resolution 3D ground-truth shapes, necessary to achieve high-quality reconstruction, are unavailable. We thus create the new dataset ANIM-Real with a multi-modal setup consisting of a consumer-grade RGB-D camera and a
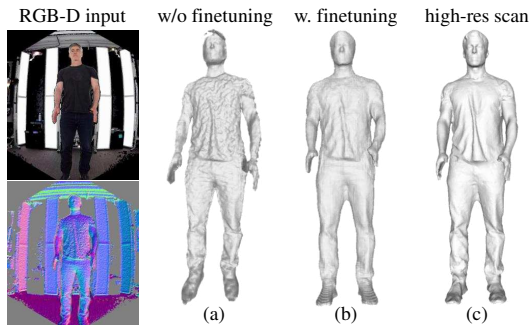
Figure 3. ANIM reconstructions from real-world capture with a consumer-grade RGB-D camera (Azure Kinect) before (a), and after (b) fine-tuning on the proposed dataset ANIM-Real, which quality is closer to a high-res scan capture (c).

high-resolution 4D scanner, as explained in Sec. 4. We then fine-tune ANIM parameters on ANIM-Real to learn typical consumer-grade sensor noise, resulting in a significant improvement in the 3D shape (see Fig. 3b).

4. Datasets

ANIM-Real dataset. We present ANIM-Real, a new multi-modal dataset that can be used to obtain high-quality reconstructions from consumer-grade RGB-D camera for real-world applications. Consumer-grade monocular RGB-D cameras (*e.g.* Azure Kinect) while ubiquitous produce less accurate and incomplete reconstructions than high-end capture systems such as full-body 3D scanners that are based on multi-view capture [24]. Depth sensors from consumer-grade cameras contain noise that directly affects surface reconstruction accuracy [21, 51]. To date, it is not possible to generate high-fidelity full-body 3D reconstruction from a one-shot capture with a consumer-grade RGB-D sensor. Hence, we create ANIM-Real as follows:

- We acquire high-quality 3D scans with a high-resolution camera system that uses active stereo and multi-view cameras [24]. It comprises 16 high-resolution RGB cameras and stereo pairs. However, raw scan reconstructions can contain holes from self-occlusions, lack of coverage, or challenging regions (*e.g.* hair), leading to incorrect sampling and SDF estimation. We therefore apply the Fast Winding Numbers algorithm [7], hence producing high-quality watertight shapes (see Fig.3c).
- RGB-D data is captured with a consumer-grade camera (Azure Kinect), to allow ANIM to learn the sensor noise introduced in the input RGB-D images (see Fig.3).
- Intrinsics and extrinsics camera calibrations and capture synchronization are crucial to align the 3D shape with the corresponding RGB-D input. The extrinsics calibration between capture systems is obtained using a generic calibration object with salient shapes, while synchronization is obtained using Sync I/O and generated trigger signals. The transformation matrix obtained with the calibration is used to project the 3D points sampled in the 3D shape to the 2D image feature, achieving pixel-alignment. The
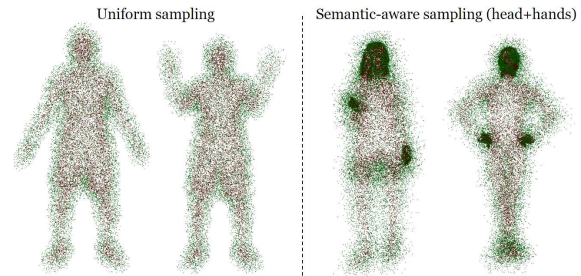

Figure 4. Semantic-aware sampling. Compared to uniform sampling (left), semantic-aware sampling (right) enables finer learning of human features on specific regions such as the head and hands.

input depth map is also aligned with the 3D ground-truth shape, ensuring voxel-alignment.

- For evaluation, we simultaneously capture 28 subjects in motion using the 2 systems. We fine-tune ANIM with an additional 16k frames, consisting of around 800 frames on average from a single view of 21 subjects.

