PINA: Learning a Personalized Implicit Neural Avatar from a Single RGB-D Video Sequence

Zijian Dong*[1] Chen Guo*[1] Jie Song[1] Xu Chen[1,2] Andreas Geiger[2,3] Otmar Hilliges[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen
[3]University of Tübingen

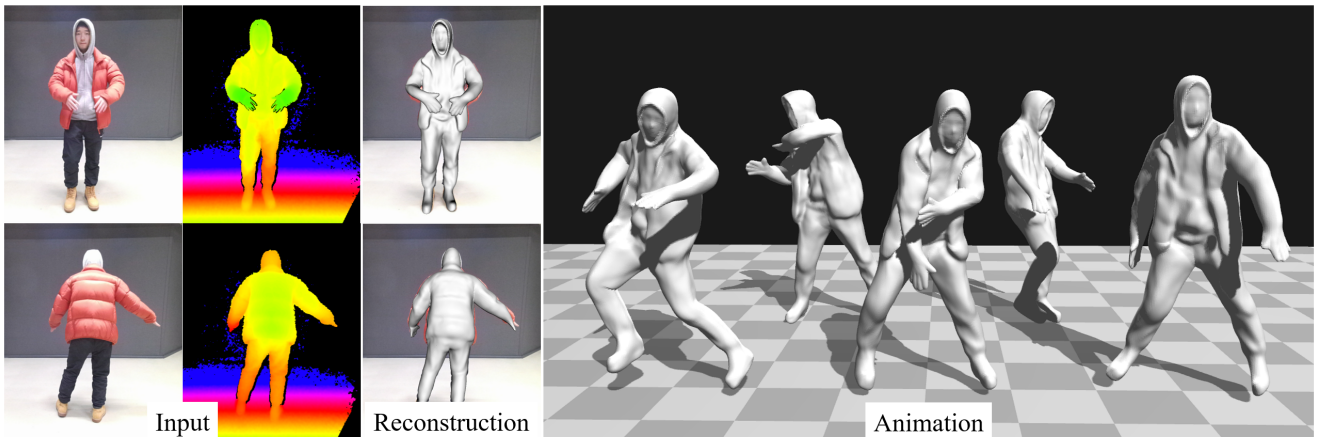

Figure 1. We propose PINA, a method to acquire personalized and animatable neural avatars from RGB-D videos. *Left*: our method uses only a single sequence, captured via a commodity depth sensor. The depth frames are noisy and contain only partial views of the body. *Middle*: Using global optimization, we fuse these partial observations into an implicit surface representation that captures geometric details, such as loose clothing. The shape is learned alongside a pose-conditioned skinning field, supervised only via depth observations. *Right*: The learned avatar can be animated with realistic articulation-driven surface deformations and generalizes to novel unseen poses.

Abstract

*We present a novel method to learn **P**ersonalized **I**mplicit **N**eural **A**vatars (PINA) from a short RGB-D sequence. This allows non-expert users to create a detailed and personalized virtual copy of themselves, which can be animated with realistic clothing deformations. PINA does not require complete scans, nor does it require a prior learned from large datasets of clothed humans. Learning a complete avatar in this setting is challenging, since only few depth observations are available, which are noisy and incomplete (i.e. only partial visibility of the body per frame). We propose a method to learn the shape and non-rigid deformations via a pose-conditioned implicit surface and a deformation field, defined in canonical space. This allows us to fuse all partial observations into a single consistent canonical representation. Fusion is formulated as a global optimization problem over the pose, shape and skinning parameters. The method can learn neural avatars from real noisy RGB-D sequences for a diverse set of people and clothing styles and these avatars can be animated given unseen motion sequences.*

*equal contribution

1. Introduction

Making immersive AR/VR a reality requires methods to effortlessly create personalized avatars. Consider telepresence as an example: the remote participant requires means to simply create a detailed scan of themselves and the system must then be able to re-target the avatar in a realistic fashion to a new environment and to new poses. Such applications impose several challenging constraints: i) to generalize to unseen users and clothing, no specific prior knowledge such as template meshes should be required ii) the acquired 3D surface must be animatable with realistic surface deformations driven by complex body poses iii) the capture setup must be unobtrusive, ideally consisting of a single consumer-grade sensor (e.g. Kinect) iv) the process must be automatic and may not require technical expertise, rendering traditional skinning and animation pipelines unsuitable. To address these requirements, we introduce a novel method for learning Personalized Implicit Neural Avatars

(PINA) from only a sequence of monocular RGB-D video.

Existing methods do not fully meet these criteria. Most state-of-the-art dynamic human models [6, 9, 30, 31] represent humans as a parametric mesh and deform it via linear blend skinning (LBS) and pose correctives. Sometimes learned displacement maps to capture details of tight-fitting clothing are used [9]. However, the fixed topology and resolution of meshes limit the type of clothing and dynamics that can be captured. To address this, several methods [12, 48] propose to learn neural implicit functions to model static clothed humans. Furthermore, several methods that learn a neural avatar for a specific outfit from watertight meshes [10,14,21,29,44,50,52] have been proposed. These methods either require complete full-body scans with accurate surface normals and registered poses [10, 14, 50, 52] or rely on complex and intrusive multi-view setups [21,29,44].

