ARCH++: Animation-Ready Clothed Human Reconstruction Revisited

Tong He[1,2]*, Yuanlu Xu[1]*, Shunsuke Saito[1], Stefano Soatto[2], Tony Tung[1]
[1]Facebook Reality Labs Research, USA [2]University of California, Los Angeles, USA
{simpleig,soatto}@cs.ucla.edu, {merayxu,shunsuke.saito16}@gmail.com, tony.tung@fb.com

Abstract

We present ARCH++, an image-based method to reconstruct 3D avatars with arbitrary clothing styles. Our reconstructed avatars are animation-ready and highly realistic, in both the visible regions from input views and the unseen regions. While prior work shows great promise of reconstructing animatable clothed humans with various topologies, we observe that there exist fundamental limitations resulting in sub-optimal reconstruction quality. In this paper, we revisit the major steps of image-based avatar reconstruction and address the limitations with ARCH++. First, we introduce an end-to-end point based geometry encoder to better describe the semantics of the underlying 3D human body, in replacement of previous hand-crafted features. Second, in order to address the occupancy ambiguity caused by topological changes of clothed humans in the canonical pose, we propose a co-supervising framework with cross-space consistency to jointly estimate the occupancy in both the posed and canonical spaces. Last, we use image-to-image translation networks to further refine detailed geometry and texture on the reconstructed surface, which improves the fidelity and consistency across arbitrary viewpoints. In the experiments, we demonstrate improvements over the state of the art on both public benchmarks and user studies in reconstruction quality and realism.

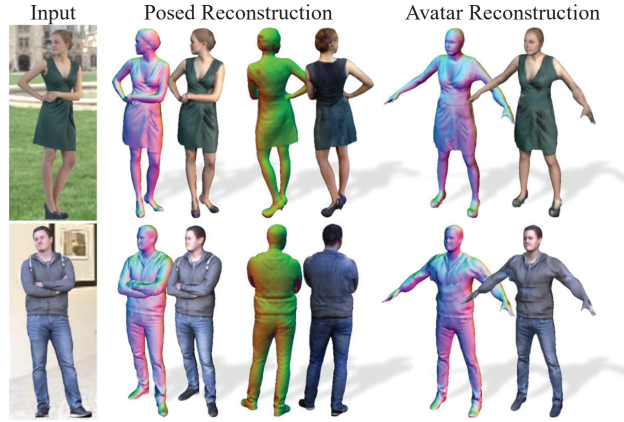

Figure 1. Given an image of a subject in arbitrary pose (left), our method could generate photorealistic avatars in both the posed input space (middle) as well as auto-rigged canonical space (right).

1. Introduction

Digital humans have become an increasingly important building block for numerous AR/VR applications, such as video games, social telepresence [48, 39] and virtual try-on. Towards truly immersive experiences, it is crucial for these avatars to obtain higher level of realism that goes beyond the uncanny valley [45]. Building a photorealistic avatar involves many manual works by artists or expensive capture systems under controlled environments [14, 21, 49], limiting access and increasing cost. Therefore, it is vital to revolutionize reconstruction techniques with minimal prerequisite (*e.g.*, a selfie) for future digital human applications.

Recent human models reconstructed from a single image combine category-specific data prior with image observations [72, 31, 66]. Among which, template-based approaches [32, 34, 67, 3, 9] nevertheless suffer from lack of fidelity and difficulty supporting clothing variations; while non-parametric reconstruction methods [55, 75, 56, 23], *e.g.*, using implicit surface functions, do not provide intuitive ways to animate the reconstructed avatar despite impressive fidelity. In the recent work ARCH [26], the authors propose reconstructing non-parametric human model using pixel-aligned implicit functions [55] in a canonical space, where all reconstructed avatars are transformed to a common pose. To do so, a parametric human body model is exploited to determine the transformations. By transferring skinning weights, which encode how much each vertex is influenced by the transformation of each body joint, from the underling body model, the reconstruction results are ready to animate. However, we observe that the advantages of a parametric body model and pixel-aligned implicit functions are not fully exploited.

In this paper we introduce ARCH++, which revisits the major steps of animatable avatar reconstruction from images and addresses the limitations in the formulation and representation of the prior work. First, current implicit function based methods mainly use hand-crafted features as the 3D space representation, which suffers from depth ambiguity and lacks human body semantic information. To address this, we propose an end-to-end geometry encoder

* This work was done as part of Tong He's internship at Facebook, Sausalito, CA, USA. The corresponding author is Yuanlu Xu.

based on PointNet++ [52, 53], which expressively describes the underlying 3D human body. Second, we find the unposing process to obtain the canonical space supervision causes topology change (*e.g*., removing self-intersecting regions) and consequently the articulated reconstruction fails to obtain the same level of accuracy in the original posed space. Therefore, we present a co-supervising framework where occupancy is jointly predicted in both the posed and canonical spaces, with additional constraints on the cross-space consistency. This way, we benefit from both: supervision in the posed space allows the prediction to retain all the details of the original scans; while canonical space reconstruction can ensure the completeness of a reconstructed avatar. Last, image-based avatar reconstruction often suffers from degraded geometry and texture in the occluded regions. To make the problem more tractable, we first infer surface normals and texture of the occluded regions in the image domain using image translation networks, and then refine the reconstructed surface with a moulding-inpainting scheme.

In the experiments, we evaluate ARCH++ on photorealistically rendered synthetic images as well as in-the-wild images, outperforming prior works based on implicit functions and other design choices on public benchmarks.

The contributions of ARCH++ include: 1) a point-based geometry encoder for implicit functions to directly extract human shape and pose priors, which is efficient and free from quantization errors; 2) we are the first to point out and study the fundamental issue of determining target occupancy space: posed-space fidelity vs. canonical-space completeness. Albeit ignored before, we outline the pros and cons of different spaces, and propose a co-supervising framework of occupancy fields in joint spaces; 3) we discover image-based surface attribute estimation could address the open problem of view-inconsistent reconstruction quality. Our moulding-inpainting surface refinement strategy generates 360° photorealistic 3D avatars. 4) our method demonstrates enhanced performance on the brand new task of image-based animatable avatar reconstruction.