Synthetic datasets. For additional quantitative and qualitative evaluations, we use large public synthetic datasets of 3D humans in various poses and clothing, providing a comprehensive assessment of the effectiveness of the proposed approach. We use 909 RenderPeople [1] scans and split them into 800 for training and 109 for testing. To evaluate the generalization power of ANIM, we use 200 human models THuman2.0 dataset [60] as another test set. We evaluate the reconstruction accuracy with 3 quantitative metrics: the average point-to-surface Euclidean distance (P2S), the normal reprojection [46], and the Chamfer distance (CD), cm.

5. Experiments

We quantitatively and qualitatively evaluate the proposed approach to the task of reconstructing 3D human shapes from a single RGB-D image. We conduct in-depth ablation studies where the proposed modules are removed. Finally, we show reconstruction examples from real-world RGB-D data where ANIM faithfully reconstructs the clothed geometry despite the presence of sensor noise.

5.1. Implementation details

We render the training RGB-D images and surface normals at 512x512 pixel resolution with a virtual camera that rotates around the 3D model with a step of 2°.

To create the ground-truth SDF s_{HF}, state-of-the-art works [46, 47, 64] label a set of 3D points that are sampled around the surface with a mixture of uniform sampling and randomly added offset following normal distribution $\mathcal{N}(0, \sigma_{LR})$. Since the sampling is homogeneous, smaller parts of the body have a lower number of 3D sampled points and cannot be reconstructed by these methods. We propose to augment the sampling in a semantic-aware manner.

Semantic-Aware Body Surface Sampling. We augment the number of points sampled in the face and hands regions by selecting the points that reproject in the same regions

Design	RenderPeople			THuman2.0		
	CD	Normal	P2S	CD	Normal	P2S
2D LR only	2.653	.5611	2.682	1.998	.3734	1.939
2D HR only	2.170	.5131	1.973	1.147	.3074	1.008
3D only	3.021	.6302	2.715	1.961	.5179	1.631
w. rand. feat.	2.136	.5081	1.947	1.052	.2823	0.954
ANIM (Ours)	**2.075**	**.4990**	**1.936**	**0.913**	**.2545**	**0.878**

Table 1. Quantitative results obtained by modifying the architecture of the network.

of a semantic mask, estimated from the RGB image using body-part segmentation [26]. We project $X_b = 48,000$ 3D points sampled around the ground-truth surface on the 2D mask. If the point is projected onto the face or hand regions of the mask, we sample additional points around that point:

$$X_t = \begin{cases} X_b + Add(X_{hh}), & \text{if } N_{X_{hh}} <= N_{X_b}/2, \\ X_b + Add(X_{hh})[0 : N_{X_b}/2], & \text{otherwise} \end{cases} \quad (6)$$

where $X_t = 36,000$ is the final number of sampled points, X_{hh} are the semantic points corresponding to hands and face and Add is defined as an N_K-steps recursion addition: $X_{hh} = X_{hh} + Add(X_{hh} + \mathcal{N}(0, \sigma_{HR}))$ with $\sigma_{HR} = 0.07$ to sample closer to the surface where fine details lie while $\sigma_{LR} = 5$. As shown in Fig. 4, the hands and face regions contain more points and fine details can be represented in those regions by the learned implicit function. As sampling is done as a pre-processing step during training, this strategy is more tractable in terms of training time compared to training separate networks for each body region, whose merging is non-trivial. For the depth-supervision strategy, we select $N_{pc} = 15000$ sub-points of the input point clouds. See supplementary for additional details about the implementation.

5.2. Ablation Studies

Network architecture. To prove the effectiveness of the proposed network architecture, we modify it as follows:
• **2D feature only**: Only RGB and normals are exploited by either the HR-FE (2D HR only) or the LR-FE (2D LR only) networks. The VFE is not implemented.
• **3D feature only**: The final estimation is obtained by processing only the depth map. HR-FE and LR-FE are not implemented and RGB and surface normals are not used.
• **w. random feature**: Similar to previous work [44], we link random features to the voxel as input to the SparseConvNet instead of using the LR feature. This shows whether linking RGB features retrieved from the input 2D image and surface normals to the voxels improves the performance.
Our proposed configuration obtains the lowest errors (see Tab. 1). The highest errors are obtained when either the 3D or the 2D encoders are not used, proving the effectiveness of using them together. Linking the LR feature to the voxel for the VFE improves ANIM performance.
High quality details. Our framework can reproduce significantly high-quality details on the final shape. We want to demonstrate the role of each component of the framework

