Received: 23 September 2022 | Revised: 3 November 2022 | Accepted: 24 November 2022

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12193

ORIGINAL RESEARCH

Skeleton-aware implicit function for single-view human reconstruction

Pengpeng Liu[1] | Guixuan Zhang[1] | Shuwu Zhang[1,2] | Yuanhao Li[3] | Zhi Zeng[1,2]

[1]Key Laboratory of Digital Rights Services, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]Beijing University of Posts and Telecommunications, Beijing, China

[3]Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan

Correspondence

Zhi Zeng, Beijing University of Posts and Telecommunications, Beijing 100876, China.
Email: zhi.zeng@ia.ac.cn

Funding information

National Key R&D Program of China, Grant/Award Number: 2022YFF0901902

Abstract

The aim is to reconstruct a complete and detailed clothed human from a single-view input. Implicit function is suitable for this task because it represents fine shape details and varied topology. Current methods, however, often suffer from artefacts such as broken or disembodied body parts, missing details, or depth ambiguity due to the ambiguity and complexity of human articulation. The main issue observed by the authors is structure-agnostic. To address these problems, the authors fully utilise the skinned multi-person linear (SMPL) model and propose a method using the Skeleton-aware Implicit Function (SIF). To alleviate the broken or disembodied body parts, the proposed skeleton-aware structure prior makes the skeleton awareness into an implicit function, which consists of a bone-guided sampling strategy and a skeleton-relative encoding strategy. To deal with the missing details and depth ambiguity problems, the authors' body-guided pixel-aligned feature exploits the SMPL to enhance 2D normal and depth semantic features, and the proposed feature aggregation uses the extra geometry-aware prior to enabling a more plausible merging with less noisy geometry. Additionally, SIF is also adapted to the RGB-D input, and experimental results show that SIF outperforms the state-of-the-arts methods on challenging datasets from Twindom and Thuman3.0.

KEYWORDS

3D human reconstruction, deep learning, neural network

1 | INTRODUCTION

Realistic human reconstruction from a single-view image is the key to a myriad of applications from virtual reality to medical imaging. Though deep learning models have shown great promise in single-view human reconstruction [1–4], it is still an extremely challenging problem due to the various poses, shapes and cloths. Existing parametric body models regress a minimally clothed human model with a consistent topology, but lacking important details like hair and clothing. Methods based on deep implicit function can represent fine shape details and varied topology; however, they often produce 3D human with broken or disembodied body parts, especially for the unseen region. In this paper, we base on deep implicit function, aimed to achieve a complete and detail-preserving clothed human reconstruction from a single-view input, as shown in Figure 1.

We observe that the main issue with these methods based on implicit function is the lack of structure prior. Implicit function is a data-driving method, which heavily relies on the learnt knowledge from data. Due to the lack of 3D data, only fed with more prior to promoting initial learning will make it possible and practicable for implicit function to reduce dependency on the data. Recent methods [2, 5] only condition on 2D image features, which is obviously insufficient for the above-mentioned problems due to the ambiguity and complexity of human articulation. Methods in refs. [6, 7] combine implicit function with the parametric model (skinned multi-person linear [SMPL]); however, they mainly focus on surface- or volume-related information from the parametric model, which leads to SMPL-like reconstruction. In contrast, we pay more attention to skeleton-related prior and propose to introduce skeleton awareness into the implicit function, which

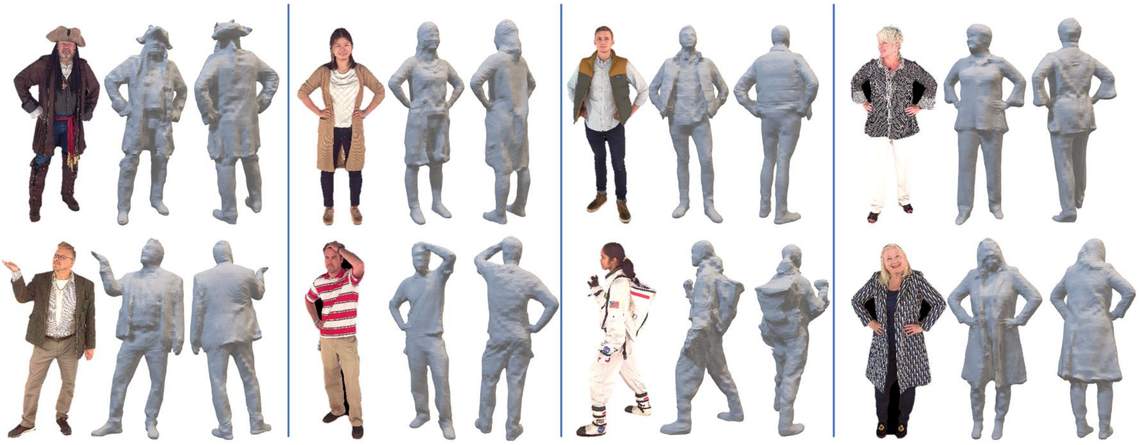

FIGURE 1 Results of SIF on various challenging shapes and cloths. We show the geometry in front and side views. SIF, Skeleton-aware Implicit Function.

is less influenced by SMPL and has the potential to handle more challenging data like loosing cloth.

Our method SIF, which stands for Skeleton-aware Implicit Function, consists of three modules: skeleton-aware structure prior, body-guided pixel-aligned feature and feature aggregation (FA). To alleviate the artefact of broken or disembodied parts, our skeleton-aware structure prior takes two strategies: one is the bone-guided sampling strategy to add extra training points sampled around the joints and bones, implicitly guiding the networks to capture the joints and the connectivity of bones and the other is the skeleton-relative encoding strategy which encodes the connectivity and bone lengths related to 3D joint locations, more explicitly guiding the networks to sense the human skeleton structure. Compared to surface- or volume-related information from the parametric model, our skeleton-aware structure prior has a higher tolerance to small disturbances to SMPL, which is demonstrated more robustly for inaccurate SMPL.

To relieve the artefacts of missing details and depth ambiguity problems, our body-guided pixel-aligned feature utilises SMPL to enhance the 2D semantic features. Besides the referred normal features from ref. [8], we also introduce the depth-relative prior from rendered SMPL depth inspired from ref. [3]. Our simple FA has demonstrated a better performance with less noisy geometry than the strategy in ref. [8], because the attached extra geometry-aware information can distinguish the query points and help to learn the difference of the reference confidence between viewpoints.

We evaluate SIF quantitatively and qualitatively on the challenging datasets, which contains various poses, shapes and cloths from the Twindom and Thuman3.0 datasets. Results show that our SIF outperforms the state-of-the-arts (SOTA) methods. SIF is also adapted to RGB-D input, which has achieved better performance than RGB-D baseline [3].

