High-Fidelity Clothed Avatar Reconstruction from a Single Image

Tingting Liao[1,2*] Xiaomei Zhang[1,2*] Yuliang Xiu[3] Hongwei Yi[3] Xudong Liu[4]
Guo-Jun Qi[4,5] Yong Zhang[6] Xuan Wang[6] Xiangyu Zhu[1,2] Zhen Lei[1,2,7†]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[4]OPPO Research [5]Westlake University [6]Tencent AI Lab [7]CAIR, HKISI, CAS
{tingting.liao, xiaomei.zhang, xiangyu.zhu, zlei}@nlpr.ia.ac.cn
{yuliang.xiu, hongwei.yi}@tuebingen.mpg.de
{yongzhang201303, xwang.cv, guojunq}@gmail.com
{xudong.liu}@oppo.com

Abstract

This paper presents a framework for efficient 3D clothed avatar reconstruction. By combining the advantages of the high accuracy of optimization-based methods and the efficiency of learning-based methods, we propose a coarse-to-fine way to realize a high-fidelity clothed avatar reconstruction (CAR) from a single image. At the first stage, we use an implicit model to learn the general shape in the canonical space of a person in a learning-based way, and at the second stage, we refine the surface detail by estimating the non-rigid deformation in the posed space in an optimization way. A hyper-network is utilized to generate a good initialization so that the convergence o f the optimization process is greatly accelerated. Extensive experiments on various datasets show that the proposed CAR successfully produces high-fidelity avatars for arbitrarily clothed humans in real scenes. The codes will be released in https://github.com/TingtingLiao/CAR.

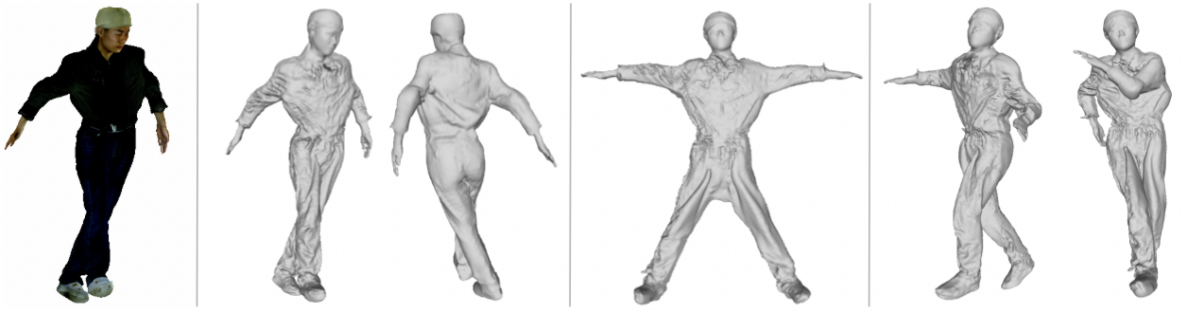
1. Introduction

Clothed avatar reconstruction is critical to a variety of applications for 3D content creations such as video gaming, online meeting [54,55], virtual try-on and movie industry [10, 21, 39]. Early attempts are based on expensive scanning devices such as 3D and 4D scanners, or complicated multi-camera studios with carefully capturing processes. While highly accurate reconstruction results can be obtained from these recording equipment, they are inflexible and even not feasible in many applications. An alternative is to collect data using depth sensors [31, 42], which is however still less ubiquitous than RGB cameras. A more practical and low-cost way is to create an avatar from an image by RGB cameras or mobile phones.

Monocular RGB reconstruction [19, 37, 51, 59] has been extensively investigated and shows promising results. ARCH [22] is the first method that reconstructs a clothed avatar from a monocular image. Due to the disadvantage of depth ambiguity, a number of methods that create an avatar from a video are proposed to resolve the problem. Most existing monocular video-based methods [2, 3, 7, 9, 10, 14, 15, 46] are typically restricted to parametric human body prediction, which lacks geometry details like cloth surface. How to create a high-fidelity avatar from an in-the-wild image, with consistent surface details is still a great challenge.

In this work, we focus on the shape recovery and propose an efficient high-fidelity clothed human avatar creation method from a single image in a coarse-to-fine way. The method consists of a learning-based canonical implicit model and an optimization-based normal refinement process. The canonical implicit model uses the canonical normal inverse transformed from original space as geometric feature to help grasp clothing detail of the general shape in canonical space. Unlike occupancy-based methods [22, 36, 37], we adopt a Signed Distance Function (SDF) to approximate the canonical human body surface, which gains advantages in learning the human body in the surface level instead of the point level, so that the reconstruction accuracy is improved. In the normal refinement process, a SDF is learned to approximates the target surface in the posed space by enforcing its surface normal closed to the predicted normal image. Compared with mesh-based refinement, our method can obtain

*Equal contribution.
†Corresponding author.



(a) Input (b) Posed Reconstruction (c) Canonical Reconstruction (d) Reposed

Figure 1. Images to avatars. Given an image of a person in an unconstrained pose (a), our method reconstructs 3D clothed avatars in both original posed space (b) and canonical space (c) and can repose the human body from the canonical mesh (d).

more realistic results without artifacts due to the flexibility of implicit representation. Moreover, to learn the SDF of the normal refinement process efficiently, we propose a meta-learning-based hyper network for parameter initialization to accelerate the convergence of the normal refinement process.