Design	RenderPeople			THuman2.0		
	CD	Normal	P2S	CD	Normal	P2S
w/o normals	2.271	.5156	2.123	1.376	.3126	0.940
w/o LR feature	2.453	.5611	2.282	1.653	.2954	2.554
w/o HR feature	2.605	.5320	3.176	2.649	.2599	3.323
w/o SA sampl.	2.636	.5328	2.238	0.993	.2710	0.908
w/o L_{depth}	**2.060**	.5647	1.956	0.947	.2689	0.915
ANIM (Ours)	2.075	**.4990**	**1.936**	**0.913**	**.2545**	**0.878**

Table 2. Quantitative evaluation to demonstrate the influence of the adopted configuration to create fine details in the shapes.

in learning details by setting up the following baselines:
• **w./o. normals**: The normals are not concatenated with the input RGB image and not considered in the reconstruction.
• **w./o. LR feature**: To show the importance of having a large receptive field, we test the approach without the LR-FE. The HR features are linked to the voxel as input to VFE.
• **w./o. HR feature**: The HR-FE is not implemented to demonstrate the effect of exploiting local HR features in addition to just normals and LR features. The output of LR-FE replaces the HR embedding of the original approach.
• **w./o. SA sampling**: To show how the semantic-aware sampling approach is essential to retrieve more details on the face and hands, we trained the approach without augmenting the sampling points on the face and hand regions.
• **w./o. L_{depth}**: To demonstrate the effectiveness of the depth-supervision, we train ANIM with only the L_{sdf} loss. As shown in Tab. 2, we conclude that each component in our proposed framework is fundamental to improving the quality of reconstruction results. We observed performance drops if any component is omitted, with the highest accuracy obtained when ANIM leverages all its components. See supplementary for a qualitative evaluation.

5.3. Comparisons to the State of the Art

Our goal is to demonstrate the advantages of using RGB-D data over other inputs by comparing ANIM with methods that rely on different single-input data, such as only RGB image (SuRS [45],PHORHUM [5]), surface normals and parametric models (PaMIR [64], PIFuHD [47], ICON [57], ECON [58]) or point cloud (IF-Net [12]). Additionally, we want to highlight the superiority of ANIM against related works that infer the 3D shape from a single RGB-D input (NormalGAN [54], OcPlanes [62]). We further adapt PIFu [46] to RGBD-based reconstruction (PIFu+D) to establish fair comparisons, by concatenating depth with the RGB inputs. To demonstrate the effectiveness of the multi-resolution image extractor and of the VFE, we modify the architecture of PIFu and IF-Net by adding the VFE to PIFu and the HR-FE to IF-Net. PIFu+VFE processes the depth map with the VFE to extract geometric features, which are concatenated with the features of PIFu 2D encoder. IF-Net+HR processes RGB images with HR-FE and concatenates the feature with the output of the IF-Net 3D encoder. We repeat all aforementioned modifications by adding sur-

	Methods	RenderPeople			THuman2.0		
		CD	Normal	P2S	CD	Normal	P2S
Other input	IF-Net [12]	4.546	.7732	4.375	1.924	.4181	1.847
	PaMIR [64]	3.944	.6562	3.261	2.602	.3721	2.727
	PIFuHD [47]	2.415	.5495	2.381	3.625	.2730	3.462
	ICON [57]	2.330	.5886	2.301	2.093	.2791	1.112
	PHORHUM [5]	2.390	.5341	2.349	3.199	.2634	2.988
	ECON [58]	2.261	.5536	2.269	1.339	.2736	1.412
	SuRS [45]	2.810	.5909	2.884	1.290	.2945	1.695
Single RGB-D input	PIFu+D	4.650	.7567	4.314	4.441	.3704	3.835
	PIFu+D+S_N	3.544	.6754	3.653	4.379	.4302	3.983
	PIFu+VFE	2.718	.5747	2.089	4.444	.3240	3.834
	PIFu+VFE+S_N	2.242	.5218	2.076	0.924	.2548	0.880
	IF-Net+HR	2.352	.5304	1.962	1.403	.3754	1.322
	IF-Net+HR+S_N	2.164	.4995	1.953	1.079	.2875	0.993
	NormalGAN [54]	3.924	.7912	3.224	2.830	.5914	2.658
	OcPlane [62]	5.619	.5324	4.188	3.734	.3303	3.728
	ANIM (Ours)	**2.075**	**.4990**	**1.936**	**0.913**	**.2545**	**0.878**

Table 3. Quantitative comparisons with state-of-the-art approaches in 3D human reconstruction from a single input.

face normals as input (indicated by + S_N).