Our contributions can be summarised as follows:

1) We firstly propose to incorporate skeleton awareness into implicit function with a bone-guided sampling strategy and skeleton-relative encoding, which reduces the artefacts like broken body parts.
2) We propose the SIF method to fully utilise SMPL from 2D feature to 3D skeleton structure feature, which consists of body guided pixel-aligned feature, skeleton-aware structure prior and FA, and achieves a complete and detailed human geometry reconstruction.
3) Our SIF can be also adapted to RGB-D input and experimental results show that our SIF outperforms RGB and RGB-D baselines. Compared to SOTA methods which extract surface- or volume-related information from the parametric model, our SIF is more robust for inaccurate SMPL and has the potential to handle more challenging data like loosing cloth due to less influence by SMPL.

2 | RELATED WORK

In the following, we focus on 3D human reconstruction and classify the existing methods into three categories according to their underlying techniques.

2.1 | Parametric body estimation from a single image

With the advent of human statistical models like SCAPE [9], SMPL [10], SMPLX [11] and Star [12], parametric body estimation from a single image has attracted a lot of attention recently. Methods optimise the shape and pose parameters by fitting the SMPL model to the 2D keypoint detections [13, 14] and other dense shape cues [15]. Recently, deep learning methods have become a trend. In refs. [1, 16–18], the authors utilise neural networks to directly regress the 3D shape and pose parameters. The authors in ref. [19] use the Graph-CNN [20] to regress the vertices of SMPL instead of shape and pose parameters. However, these methods cannot reconstruct a detailed human.

2.2 | Deep implicit function for human reconstruction

Implicit function defines a surface as a level set of an occupancy probability function. Compared to explicit representations, such as point clouds [21, 22], voxel grids [23], and meshes [24, 25], implicit function can represent detailed 3D shapes with an arbitrary topology, not limited by the output resolution. PIFu [2] for the first time utilises the pixel-aligned implicit function for clothed human reconstruction, after that PIFuHD [5] improves geometry details significantly by introducing normal feature, and Monoport [26] proposes an efficient volumetric sampling scheme to speed up the inference time. Besides RGB input, methods extend the input to point clouds [27] and RGB-D [28]. Recently, Function4D [3] fully explores the depth information and achieves a real-time and detail-preserving human reconstruction. However, implicit shapes cannot be posed and animated due to lack of a consistent mesh topology, and shape reconstructions with single-view input often produce artefacts like broken or disembodied parts or geometry noise.

2.3 | Parametric models and implicit function

Parametric models are well regularised with a consistent topology, while implicit function models are more expressive. Recent methods combine the two representations to achieve a field complementation. PaMIR [6] and Deephuman [7] use a heavy 3D deep neural network to extract features from voxelised SMPL volume, which is hard to train and requires larger datasets for a good generalisation. For the sparse point clouds input, IPNet [4] infers an occupancy filed to jointly represent two surfaces with body/clothing layers and then registers SMPL/SMPL + D to the two layers. Recently, ICON [8] proposes a body-guided normal estimation and a visibility-aware implicit regressor with a local feature, which is robust to a large pose; however, it is prone to produce a noisy and tight clothing geometry, heavily relying on the signed distance function (SDF) field of SMPL. Our work is similar to ICON, but much different from the following points.

The differences to ICON: 1) ICON mainly focuses on surface- or volume-related information from SMPL (i.e. SDF field), leading to SMPL-like reconstruction, while our SIF pays more attention to skeleton-related prior and proposes the skeleton awareness into implicit function to make the networks "sense" human articulation. To enhance the space information, we also introduce a depth-relative prior of rendered depth from SMPL. **2)** ICON proposes a visibility-aware FA based on SMPL, while our geometry-aware FA leads to less noisy geometry. **3)** Our SIF has a higher tolerance to small disturbances to SMPL, which is more robust for inaccurate SMPL as shown in Table 1D. Additionally, considering the nature of sparsity of joints, our SIF can be less influenced by SMPL, which has the potential to better handle loosing cloth as showed in Figure 1.

TABLE 1 Quantitative errors (mm) based on RGB input.

		Twindom + THuman3.0		
	Methods	**Chamfer ↓**	**P2S ↓**	**Normals ↑**
A	SIF **w.** *PE*	**3.6258**	**3.5078**	**0.8793**
	SIF	3.6775	3.5649	0.8784
B	PIFu	8.5336	8.4515	0.7975
	PIFuHD	5.5912	5.5553	0.8536
	PaMIR	5.5982	5.3239	0.8573
	ICON	4.5096	4.6425	0.8194
	SMPL	10.3470	9.6107	0.7954
C	SIF **w/o.** $\mathcal{F}_J$	4.3151	4.1625	0.8724
	SIF **w/o.** *JtsEncode*	3.9316	3.9778	0.8421
	SIF **w/o.** *JtsSample*	3.7282	3.5981	0.8791
	SIF **w/o.** $\mathcal{F}_N$	6.1333	6.0628	0.8117
	SIF **w/o.** *Depth*	4.2072	4.0701	0.8672
	SIF **w/o.** *FA*	3.8593	3.7032	0.8674
	ICON **w.** *FA*	4.0898	4.1057	0.8468
D	SIF **w.** $\mathcal{N}$	3.8185	3.6992	0.8770
	ICON **w.** $\mathcal{N}$	5.2072	5.7652	0.8099
	SMPL **w.** $\mathcal{N}$	15.2560	13.6187	0.6752

Note: The best results are highlighted with bold numbers. (A) Our methods with different cutoff strategies: simple truncation and positional encoding strategy (*PE*); (B) performance with respect to SOTA and the errors of estimated SMPL; (C) ablation study about three modules; (D) SIF and ICON with perturbed SMPL ($\mathcal{N}$).

Abbreviations: SIF, skeleton-aware implicit function; SMPL, Skinned Multi-Person Linear; SOTA, state-of-the-arts.

3 | PRELIMINARY

A deep implicit function F is usually represented by multi-layer perceptrons (MLPs) [2], which predicts the continuous inside/outside probability field of a 3D model. It can define a surface as a level set of an occupancy probability function, e.g. $F(q) = 0.5$ where $q \in \mathbb{R}^3$ represents a 3D point. To represent a specific object surface, F usually takes the conditioned features (e.g. image feature of the object) as input and can be written as follows:

$$F(C(q)) : \mathbb{R}^3 \mapsto [0, 1] \tag{1}$$

The work PIFu [2] combines the pixel-aligned feature with the point coordinate and formulates the $C(q)$ as follows:

$$C(q) = (\mathcal{S}(F_I, \pi(p)), Z(p)) \tag{2}$$

where F_I represents the image feature maps from the deep encoder, $\pi(p)$ represents the 2D projection on the input image, $\mathcal{S}(F_I, \pi(p))$ is the sampling function using bilinear interpolation to sample the value on the feature maps F_I at pixel $\pi(p)$, $Z(p)$ is the z-value of p in the camera coordinate space. With

256-dimension pixel-aligned feature and 1-dimension z-value, PIFu [2] can reconstruct a detailed clothed human surface aligned to the input image. However, heavily conditioned on the 2D feature is not enough to deal with large pose variation and occlusion problems, which leads to artefacts, such as broken body parts, lacking of details, or non-human shape, especially for the unseen region. Considering that implicit function is structure-agnostic, in this paper, we propose the skeleton awareness into implicit function, to make the networks "sense" human articulation.