Extensive experiments have been conducted on MVP-Human [60] and RenderPeople [1] datasets. Both qualitative and quantitative results demonstrate that our proposed method outperforms related avatar reconstruction methods. The main contributions are summarized as follows:

- We propose a coarse-to-fine framework for efficient clothed avatar reconstruction from a single image. Thanks to the integration of image and geometry features, as well as the meta-learning, it achieves high-fidelity clothed avatar reconstruction efficiently.
- We design the canonical implicit regression model and the normal refinement process. The former fuses all observations to the canonical space where the general shape of a person is depicted, and the latter learns pose dependent deformation.
- Results validate that our method could reconstruct high-quality 3D humans in both posed and canonical space from a single in-the-wild image.

2. Related Work

3D Clothed Human Reconstruction. 3D clothed human reconstruction [4, 24–27, 44, 51] from multi-view or even a single image has achieved great progress in recent years. Saito et al. [36] introduce Pixel-Aligned Implicit Function (PIFu), which formulates an implicit function using pixel-aligned image features and points the depth to obtain the human body's occupancy field for the first time. However, PIFu cannot preserve high-frequency details like cloth wrinkles and then the generated surfaces tend to be smooth. To address this issue, PIFuHD [37] proposes a multi-level framework to reconstruct high-fidelity clothed humans from high-resolution normal images. Despite impressive results, both PIFu and PIFuHD show poor robustness on in-the-wild images with out-of-distribution poses. Some works [8, 19, 50, 51, 58] try to tackle this problem by utilizing the human body prior to learn 3D semantic information. The combination of explicit parametric models and implicit representations enables the model to be more robust to out-of-distribution poses. For example, GeoPIFu [19] and PaMIR [58] extract voxel-aligned semantic features from SMPL [29] body to make the model more robust to pose variation. Recently, Xiu et al. [51] find that these methods are sensitive to the global pose, due to their 3D convolutional encoders. ICON [51] uses local features including normal and signed distance to estimate the occupancy value of a query point. PHORHUM [5] additionally estimates the albedo and global scene illumination, hence enabling relighting. While impressive results can be obtained from existing methods, such approaches reconstruct static 3D humans which cannot be animated.

Avatar Reconstruction. Many works [11, 12, 38, 47] use scanning devices to obtain 4D scans and fuse them into an animatable avatar. Similarly, human performance capture approaches [17, 18, 28, 53, 56] use a pre-scanned template and track per-frame shape deformations. Nevertheless, all these methods require expensive and unportable capture devices. In contrast, RGB monocular camera-based avatar reconstruction gains more popularity in recent years due to its low cost and convenience. The methods can be roughly categorized into optimization-based and learning-based ones.

The optimization-based methods focus on overfitting an avatar from a video of a specific moving subject. Early works [13, 14, 43, 48] are based on visual hulls using silhouettes from multiple views to obtain the visible areas of the captured person. The concavity problem is changeable and difficult to handle. After that, researchers [2, 3, 7, 10, 34, 39, 49] attempt to model the geometry on top of parametric models with vertex offsets. The mesh representation has a fixed topology which is insufficient to recover high-quality results, especially on loose clothes such as skirts and dresses. Unlike meshes, implicit representations are more powerful to help