Quantitative comparisons. In Tab. 3 we report a plethora of quantitative comparisons of ANIM against other works on the RenderPeople [1] and THuman2.0 [60] datasets and showcase that our approach outperforms top performing competing methods by a large margin in both fidelity and accuracy. Extracting geometric information from the input depth map achieves better results compared to methods that estimate surface normals and parametric models. Moreover, the complete information extracted from the combination of RGB, normals and depth allows ANIM to outperform all the methods that rely on a single input. The novel architecture of ANIM guarantees the highest performance compared to methods that reconstruct 3D shapes from a single RGB-D.

Qualitative comparisons with works that reconstruct the 3D shape from input different than RGB-D for RenderPeople [1] are shown in Fig. 6, whilst Fig. 7 shows reconstruction from RGB-D for THuman2.0 [60]. The quality of the 3D shapes reconstructed by ANIM is significantly higher than all the other methods. ANIM outperforms methods that do not process RGB-D data because the integration of depth information, alongside RGB, is essential for achieving high-quality and accurate estimations. By learning spatial relationships from the geometric information extracted from the depth maps, ANIM effectively avoids depth ambiguity, resulting in more accurate reconstructions compared to methods that rely on estimating parametric models. The multi-resolution feature extractors employed by ANIM ensure the reproduction of finer details in contrast to other approaches. RGB-D methods are outperformed thanks to the depth supervision strategy and the combination of 2D multi-resolution and 3D geometric features, leveraging pixel-voxel-aligned properties inherent in the implicit representation. Implementing the introduced modules within benchmark works (PIFu, IF-Net) significantly improves their quantitative and qualitative performance, but
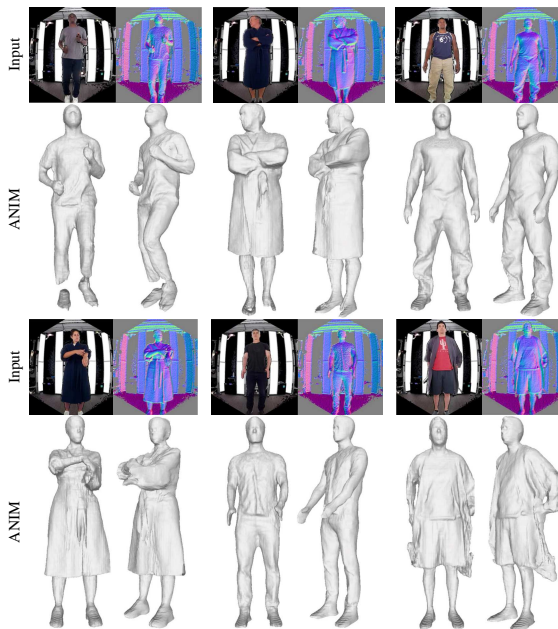

Figure 5. ANIM reconstructs fine-level cloth details such as wrinkles on the cloth and body with high accuracy even when the input is a consumer-grade RGB-D camera (Azure Kinect).

the highest accuracy is still obtained by ANIM.

Real-world capture. We provide qualitative results of 3D human shape reconstructed by ANIM after being fine-tuned with the ANIM-Real dataset (Fig. 5). The RGB-D input acquired with the Azure Kinect exhibits significant noise in the depth maps and surface normals. However, thanks to the fine-tuning, ANIM faithfully reconstructs the geometry of the clothed human, capturing fine-level details such as wrinkles and specific body features like the face and hands. See supplementary materials for additional results.

6. Conclusion

We introduce ANIM, a novel neural implicit model that reconstructs accurate and high-fidelity 3D humans from single RGB-D images, outperforming existing methods over other kinds of input and proving the benefit of leveraging RGB-D data. We demonstrate the effectiveness of combining both multi-resolution pixel-voxel-aligned features and a novel depth-supervision strategy to address depth ambiguity and reconstruct high-quality 3D human shapes. We also present the multi-modal dataset ANIM-Real consisting of high-quality 3D scans and real-world captures, obtained with a high-resolution camera system paired with a consumer-grade RGB-D camera. ANIM-Real significantly leverages ANIM for human reconstruction, and can be valuable to the community for neural implicit 3D human reconstruction. Future work includes exploring temporal information for fusion of body pose and appearance across time.