4 | METHOD

The proposed method SIF is a deep learning model aimed to output a complete and detail-preserving clothed 3D human from the single view input, which takes a segmented monocular image as input, along with an estimated parametric human shape (SMPL). Figure 2 gives an overview of our SIF architecture. We will describe the major modules in detail in the following : (1) body-guided pixel-aligned feature, (2) skeleton-aware structure prior and (3) FA.

4.1 | Body-guided pixel-aligned feature

The pixel-aligned feature from 2D feature maps plays a leading role as shown in Equation (2). Introducing an extra depth input [3, 28] or referred normal [5] can effectively alleviate the artefacts of over-smoothing. To enrich the geometric and spatial features, we exploit the body model SMPL to combine the normal and depth prior. We followed ref. [8] for a more plausible normal estimation. But different from ref. [3], we use the rendered depth from SMPL as input. Given the RGB image I, the depth and normal images can be obtained through the following steps, described as follows:

$$\mathcal{DR}(\mathcal{M}) \rightarrow \mathcal{N}^{\mathrm{b}}, \mathcal{D}^{\mathrm{b}} \tag{3}$$

$$\mathcal{G}_N\left(\mathcal{N}^{\mathrm{b}}, I\right) \rightarrow \mathcal{N}^{\mathrm{c}} \tag{4}$$

where $\mathcal{M}$ denotes the estimated SMPL body mesh and $\mathcal{DR}(\cdot)$ represents the differentiable render. We use $\mathcal{DR}$ to render $\mathcal{M}$ from the given view and the opposite view to obtain the SMPL-body normal maps $\mathcal{N}^{\mathrm{b}} = \left\{\mathcal{N}^b_{front}, \mathcal{N}^b_{back}\right\}$ and SMPL-body depth $\mathcal{D}^{\mathrm{b}} = \left\{\mathcal{D}^b_{front}, \mathcal{D}^b_{back}\right\}$. The network $\mathcal{G}_N$ predicts the clothed-human normal $\mathcal{N}^{\mathrm{c}} = \left\{\mathcal{N}^c_{front}, \mathcal{N}^c_{back}\right\}$ with image I and body normal $\mathcal{N}^{\mathrm{b}}$ as input. Then we can obtain the feature maps from these images $\left(I, \mathcal{N}^{\mathrm{c}}, \mathcal{D}^{\mathrm{b}}\right)$.

To fully utilise the rendered depth information, inspired from ref. [3], we calculated the relative depth between z-value of the 3D query point p and the projection value on the rendered depth, which can be written as follows:

$$\Delta z(p) = T\left(\mathcal{S}\left(\mathcal{D}^b, \pi(p)\right) - Z(p)\right) \tag{5}$$

where $Z(p)$ is the z-value of p in the camera coordinate space, $\mathcal{S}(\cdot)$ is used to fetch pixel-aligned values, $T(\cdot)$ is used to truncate the relative depth values in $[-\tau, \tau]$, which avoids misleading guidance for the invisible region. Compared to the

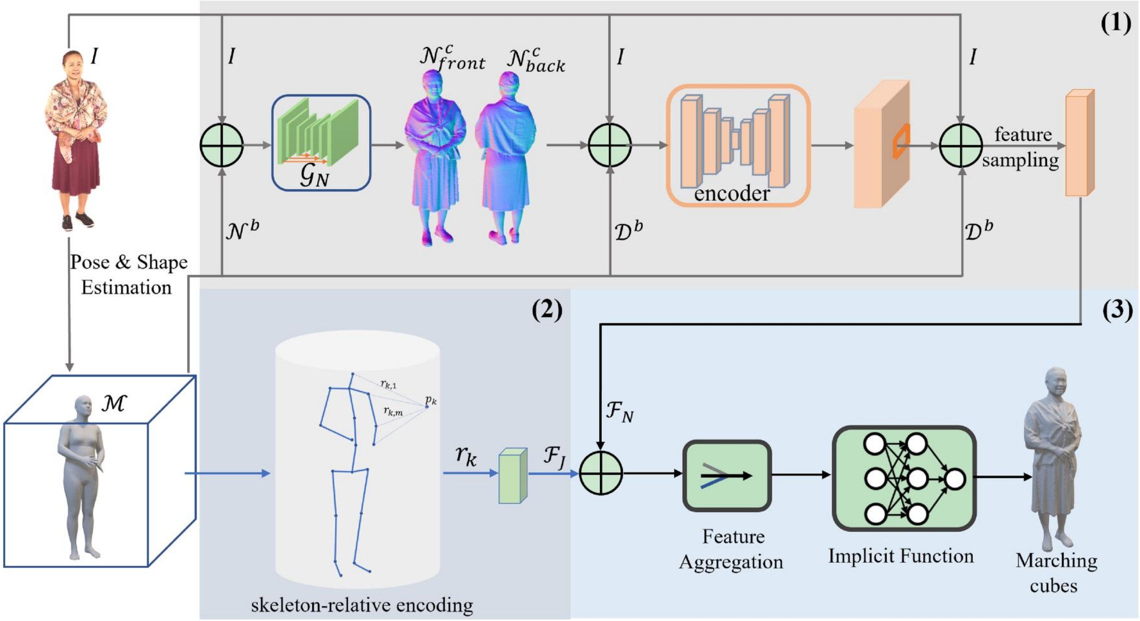

FIGURE 2 Overline of our SIF architecture. Given an input image and estimated SMPL model, SIF contains three main modules for (1) body-guided pixel-aligned feature, (2) skeleton-aware structure prior and (3) feature aggregation (FA). SIF, Skeleton-aware Implicit Function; SMPL, skinned multi-person linear.

absolute depth, the depth-relative value provides a view-dependent spatial semantics. Additionally, we introduce a local feature as demonstrated in ref. [8] that only feeding global features are sensitive to varied poses. Thus, our body-guided pixel-aligned feature can be formulated as follows:

$$\mathcal{F}_N(p) = [\mathcal{S}((F_*), \pi(p)), \mathcal{S}(I, \pi(p)), \Delta z(p)] \qquad (6)$$

where F_* represents the feature map extracted from $\left(I, \mathcal{N}^c, \mathcal{D}^b\right)$ by neural networks. Our pixel-aligned feature combines a global feature sampled from the feature maps F_*, a local feature sampled from the images I, and a depth-relative clue Δz.