Learning an animatable avatar from a monocular RGB-D sequence is challenging since raw depth images are noisy and only contain partial views of the body (Fig. 1, left). At the core of our method lies the idea to fuse partial depth maps into a single, consistent representation and to learn the articulation-driven deformations at the same time. To do so, we formulate an implicit signed distance field (SDF) in canonical space. To learn from posed observations, the inverse mapping from deformed to canonical space is required. We follow SNARF [10] and locate the canonical correspondences via optimization. A key challenge brought on by the monocular RGB-D setting is to learn from incomplete point clouds. Inspired by *rigid* learned SDFs for objects [17], we propose a point-based supervision scheme that enables learning of articulated *non-rigid* shapes (*i.e.* clothed humans). Transforming the spatial gradient of the SDF into posed space and comparing it to surface normals from depth images leads to the learning of fine geometric details. Training is formulated as a global optimization that jointly optimizes the canonical SDF, the skinning fields and the per-frame pose. PINA learns animatable avatars without requiring any additional supervision or priors extracted from large datasets of clothed humans.

In detailed ablations, we shed light on the key components of our method. We compare to existing methods in the reconstruction and animation tasks, showing that our method performs best across several datasets and settings. Finally, we demonstrate the ability to capture and animate different humans in a variety of clothing styles qualitatively.

In summary, our contributions are:

- a method to fuse partial RGB-D observations into a canonical, implicit representation of 3D humans; and
- to learn an animatable SDF representation directly from partial point clouds and normals; and
- a formulation which jointly optimizes shape, per-frame pose and skinning weights.

Code and data will be made available for research purposes.

2. Related Work

Parametric Models for Clothed Humans A large body of literature utilizes explicit surface representations (particularly polygonal meshes) for human body modeling [5, 22, 25, 39, 43, 55]. These works typically leverage parametric models for minimally clothed human bodies [7, 15, 16, 26, 43, 51] (e.g. SMPL [30]) and use a displacement layer on top of the minimally clothed body to model clothing [3, 4, 32, 36, 56]. Recently, DSFN [9] proposes to embed MLPs into the canonical space of SMPL to model pose-dependent deformations. However, such methods depend upon SMPL learned skinning for deformation and are upper-bounded by the expressiveness of the template mesh. During animation or reposing, the surface deformations of parametric human models either solely rely on the skinning weights trained from minimally clothed body scans [30,32,59] or are learned from synthetically simulated data [13, 18, 20, 42]. Methods that drape garments onto the SMPL body suffer from artifacts during reposing as they typically rely on skinning weights of a naked body for animation which may be incorrect for points on the surface of the garment. In contrast, our method represents clothed humans as a flexible implicit neural surface and jointly learns shape and a neural skinning field from depth observations.

Implicit Human Models from 3D Scans Implicit neural representations [11, 34, 40] can handle topological changes better [8, 41] and have been used to reconstruct clothed human shapes [23, 24, 27, 45, 46, 48, 49]. Typically, based on a learned prior from large-scale datasets, they recover the geometry of clothed humans from images [48, 49, 60] or point clouds [12]. However, these reconstructions are static and cannot be reposed. Follow-up work [6, 24] attempts to endow static reconstructions with human motion based on a generic deformation field which tends to output unrealistic animated results. To model pose-dependent clothing deformations, SCANimate [50] proposes to transform scans to canonical space in a weakly supervised manner and to learn the implicit shape model conditioned on joint-angle rotations. Follow-up works further improve the generalization ability to unseen poses and accelerate the training process via a displacement network [52], deform the shape via a forward warping field [10] or leverage prior information from large-scale human datasets [54]. However, all of these methods require complete and registered 3D human scans for training, even if they sometimes can be fine-tuned on RGB-D data. In contrast, PINA is able to learn a personalized implicit neural avatar directly from a short monocular RGB-D sequence without requiring large-scale datasets of clothed human 3D scans or other priors.

Reconstructing Clothed Humans from RGB-D Data One straightforward approach to acquiring a 3D human model from RGB-D data is via per-frame reconstruction [6,