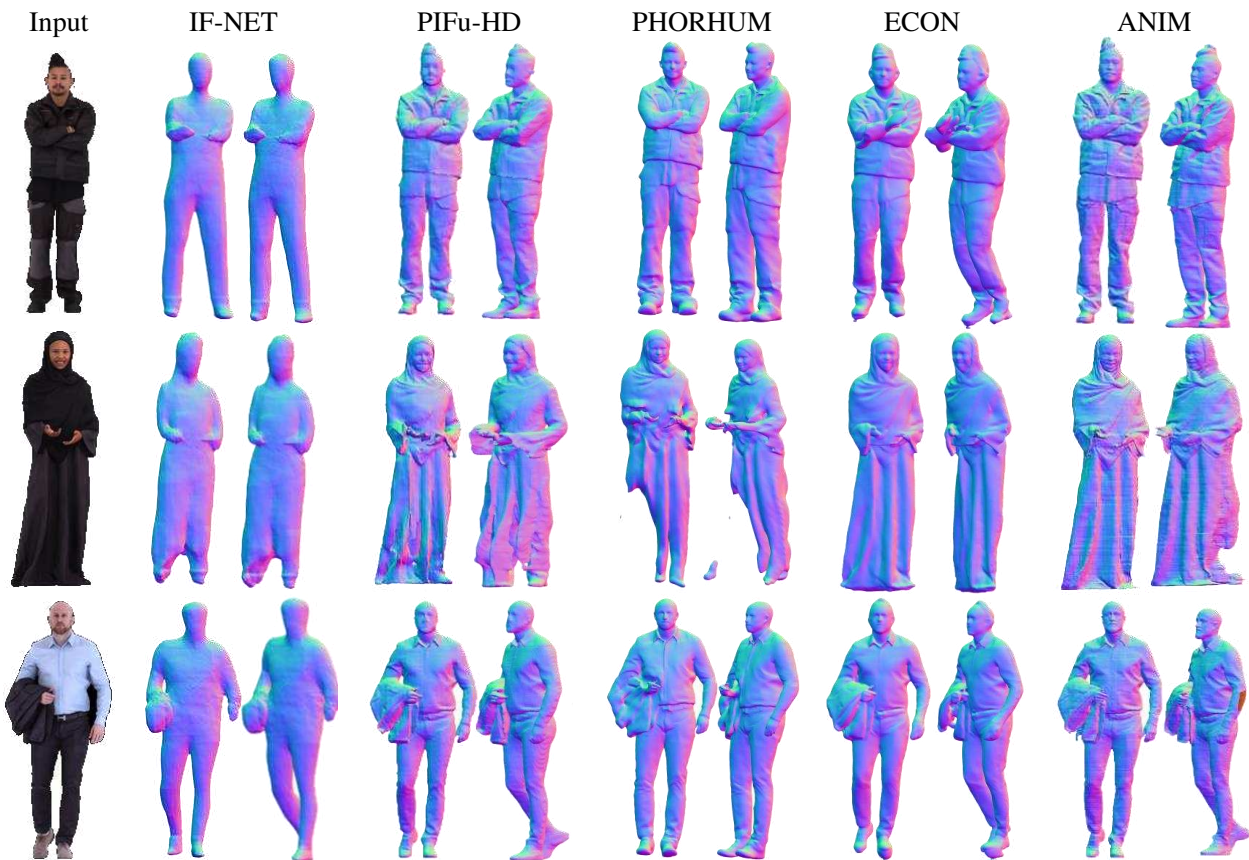| Input	IF-NET	PIFu-HD	PHORHUM	ECON	ANIM

Figure 6. Qualitative comparisons with state-of-the-art approaches on RenderPeople dataset given different kinds of input.