4.2 | Skeleton-aware structure prior

Relying on clues from 2D images cannot make implicit function structure-aware, leading to artefacts such as broken parts. To make the networks "sense" inherent human articulation, we propose to incorporate the skeleton awareness into the deep implicit function, which consists of two main strategies: a bone-guided sampling strategy for a more implicit guidance, and a skeleton-relative encoding strategy for a much more explicit induction. In the following, we describe our joints estimation firstly and then our two strategies at length.

4.2.1 | Skeleton estimation

Given the estimated SMPL body mesh $\mathcal{M}$, the skeleton joints can be easily obtained by $J = \mathcal{J}(v)$, where v is the vertex set of $\mathcal{M}$, $\mathcal{J}$ is a matrix that transforms vertices v into joints. Since SMPL does not model hand and face motion, we filter two inaccurate hand joints and add some facial joints [29] and get our 30 body skeleton joints finally as shown in Figure 3.

Although our skeleton joints are obtained from SMPL for convenience, much more pose estimators can be adapted to our method, which can be discussed in future.

4.2.2 | Bone-guided sampling strategy

The sampling strategy for training data plays a central role in achieving expressiveness and accuracy of implicit function. To model the clothed human surface, PIFu [27] samples training data around the human surface, and similarly, [30] samples around the hands for hand reconstruction. That is to say that implicit function pays more attention to what data gives. Inspired from this, we propose the bone-guided sampling strategy, to guide the implicit function "sense" to the bone (the connectivity of two adjacent joints).

Compared to the surface-guided sampling strategy, which mainly focuses on sampling points around the surface, our strategy enhances the expressiveness of human joints and articulation skeleton. Specifically, we extra sample K_0 training points around the bones with a small Gaussian noise σ_0 and K_1 random sampling points around the joints with a small Gaussian noise σ_1. These bones and joints related sampling points share the same label (inside the surface), which guides the network to give the same predicted results for points around the bones, thus making the network "sense" the skeleton and the connectivity of bones during the learning procedure. To ensure all the bone-guided sampling points are inside the surface, we filter out some error prone bones and joints, such as joints on the ears, nose and eyes, as depicted in Figure 3.

4.2.3 | Skeleton-relative encoding

To more explicitly incorporate domain knowledge of how the body parts are linked and transformed to each other, we propose a skeleton-relative representation into implicit function. Our original intention is to transform the query point p relative to skeleton before determining the occupancy probability. For the purpose, we encode the connectivity and bone lengths via 3D joint locations. Inspired from A-nerf [31], our skeleton-relative representation can be written as follows:

$$r_k = \left[r_{k,1}, \ldots, r_{k,m}\right], r_{k,j} = \|p_k - J_j\|_2 \in \mathbb{R} \qquad (7)$$

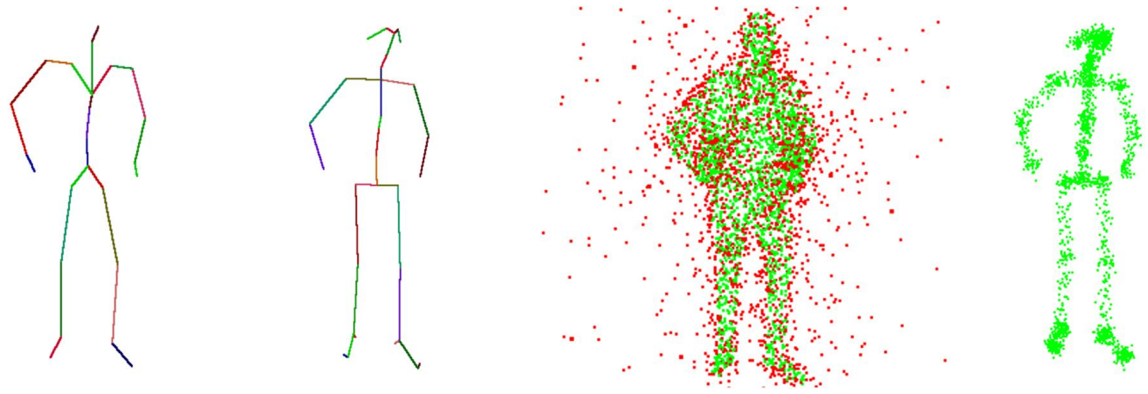

FIGURE 3 Visualisation of joints and sampling data. From left to right: 24 joints of SMPL, 30 joints of our SIF, surface-guided sampling data and bone-guided sampling data (green for points inside the surface while red for points outside the surface). SIF, Skeleton-aware Implicit Function.

where the subscript m denotes the number of joints, $r_{k,j}$ denotes the Euclidean distance between the 3D query point p_k and joint J_j. Thus, a 3D point p_k can map into a vector r_k relative to all the m skeleton joints. Specially, a point should not be influenced by all but only nearby bones to reduce the impact of irrelevant skeletons, a cutoff strategy is required to filter the far distant values. We give two alternative cutoff strategies as follows:

Truncation. We set a distance threshold δ and use a truncation function $T(\cdot)$ which truncates the distance encoding r_k into $[0, \delta]$, described as follows: $r'_k = T(r_k)$. This strategy is intuitive and lower-dimensional.

Positional Encoding. Inspired from A-nerf [31], we firstly use a positional encoding λ which consists of high frequency functions, mapping r_k to higher dimensional space, described as follows:

$$r^p_k = (\delta - r_k) \circ \lambda \tag{8}$$

$$\lambda(p) = \left[\sin\left(2^0 \pi p\right), \cos\left(2^0 \pi p\right), \ldots, \sin\left(2^{L-1} \pi p\right), \cos\left(2^{L-1} \pi p\right)\right. \tag{9}$$

In our experiments, we set $L = 3$. Then, we use sigmoid step function S to learn a weight relative to bone m by $w(k, m) = 1 - S(\tau(r(k, m) - \delta))$, where τ denotes the sharpness, we set $\tau = 20$ empirically. Thus, the weighted skeleton-relative encoding can be described as follows: $r'_k = w_k * r^p_k$. However, this strategy is much higher-dimensional and also requires a larger pixel-aligned feature for balance.

Compared to the truncation strategy, the positional encoding strategy [31] is much more well-designed which introduces a windowed version of positional encoding by multiplying the weight with respect to the bones. However, we find that the simper truncation cutoff strategy has a comparable result to the positional encoding strategy, as shown in Table 1, for that our depth-relative prior Δz can also provide a space information which lightens the burden of our skeleton encoding. Although we choose this simple truncation for our SIF considering the tradeoff between efficiency and effectiveness, we do not limit the special cutoff strategy and more ingenious alternatives can be discussed in future work. Thus, our skeleton-relative encoding can be written as follows:

$$\mathcal{F}_J(p_k) = T(r_k), k = 1, .., N + K \tag{10}$$

where K denotes the extra training points sampled around the bones and joints.