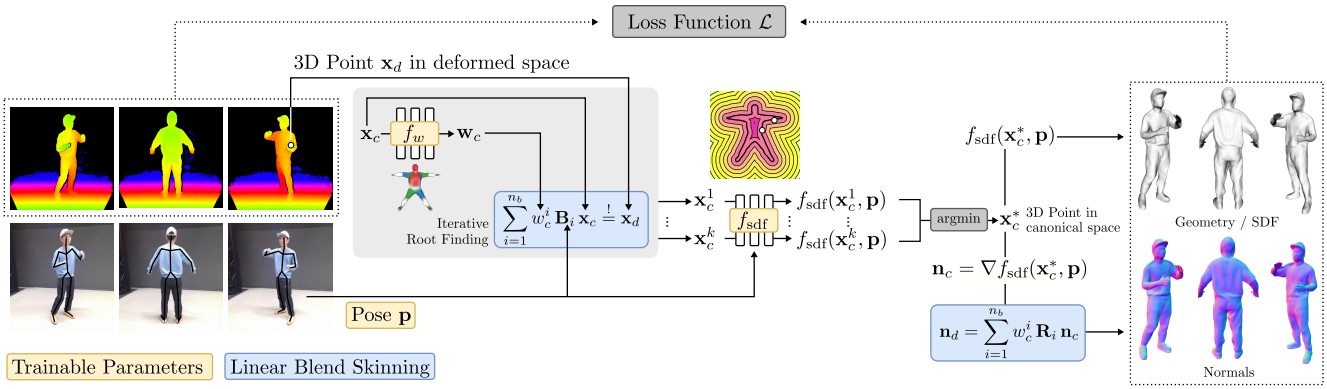

Figure 2. **Method Overview.** Given input depth frames and human pose initializations inferred from RGB-D images, we first sample 3D points $\mathbf{x}_d$ on and off the human body surface in deformed (posed) space. Their corresponding canonical location $\mathbf{x}_c^*$ is calculated via iterative root finding [10] of the linear blend skinning constraint (here $\stackrel{!}{=}$ denotes that we seek the top-k roots $\mathbf{x}_c^{1:k}$, indicating the possible correspondences) and minimizing the SDF over these k roots. Given the canonical location $\mathbf{x}_c^*$, we evaluate the SDF of $\mathbf{x}_c^*$ in canonical space, obtain its normal as the spatial gradient of the signed distance field and map it into deformed space using learned linear blend skinning. We minimize the loss $\mathcal{L}$ that compares these predictions with the input observations. Our loss regularizes off-surface points using proxy geometry and uses an Eikonal loss to learn a valid signed distance field f_{sdf}.

12]. To achieve this, IF-Net [12] learns a prior to reconstruct an implicit function of a human and IP-Net [6] extends this idea to fit SMPL to this implicit surface for articulation. However, since input depth observations are partial and noisy, artifacts appear in unseen regions. Real-time performance capture methods incrementally fuse observations into a volumetric SDF grid. DynamicFusion [37] extends earlier approaches for static scene reconstruction [38] to non-rigid objects. BodyFusion [57] and DoubleFusion [58] build upon this concept by incorporating an articulated motion prior and a parametric body shape prior. Follow-up work [8, 9, 28] leverages a neural network to model the deformation or to refine the shape reconstruction. However, it is important to note that such methods only reconstruct the surface, and sometimes the pose, for tracking purposes but typically do not allow for the acquisition of skinning information which is crucial for animation. In contrast, our focus differs in that we aim to acquire a detailed avatar including its surface and skinning field for reposing and animation.

3. Method

We introduce PINA, a method for learning personalized neural avatars from a single RGB-D video, illustrated in Fig. 2. At the core of our method lies the idea to fuse partial depth maps into a single, consistent representation of the 3D human shape and to learn the articulation-driven deformations at the same time via global optimization.

We parametrize the 3D surface of clothed humans as a pose-conditioned implicit signed-distance field (SDF) and a learned deformation field in canonical space (Sec. 3.1). This parametrization enables the fusion of partial and noisy depth observations. This is achieved by transforming the canonical surface points and the spatial gradient into posed space, enabling supervision via the input point cloud and its normals. Training is formulated as global optimization (Sec. 3.2) to jointly optimize the per-frame pose, shape and skinning fields without requiring prior knowledge extracted from large datasets. Finally, the learned skinning field can be used to articulate the avatar (Sec. 3.3).

3.1. Implicit Neural Avatar

Canonical Representation We model the human avatar in canonical space and use a neural network f_{sdf} to predict the signed distance value for any 3D point $\mathbf{x}_c$ in this space. To model pose-dependent local non-rigid deformations such as wrinkles on clothes, we concatenate the human pose $\mathbf{p}$ as additional input and model f_{sdf} as:

$$f_{\text{sdf}} : \mathbb{R}^3 \times \mathbb{R}^{n_p} \rightarrow \mathbb{R}. \tag{1}$$

The pose parameters ($\mathbf{p}$) are defined consistently to the SMPL skeleton [30] and n_p is their dimensionality. The canonical shape $\mathcal{S}$ is then given by the zero-level set of f_{sdf}:

$$\mathcal{S} = \{\, \mathbf{x}_c \mid f_{\text{sdf}}(\mathbf{x}_c, \mathbf{p}) = 0 \,\} \tag{2}$$

In addition to signed distances, we also compute normals in canonical space. We empirically find that this resolves high-frequency details better than calculating the normals in the posed space. The normal for a point $\mathbf{x}_c$ in canonical space is computed as the spatial gradient of the signed distance function at that point (attained via backpropagation):

$$\mathbf{n}_c = \nabla_{\mathbf{x}} f_{\text{sdf}}(\mathbf{x}_c, \mathbf{p}). \tag{3}$$

To deform the canonical shape into novel body poses we additionally model deformation fields. To animate implicit

human shapes in the desired body pose $\mathbf{p}$, we leverage linear blend skinning (LBS). The skeletal deformation of each point in canonical space is modeled as the weighted average of a set of bone transformations $\mathbf{B}$, which are derived from the body pose $\mathbf{p}$. We follow [10] and define the skinning field in canonical space using a neural network f_w to model the continuous LBS weight field:

$$f_w : \mathbb{R}^3 \rightarrow \mathbb{R}^{n_b}. \tag{4}$$

Here, n_b denotes the number of joints in the transformation and $\mathbf{w}_c = \{w_c^1, ..., w_c^{n_b}\} = f_w(\mathbf{x}_c)$ represents the learned skinning weights for $\mathbf{x}_c$.

Skeletal Deformation Given the bone transformation matrix $\mathbf{B}_i$ for joint $i \in \{1, ..., n_b\}$, a canonical point $\mathbf{x_c}$ is mapped to the deformed point $\mathbf{x}_d$ as follows:

$$\mathbf{x}_d = \sum_{i=1}^{n_b} w_c^i \mathbf{B}_i \, \mathbf{x}_c \tag{5}$$

The normal of the deformed point $\mathbf{x_d}$ in posed space is calculated analogously:

$$\mathbf{n}_d = \sum_{i=1}^{n_b} w_c^i \, \mathbf{R}_i \, \mathbf{n}_c \tag{6}$$

where $\mathbf{R}_i$ is the rotation component of $\mathbf{B}_i$.