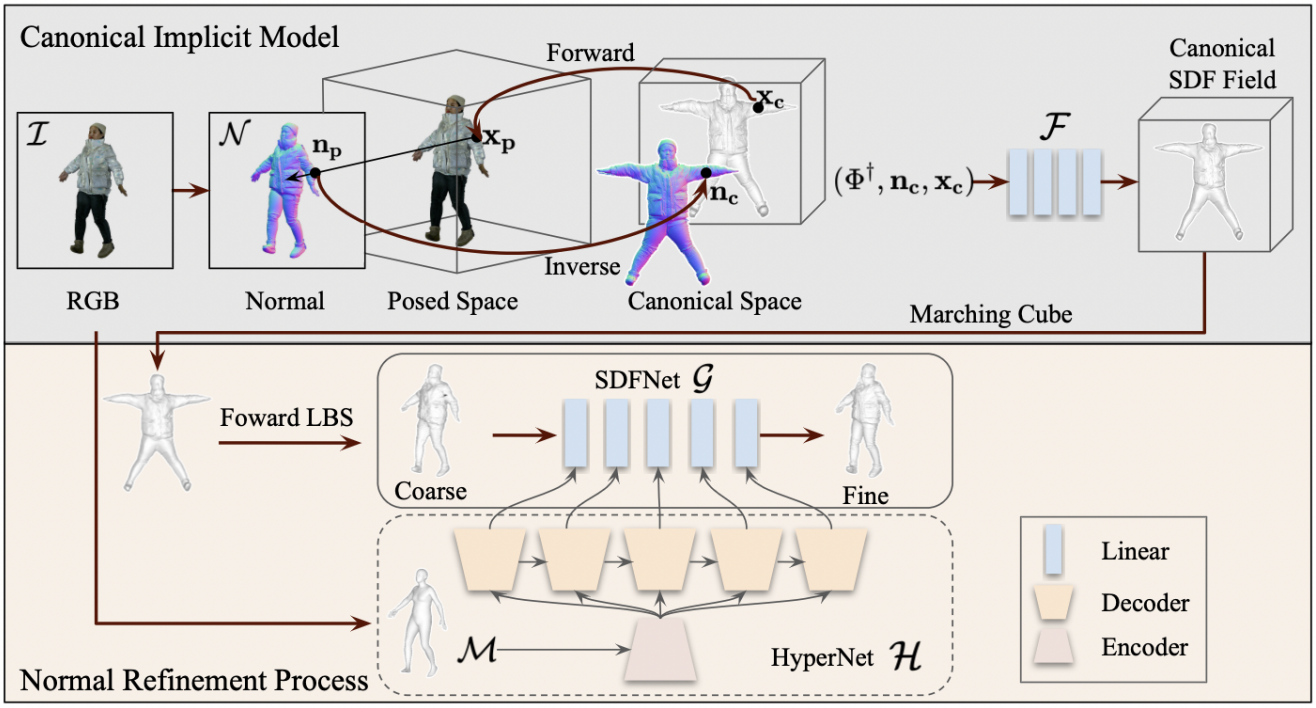


Figure 2. Framework of CAR. Given an RGB image $\mathcal{I}$, we first estimate its SMPL body $\mathcal{M}$ and the normal map $\mathcal{N}$. The canonical implicit model then takes $\mathcal{N}$ and body pose as input to estimate a canonical SDF field. Then, the normal refinement process warps it to the posed space and generates a high-fidelity clothed avatar reconstruction.

recover detailed 3D shapes with arbitrary topology. NeRF-based methods [33, 52, 57] optimize a goal using conditions on articulated cues. Jiang et al. [23] further combine the explicit and implicit representations to reconstruct high-fidelity geometry. However, the optimization process in these methods is time-consuming and inefficient in real applications.

Comparatively, the learning-based methods are more efficient in the prediction process. ARCH [22] is the first work to propose an end-to-end learning-based framework to estimate a canonical avatar from a single image. It computes the Radial Basis Function (RBF) distance between query points to body landmarks as geometric features. ARCH++ [20] employs Pointnet++ [35] as a geometry encoder to extract human body prior information. However, the geometric features leveraged from a naked body make the recovered surface lose details. The human body prior provides pose information which is helpful for reconstructing 3D humans in the original posed space (e.g. ICON), but has a minor effect for canonical shape recovery. The geometry cues such as normal are supposed to be more important for clothed avatar reconstruction. Different from ARCH and ARCH++, We utilize the canonical normal transformed from the original space to help preserve high-frequency details.

3. Method

Figure 2 shows the framework of the clothed avatar reconstruction (CAR) method. Given an image, the front and back normal images and a SMPL body are simultaneously obtained by the body-guided normal prediction model described in [51]. In the first stage, the canonical implicit model takes the predicted normal image as input and recover a coarse human body in canonical space. In the second stage, the coarse mesh surface is implicitly refined by a SDF Network.

3.1. Canonical Implicit Model

The canonical implicit model aims to reconstruct the general shape of a subject in the canonical space. Previous methods [20, 22] estimate an occupancy field by learning a classification task that whether a 3D point is inside or outside a target human body. This scheme is unfriendly for animatable avatar reconstruction tasks where the mapping between posed space and canonical space is required. While in practice, imperfect mapping is unavoidable and there is always a gap between estimated poses and ground truth poses. As a result, the occupancy based methods tend to classify the points inside a human body as background when they are erroneously projected to the background area in a 2D image.

Unlike the occupancy based methods which are point level, our method adopts a signed distance function (SDF) to approximate the target human body in a surface level, which is more robust to local mapping noises. Unlike occupancy, SDF aims to find an optimal surface where the surface nor-

mal is closest to the target surface normal. Instead of using the classification loss, CAR constrains points' gradients and surface normal as in [16]. Table 1 lists a comparison of the implicit functions, which mainly use image features and geometric features to estimate a points' signed distance or occupancy value.

There are three kinds of features used in our implicit function $\mathcal{F}$ to predict a point's signed distance to a target surface, which are pixel-aligned image feature $\Phi^{\dagger}$, canonical normal $\mathbf{n_c}$ and the canonical location $\mathbf{x_c} \in \mathbb{R}^3$. The zero-level surface is formulated as:

$$\mathcal{S}_\eta = \{\mathbf{x_c} \in \mathbb{R}^3 | \mathcal{F}(\Phi^{\dagger}, \mathbf{n_c}, \mathbf{x_c}; \eta) = 0\}, \tag{1}$$

where η is the network parameters.

Linear Blend Skinning. LBS [29] is widely used to control the large-scale movements of a human body, by transforming the skin according to the motion of the skeleton. Let $\mathbf{X} = \{\mathbf{x}_\mathbf{c}^i \in \mathbb{R}^3\}_{i=1}^{N_\mathbf{v}}$ be the body vertices in the canonical space and $W = \{w^i \in \mathbb{R}^{N_J}\}_{i=1}^{N_\mathbf{v}}$ be the vertex-to-bone skinning weights, where $N_\mathbf{v}$ and N_J are the number of vertices and joints respectively. For simplicity, we omit index i for $\mathbf{x}_\mathbf{c}^i$ and w^i. Given pose parameters $\theta \in \mathbb{R}^{N_J \times 3}$ and joints J, the LBS function $\mathcal{W}$ maps a canonical vertex $\mathbf{x_c}$ with its skinning weight $w \in \mathbb{R}^{N_J}$ to the target posed space as follow:

$$\mathbf{x_p} = \mathcal{W}(\mathbf{x_c}, w, \theta, J) = \sum_{j=1}^{N_J} w_j \mathbf{B}_j(\theta, J)\mathbf{x_c}, \tag{2}$$

where $\mathbf{B}_j(\theta, J)$ is the bone transformation applied on a body part j. We define $\mathcal{W}^{-1}$ as the inverse LBS function mapping vertices from original space to canonical space.