2. Related Work

Template-based reconstruction utilizes parametric human body models, *e.g*., SCAPE [4] and SMPL [40] to provide strong prior on body shape and pose to address ill-posed problems including body estimation under clothing [69, 73] and image-based human shape reconstruction [11, 37, 32, 20, 35, 65, 34, 67]. While these works primarily focus on underling body shapes without clothing, the template-based representations are later extended to modeling clothed humans with displacements from the minimal body [51], or external clothing templates [10], from 3D scans [68, 51], videos [2, 22], and a single image [1, 10, 29]. As these approaches build clothing shapes on a body template mesh, the reconstructed models can be easily driven by pose parameters of the parametric body model. To address the lack of details with limited mesh resolutions, recent works propose to utilize 2D UV maps [36, 3]. However, as a clothing topology can significantly deviate from the underling body mesh and its variation is immense, these template-based solutions fail to capture clothing variations in the real world.

Non-parametric capture is widely used to capture highly detailed 3D shapes with an arbitrary topology from multi-view systems under controlled environments [43, 5, 61, 58, 18, 62, 16, 64, 59, 42]. Recent advances of deep learning further push the envelope by supporting sparse view inputs [19, 25], and even monocular input [38]. For single-view clothed human reconstruction, direct regression methods demonstrate promising results, supporting various clothing types with a wide range of shape representations including voxels [60, 28], two-way depth maps [17, 57], visual hull [46], and implicit functions [55, 56, 23]. In particular, pixel-aligned implicit functions (PIFu) [55] and its follow-up works [56, 23] demonstrate impressive reconstruction results by leveraging neural implicit functions [44, 12, 50] and fully convolutional image features. Unfortunately, despite its high-fidelity results, non-parametric reconstructions are not animation-ready due to missing body part separation and articulation. Recently, IF-Net [13] exploits partial point cloud inputs and learns implicit functions using latent voxel features. Compared with image-based avatar reconstruction, completion from points can leverage directly provided strong shape and pose cues, and thus skip learning them from complex images.

Hybrid approaches combine template-based and non-parametric methods and allow us to leverage the best of both worlds, namely structural prior and support of arbitrary topology. Recent work [8] shows that using SMPL model as guidance significantly improves robustness of non-rigid fusion from RGB-D inputs. For single-view human reconstruction, Zheng et al. first introduce a hybrid approach of a template-model (SMPL) and a non-parametric shape representation (voxel [75] and implicit surface [74]). These approaches, however, choose an input view space for shape modeling with reconstructed body parts potentially glued together, making the reconstruction difficult to animate as in the aforementioned non-parametric methods. The most relevant work to ours is ARCH [26], where the reconstructed clothed humans are ready for animation as pixel-aligned implicit functions are modeled in an unposed canonical space. However, such framework fundamentally leads to sub-optimal reconstruction quality. We achieve significant improvement on accuracy and photorealism by addressing the hand-crafted spatial encoding for implicit functions, the lack of supervision in the original posed space, and the limited fidelity of occluded regions.

3. Proposed Methods

Our proposed framework, ARCH++, uses a coarse-to-fine scheme, *i.e*., initial reconstruction by learning joint-space implicit surface functions (see Fig. 2), and then mesh refinement in both spaces (see Fig. 3).

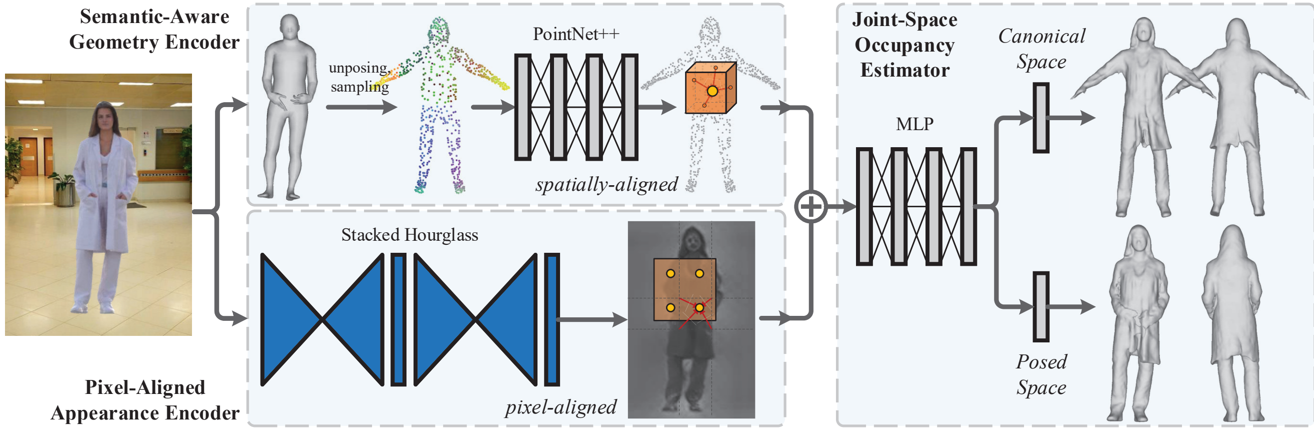

Figure 2. *Overview of the initial joint-space implicit surface reconstruction.* This procedure includes three components: i) semantic-aware geometry encoder, ii) pixel-aligned appearance encoder and iii) joint-space occupancy estimator. See text for detailed explanation.

3.1. Joint-Space Implicit Surface Reconstruction

Semantic-Aware Geometry Encoder. The spatial feature representation of a query point is critical for deep implicit function. While the pixel-aligned appearance feature via Stack Hourglass Network [47] has already demonstrated its effectiveness in detailed clothed human reconstruction by prior works [55, 56, 26, 23], an effective design of point-wise spatial encoding has not yet been well studied. The extracted geometry features should be informed of the semantics of the underlying 3D human body, which provide strong priors to regularize the overall dressed people shape.