To compute the signed distance field $SDF(\mathbf{x}_d)$ in deformed space, we need the canonical correspondences $\mathbf{x}_c^*$.

Correspondence Search For a deformed point $\mathbf{x}_d$, we follow [10] and compute its canonical correspondence set $\mathcal{X}_c = \{\mathbf{x}_c^1, ..., \mathbf{x}_c^k\}$, which contains k canonical candidates satisfying Eq. 5, via an iterative root finding algorithm. Here, k is an empirically defined hyper-parameter of the root finding algorithm (see Supp. Mat for more details).

Note that due to topological changes, there exist one-to-many mappings when retrieving canonical points from a deformed point, *i.e.*, the same point $\mathbf{x}_d$ may correspond to multiple different valid $\mathbf{x}_c$. Following Ricci et al. [47] , we composite these proposals of implicitly defined surfaces into a single SDF via the union (minimum) operation:

$$SDF(\mathbf{x}_d) = \min_{\mathbf{x}_c \in \mathcal{X}_c} f_{\text{sdf}}(\mathbf{x}_c) \tag{7}$$

The canonical correspondence $\mathbf{x}_c^*$ is then given by:

$$\mathbf{x}_c^* = \arg\min_{\mathbf{x}_c \in \mathcal{X}_c} f_{\text{sdf}}(\mathbf{x}_c) \tag{8}$$

3.2. Training Process

Defining our personalized implicit model in canonical space is crucial to integrating partial observations across all depth frames because it provides a common reference frame. Here, we formally describe this fusion process. We train our model jointly w.r.t. body poses and the weights of the 3D shape and skinning networks.

Objective Function Given an RGB-D sequence with N input frames, we minimize the following objective:

$$\mathcal{L}(\mathbf{\Theta}) = \sum_{i=1}^{N} \mathcal{L}_{\text{on}}^i(\mathbf{\Theta}) + \lambda_{\text{off}} \mathcal{L}_{\text{off}}^i(\mathbf{\Theta}) + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}}^i(\mathbf{\Theta}) \tag{9}$$

$\mathcal{L}_{\text{on}}^i$ represents an on-surface loss defined on the human surfaces for frame i. $\mathcal{L}_{\text{off}}^i$ represents the off-surface loss which helps to carve free-space and $\mathcal{L}_{\text{eik}}^i$ is the Eikonal regularizer which ensures a valid signed distance field. $\mathbf{\Theta}$ is the set of optimized parameters which includes the shape network weights $\mathbf{\Theta}_{\text{sdf}}$, the skinning network weights $\mathbf{\Theta}_w$ and the pose parameters $\mathbf{p}_i$ for each frame.

To calculate $\mathcal{L}_{\text{on}}^i$, we first back-project the depth image into 3D space to obtain partial point clouds $\mathcal{P}_{\text{on}}^i$ of human surfaces for each frame i. For each point $\mathbf{x}_d$ in $\mathcal{P}_{\text{on}}^i$, we additionally calculate its corresponding normal $\mathbf{n}_d^{\text{obs}}$ from the raw point cloud using principal component analysis of points in a local neighborhood. $\mathcal{L}_{\text{on}}^i$ is then defined as

$$\begin{aligned} \mathcal{L}_{\text{on}}^i &= \lambda_{\text{sdf}} \, \mathcal{L}_{\text{sdf}}^i + \lambda_n \, \mathcal{L}_{\text{n}}^i \\ &= \lambda_{\text{sdf}} \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}^i} |SDF(\mathbf{x}_d)| + \lambda_n \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{on}}^i} \|NC(\mathbf{x}_d)\| \end{aligned} \tag{10}$$

Here, $NC(\mathbf{x}_d) = \mathbf{n}_d^{\text{obs}}(\mathbf{x}_d) - \mathbf{n}_d(\mathbf{x}_d)$.

We add two additional terms to regularize the optimization process. $\mathcal{L}_{\text{off}}^i$ complements $\mathcal{L}_{\text{on}}^i$ by randomly sampling points $\mathcal{P}_{\text{off}}^i$ that are far away from the body surface. For any point $\mathbf{x}_d$ in $\mathcal{P}_{\text{off}}^i$, we calculate the signed distance between this point and an estimated body mesh (see initialization section below). This signed distance $SDF_{body}(\mathbf{x}_d)$ serves as pseudo ground truth to force plausible off-surface SDF values. $\mathcal{L}_{\text{off}}^i$ is then defined as:

$$\mathcal{L}_{\text{off}}^i = \sum_{\mathbf{x}_d \in \mathcal{P}_{\text{off}}^i} |SDF(\mathbf{x}_d) - SDF_{body}(\mathbf{x}_d)| \tag{11}$$

Following IGR [17], we leverage $\mathcal{L}_{\text{eik}}^i$ to force the shape network f_{sdf} to satisfy the Eikonal equation in canonical space:

$$\mathcal{L}_{\text{eik}}^i = \mathbb{E}_{\mathbf{x}_c} \left(\|\nabla f_{\text{sdf}}(\mathbf{x}_c)\| - 1 \right)^2 \tag{12}$$

Implementation The implicit shape network and blend skinning network are implemented as MLPs. We use positional encoding [35] for the query point $\mathbf{x}_c$ to increase the expressive power of the network. We leverage the implicit differentiation derived in [10] to compute gradients during iterative root finding.