4.3 | Feature aggregation

For single-view input, we lighten the invisible region by referring the backside normal. Thus, for a 3D query point, it corresponds to two pixel-aligned features from the frontside and backside views, which requires to aggregate these features before the final discrimination. PIFuHD [5] directly concatenates normal features as the input to implicit function, due to the orthogonal projection assumption. ICON [8] calculates the visibility of the closet point in SMPL and determines which view pixel-aligned feature to use. However, this strategy in ref. [8] is more similar to the 0-1 decision that often produces artefacts at the boundary region of two views (see Figure 4). We observe that the depth-relative value Δz and skeleton-relative encoding F_J can distinguish the query points with space and structure information. Thus, we attach these geometry-aware information to the pixel-aligned feature maps and simply use two layers perceptrons followed by a pooling operation for FA. We find that with these extra geometry-aware prior, our simple architecture achieves better results than the strategy in ref. [8], especially in the boundary region as shown in Figure 4.

5 | EXPERIMENTS

5.1 | Datasets

Most of the existing 3D clothed-human datasets are commercialised and few are public. High-fidelity 3D geometry scans with corresponding SMPL fits are required for training our detail-preserving human reconstruction, we use realistic Twindom data and THuman3.0 [32] datasets for experiments. For the Twindom, to get the ground-truth SMPL, we firstly use MuVS [33] to the multi-view images for pose calculation and then register the estimated SMPL to the scans followed in ref. [6]. Finally, we choose 500 high-quality scans from the two datasets for training and 100 subjects for testing, which contain various cloths, poses and human–object interactions. For realistic images, we use the PRT-based renderer as in ref. [2] with a resolution of 512 × 512. Each scan and SMPL fit is rendered from every 6° in yaw axis, resulting in 500 × 60 = 30, 000 for training. For points sampling during training, we firstly sample 8000 points around the surface followed in ref. [2], which uses the mixture of uniform sampling and importance sampling per subject, and then sample 1000 points around joints and bones with $K_0 = 800$, $\sigma_0 = 0.05$, $K_1 = 200$, $\sigma_1 = 0.1$. For testing, we select 10 uniformly distributed rendered images for each scan (result in 100 × 10 = 1, 000 images), and for occupancy querying, we follow the embree algorithm [34].

5.2 | Implementation details

For the image encoders, we use the backbone of HRNetV2-W18-Small-v2 [26], which takes as input an image of 512 × 512 resolution and outputs a 32-channel feature map with resolution of 64×64. For the FA, we use a two-layer fully convolutional networks with hidden neurons (256,128) followed by a pooling operation. For the feature decoders, the implicit function is implemented as MLPs with the skip connections, where the hidden neurons are (128,128,128 and 1).

For training, we use Adam optimiser with learning rate of 2.0×10^{-4} which is decayed by the factor of 0.1 at every 10 epochs, the batch size of nine, the number of epochs of 25.

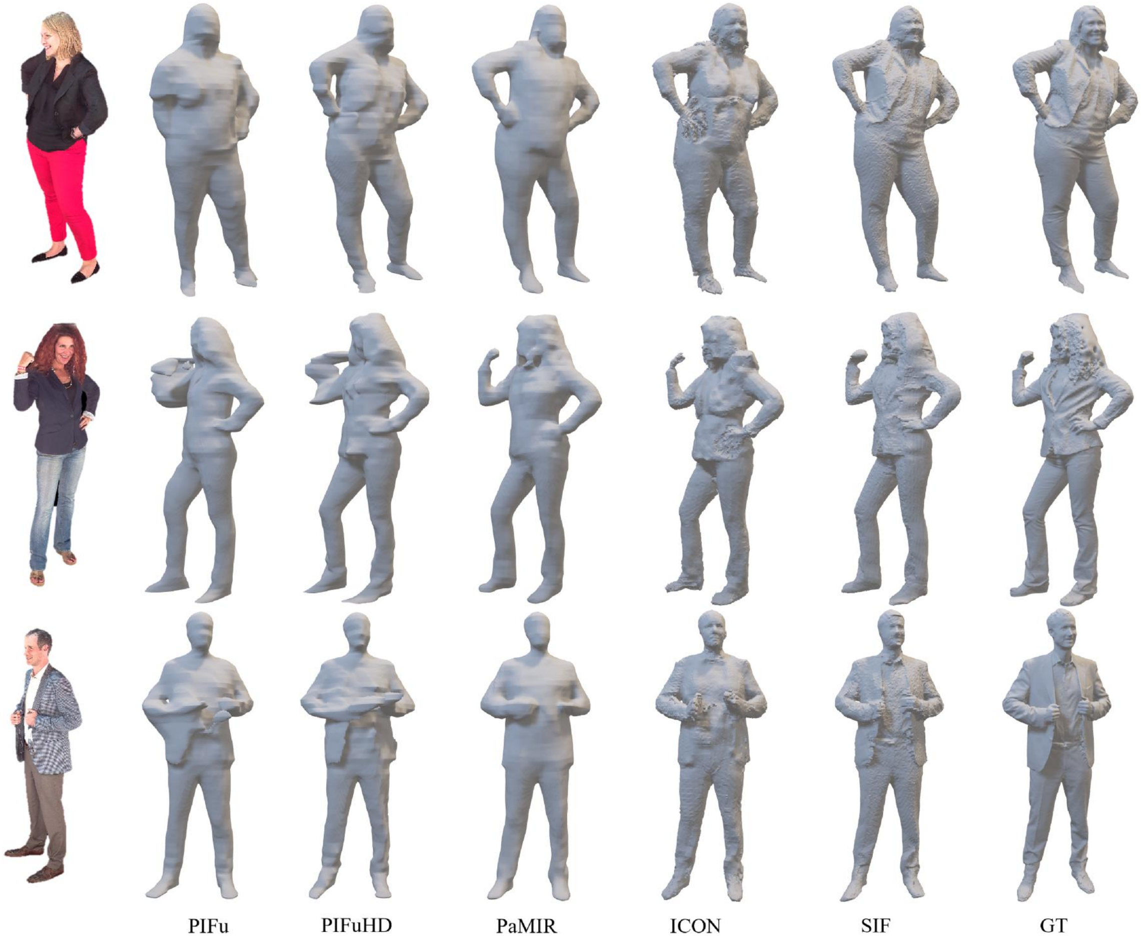

FIGURE 4 Qualitative comparison. We compare our method SIF mainly with PIFu [2], PIFuHD [5], PaMIR [6] and ICON [8]. SIF, skeleton-aware implicit function.