The spatial encoding methods used previously include hand-crafted features (*e.g.*, RBF [26]) and latent voxel features [13, 23, 74]. The former is constructed based on Euclidean distances between a query point and the body joints, ignoring the shapes. The voxel-based features capture both shape and pose priors of a parametric body mesh. Compared with the hand-crafted features, the end-to-end learned voxel features are better informed of the underlying body structures but often constrained by GPU memory sizes and suffer from quantization errors due to low spatial resolution. To effectively encode the shape and pose priors without losing any precision, we propose a novel semantic-aware geometry encoder that extracts point-wise spatial encodings. Essentially a parametric body mesh can be sampled into a point cloud and fed into PointNet++ [52, 53] to learn point-based spatial features, which have several advantages over both hand-crafted RBF features and voxel-based ones. Our method encodes both shape and pose priors from parametric shapes without computation overhead and quantization errors caused by the mesh voxelization process. Additional detailed statistical comparisons on points v.s. voxels in representing 3D shapes are reported in [17].

Given a parametric body mesh estimated and deformed by [67, 26], we use a PointNet++ [52, 53] based semantic-aware geometry encoder to learn the underlying 3D human body prior. We sample N_0 (*e.g.*, 7324) points from the body mesh surfaces and feed them into the geometry encoder for spatial feature learning, that is,

$$f_{pn}: \{x_0^i\}_{i=1}^{N_0} \mapsto \{x_1^j, h_1^j\}_{j=1}^{N_1}, \{x_2^k, h_2^k\}_{k=1}^{N_2}, \{x_3^l, h_3^l\}_{l=1}^{N_3}, \quad (1)$$

where $x_0^i \in \mathbb{R}^3$ is a point sampled from the parametric body mesh. The PointNet++ based encoder utilizes fully-connected layers and neighborhood Max-Pooling to extract semantic-aware geometry features $h \in \mathbb{R}^{32}$ of a point. It also applies Furthest Point Sampling to progressively down sample the points $N_1 = 2048, N_2 = 512, N_3 = 128$ to extract latent features with increasing receptive fields. For example $\{x_1^j\}$ is a down sampled point set with size N_1, and $h_1^j \in \mathbb{R}^{32}$ is the learned feature w.r.t. each point.

As illustrated in Fig. 2, for any query point $p_a \in \mathbb{R}^3$ in the canonical space we obtain its point-wise spatial encoding $f_g \in \mathbb{R}^{96}$ via inverse $L2$-norm kernel based feature interpolation, followed by query coordinates concatenated Multi-layer Perceptrons (MLP). Particularly, we extract these features from different point set densities-j, k, l to construct concatenated features $f_g = (f_g^j \oplus f_g^k \oplus f_g^l)$ that are informed of multi-scale structures. For example, $f_g^j \in \mathbb{R}^{32}$ is defined as:

$$\begin{aligned} f_g^j(p_a, \{x_1^j, h_1^j\}) &= \text{MLP}(p_a \oplus \sum_m \frac{\|p_a - x_1^m\|^{-2}}{S(p_a, \{x_1^j, h_1^j\})} h_1^m), \\ S(p_a, \{x_1^j, h_1^j\}) &= \sum_m \|p_a - x_1^m\|^{-2}, \end{aligned} \quad (2)$$

where the index m is determined by finding the K nearest neighbors among the point set $\{x_1^j\}$ *w.r.t.* the query point. Empirically we found setting $K = 3$ obtains fair performance. The features extracted at other point set densities $f_g^k, f_g^l \in \mathbb{R}^{32}$ are obtained similarly leveraging $\{x_2^k, h_2^k\}$ and $\{x_3^l, h_3^l\}$, respectively.

Pixel-Aligned Appearance Encoder. We share the same architecture design as [55, 56, 26, 23] to map an input image $I \in \mathbb{R}^{512\times512\times3}$ into the latent feature maps $\psi_\mu(I) \in \mathbb{R}^{128\times128\times256}$ via a Stacked Hourglass Network [47] with weights μ. To obtain appearance encoding $f_a \in \mathbb{R}^{256}$ of any query point $p_b \in \mathbb{R}^3$ in the posed space, we project it back to the image plane based on a camera model of weak perspective projection, and bilinearly interpolate the latent image features:

$$f_a(p_b, I) = \mathcal{B}(\psi_\mu(I), \pi(p_b)), \quad (3)$$

where $\mathcal{B}(\cdot)$ indicates the differentiable bilinear sampling operation, and $\pi(\cdot)$ means weak perspective camera projection from the query point p_b to the image plane of I.

Joint-Space Occupancy Estimator. While most non-parametric and hybrid methods use the posed space as the learning and inference target space, ARCH instead reconstructs the clothed human mesh directly in a canonical space where humans are in a normalized A-shape pose. Different choices of the target space have pros and cons. The posed space is naturally aligned with the input pixel evidence and therefore the reconstructions have high data fidelity leveraging the direct image feature correspondences. Thus, many works choose to reconstruct a clothed human mesh in its original posed space (*e.g.*, PIFu(HD) [55, 56], Geo-PIFu [23], PaMIR [74]). However, in many situations the human can demonstrate complex poses with self-intersection (*e.g.*, hands in the pocket, crossed arms) and cause a "glued" mesh that is difficult to articulate. Meanwhile, canonical pose reconstruction offers us a rigged mesh that is animation ready (via its registered A-shape parametric mesh [26]). The problem of using the canonical space as the target space is that when we warp the mesh into its posed space there could be artifacts like intersecting surfaces and distorted body parts (see Fig. 6). Thus, the reconstruction fidelity of the warping obtained canonical-to-posed space mesh will degenerate. To maintain both input image fidelity and reconstruction surface completeness, we propose to learn the joint-space occupancy distributions.

We use a joint-space defined occupancy map O to implicitly represent the 3D clothed human under both its original posed space and a rigged canonical space:

$$O = \{(p_a, p_b, o_a, o_b) : \ p_a, p_b \in \mathbb{R}^3, \ -1 \leq o_a, o_b \leq 1\}, \quad (4)$$

where o_a, o_b denote the occupancy for points p_a and p_b. A point in the posed space is p_b and its mapped counterpart in the canonical space is $p_a = \text{SemDF}(p_b)$. The semantic deformation mapping (SemDF) between the original posed and the rigged canonical spaces is enabled by nearest neighbor-based skinning weights matching between p_b and the estimated underlying parametric body mesh [26].