Initialization We initialize body poses by fitting SMPL model [30] to RGB-D observations. This is achieved by minimizing the distances from point clouds to the SMPL mesh and jointly minimizing distances between the SMPL mesh and the corresponding surface points obtained from a DensePose [19] model. Please see Supp. Mat. for details.

Optimization Given a sequence of RGB-D video, we deform our neural implicit human model for each frame based on the 3D pose estimate and compare it with its corresponding RGB-D observation. This allows us to jointly optimize both shape parameters $\mathbf{\Theta}_{\text{sdf}}$, $\mathbf{\Theta}_w$ and pose parameters $\mathbf{p}_i$ of each frame and makes our model robust to noisy initial pose estimates. We follow a two-stage optimization protocol for faster convergence and more stable training: First, we pretrain the shape and skinning networks in canonical space based on the SMPL meshes obtained from the initialization process. Then we optimize the shape network, skinning network and poses jointly to match the RGB-D observations.

3.3. Animation

To generate animations, we discretize the deformed space at a pre-defined resolution and estimate $SDF(\mathbf{x}_d)$ for every point $\mathbf{x}_d$ in this grid via correspondence search (Sec 3.1). We then extract meshes via MISE [34].

4. Experiments

We first conduct ablations on our design choices. Next, we compare our method with state-of-the-art approaches on the reconstruction and animation tasks. Finally, we demonstrate personalized avatars learned from only a single monocular RGB-D video sequence qualitatively.

4.1. Datasets

We first conduct experiments on two standard datasets with clean scans projected to RGB-D images to evaluate our performance on both **reconstruction** and **animation**. To further demonstrate the robustness of our method to real-world sensor noise, we collect a dataset with a single Kinect including various challenging garment styles.

BUFF Dataset [59]: This dataset contains textured 3D scan sequences. Following [9], we obtain monocular RGB-D data by rendering the scans and use them for our **reconstruction** task, comparing to the ground truth scans.

CAPE Dataset [31]: This dataset contains registered 3D meshes of people wearing different clothes while performing various actions. It also provides corresponding ground-truth SMPL parameters. Following [50], we conduct **animation** experiments on CAPE. To adapt CAPE to our monocular depth setting, we acquire single-view depth inputs by rendering the meshes. The most challenging subject (blazer) is used for evaluation where 10 sequences are used for training and 3 unseen sequences are used for evaluating the **animation** performance. Note that our method requires RGB-D for initial pose estimation since CAPE does not provide texture, we take the ground-truth poses for training (same for the baselines).

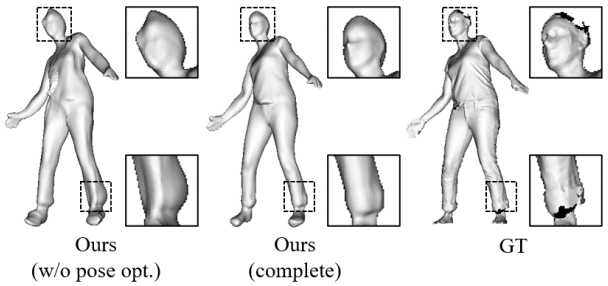

Figure 3. **Qualitative ablation (BUFF)**. Joint optimization corrects pose estimates and achieves better reconstruction quality.

Method	IoU ↑	C − ℓ_2 ↓	NC ↑
Ours w/o pose opt.	0.850	1.6	0.887
Ours	**0.879**	**1.1**	**0.927**

Table 1. **Importance of pose optimization on BUFF.** We evaluate the reconstruction results of our method without jointly optimizing pose and shape.

Real Data: To show robustness and generalization of our method to noisy real-world data, we collect RGB-D sequences with an Azure Kinect at 30 fps (each sequence is approximately 2-3 minutes long). We use the RGB images for pose initialization. We learn avatars from this data and animate avatars with unseen poses [31, 33, 53].

Metrics: We consider volumetric IoU, Chamfer distance (cm) and normal consistency for evaluation.

4.2. Ablation Study

Joint Optimization of Pose and Shape: The initial pose estimate from a monocular RGB-D video is usually noisy and can be inaccurate. To evaluate the importance of *jointly* optimizing pose and shape, we compare our full model to a version without pose optimization. **Results**: Tab. 1 shows that joint optimization of pose and shape is crucial to achieve high reconstruction quality and globally accurate alignment (Chamfer distance and IoU). It is also important to recover fine details (normal consistency). As shown in Fig. 3, unnatural reconstructions such as the artifacts on the head and trouser leg can be corrected by pose optimization.

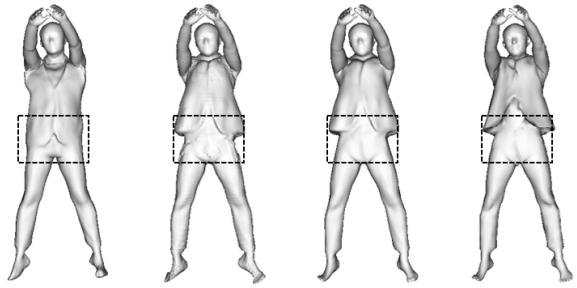

Figure 4. **Qualitative ablation (CAPE).** Without conditioning our shape network on poses, the static network cannot represent dynamically changing surface details including deformations on necklines and bottom hems. Further, the lack of learned skinning weights results in noisy surfaces.

Method	IoU ↑	C $-\ell_2$ ↓	NC ↑
Ours (w/o pose cond.)	0.936	0.911	0.884
Ours (w/ SMPL weights)	0.945	0.612	0.887
Ours (complete)	**0.955**	**0.553**	**0.912**

Table 2. **Importance of pose-dependent deformations and learned skinning weights on CAPE.** We evaluate the animation results of our method without pose conditioning and driven by SMPL skinning weights.