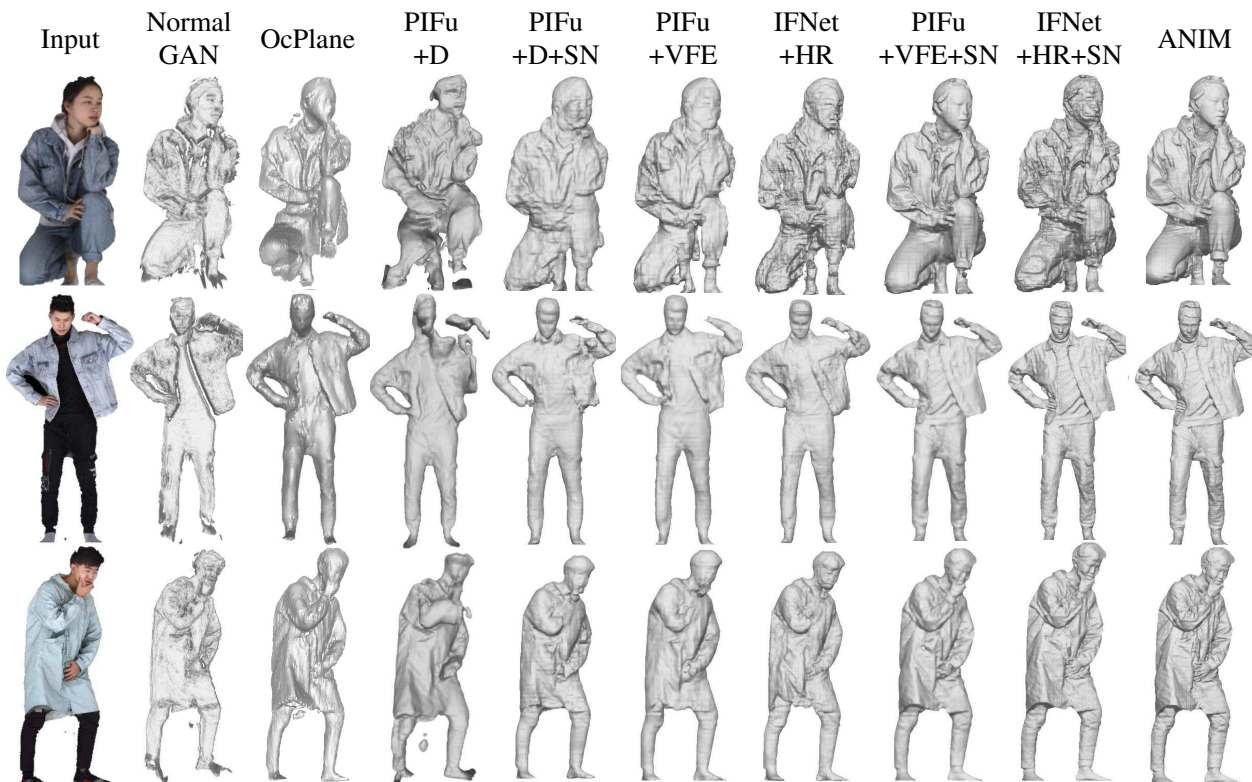| Input	Normal GAN	OcPlane	PIFu +D	PIFu +D+SN	PIFu +VFE	IFNet +HR	PIFu +VFE+SN	IFNet +HR+SN	ANIM

Figure 7. Qualitative comparisons with state-of-the-art approaches on THuman2.0 dataset given an RGB-D image as input.

References

[1] Renderpeople. https://renderpeople.com/. Accessed: 2020-07-26. 5, 7

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision*, 2019. 2

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 6, 7

[6] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *European Conference on Computer Vision Workshops*, 2012. 3

[7] Gavin Barill, Neil Dickson, Ryan Schmidt, David I.W. Levin, and Alec Jacobson. Fast winding numbers for soups and clouds. *ACM Transactions on Graphics*, 2018. 5

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, 2020. 1

[9] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, 2020. 2

[10] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, 2016. 2

[11] A. Burov, M. Niesner, and J. Thies. Dynamic surface function networks for clothed human bodies. In *IEEE International Conference on Computer Vision*, 2021. 2

[12] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 6, 7

[13] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 3

[14] C. Coppola, D. Faria, U. Nunes, and N. Bellotto. Social activity recognition based on probabilistic merging of skeleton features with proximity priors from rgb-d data. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016. 3