Pixel-Aligned Image Feature. We use Stacked Hourglass (SHG) as the normal image encoder which is the same as in [20, 22, 36, 37]. Given an RGB image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, we predict the normal image $\mathcal{N}$ using the normal prediction model in [51]. Then, the normal image encoder takes $\mathcal{N}$ as input and outputs a feature map $G(\mathcal{N}) \in \mathbb{R}^{H' \times W' \times C'}$. By projecting a point in posed space onto a normal image plane, the pixel-aligned image features can be obtained using the bilinear interpolation as follow:

$$\Phi^{\dagger} = \mathcal{B}(G(\mathcal{N}), \pi(\mathbf{x_p})), \tag{3}$$

where $\mathbf{x_p}$ is the deformed point in the posed space obtained by Eqn. 2, $\pi(\cdot)$ indicates the weak orthogonal camera projection, and $\mathcal{B}$ denotes the bilinear interpolation operation.

Geometric Feature. CAR leverages canonical normal as the geometric feature. Inspired by [40], the canonical normal $\mathbf{n_c}$ can be obtained by an inverse transform of the predicted normal $\mathbf{n_p}$ in original space, using the Jacobian matrix of a forward transform of $\mathbf{x_c}$.

$$\mathbf{n_c} = unit(\nabla_{\mathbf{x_c}} \mathcal{W}^{-1} \cdot \pi^{-1} \cdot \mathbf{n_p}), \tag{4}$$

Table 1. Comparison of implicit functions of different human body reconstruction methods. $\Phi^{\dagger}$ denotes the pixel-aligned image feature, $\mathcal{N}^{\dagger}$ denotes the pixel-aligned normal predicted from an RGB image, Ψ denotes the geometric feature leveraging the human body prior $\mathcal{M}$, $\mathbf{x} = (x, y, z) \in \mathbb{R}^3$ is a 3D point, z is the depth of $\mathbf{x}$ and $\mathbf{n}$ is the normal of $\mathbf{x}$ in the canonical space.

Method	Implicit Function
PIFu [36], PIFuHD [37]	$\mathcal{F}(\Phi^{\dagger}, z)$
PHORHUM [4]	$\mathcal{F}(\Phi^{\dagger}, \mathbf{x})$
PaMIR [58], GeoPIFu [19],ARCH [22]	$\mathcal{F}(\Phi^{\dagger}, \Psi(\mathcal{M}))$
ICON [51]	$\mathcal{F}(\mathcal{N}^{\dagger}, \Psi(\mathcal{M}))$
CAR (ours)	$\mathcal{F}(\Phi^{\dagger}, \mathbf{n}, \mathbf{x})$

where π^{-1} means the inverse camera projection matrix, $\mathbf{n_p} = \mathcal{B}(\mathcal{N}, \pi(\mathbf{x_p}))$ is the pixel-aligned normal indexed from predicted normal image $\mathcal{N}$, and $unit(\cdot)$ means normalizing the input vector. Ablation study on different types of geometric features shows that the normal feature is better than other methods.

Point Position. Specifically, we use the basic position encoding [30] of the canonical point as an additional feature. This term is maintained for the sake of computing gradients and surface normal $\nabla_\mathbf{x}\mathcal{F}$.

Training Loss. Following [16], our training loss contains three terms: surface reconstruction loss $\mathcal{L}_\mathrm{I}$, geometric regularization loss $\mathcal{L}_\mathrm{eik}$, and off-surface regularization $\mathcal{L}_\mathrm{o}$.

$$\mathcal{L} = \lambda_\mathrm{I}\mathcal{L}_\mathrm{I} + \lambda_\mathrm{eik}\mathcal{L}_\mathrm{eik} + \lambda_\mathrm{o}\mathcal{L}_\mathrm{o}, \tag{5}$$

Empirically, we set $\lambda_\mathrm{I} = 1, \lambda_\mathrm{eik} = 0.1, \lambda_\mathrm{o} = 0.1$.

Reconstruction Loss. $\mathcal{L}_\mathrm{I}$ is a reconstruction loss, which ensures the Signed Distance Function vanish on surface points and its normal is consistent with the ground truth surface normal.

$$\mathcal{L}_\mathrm{I} = \frac{1}{|\Omega_\mathrm{I}|} \sum_{\mathbf{x} \in \Omega_\mathrm{I}} (|\mathcal{F}_\mathbf{x}| + ||\mathbf{n_x} - \hat{\mathbf{n}}_\mathbf{x}||), \tag{6}$$

where $\mathbf{n_x} = \nabla_\mathbf{x}\mathcal{F}_\mathbf{x}$ is the differential normal at $\mathbf{x}$, $\hat{\mathbf{n}}_\mathbf{x}$ is the target normal and Ω_I is a set of points which are randomly sampled from surface points.

Eikonal Loss. The formulated Eikonal loss [6] is a regular loss commonly used to constrain $\mathcal{G}$ and $\mathcal{F}$ to be a SDF, by enforcing the implicit function to have a unit gradient:

$$\mathcal{L}_\mathrm{eik} = \frac{1}{|\Omega_\mathrm{D}|} \sum_{\mathbf{x} \in \Omega_\mathrm{D}} (||\mathbf{n_x}|| - 1), \tag{7}$$

where Ω_D is a point set sampled from a uniform distribution within the bounding box.