Deformation Model: The deformation of the avatar can be split into **pose-dependent deformation** and skeletal deformation via LBS with the learned skinning field. To model pose-dependent deformations such as cloth wrinkles, we leverage a pose-conditioned shape network to represent the SDF in canonical space. **Results**: Fig. 4 shows that without pose features, the network cannot represent dynamically changing surface details of the blazer, and defaults to a smooth average. This is further substantiated by a 70% increase in Chamfer distance, compared to our full method.

To show the importance of **learned skinning weights**, we compare our full model to a variant with a fixed shape network (w/ SMPL weights). Points are deformed using SMPL blend weights at the nearest SMPL vertices. **Results**: Tab. 2 indicates that our method outperforms the baseline in all metrics. In particular, the normal consistency improved significantly. This is also illustrated in Fig. 4, where the baseline (w/ SMPL weights) is noisy and yields artifacts. This can be explained by the fact that the skinning weights of SMPL are defined only at the mesh vertices of the naked body, thus they can't model complex deformations.

4.3. Reconstruction Comparisons

Baselines: Although not our primary goal, we also compare to several reconstruction methods, including IP-Net [6], CAPE [31] and DSFN [9]. The experiments are conducted on the BUFF [59] dataset. The RGB-D inputs are rendered from a sequence of registered 3D meshes. IP-Net relies on a learned prior from [1, 2]. It takes the partial 3D point cloud from each depth frame as input and predicts the implicit surface of the human body. The SMPL+D model is then registered to the reconstructed surface. DSFN models per-vertex offsets to the minimally-clothed SMPL body via pose-dependent MLPs. For CAPE, we follow the protocol in DSFN [9] and optimize the latent codes based on the RGB-D observations (see Supp. Mat for details).

Method	IoU ↑	C $-\ell_2$ ↓	NC ↑
IP-Net [6]	0.783	2.1	0.861
CAPE [31]	0.648	2.5	0.844
DSFN [9]	0.832	1.6	0.916
Ours	**0.879**	**1.1**	**0.927**

Table 3. **Quantitative evaluation on BUFF.** We provide rendered depth maps as input for all methods. Our method consistently outperforms all other baselines in all metrics (see Fig. 5 for qualitative comparison).

Results: Tab. 3 summarizes the reconstruction comparison on BUFF. We observe that our method leads to better reconstructions in all three metrics compared to current SOTA methods. A qualitative comparison is shown in Fig. 5. Compared to methods based on implicit reconstruction, i.e., IP-Net, our method reconstructs person-specific details better and generates complete human bodies. This is due to the fact that IP-Net reconstructs the human body frame-by-frame and can't leverage information across the sequence. In contrast, our method solves the problem via global optimization. Compared to methods with explicit representations, i.e., CAPE and DSFN, our method reconstructs details (hair, trouser leg) that geometrically differ from the minimally clothed human body better. We attribute this to the flexibility of implicit shape representations.

4.4. Animation Comparisons

Baselines: We compare animation quality on CAPE [31] with IP-Net [6] and SCANimate [50] as baselines. IP-Net does not natively fuse information across the entire depth sequence (discussed in Sec 4.3). For a fair comparison, we feed one complete T-pose scan of the subject as input to IP-Net and predict implicit geometry and leverage the registered SMPL+D model to deform it to unseen poses. For SCANimate, we create two baselines. The first baseline (SCANimate 3D) is learned from complete meshes and follows the original setting of SCANimate. Note that in this comparison ours is at a disadvantage since we only assume monocular depth input without accurate surface normal information. Therefore, we also compare to a variant (SCANimate 2.5D) which operates on equivalent 2.5D inputs. For

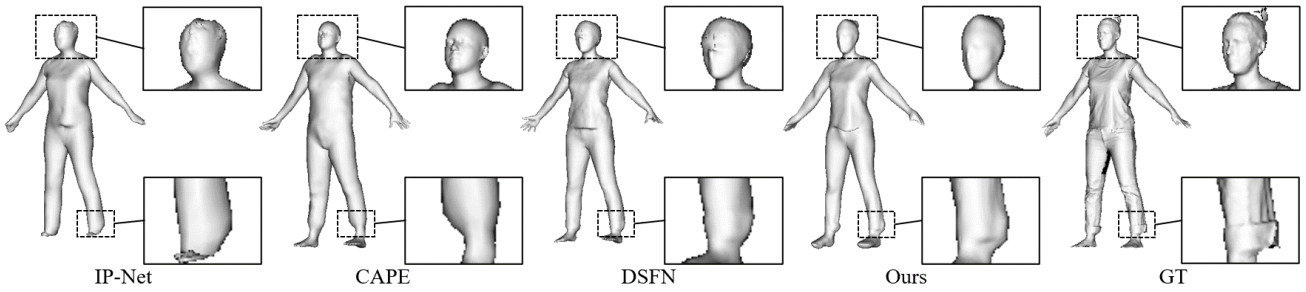

Figure 5. **Qualitative reconstruction comparisons on BUFF.** Our method reconstructs better details and generates less artifacts compared to IP-Net. The implicit shape representation enables accurate reconstruction of complex geometry (hair, trouser heel) compared to methods with explicit representations, i.e., CAPE and DSFN.