And we set $\tau = 0.05\ m$ and $\delta = 0.175\ m$ empirically. We train the network with ground-truth SMPL and normal images with random noise. For the training loss, we calculate the L_2 loss between the predicted labels and ground-truth labels for the training sampling points followed in ref. [5]. For training normal images, we follow ref. [8] with a pix2pixHD [35] network as the backbone and use the Adam optimiser with the learning rate of 2.0×10^{-4} until convergence at 100 epochs. The frontside and backside normal prediction networks are trained individually with batch size of six under the following objective function defined as follows:

$$\mathcal{L}_N = \mathcal{L}_{VGG} + \lambda \mathcal{L}_{l1} \tag{11}$$

Where $\mathcal{L}_{l1}$ is the L_1 distance between the ground truth and the prediction, $\mathcal{L}_{VGG}$ is the perceptual loss [36]. The weight λ is set to 5.0 in our experiments. We use the Adam optimiser with a learning rate of 2.0×10^{-4} until convergence at 100 epochs.

For reference, we evaluate the implicit fields with the resolution of 256 × 256 × 256 and use the Marching Cubes [37] with iso-surface threshold at 0.5 to extract the meshes. During the test, it is required to estimate SMPL and refine the SMPL and estimated normal images first. We use the PARE [38] as a SMPL estimator and follow ref. [8] to use a feedback loop between refining the SMPL mesh and normal maps for 2K iterations. For details, please refer the original papers [5, 8].

5.3 | Comparisons

We quantitatively evaluate our reconstruction with three different metrics, described in the following: **Chamfer distance**: For the ground truth scans and estimated meshes, we firstly sample 15,000 points uniformly on the scans/meshes and calculate the average bi-directional point-to-surface distance. This metric can capture the geometry difference, but misses very small geometric details; **P2S distance**: We additionally report the point-to-surface distance from scan points to the closet predicted surface points. This metric can be regarded as a single-directional version of chamfer distance; and **Normal consistency**: We compute the normal of these sample points and measures the accuracy and completeness of the shape normal (higher is better), followed [27].

We firstly qualitatively and quantitatively compare our method with the SOTA methods, including PIFu [2], PIFuHD [5], PaMIR [6], and ICON [8]. We retrain their networks using our training dataset and evaluate the performance on our test dataset. ICON provides two types of networks: without or with encoder, considering our more challenging datasets with various clothing, we compare to the type of ICON with encoder for a fairer comparison. Considering that parametric model based methods rely on an accurate estimated SMPL, we evaluate the robustness to small SMPL noise of our SIF compared to PaMIR and ICON. Finally, we extend out SIF to the RGB-D input and compare our SIF* with RGB-D baseline method from ref. [3].

5.3.1 | Quantitative comparison

As shown in Table 1B, our SIF with much lower-dimensional (32-dim) feature maps still performs much better than PIFu and PIFuHD with higher-dimensional (256-dim) feature maps, since we find that they heavily rely on higher-dimensional image feature to synthesise human structure and shape, especially for PIFu even using much heavier encoder networks (4 stacks Hour Glass [39]). For PaMIR and ICON, they extract volume- or surface-related information from SMPL and produce tight-clothing reconstruction in our challenging dataset due to their heavy dependence on SMPL, while our SIF still achieves better results, thanks to our richer pixel-aligned feature, more plausible FA and skeleton-aware structure prior.

5.3.2 | Qualitative comparison

Figure 5 shows the qualitative comparison with SOTA methods. Given the single view input, we can see that our SIF gets a more complete and detailed geometry. PIFu and PaMIR generate over-smooth results missing details due to lack of normal information, and ICON is prone to producing noisy surface while our SIF produces a more detailed surface, thanks to our rich pixel-aligned feature and plausible FA. For the unseen region, PIFu and PIFuHD produce broken or disembodied limbs due to lack of the structure prior, especially for PIFu with an obvious depth ambiguity problem, while our SIF performs better than PaMIR and ICON with a more complete and plausible geometry thanks to our more explicit skeleton-aware structure modelling.

5.3.3 | Robustness to SMPL noise

The estimated SMPL from an image might not be perfectly aligned with ground truth. Thus, SIF, ICON and PaMIR need to be robust against noise SMPL. As demonstrated in [8], ICON performs more robust than PaMIR, since PaMIR heavily relies on 3D features extracted from SMPL. To evaluate our method, we add small noise to SMPL pose and shape followed in ref. [8], and compare SIF with ICON without any optimisation strategy. Table 1D shows that SIF gets slightly affected by small SMPL noise (the result drops 5.4%) while ICON gets bigger errors (the result drops 19.8%). We argue that SIF is more robust to small noise SMPL than ICON and PaMIR because SIF mainly relies on skeleton joints compared to ICON with SDF of SMPL and PaMIR with volume of holistic SMPL. The skeleton-relative encoding of SIF bases on the assumption that all joints should be inside the body and are not required to remain in the exact position. Thus, SIF performs a higher fault tolerance to small noise SMPL than ICON and PaMIR.

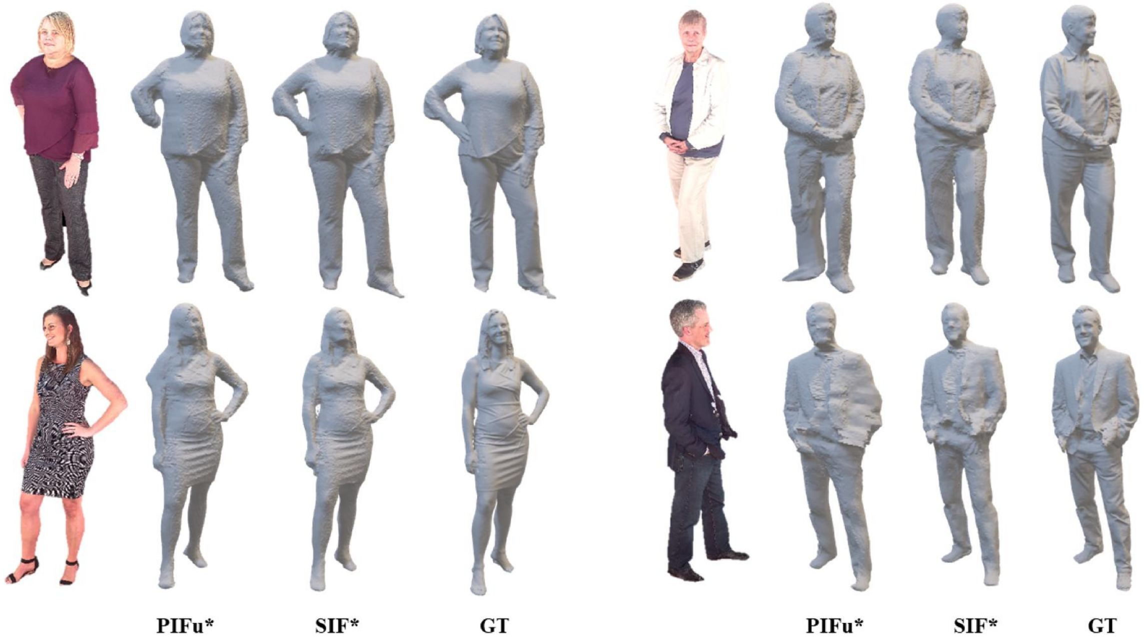

FIGURE 5 We compare our RGB-D extension SIF* with the baseline PIFu* [3]. From left to right: the input, PIFu* results, SIF* results and the ground truth. Our SIF* can deal with artefacts, including broken or disembodied parts, and missing details. SIF, Skeleton-aware Implicit Function.