Off-surface Regularization. $\mathcal{L}_\mathrm{o}$ enforces the sign distance of points far from the surface as large as possible as in [41].

$$\mathcal{L}_\mathrm{o} = \frac{1}{|\Omega_\mathrm{D}|} \sum_{\mathbf{x} \in \Omega_\mathrm{D}} \exp(-\alpha \cdot |\mathcal{F}_\mathbf{x}|), \tag{8}$$

where $\alpha \gg 1$ is the sharpness of decision boundary.

3.2. Normal Refinement Process

The canonical implicit model tends to ignore surface details and focuses on the general shape of the captured subject due to the the bad mapping between canonical space and posed space. To solve this problem, we refine the predicted canonical surface and enforce it to be consistent with the input image. Like [51, 59], we adopt the normal image in the posed space to refine the reconstructed results. Different from previous methods, we use an implicit function, i.e., a Signed Distance Function instead of learning the per-vertex displacement which always produce artifacts because the topology of a mesh is fixed. The normal refinement process mainly consists of two parts: a surface reconstruction network $\mathcal{G}_\varphi$ to generate a high-quality human body by a SDF, and a meta hyper-network $\mathcal{H}_\phi$ generating the initial parameters of the reconstruction network for fast optimization.

3.3. Hyper Network Training

Optimizing the reconstruction network $\mathcal{G}_\varphi$ from scratch is inefficient and not necessary, since a coarse mesh is already obtained in the first stage and it can be used as an initialization. Then the problem becomes how to estimate the parameters φ_0 that $\mathcal{G}_{\varphi_0}$ can approximate a known mesh. Our solution is to leverage a hyper network $\mathcal{H}_\phi$ takes a mesh as input and output a set of parameters. We condition the hyper network on a SMPL body mesh instead of the reconstruction in the first stage. The are two main reasons. The first is that SMPL body mesh is naked without clothing variation, and the topology is simple and easy for learning. The second reason is that the real 3D human data is limited, while a large scale SMPL data with various poses and shapes can be synthesized to train the hyper network, thus improving the networks' generalization to unseen data. Note the hyper-network is only trained once in our method.

Given a SMPL body $\mathcal{M}(\theta, \beta)$, the hyper network $\mathcal{H}_\phi$ generates a set of parameters $\varphi_0 = \mathcal{H}(\mathcal{M}; \phi)$, which are used to parameterize the SDF reconstruction network $\mathcal{G}$ to reproduce $\mathcal{M}$. The zero-level surface can be represented as follow:

$$\mathcal{M}^* = \{\mathbf{x} \in \mathbb{R}^3 | \mathcal{G}(\mathbf{x}; \mathcal{H}(\mathcal{M}; \phi)) = 0\} \tag{9}$$

where ϕ are learnable parameters of the hyper network. During training, ϕ is updated by enforcing $\mathcal{M}^*$ closed to $\mathcal{M}$. The training loss is the same as equation (5). Once the hyper network is trained, the output parameters are used to initialize the SDF reconstruction network as $\mathcal{G}_{\varphi_0}$.

3.4. SDF Network Optimization

After obtaining the canonical mesh by the canonical implicit model described in section 3.1, it is warped to the posed space and refined to improve the quality of surface geometry. The zero isosurface in the posed space is formulated as:

$$\mathcal{S}_\varphi = \{\mathbf{x} \in \mathbb{R}^3 | \mathcal{G}(\mathbf{x}; \varphi) = 0\}, \tag{10}$$

where φ is trainable. Starting from φ_0, φ is updated by optimizing the surface normal supervised by the predicted normal image $\mathcal{N}$. The optimization loss is similar to equation (5). For the second term in equation (6), the target normal $\hat{\mathbf{n}}_\mathbf{x} = \mathcal{B}(\mathcal{N}, \pi(\mathbf{x_p}))$ can be obtained by projecting a point to either front or back normal image. In practice, we use both front and back normal images to optimize the surface normal.

4. Experiments

In this section, we evaluate CAR with state-of-the-art methods on MVP-Human [60] and RenderPeople [1] datasets. The ablation study and discussion are also conducted to show the effectiveness of the proposed method.

Dataset Description. The training set contains 100 subjects from MVP-Human dataset and 50 subjects from RenderPeople dataset. The testing set includes 50 scans from MVP-Human, 11 scans from RenderPeople, and 2D real images from the internet. There is no intersection of training and testing sets. In the training phase, we fit a rigged 3D body template in the canonical pose with corresponding skinning weights to the scan mesh for each subject. We generate a motion sequence for each subject by warping the canonical mesh with the poses provided in AIST++ dataset [45]. All generated meshes are rendered by rotating a camera around the vertical axis with intervals of 3 degrees.

Implementation Details. We use stacked hourglass network [32] as the normal image encoder which has the same architecture with [19, 22, 37]. The MLP of $\mathcal{F}$ has the number of neural layers (262, 512, 512, 512, 512, 512, 512, 1) with a skip connection at the fourth layer. We use the geometric initialization proposed in [6] for $\mathcal{F}$. The SMPL encoder has the number of neural layers (6, 256, 256, 256, 256, 256) with skip connections at the second, the third and the fourth layer, while the decoder contains 5 blocks, each has 3 hidden layers with 256 neural layers and an output layer whose number of neural is the same as the parameters' number of the corresponding layer of SDF network. The SDF network has the number of neural layers (3, 1024, 512, 256, 128, 1) and the parameters of each layer are initialized by the output of each decoder block. The canonical implicit model is trained with a batch size of 4 and a random window crop of 512×512 sizes. We use Adam optimizer and learning rate 1e-3 with decay by a factor of 0.1 every 3 epochs. In each iteration, we sample 8192 points on the surface, 8192 points around the surface with a normal distribution sigma of 0.1, and 2048 points uniformly sampled in a bounding box. We train the

Table 2. Quantitative comparisons of different methods in both canonical and posed space on MVP-Human (MVP) and RenderPeople (RP).