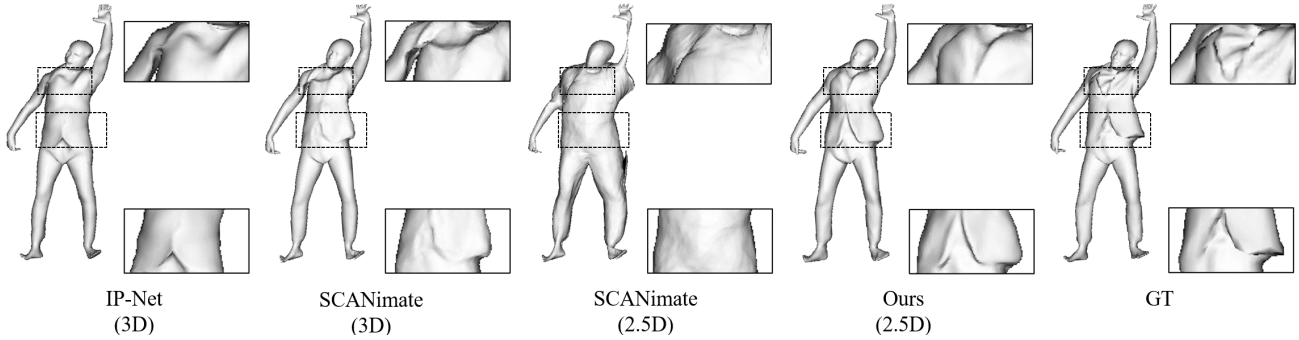

Figure 6. **Qualitative animation comparison on CAPE.** IP-Net produces unrealistic animation results potentially due to overfitting and wrong skinning weights. The deformation field of SCANimate is defined in the deformed space and thus limits its generalization to unseen poses. This is made worse in SCANimate (2.5D) which only uses partial point clouds as input. In contrast, our method solves this problem naturally via joint optimization of skinning field and shape in canonical space.

details on the variants we refer to Supp. Mat.

Results: Tab. 4 shows the quantitative results. Our method outperforms IP-Net and SCANimate (2.5D), and achieves comparable results to SCANimate (3D) which is trained on complete and noise-free 3D meshes. Fig. 6 shows that the clothing deformation of the blazer is unrealistic when animating IP-Net. This may be due to overfitting to the training data. Moreover, the animation is driven by skinning weights that are learned from minimally-clothed human bodies. As seen in Fig. 6, SCANimate also leads to unrealistic animation results for unseen poses. This is because the deformation field in SCANimate depends on the pose of the *deformed* object, which limits generalization to unseen poses [10]. Furthermore, we find that this issue is amplified in SCANimate (2.5D) with partial point clouds. In contrast, our method solves this problem well via jointly learned skinning field and shape in canonical space.

4.5. Real-world Performance

To demonstrate the performance of our method on noisy real-world data, we show results on additional RGB-D sequences from an Azure Kinect in Fig. 7. More specifically, we learn a neural avatar from an RGB-D video and drive the animation using unseen motion sequences from [31,33,53]. Our method is able to reconstruct complex cloth geometries like hoodies, high collar and puffer jackets. Moreover, we demonstrate reposing to novel out-of-distribution motion sequences including dancing and exercising. **Please refer to Supp. Mat for real-world performance demos.**

Method	Input	IoU ↑	C $-\ell_2$ ↓	NC ↑
IP-Net [6]	3D	0.916	0.735	0.843
SCANimate [50]	3D	**0.941**	**0.560**	**0.906**
SCANimate [50]	2.5D	0.665	4.710	0.785
Ours	2.5D	**0.946**	**0.621**	**0.906**

Table 4. **Quantitative evaluation on CAPE.** Our method outperforms IP-Net and SCANimate (2.5D) by a large margin and achieves comparable result with SCANimate (3D) which is trained on complete 3D meshes, a significantly easier setting compared to using partial 2.5D data as input.

5. Conclusion

In this paper, we presented PINA to learn personalized implicit avatars for reconstruction and animation from noisy