To enable mesh reconstruction in joint spaces, we use both point-wise spatial features $f_g \in \mathbb{R}^{96}$ that are informed of semantic full-body structures, and pixel-aligned features $f_a \in \mathbb{R}^{256}$ that encode human front-view appearances:

$$o_a = \mathcal{F}_\theta(f_g \oplus f_a), \ o_b = \mathcal{F}_\beta(f_g \oplus f_a), \quad (5)$$

where θ, β are network weights of the MLP-based deep implicit surface functions. To reconstruct avatars from the dense occupancy estimations in two spaces, we use Marching Cube [41] to extract the isosurface at $o_a = \tau$ and $o_b = \tau$ (*i.e.*, $\tau = 0$), respectively.

The network outputs o_a, o_b are supervised by the ground truth joint-space occupancy $\hat{o}_a, \hat{o}_b$, depending on whether a posed space query point p_b and its corresponding canonical space point p_a are inside the clothed human meshes or not. Though p_a, p_b are a pair of mapped points their ground-truth occupancy values are not the same in all cases. For example, a point outside and close to the hand of a parametric body could has $\hat{o}_b > 0$ and $\hat{o}_a < 0$ if the original mesh in posed space has self-contact (*e.g.*, hands in the pocket). Namely, the SemDF defines a dense correspondence mapping between the two spaces but their occupancy values are not necessarily the same. Therefore, naively learning the distribution in one space and then warping the reconstruction into another pose can cause mesh artifacts (see Fig. 6). This motivates us to model two space occupancy distributions jointly in order to maintain both canonical space mesh completeness and posed space reconstruction fidelity.

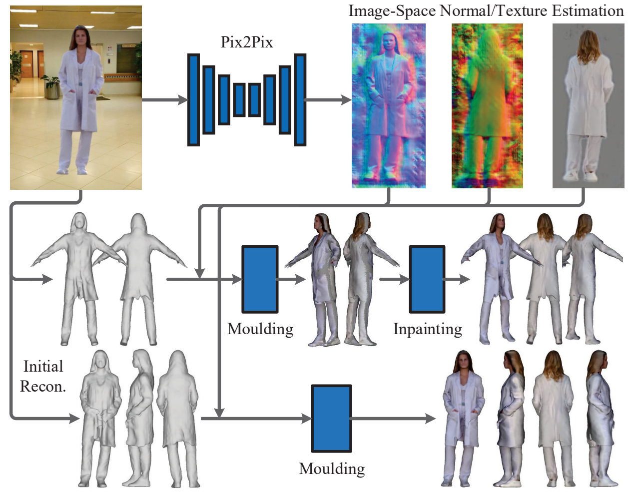

Figure 3. *Overview of the mesh refinement steps*. Our approach refines the initially estimated joint-space meshes from Fig. 2 using estimated normals and textures.

3.2. Mesh Refinement

We further refine the reconstructed meshes in joint spaces by adding geometric surface details and photorealistic textures. As illustrated in Fig. 3, we propose a moulding-inpainting scheme to utilize the front and back side normals and textures estimated in the image space. This is based on the observation that direct learning and inference of dense normal/color fields using deep implicit functions as [26] usually leads to over-smooth blur patterns and block artifacts (see Fig. 5). In contrast, image space estimation of normal and texture maps produces sharp results with fine-scale details, and is robust to human pose and shape changes. These benefits are from well-designed 2D convolutional deep networks (*e.g.*, Pix2Pix [27, 63]) and advanced (adversarial) image generation training schemes like GAN, with perceptual losses. The image-space estimated normal (and texture) maps could be used in two different ways. They can be used as either direct inputs into the Stack Hourglass as additional channels of the single-view image, or moulding-based front and back side mesh refinement sampling sources. In the experiments, we conduct ablation studies on these two schemes (*i.e.*, early direct input, late surface refinement) and demonstrate that our moulding-based refinement is better at maintaining fine-scale surface details across different views (see Fig. 8).

Posed Space. For the clothed human mesh obtained by Marching Cube in the original posed space, we conduct visibility tracing to determine if a vertex $V \in \mathbb{R}^3$ should be projected onto the front or the back side to bilinearly sample the normal/texture maps. Essentially, this is a moulding-based mesh refinement process for surface details and textures enhancement. We first conduct normal refinement. Note that for vertices whose unrefined normals $\mathbf{n} \in \mathbb{R}^3$ are near parallel (*i.e.*, within ε degrees) to the input image plane, we project them onto both the front and the back side normal maps $I_n^f, I_n^b \in \mathbb{R}^{512\times512\times3}$. We could then compute the refined surface normals $\mathbf{n}' \in \mathbb{R}^3$ via a linear blend fusion:

$$\begin{aligned} \mathbf{n}' &= \chi(1-\alpha')\,\mathcal{B}(I_n^f, \pi(V)) + \chi(\alpha')\,\mathcal{B}(I_n^b, \pi(V)), \\ \alpha' &= (90^\circ + \varepsilon - \alpha)/(2\varepsilon), \end{aligned} \tag{6}$$

where α is the angle between the unrefined normal and the forward camera raycast, and α' is the normalized value of α. Again, $\mathcal{B}(\cdot)$ indicates the bilinear sampling operation. The indicator function $\chi(\cdot)$ determines the blending weights of sampled normals from the front and the back sides:

$$\chi(\alpha') = \min(\max(\alpha', 0), 1) \tag{7}$$

This simple yet effective fusion scheme creates a normal-refined mesh with negligible blending boundary artifacts. With the refined surface normals we can further apply Poisson Surface Reconstruction [33] to update the mesh topology but in practice we find this unnecessary since the moulding-refined avatar can already satisfy various AR/VR and novel-view rendering applications. This bump rendering idea is also used in DeepHuman [75] but they only refine meshes using the front views. We further conduct the texture refinement in a similar manner but use the refined normals to help determine the linear blending weights of boundary vertices. Our moulding-based front/back normal and texture refinement method yields clothed human meshes that look photorealistic at different viewpoints with full-body surface details (*e.g.*, clothes wrinkles, hairs).