Methods	MVP-Canonical			MVP-Posed			RP-Canonical			RP-Posed		
	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓
PIFu [36]	-	-	-	4.9638	6.1931	0.8013	-	-	-	4.8884	5.1182	0.7089
PIFuHD [37]	-	-	-	3.9068	4.3833	0.8247	-	-	-	5.2701	5.3971	0.7375
ICON [51]	-	-	-	3.9583	4.3886	0.1957	-	-	-	4.9126	5.1269	0.7610
ARCH [22]	1.5894	1.8044	**0.0942**	3.8274	4.3614	0.1819	2.3916	2.1424	0.1178	2.3225	2.0506	0.1543
ARCH++ [20]	2.3906	2.0035	0.1849	4.0438	3.9825	0.2523	1.9046	1.8306	0.0971	1.8805	1.7720	0.1065
CAR (ours)	**1.0572**	**1.0811**	0.1287	**1.0771**	**1.0654**	**0.0902**	**1.5401**	**1.4963**	**0.0821**	**1.5142**	**1.4147**	**0.0871**

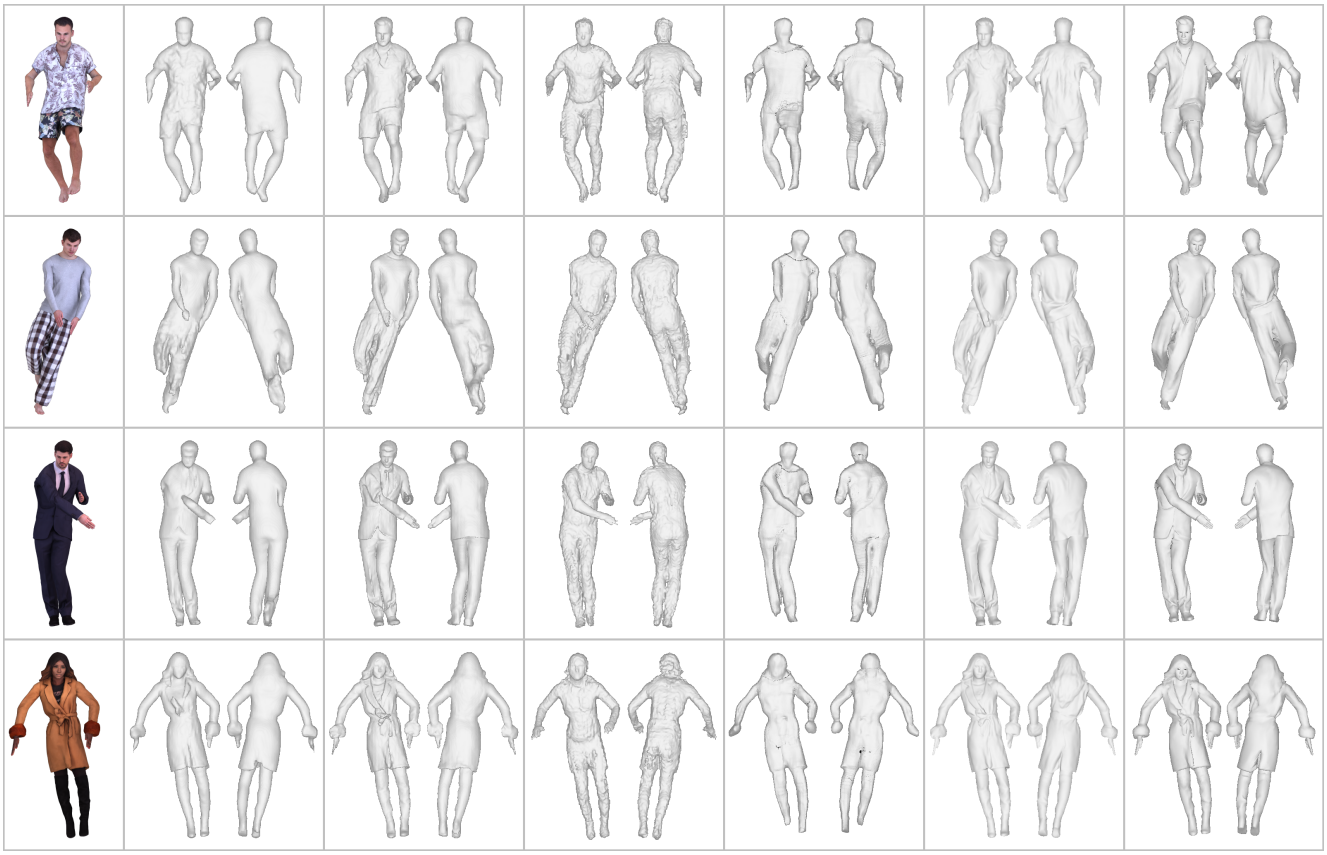


Figure 3. Qualitative comparisons against the state-of-the-art methods on RenderPeople testing set.

hyper-network using SMPL data generate by shape and pose parameters randomly sampled from SMPL data distribution. The normal prediction network and SMPL estimation are followed from [51].