5.3.4 | RGB-D extension

Our SIF can be easily adapted to RGB-D input, since we originally use the depth information rendered from SMPL body model. For the RGB-D input, we share the same network architecture but for more efficiency and effectiveness, we only need to render the backside depth from SMPL and set the different τ (1 cm for frontside and 5 cm for backside) in Equation (5). Additionally, we can filter out empty voxels with the depth image, which is faster for our inference. To create realistic depth data, we render ground truth depth and then add the synthesis depth sensor [40] on top of depth maps followed in ref. [3]. From Function 4d [3]we compare the RGB-D baseline (denoted as PIFu*), as demonstrated in Table 2A and Figure 6. Our extension SIF (denoted as SIF*) outperforms the baseline and performs better for the depth ambiguity problem and artefacts such as broken parts in the unseen region.

5.4 | Ablation study

We make the ablation study to verify our proposed modules based on the RGB input, since our extension SIF* shares the same network architecture as SIF; for a more comprehensive evaluation, we also conduct some experiments based on the RGB-D input considering the efficiency for faster convergence and reference.

TABLE 2 Quantitative errors (mm) with RGB-D input.

		Twindom + THuman3.0		
	Methods	**Chamfer ↓**	**P2S ↓**	**Normals ↑**
A	PIFu*	4.5058	4.4950	0.8525
	SIF*	**3.0672**	**3.0166**	**0.8712**
B	SIF***w/o.**$\mathcal{F}_J$	4.2427	4.2706	0.8456
	SIF***w/o.**$\mathcal{F}_N$	3.8663	3.6204	0.8680
	SIF***w/o.***FA*	3.3042	3.5229	0.8274
	PIFu* **w.** *Depth*	4.3363	4.3720	0.8359
	PIFu***w.** *JtsEncode*	3.9337	3.7501	0.8575
	PIFu***w.** *JtsEncode* + *JtsSample*	3.8663	3.6204	0.868
C	SMPL	10.3470	9.6107	0.7954

Note: The best results are highlighted with bold numbers. (A) Our SIF* compares to the baseline PIFu*, (B) ablation study and (C) the errors of estimated SMPL.

Abbreviations: SIF, Skeleton-aware Implicit Function; SMPL, skinned multi-person linear.

5.4.1 | Body-guided pixel-aligned feature

We quantitatively evaluate our pixel-aligned feature based on RGB and RGB-D input. As shown in Tables 1C and 2B, without our bone-guided pixel-aligned feature ($\mathcal{F}_N$), the results of SIF and SIF* considerably drop . Specially, to further evaluate our proposed depth-related feature (denoted as *Depth*), we further conduct two experiments in Tables 1C and 2B. We can see that SIF without *Depth* performs much worse, which demonstrates the effectiveness of our depth-related feature, while PIFu* with *Depth* shows slightly better results for RGB-D input that has provided space information for PIFu*.

5.4.2 | Skeleton-aware structure prior

We qualitatively and quantitatively evaluate our skeleton-aware module F_J, which consists of bone-guided sampling strategy and skeleton-relative encoding. As shown in Tables 1C and 2B, without our F_J, the result declines considerably. To further evaluate our two strategies, we conduct two experiments based on the RGB input as shown in Table 1C, where *JtsEncode* represents the skeleton-relative encoding strategy and *JtsSample* represents the bone-guided sampling strategy. Additionally, we also add the two strategies to PIFu* to verify the effectiveness as shown in Table 2B. From the experiments, we can see that SIF without each strategy achieves much worse results, while PIFu* with these two strategies both significantly improve the results. Meanwhile, for a more obvious comparison, we qualitatively show the results based on PIFu* as demonstrated in Figure 7. For the unseen right arm, our skeleton-relative encoding can capture a coarse profile, but still existing broken parts; while adding our sampling strategy, the networks can sense the connectivity of the skeleton and produce a complete surface, which demonstrates the effectiveness of our two strategies.

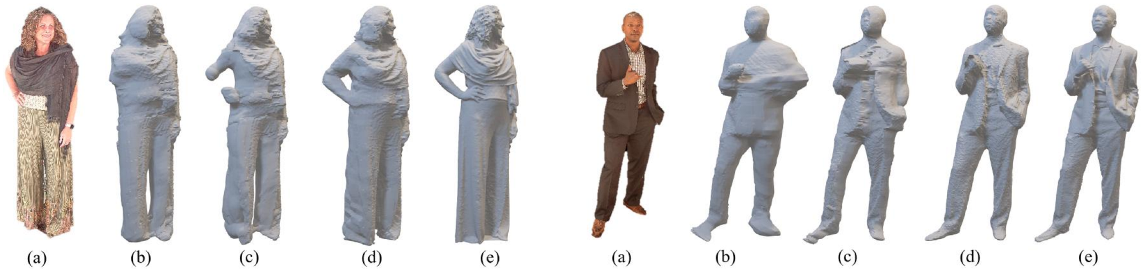

FIGURE 6 Ablation study about skeleton-aware structure prior. (a) the input, (b) PIFu* results, (c) PIFu* with only skeleton-relative encoding strategy, (d) PIFu* with our two strategies, (e) ground truth. Our two strategies can make the implicit function sense of the connectivity of the skeleton and produce a complete surface

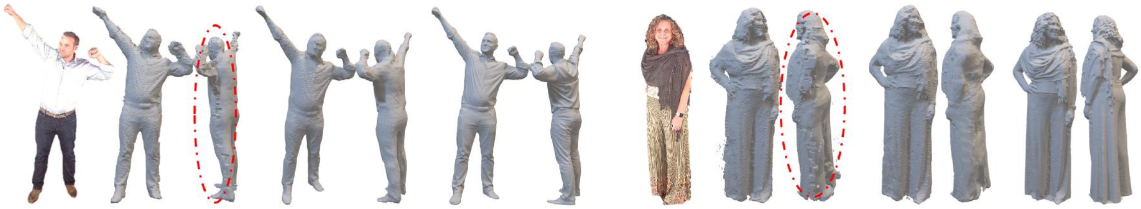

FIGURE 7 Ablation study about feature aggregation (FA). We show two views of mesh. From left to right: the input, results of SIF* without our FA, results of our SIF*, ground truth. Our FA can produce less noisy geometry. SIF, Skeleton-aware Implicit Function.

5.4.3 | Feature aggregation

To evaluate the effectiveness of our FA, we compare it with the module in ICON, which relies on the visibility of the point to choose a feature. As shown in Tables 1C and 2B, using the same module as ICON (denoted as **w/o.***FA*), our SIF (**w/o.** *FA*) and SIF* (**w/o.***FA*) gets a slightly worse result, while the results of ICON (**w.***FA*) improve considerably, for which our geometry-aware FA can guide the networks to distinguish the subtle difference between viewpoints. Meanwhile, Figure 4 demonstrates that our FA gets a less noisy geometry especially in the boundary region as explained in Section 4.3.