[15] Zheng Dong, Ke Xu, Ziheng Duan, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Geometry-aware two-scale pifu representation for human reconstruction. In *Advances in Neural Information Processing Systems*, 2021. 1

[16] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE International Conference on Computer Vision*, 2019. 2

[18] Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana. Human activity recognition process using 3-d posture data. *IEEE Transactions on Human-Machine Systems*, 45(5):586–597, 2014. 3

[19] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[20] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[21] Azmi Haider and Hagit Hel-Or. What can we learn from depth camera sensor noise? *Sensors*, 22(14):5448, 2022. 5

[22] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Annual Conference on Neural Information Processing Systems*, 2020. 2, 4

[23] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *IEEE International Conference on Computer Vision*, 2021. 2

[24] https://3dmd.com/. 3dmd 4d scanner. 5

[25] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable reconstruction of clothed humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[26] Anastasia Ianina, Nikolaos Sarafianos, Yuanlu Xu, Ignacio Rocco, and Tony Tung. Bodymap: Learning full-body dense correspondence map. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 6

[27] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 3

[28] Aaron S. Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. *European Conference of Computer Vision Workshops*, 2018. 2

[29] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[31] Dongping Li, Tianjia Shao, Hongzhi Wu, and Kun Zhou. Shape completion from a single rgbd image. *IEEE Transactions on Visualization and Computer Graphics*, 23(7):1809–1822, 2017. 2

[32] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision*, 2020. 2

[33] Xing Li, Yangyu Fan, Di Xu, Wenqing He, Guoyun Lv, and Shiya Liu. Sfnet: Clothed human 3d reconstruction via single side-to-front view rgb-d image. In *International Conference on Virtual Reality*, 2022. 1, 2

[34] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6): 248, 2015. 2

[36] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987. 4

[37] Yang Lu, Han Yu, Wei Ni, and Liang Song. 3d real-time human reconstruction with a single rgbd camera. *Applied Intelligence*, pages 1–11, 2022. 1

[38] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *IEEE International Conference on Computer Vision*, 2021. 2

[39] Qianli Ma, Jinlong Yang, Michael J. Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision*, 2022. 2

[40] Aihua Mao, Hong Zhang, Yuxin Liu, Yinglong Zheng, Guiqing Li, and Guoqiang Han. Easy and fast reconstruction of a 3d avatar with an rgb-d sensor. *Sensors*, 17(5), 2017. 2

[41] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. *Person Re-Identification*, pages 161–181, 2014. 3

[42] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[44] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 4, 6

[45] Marco Pesavento, Marco Volino, , and Adrian Hilton. Super-resolution 3d human shape from a single low-resolution image. In *European Conference on Computer Vision*, 2022. 1, 2, 6, 7

[46] 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 *IEEE International Conference on Computer Vision*, 2019. 2, 3, 4, 5, 6

[47] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 5, 6, 7

[48] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *IEEE International Conference on Computer Vision*, 2019. 2

[49] Dae-Young Song, HeeKyung Lee, Jeongil Seo, and Donghyeon Cho. Difu: Depth-guided implicit function for clothed human reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[50] Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang. Robustfusion: Robust volumetric performance reconstruction under human-object interactions from monocular rgbd stream. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[51] Chris Sweeney, Greg Izatt, and Russ Tedrake. A supervised approach to predicting noise in depth images. In *International Conference on Robotics and Automation*, 2019. 5

[52] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *IEEE International Conference on Computer Vision*, 2021. 1

[53] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision*, 2018. 2

[54] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *European Conference on Computer Vision*, 2020. 1, 2, 6, 7

[55] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014. 3

[56] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012. 3

[57] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. Icon: Implicit clothed humans obtained from normals. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 6, 7

[58] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6, 7

[59] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human pose, shape and texture from low-resolution images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4490–4504, 2021. 2

[60] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 5, 7

[61] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single rgb-d image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[62] Xiaoming Zhao, Yuan-Ting Hu, Zhongzheng Ren, and Alexander G Schwing. Occupancy planes for single-view rgb-d human reconstruction. *arXiv preprint arXiv:2208.02817*, 2022. 2, 6, 7

[63] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE International Conference on Computer Vision*, 2019. 2

[64] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3170–3184, 2021. 1, 2, 3, 4, 5, 6, 7