Canonical Space. The reconstructed canonical space avatar is rigged and thus can be warped back to its posed space and then refined via the same pipeline described above. However, a unique challenge for canonical avatar refinement is that mesh reconstructions in this space might contain unseen surfaces under the posed space. For example, in the third row of Fig. 5, the folded arm is in contact with the chest in the posed space but unfolded in the canonical space. Therefore, we do not have direct normal/texture correspondences of the chest regions of the canonical mesh. To address this problem, we render the front and the back side images of the canonical mesh with incomplete normal and texture, and treat it as an inpainting task. This problem has been well studied using deep neural networks [70, 71] and patch matching based methods [7, 6, 24]. We use PatchMatch [6] for its robustness. As demonstrated in the last two columns of Fig. 5, compared to directly regressing point-wise normal and texture, our inpainting-based results obtain sharper details and fewer artifacts.

4. Training Losses

The training process involves learning deep networks for two goals: joint-space occupancy estimation with $\mathcal{L}_o$, and normal/texture estimation with $\mathcal{L}_n$ and $\mathcal{L}_t$. Specifically, $\mathcal{L}_o$ is the occupancy regression loss of our joint-space deep implicit functions, and $\mathcal{L}_n, \mathcal{L}_t$ are image translation losses of the normal, texture estimation networks.

Joint-space Occupancy Estimation. The deep implicit function training is based on query point sampling and supervised occupancy regression with Tanh output layers. We randomly sample mesh points p_a, p_b in two spaces and then add diagonal Gaussian perturbation with a standard deviation of $5\,\mathrm{cm}$ to increase the sample coverage of close-to-surface regions in space. In each training iteration we sample 20480 pairs of query points (p_a, p_b), with predicted occupancy (o_a, o_b). The joint-space occupancy regression loss contains three terms:

$$\mathcal{L}_o(o_a, o_b) = \mathcal{L}_o^{occ}(o_a) + \mathcal{L}_o^{occ}(o_b) + \mathcal{L}_o^{con}(o_a, o_b), \tag{8}$$

where $\mathcal{L}_o^{occ}(o_a), \mathcal{L}_o^{occ}(o_b)$ denote the Smooth $L1$-Loss between the estimated occupancy values and their ground truth in the canonical and the posed spaces, respectively. $\mathcal{L}_o^{con}(o_a, o_b)$ is a contrastive loss regularizing the occupancy consistency between the two spaces, that is,

$$\mathcal{L}_o^{con}(o_a, o_b) = \begin{cases} |o_a - o_b|, & \text{if } \hat{o}_a = \hat{o}_b, \\ \lambda_1 \max(\lambda_2 - |o_a - o_b|, 0), & \text{otherwise,} \end{cases} \tag{9}$$

where λ_1 and λ_2 are two parameters to adjust the penalty of inconsistent joint-space groundtruth pairs. Those pairs usually exist around the self-intersecting regions and need to be down-weighted due to the errors in canonical space supervision. Empirically, we set $\lambda_1 = 0.1$ and $\lambda_2 = 0.3$.

Mesh Refinement. We consider the image-space normal and texture estimation as an image-to-image translation task. Given an input image I, our task is to learn the front normal map I_n^f, the back normal map I_n^b and the back side texture map $I_t^b \in \mathbb{R}^{512\times512\times3}$. Note we assume the input image can be directly used as the front texture map. Inspired by the demonstrated superior results of Pix2Pix [27, 63], we define the training losses as:

$$\begin{aligned} \mathcal{L}_n(I_n^f, I_n^b) &= \mathcal{L}_n^{rec}(I_n^f) + \mathcal{L}_n^{rec}(I_n^b) + \mathcal{L}_n^{vgg}(I_n^f) + \mathcal{L}_n^{vgg}(I_n^b), \\ \mathcal{L}_t(I_t^b) &= \mathcal{L}_t^{rec}(I_t^b) + \mathcal{L}_t^{vgg}(I_t^b) + \mathcal{L}_t^{adv}(I_t^b), \end{aligned} \tag{10}$$

where $\mathcal{L}^{rec}(\cdot)$ denotes the $L1$ distance reconstruction loss, $\mathcal{L}^{adv}(\cdot)$ means the generative adversarial loss and $\mathcal{L}^{vgg}(\cdot)$ is the VGG-perceptual loss proposed by [30]. In the experiments, we found that the generative adversarial loss $\mathcal{L}^{adv}(\cdot)$ counteracts to performance in the normal map estimation task and thus we only enforce this loss term upon the back side texture map. One explanation is that the normal map space is more constrained and has fewer variations than

Components	Posed Space			Canonical Space			Mean		
	Normal ↓	P2S ↓	Chamfer ↓	Normal ↓	P2S ↓	Chamfer ↓	Normal ↓	P2S ↓	Chamfer ↓
Posed Sup. Only	0.037	0.674	0.787	0.087	1.898	1.597	0.062	1.286	1.192
Canonical Sup. Only	0.039	0.716	0.838	0.046	0.606	0.997	0.043	0.661	0.917
Joint	0.037	0.662	0.789	0.045	0.620	0.988	0.041	0.641	0.825
Joint + GeoEnc	0.033	**0.495**	**0.614**	0.040	**0.471**	**0.819**	0.036	**0.483**	**0.717**
Joint + GeoEnc + Refine	**0.031**	**0.495**	**0.614**	**0.039**	**0.471**	**0.819**	**0.035**	**0.483**	**0.717**

Table 1. *Ablation studies on the effectiveness of ARCH++ proposed components in both spaces: posed vs. canonical*. Best scores are in **bold**. Rows are target reconstruction spaces, columns are evaluation spaces. The first row means using the posed space as the target space (*e.g.*, PIFu, PIFuHD, Geo-PIFu, PaMIR), whose reconstruction can be warped into the canonical space via a registered parametric body to compute evaluation metrics in both spaces. The second row means direct supervision and reconstruction in the canonical space, followed by warping into the posed space (*e.g.*, ARCH). The rest rows are based on our joint-space co-supervision and reconstruction scheme.