4.1. Quantitative Evaluation

We compare our method with two kinds of methods: clothed human reconstruction algorithms include PIFu [36], PIFuHD [37] and ICON [51], and avatar reconstruction approaches ARCH [22] and ARCH++ [20]. Table 2 illustrates the quantitative results in both canonical and posed space on MVP-Human and RenderPeople datasets. For PIFu, PIFuHD and ICON, we directly use the model published by original works. For ARCH and ARCH++, we train the model by ourselves in our training set and report the testing performance. From the results, although PIFu/PIFuHD usually shows good performance in visual results, they do not look so good in quantitative evaluation. That's because they do not utilize the human body prior and are not so robust to pose variations. ICON leverages SMPL to improve pose robustness. However, it relies too heavily on the naked SMPL body and is not good enough to handle loose clothes such as coats (4th column, 4th row in Figure 3). ARCH, ARCH++, and the proposed CAR introduce the process of general shape reconstruction in the canonical space so that the pose robustness is greatly improved. Our method CAR further pays more attention to the geometry detail recovery on the surface, and it achieves the best accuracy over all datasets in both

Figure 4. Images to animated avatars.

Figure 5. Qualitative results on real images from the internet. These results demonstrate that our model trained by synthetically generated data can successfully reconstruct high-fidelity 3D from humans in real world data.

Table 3. Ablation study of different modules (RenderPeople).

Methods	Canonical Space			Posed Space		
	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓
CAR,baseline	1.5760	1.5766	**0.0774**	1.5484	1.4926	0.0983
+GeoFeat	**1.5329**	1.5069	0.0927	**1.5043**	1.4239	0.1120
+Refine	1.5401	**1.4963**	0.0821	1.5142	**1.4147**	**0.0871**

canonical and posed space, validating its effectiveness for the clothed avatar reconstruction.

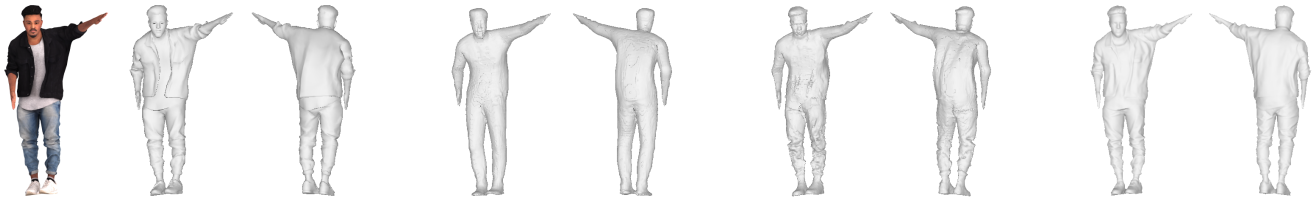
4.2. Qualitative Evaluation

Fig. 3 shows the qualitative results in RenderPeople. PIFu fails to reconstruct whole limbs since it does not use human body priors. PIFuHD captures better details than PIFu, but the backside of reconstructions is overly smooth due to the lacking of end-to-end geometry encoder. ICON suffers from surface noise since the image global encoder is removed thus extracted features are purely local. It is worth noting that PIFu, PIFuHD and ICON do not support animation. ARCH can generate animatable avatars, but the recovered surface is overly smooth or with artifacts. Our method successfully produces realistic 3D humans which can be animated, see Fig. 4 for some examples. Fig. 5 shows more results on in-the-wild images, which demonstrates that our method can reconstruct high-fidelity 3D humans, regardless of poses or clothing.

4.3. Analysis and Discussion

To evaluate the influences of different factors, we conduct three experiments including: 1) the ablation study on our method of different factors; 2) the comparison of SDF and occupancy losses; 3) different choices of geometric features.

Ablation Study. Table 3 and Figure 6 demonstrate the effectiveness of different parts of our method. The first row



Input GT Ours w/o GeoFeat Ours w/o refinement Ours

Figure 6. Ablation analysis on different modules, corresponding to Table 3.

Table 4. Reconstruction errors on different types of loss and geometric feature variants of our method on MVP-Human.

	Methods	Canonical Space			Posed Space			Mean		
		Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓	Chamfer↓	P2S↓	Normal↓
Loss	occupancy	1.3867	1.4974	0.1119	3.5136	3.9336	0.1945	2.4501	2.7155	0.1532
	sdf (ours)	1.3758	1.4081	0.0896	3.7589	3.8199	0.2006	2.5673	2.6140	0.1451
Geometric Feature	P2J distance	1.5227	1.3106	0.1288	3.8324	3.7256	0.2310	2.6775	2.5181	0.1799
	signed distance	2.0142	1.3137	0.1537	4.2412	3.7999	0.2365	6.5639	2.3932	0.3429
	surface normal (ours)	1.2463	1.3020	0.1053	3.6058	3.7560	0.2069	2.4411	2.5237	0.1570

of Table 3 is our method without the geometric feature and normal refinement. We denote it $\mathcal{F}(\Phi^{\dagger}, \mathbf{x_c})$ as the baseline method. The second row is our method described in section 3.1 without the normal refinement process. The third is our proposed method. We can see that the geometry feature improves the accuracy of reconstructed results. The normal refinement process further reduces the reconstruction errors and their qualitative results are shown in Figure 6.