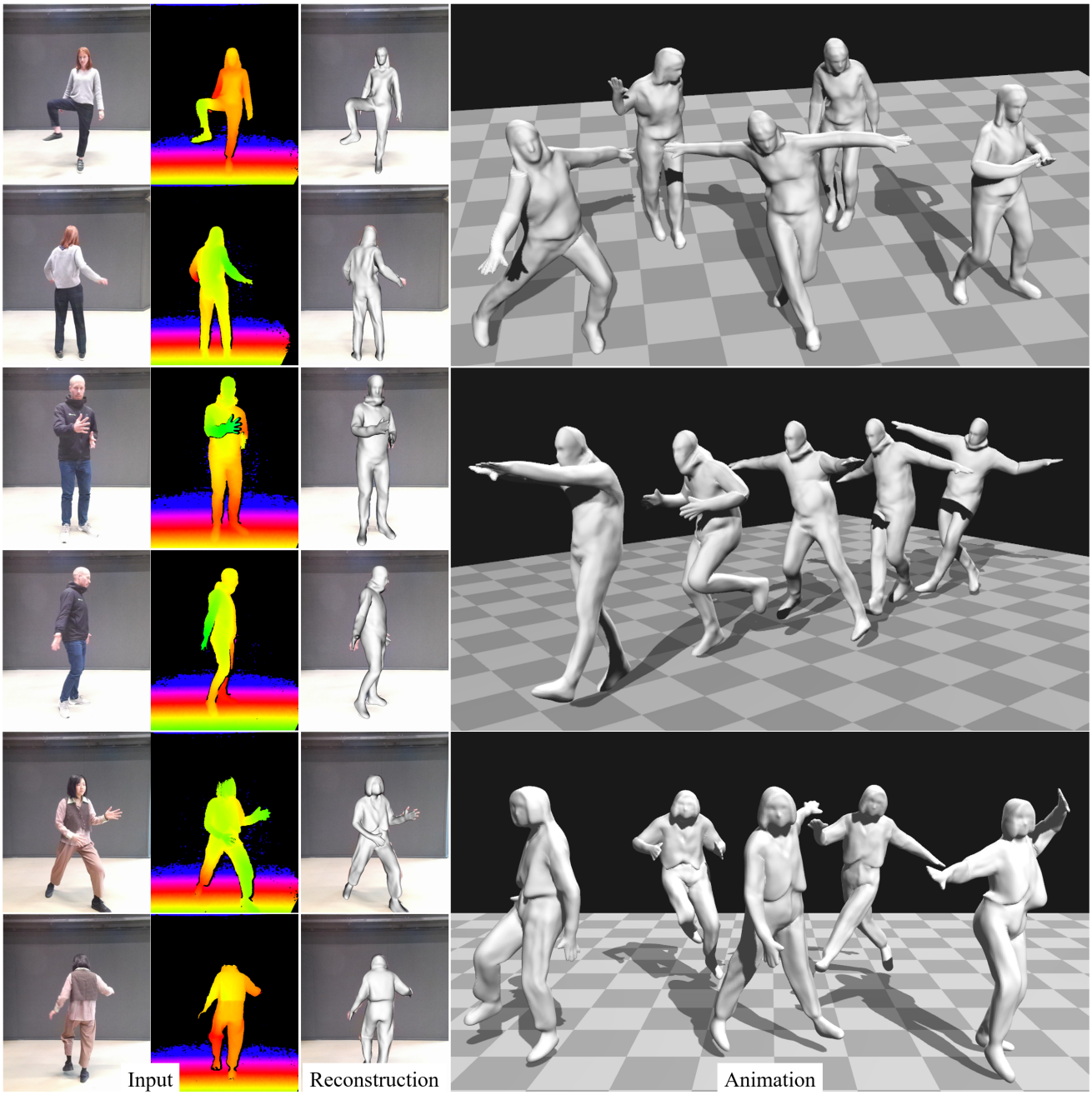

Figure 7. **RGB-D results.** We show qualitative results of our method from real RGB-D videos. Each subject has been recorded for 2-3 min (left). From noisy depth sequences (cf. head region bottom row), we learn shape and skinning weights (reconstruction), by jointly fitting the parameters of the shape and skinning network and the poses. We use unseen poses from [31, 33, 53] to animate the learned character.

and partial depth maps. The key idea is to represent the implicit shape and the pose-dependent deformations in canonical space which allows for fusion over all frames of the input sequence. We propose a global optimization that enables joint learning of the skinning field and surface normals in canonical representation. Our method learns to recover fine surface details and is able to animate the human avatar in novel unseen poses. We compared the method to explicit and neural implicit state-of-the-art baselines and show that we outperform all baselines in all metrics. Currently, our method does not model the appearance of the avatar. This is an exciting direction for future work. We discuss potential negative societal impact and limitations in the Supp. Mat.

Acknowledgements: Zijian Dong was supported by ELLIS. Xu Chen was supported by the Max Planck ETH Center for Learning Systems.

References

[1] https://web.twindom.com/. 6

[2] https://www.treedys.com/. 6

[3] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[6] 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, 3, 6, 7

[7] 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*, pages 561–578. Springer, 2016. 2

[8] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021. 2, 3

[9] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic surface function networks for clothed human bodies. In *Proc. International Conference on Computer Vision (ICCV)*, Oct. 2021. 2, 3, 5, 6

[10] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4, 7

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

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

[13] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11875–11885, 2021. 2

[14] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 2

[15] Zijian Dong, Jie Song, Xu Chen, Chen Guo, and Otmar Hilliges. Shape-aware multi-person pose estimation from multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11158–11168, 2021. 2

[16] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12814–12823, 2021. 2

[17] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 2, 4

[18] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (TOG)*, 31(4):1–10, 2012. 2

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 5

[20] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 2

[21] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics*, 40(4), aug 2021. 2

[22] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*, volume 28, pages 337–346. Wiley Online Library, 2009. 2

[23] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072*, 2020. 2

[24] 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. 2

[25] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and

shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 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, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14172, 2021. 3

[29] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *arXiv preprint arXiv:2106.02019*, 2021. 2

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3, 5

[31] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5, 6, 7, 8

[32] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2

[33] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019. 5, 7, 8

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2, 5

[35] 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. In *European conference on computer vision*, pages 405–421. Springer, 2020. 4

[36] Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 171–178. IEEE, 2014. 2

[37] 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. 3

[38] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011. 3

[39] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 598–613. Springer, 2020. 2

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2

[42] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 2

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[44] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14314–14323, October 2021. 2

[45] 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. 2

[46] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr: Articulated neural rendering for virtual avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2021. 2

[47] Antonio Ricci. A constructive geometry for computer graphics. *The Computer Journal*, 16(2):157–160, 1973. 4

[48] 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. 2

[49] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[50] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 5, 6, 7

[51] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. 2020. 2

[52] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 2

[53] 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, 7, 8

[54] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *arXiv preprint arXiv:2106.11944*, 2021. 2

[55] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[56] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 237–253, 2018. 2

[57] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. 3

[58] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018. 3

[59] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 5, 6

[60] 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