Methods	RenderPeople			BUFF		
	Normal ↓	P2S ↓	Chamfer ↓	Normal ↓	P2S ↓	Chamfer ↓
BodyNet [60]	0.26	5.72	5.64	0.31	4.94	4.52
VRN [28]	0.12	1.42	1.60	0.13	2.33	2.48
SiCloPe [46]	0.22	3.81	4.02	0.22	4.06	3.99
IM-GAN [12]	0.26	2.87	3.14	0.34	5.11	5.32
PIFu [55]	0.11	1.45	1.47	0.13	1.68	1.76
PIFuHD [56]	0.11	1.37	1.43	0.13	1.63	1.75
ARCH [26]	0.04	0.74	0.85	0.04	0.82	0.87
ARCH++ [Ours]	**0.03**	**0.50**	**0.61**	**0.03**	**0.58**	**0.61**

Table 2. *Quantitative results and comparisons of normal, P2S and Chamfer errors between posed reconstruction and ground truth on RenderPeople and BUFF datasets*. Best scores are in **bold**.

the texture map, and therefore adversarial training does not fully show its effectiveness in this case.

5. Experiments

In this section, we present the experimental settings, result comparisons and ablation studies of ARCH++.

5.1. Implementation Details

We implement our framework using PyTorch and conduct the training with one NVIDIA Tesla V100 GPU. The proposed deep neural networks are trained with RMSprop optimizer with a learning rate starting from 1e-4. We use an exponential learning rate scheduler to update it every 3 epochs by multiplying with the factor 0.1 and terminate the training after 12 epochs.

5.2. Datasets

We adopt the dataset setting from [26, 56]. Our training dataset consists of 450 3D scans from RenderPeople dataset [54]. These watertight human meshes have various clothes styles as well as body shapes and poses. Our testing set includes 37 scans from RenderPeople dataset [54], 32 scans from AXYZ dataset [15], 26 scans from BUFF dataset [73], and 2D images from Internet public domains, representing clothed people with a large variety of complex clothes. The subjects in the training dataset are mostly in standing pose, while the subjects in the test dataset contain various poses including sitting, twisted and standing, as well as self-glued and separated limbs. We use Blender and 38 environment maps to render each scan under different natural lighting conditions. For each 3D scan, we generate 360 images by rotating a camera around the mesh with a step size of 1 degree. These RenderPeople images are used

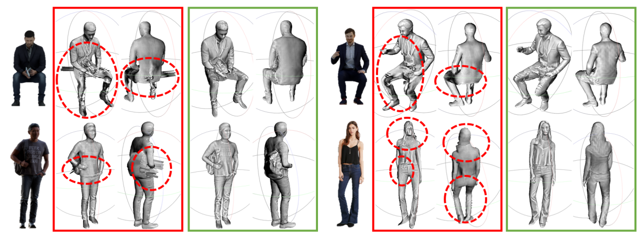

Figure 4. We intentionally hide the method names for you to have a fair comparison on your own (*please zoom in*). The answers are[2].

to train both the occupancy estimation and the image translation networks.

We generate ground truth clothed human meshes in the canonical pose using the method introduced in [26]. Note that the warping process between the posed and the canonical spaces inevitably contain model noises (*e.g.*, self-contact region artifacts, skinning weights nearest neighbor discontinuities), which motivates our joint-space co-supervision and reconstruction scheme.

5.3. Results and Comparisons

We use the same metrics as [55, 56, 26] for quantitative evaluation of the reconstructed meshes. We report the average point-to-surface Euclidean distance (P2S) and the Chamfer distance in centimeters, as well as the $L2$ normal re-projection errors. The two state of the art methods for our main comparisons are PIFuHD [56] and ARCH [26], both are built upon PIFu [55] with improvements in different aspects. PIFuHD ingests high-resolution images in a sliding window manner to achieve rich surface reconstruction details. ARCH leverages nearest neighbor-based linear blend skinning weights and hand-crafted RBF features to reconstruct animatable avatars in a canonical space. In addition to these two most related methods, we also include multiple prior methods [60, 28, 46, 12, 55] and report the benchmark results on the RenderPeople and the BUFF datasets in Tab. 2. ARCH++ [Ours] results outperform the second best method ARCH by large gaps.

The visual comparisons in Fig. 5 and Fig. 4 further explain the advantages of our improvement. PIFuHD suffers from shape distortions due to lacking shape and pose priors provided by the end-to-end geometry encoder. Note that

[2]Leveraging the semantic-aware geometric encoder, our results (green boxes) have fewer shape and pose artifacts than ARCH (red boxes).