Occupancy vs SDF. Previous methods [20, 22, 36, 37, 51] tries to estimate an occupancy value for a query point that "1" means inside the human body while "0" means outside. To this end, a regression loss (e.g., L2 loss) or classification loss (e.g., binary cross entropy loss) is always used to enforce estimations to be close to the real occupancy field. Our method, instead, predicts a SDF field and trains our network using the loss described in section 3.1. In this part, we compare the results of these two kinds of losses. We train two models using L2 occupancy loss and SDF loss respectively. For a fair comparison, all configurations are the same except the training loss. The first two rows of Table 4 show that the SDF loss performs better than occupancy loss.

Geometric Feature Evaluation. We analyse three different geometric features in Table 4: 1) spatial feature P2J distance proposed in ARCH [22] (i.e., distance from a point to SMPL joints) 2) signed distance proposed in ICON [51] (i.e., distance from a point to the nearest point on SMPL surface) 3) ours with canonical normal. The bottom three rows in Table 4 show that the canonical normal performs the best, which achieves the lowest errors.

4.4. Inference Speed

The proposed CAR consists of two stages. The first stage adopts a learning-based way so that the shape recovery in canonical space is efficient. The second stage is an optimized way. Fortunately, with the initialization of the hyper network, the normal refinement process converges faster. It takes about 1500 iterations to output an optimal result (compared to 3000 iterations with random initialization). For a single subject, these two stages require an average time of about 5 minutes on a single TITAN X GPU, while optimization-based methods usually takes several hours to construct a subject.

5. Conclusion

We present a method for clothed avatar reconstruction from a single image in free viewpoints and unconstrained poses. The person image is decomposed into a canonical mesh describing its general shape as well as a pose-dependent non-rigid deformation. By incorporating the normal information in both canonical mesh learning and the non-rigid deformation refinement process, we successfully reconstruct high-fidelity avatars which preserve surface details like cloth wrinkles. With a hyper network for parameter initialization, it further accelerates the convergent process and improves the optimization efficiency. Our method can be easily extended to multiple image settings and the avatar reconstruction results are expected to be improved. How to utilize consistent information across temporal images or monocular video of a dynamic human to reconstruct a complete avatar is one of our future works.

Acknowledgement

This work was supported in part by Chinese National Natural Science Foundation Projects #62176256, #62206280, #62276254, the Tencent AI Lab Rhino-Bird Focused Research Program RBFR2022010, the OPPO Research Fund and the InnoHK program. Yuliang Xiu has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.860768 (CLIPE project). Hongwei Yi is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B.

References

[1] Renderpeople. renderpeople.com. 2, 5

[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*. CVPR Spotlight Paper. 1, 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, Sep 2018. 1, 2

[4] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 2

[6] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 4, 5

[7] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1, 2

[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 (ECCV)*. Springer, August 2020. 2

[9] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020. 1

[10] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5420–5430, 2019. 1, 2

[11] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 2

[12] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pages 612–628. Springer, 2020. 2

[13] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004. 2

[14] Yasutaka Furukawa and Jean Ponce. Carved visual hulls for image-based modeling. In *European Conference on Computer Vision*, pages 564–577. Springer, 2006. 1, 2

[15] 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 *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. Ieee, 2009. 1

[16] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 4

[17] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019. 2

[18] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2

[19] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 4, 5

[20] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 3, 4, 6, 8

[21] Pengpeng Hu, Edmond Ho, and Adrian Munteanu. 3dbodynet: Fast reconstruction of 3d animatable human body shape from a single commodity depth camera. *IEEE Transactions on Multimedia*, PP:1–1, 04 2021. 1

[22] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 1, 3, 4, 5, 6, 8

[23] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 3

[24] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*, pages 18–35. Springer, 2020. 2

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

[26] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, pages 1–1. 2020. 2

[27] 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*, pages 49–67. Springer, 2020. 2

[28] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *ECCV*, 2022. 2

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 4

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[31] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 5

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 3

[34] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 2

[35] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[36] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 4, 6, 8

[37] 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. 1, 2, 4, 5, 6, 8

[38] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 2

[39] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. *Computer Graphics Forum (CGF)*, 38(2):355–366, 2019. 1, 2

[40] Dario Seyb, Alec Jacobson, Derek Nowrouzezahrai, and Wojciech Jarosz. Non-linear sphere tracing for rendering deformed signed distance fields. *ACM Transactions on Graphics*, 38(6), 2019. 4

[41] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 4

[42] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017. 1

[43] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE computer graphics and applications*, 27(3):21–31, 2007. 2

[44] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7750–7759, 2019. 2

[45] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019. 5

[46] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 1

[47] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems*, 34:2810–2822, 2021. 2

[48] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2019. 2

[49] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 2

[50] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[51] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 1, 2, 3, 4, 5, 6, 8

[52] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34:14955–14966, 2021. 3

[53] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian

Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018. 2

[54] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1

[55] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1

[56] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5746–5756, 2021. 2

[57] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 3

[58] 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*, pages 1–1, 2021. 2, 4

[59] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 5

[60] Xiangyu Zhu, Tingting Liao, Jiangjing Lyu, Xiang Yan, Yunfeng Wang, Kan Guo, Qiong Cao, Stan Z. Li, and Zhen Lei. Mvp-human dataset for 3d human avatar reconstruction from unconstrained frames. *arXiv preprint arXiv:2204.11184*, 2022. 2, 5