6 | DISCUSSION

6.1 | Conclusion

In this paper, we propose the method using SIF to recover a 3D mesh of a completed and detailed person from a single-view image. Our SIF reduces the depth ambiguity problem and artefacts such as broken or disembodied parts, missing details, and frequency noise. The main technical contributions include (1) introducing the skeleton awareness into implicit function with a bone-guided sampling strategy and a skeleton-relative encoding strategy and (2) providing a method to fully dig out SMPL feature with a body-guided pixel-aligned feature, a skeleton-aware structure feature and a geometry-aware FA. The ablation study demonstrates the effectiveness of each module. Moreover, SIF can be extended to RGB-D input and experiments show that SIF improves the performance compared to RGB and RGB-D based SOTA methods, which performs more robustly for inaccurate SMPL and has the potential to handle more challenging data such as loosing cloth.

6.2 | Limitation and future work

Our method requires high-quality human scans with SMPL annotation for training. However, it is time-consuming and costly to make a large-scale high-fidelity 3D human dataset. SIF explores bone- and skeleton-related prior rather than surface- or volume-related information from SMPL, which performs robust to small SMPL noise; however, significant failure of body estimation still leads to bad reconstruction, which is a general problem for all parametric model-based methods. Although we theoretically analyse and argue that our skeleton-aware structure prior can be adapted to many pose estimators, it needs to be extensively evaluated for completeness in future. Moreover, SIF is trained with weak-perspective or orthographic projection, which is much different from the real scenes. In the near future, we will explore how to improve the generation of the model to complex projection of real scenes, such as perspective projection.

ACKNOWLEDGEMENTS

Thanks to Mr Zheng, Zhi Zeng, Guixuan Zhang, Shuwu Zhang and Yuanhao Li. They helped me a lot and supported my research. This work is supported by the National Key R&D Program of China (2022YFF0901902), which is also the research achievement of the Key Laboratory of Digital Rights Services.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

We use realistic Twindom data and THuman3.0 datasets for experiments. For THuman3.0 datasets, it is available from Tsinghua University. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://github.com/fwbx529/THuman3.0-Dataset with permission. For Twindom data, it can be purchased at https://web.twindom.com with permission.

ORCID

Pengpeng Liu https://orcid.org/0000-0003-3583-5422

REFERENCES

1. Kanazawa, A., et al.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122–7131 (2018)
2. Saito, S., et al.: Pifu: pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2304–2314 (2019)
3. Yu, T., et al.: Function4d: real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5746–5756 (2021)

4. Bhatnagar, B.L., et al.: Combining implicit function learning and parametric models for 3d human reconstruction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pp. 311–329. Springer (2020)
5. Saito, S., et al.: 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, pp. 84–93 (2020)
6. Zheng, Z., et al.: Pamir: parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. 44(6), 3170–3184 (2021). https://doi.org/10.1109/tpami.2021.3050505
7. Zheng, Z., et al.: Deephuman: 3d human reconstruction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7739–7749 (2019)
8. Xiu, Y., et al.: Implicit clothed humans obtained from normals. In: Proc. IEEE Conf. On Computer Vision and Pattern Recognition (CVPR), vol. 2 (2022)
9. Anguelov, D., et al.: Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp. 408–416 (2005)
10. Loper, M., et al.: A skinned multi-person linear model. ACM Trans. Graph. 34(6), 1–16 (2015). https://doi.org/10.1145/2816795.2818013
11. Pavlakos, G., et al.: 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, pp. 10975–10985 (2019)
12. Osman, A.A., et al.: Sparse trained articulated human body regressor. In: European Conference on Computer Vision, pp. 598–613. Springer (2020)
13. Bogo, F., et al.: Keep it smpl: automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision, pp. 561–578. Springer (2016)
14. Pishchulin, L., et al.: Deepcut: joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4929–4937 (2016)
15. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7297–7306 (2018)
16. Kolotouros, N., et al.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2252–2261 (2019)
17. Kocabas, M., et al.: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5253–5263 (2020)
18. Guler, R.A., Kokkinos, I.: Holopose: holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10884–10894 (2019)
19. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4501–4510 (2019)
20. Kipf, T.N., Welling, M.: Semi-supervised Classification with Graph Convolutional Networks (2016). arXiv preprint arXiv:1609.02907
21. Qi, C.R., et al.: Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)
22. Lin, C.-H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3d object reconstruction. Proc. AAAI Conf. Artif. Intell. 32(1) (2018). https://doi.org/10.1609/aaai.v32i1.12278
23. Maturana, D., Scherer, S.: Voxnet: a 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 922–928. IEEE (2015)
24. Wang, N., et al.: Pixel2mesh: generating 3d mesh models from single rgb images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 52–67 (2018)
25. Groueix, T., et al.: A papier-mâché approach to learning 3d surface generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 216–224 (2018)
26. Li, R., et al.: Monocular real-time volumetric performance capture. In: European Conference on Computer Vision, pp. 49–67. Springer (2020)
27. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6970–6981 (2020)
28. Li, Z., et al.: Robust 3d self-portraits in seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1344–1353 (2020)
29. Cao, Z., et al.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
30. Corona, E., et al.: Lisa: learning implicit shape and appearance of hands. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20533–20543 (2022)
31. Su, S.-Y., et al.: Articulated neural radiance fields for learning human shape, appearance, and pose. Adv. Neural Inf. Process. Syst. 34, 12278–12291 (2021)
32. Shao, R., et al.: Doublefield: bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15872–15882 (2022)
33. Huang, Y., et al.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 International Conference on 3D Vision (3DV), pp. 421–430. IEEE (2017)
34. Wald, I., et al.: Embree: a kernel framework for efficient cpu ray tracing. ACM Trans. Graph. 33(4), 1–8 (2014). https://doi.org/10.1145/2601097.2601199
35. Wang, T.-C., et al.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)
36. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp. 694–711. Springer (2016)
37. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3d surface construction algorithm. ACM siggraph computer graphics 21(4), 163–169 (1987). https://doi.org/10.1145/37402.37422
38. Kocabas, M., et al.: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11127–11137 (2021)
39. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, pp. 483–499. Springer (2016)
40. Fankhauser, P., et al.: Kinect v2 for mobile robot navigation: evaluation and modeling. In: 2015 International Conference on Advanced Robotics (ICAR), pp. 388–394. IEEE (2015)

How to cite this article: Liu, P., et al.: Skeleton-aware implicit function for single-view human reconstruction. CAAI Trans. Intell. Technol. 8(2), 379–389 (2023). https://doi.org/10.1049/cit2.12193