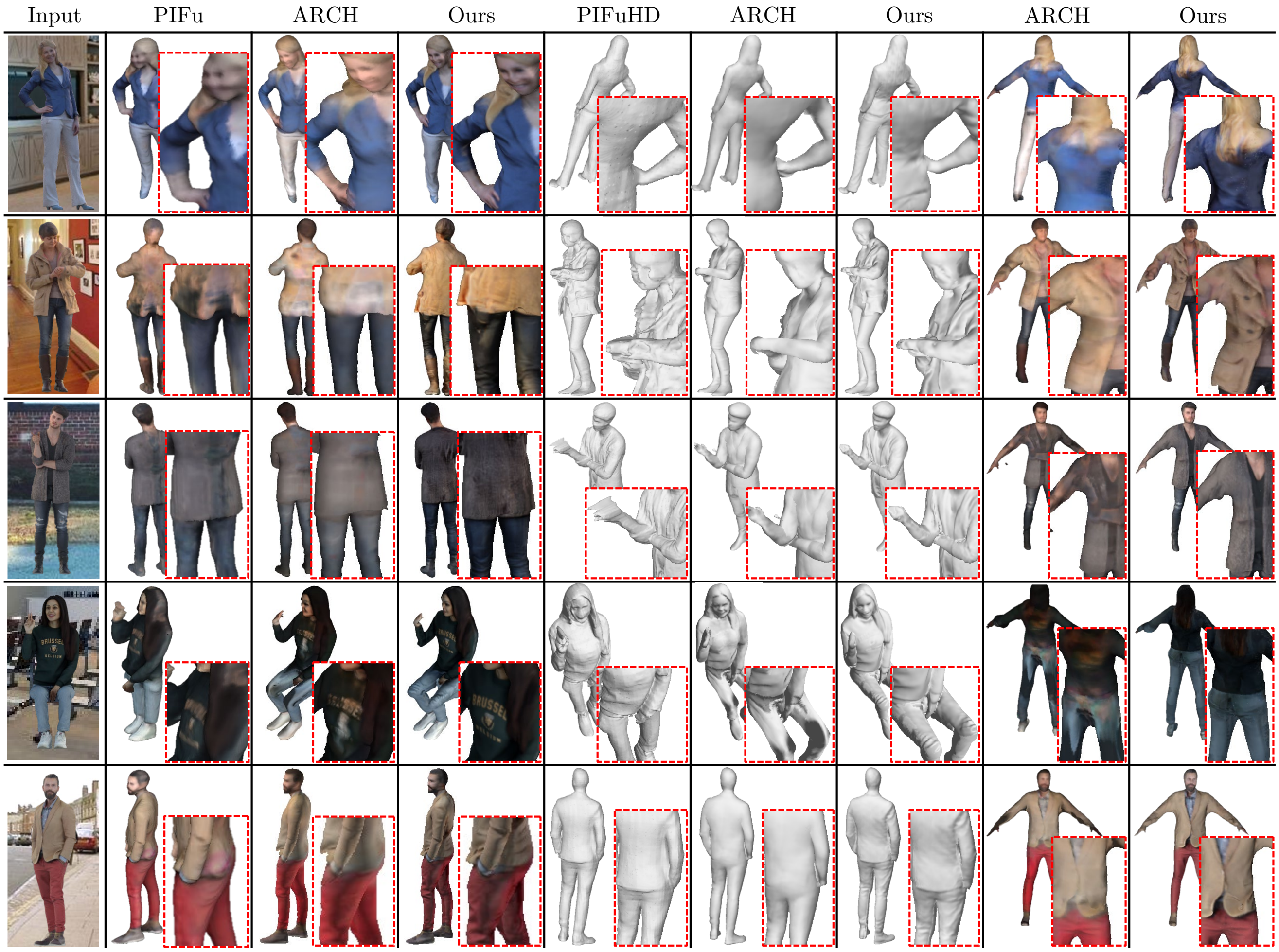

Figure 5. *Qualitative comparisons against the state-of-the-art methods [55, 56, 26].* The first column is input. Column 2-4, 5-7 are color and shape reconstruction results, respectively, in the posed space. The last two columns are canonical space avatar reconstructions. Our method handles arbitrary poses with self-contact and occlusions robustly, and reconstructs a higher level of details than existing methods.

Variants	Normal ↓	P2S ↓	Chamfer ↓
Depth [55]	0.047	0.78	0.93
RBF [26]	0.042	0.74	0.85
End-to-end Voxel [23, 74]	0.034	0.52	0.63
End-to-end Point	**0.033**	**0.50**	**0.61**

Table 3. *Ablation studies on different types of geometry encoders.*

Variants	Posed ↓	Canonical ↓	Mean ↓
Baseline	0.033	0.040	0.037
Object-space Regression [26]	0.032	0.041	0.037
Image-space Input [56]	0.032	**0.038**	**0.035**
Image-space Regression	**0.031**	0.039	**0.035**

Table 4. *Ablation studies on different ways of normal refinement.*

PIFuHD is incapable of reconstruct canonical space avatars and lacks texture estimation. ARCH reconstructions tend to be over smooth and blurry. Its recovered mesh normal and texture also have several block artifacts. Additionally, both methods fail to hallucinate plausible back-side surface details like clothes wrinkles, hairs, *etc*. In comparison, our approach achieves photorealistic and animatable reconstructions in joint spaces and across different viewpoints. We further show our results on Internet images in Fig. 9.

5.4. Ablation Studies

Joint Space Reconstruction. To further understand the impact of the proposed methods, we present ablation studies in Tab. 1. The first three rows demonstrate the effectiveness of joint-space co-supervision, achieving balanced performances on both the posed and the canonical space mesh reconstructions. Choosing the posed space as the reconstruction target space (*e.g.*, PIFu, PIFuHD, Geo-PIFu, PaMIR) can cause missing surfaces and topology distortions in the posed-to-canonical space warped meshes (see Fig. 6). Meanwhile, choosing the canonical space as the target space (*e.g.*, ARCH) can cause self-intersecting meshes with broken manifold as well as body part un-natural deformations in the canonical-to-posed space warped meshes. In contrast, our co-supervision and joint-space inference methods achieve both reconstruction fidelity in the posed space and body mesh completeness in the canonical space.

Geometry Encoding. As shown in Tab. 1, we observe further error reduction leveraging the end-to-end learned point-wise spatial encodings. The prior method ARCH uses

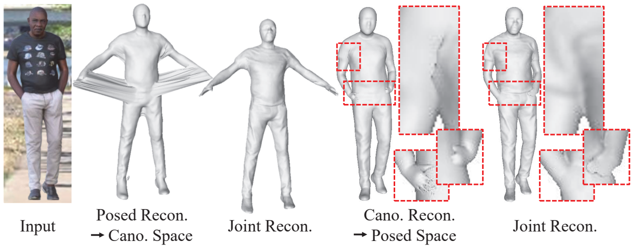

Figure 6. *Ablation studies on the reconstruction space*. Single-space reconstruction shows artifacts of either mesh surface over-stretching or intersecting surfaces when warping from one space to another. Our joint-space reconstruction obtains balanced performance of both high reconstruction completeness under the canonical space and high input image fidelity under the posed space.

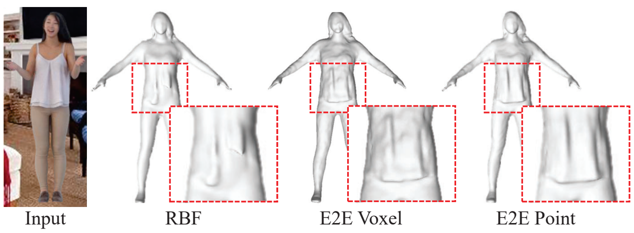

Figure 7. *Ablation studies on geometry encoding*. Learned spatial features capture both pose and shape priors of the underlying parametric models and thus enable mesh reconstruction with more surface details than the handcrafted RBF features. Meanwhile, results of the voxel-based features are noisier than the point-based ones due to mesh quantization (*i.e.*, voxelization) errors.

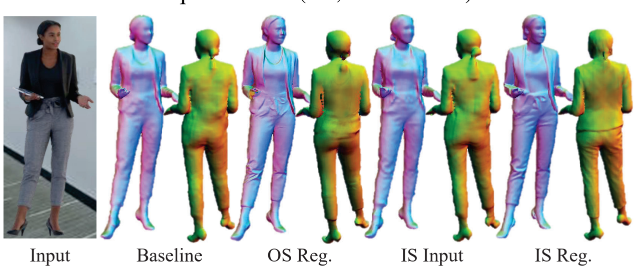

Figure 8. *Ablation studies on normal refinement*: Object-space Regression (OS Reg.), Image-space Input (IS Input) and Image-space Regression (IS Reg.). Our method IS Reg. leads to rich reconstruction details (*e.g.*, clothes wrinkles) in all views.

handcrafted RBF features that only model the pose prior of parametric body mesh skeletons, ignoring the mesh shape. In comparison, our point-based features are informed of both pose and shape priors of the underlying parametric body model w.r.t. a clothed human mesh, and thus improve the surface reconstruction quality. We further implement the learned volumetric spatial feature encodings used in Geo-PIFu and PaMIR as an alternative encoder and inject into our framework for direct comparisons. The results are shown in Tab. 3 and Fig. 7. While both types of end-to-end spatial features outperform the hand crafted RBF features, our point-based feature extraction method does not suffer from computation overhead and mesh quantization errors of the voxel-based approach.

Normal Refinement. While single-image based direct inference of human meshes with rich surface details at both the front and the back side remains an open question, some empirical observations and prior works indicate that normal estimation is a relatively easier task and can help refine the reconstructions. In Tab. 4 and Fig. 8 we experiment on three principle ways of leveraging the estimated normals for mesh reconstructions with refined surface details. Among these normal refinement methods, our front/back-side image space normal regression and moulding-based surface refinement approach outperforms other variants. Object-space normal regression is adopted in ARCH and is based on learning deep implicit functions of spatial normal fields. It fails to generate rich back side details and sometimes causes block artifacts as shown in the fourth row of Fig. 5. Image-space input is used in PIFuHD. It concatenates the color image input with estimated image-space normal maps and feeds them into Stack Hourglass for feature extraction. While this method achieves the same level of quantitative performance as our mesh refinement approach, its visual results are not as sharp as ours at both the front and the back sides. A degenerated case of our mesh refinement method is studied before in DeepHuman where they only estimate front-view normal maps and therefore lack reconstruction details at the back side.

Figure 9. *An application of digital human capture from photos*.

6. Conclusion

In this paper, we revisit the major components in existing deep implicit function based 3D avatar reconstruction. Our method ARCH++ produces results which have high-level fidelity and are animation-ready for many AR/VR applications. We conduct a series of comparisons with and analysis on the state of the art to validate our findings. For future works, we plan to incorporate environment information (e.g., lighting, affordance) to further understand the body pose and appearance, and address current limitations.

Acknowledgements. We would like to thank Minh Vo and Nikolaos Sarafianos for the discussions and synthetic data creation.

References

[1] 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. 2

[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, 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. 1, 2

[4] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. In *ACM Transactions on Graphics*, 2005. 2

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

[6] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 5

[7] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *ACM SIGGRAPH*, 2000. 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. 2

[9] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision*, 2019. 1

[10] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision*, 2019. 2

[11] 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

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 6

[13] 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. 2, 3

[14] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):1–13, 2015. 1

[15] AXYZ design S.R.L. *https://secure.axyz-design.com/*. 6

[16] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 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, 3

[18] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[19] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conference on Computer Vision*, 2018. 2

[20] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[21] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics*, 38(6):1–19, 2019. 1

[22] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[23] 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. 1, 2, 3, 4, 7

[24] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014. 5

[25] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *European Conference on Computer Vision*, 2018. 2

[26] 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. 1, 2, 3, 4, 6, 7

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 4, 5

[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, 6

[29] 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*. Springer, 2020. 2

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[31] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020. 1

[32] 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. 1, 2

[33] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3), 2013. 5

[34] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *IEEE International Conference on Computer Vision*, 2019. 1, 2

[35] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[36] Zorah Laehner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *European Conference on Computer Vision*, 2018. 2

[37] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[38] 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

[39] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics*, 37(4):1–13, 2018. 1

[40] 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

[41] William E. Lorensen and Harvey E. Cline. Differentiable monte carlo ray tracing through edge sampling. *ACM SIGGRAPH Computer Graphics*, 21(4):163–169, 1987. 4

[42] Takashi Matsuyama, Shohei Nobuhara, Takeshi Takai, and Tony Tung. *3D Video and Its Applications*. Springer, 2012. 2

[43] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Leonard McMillan, and Steven Gortler. Image-based visual hulls. In *ACM SIGGRAPH*, 2000. 2

[44] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Learning 3d reconstruction in function space. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[45] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012. 1

[46] 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, 6

[47] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016. 3

[48] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Annual Symposium on User Interface Software and Technology*, 2016. 1

[49] Mathias Parger, Chengcheng Tang, Yuanlu Xu, Christopher Twigg, Lingling Tao, Yijing Li, Robert Wang, and Markus Steinberger. UNOC: Understanding occlusion for embodied presence in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 1

[50] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

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

[52] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[53] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Annual Conference on Neural Information Processing Systems*, 2017. 2, 3

[54] RENDERPEOPLE. *http://renderpeople.com/*. 6

[55] 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. 1, 2, 3, 4, 6, 7

[56] 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, 3, 4, 6, 7

[57] 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

[58] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. Simultaneous super-resolution and 3d video using graph-cuts. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2

[59] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *IEEE International Conference on Computer Vision*, 2009. 2

[60] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Vol-

umetric inference of 3D human body shapes. In *European Conference on Computer Vision*, 2018. 2, 6

[61] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):1–9, 2008. 2

[62] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popovic, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH*, 2009. 2

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 5

[64] Chenglei Wu, Kiran Varanasi, and Christian Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *European Conference on Computer Vision*, 2012. 2

[65] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[66] Yuanlu Xu, Wenguan Wang, Tengyu Liu, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. Monocular 3d pose estimation via pose grammar and data augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[67] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *IEEE International Conference on Computer Vision*, 2019. 1, 2, 3

[68] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *European Conference on Computer Vision*, 2018. 2

[69] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, 2016. 2

[70] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 5

[71] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[72] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, 2020. 1

[73] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6

[74] 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*, 2021. 2, 3, 4, 7

[75] 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. 1, 2, 5