Image and Vision Computing 137 (2023) 104782

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Temporally consistent reconstruction of 3D clothed human surface with warp field

Yong Deng, Baoxing Li, Yehui Yang, Xu Zhao *

Department of Automation, 800 Dongchuan Road, Minhang District, ShangHai 201100, China

ARTICLE INFO

Keywords:
Warp field
Normal maps
Implicit function
Temporally consistent information

ABSTRACT

Implicit functions are widely used in 3D human surface reconstruction due to their advantage to represent details. However, human reconstruction based on implicit functions struggles to maintain the integrity (unbroken body structure) and accuracy (no non-human parts) of human models. To address these issues, we propose a method, called TCR, for temporally consistent reconstruction of 3D clothed human surface with warp field. The fact that the general shape of a person does not change largely over time inspires us to exploit the temporally consistent shape information from previous frames to refine the human model of current frame. Therefore, we construct a canonical space and then store the shape information by updating the canonical model. To align the observed space with the canonical space, a warp field is firstly estimated for the forward and inverse warping of the human model. A probabilistic fusion strategy is then used to update the canonical model. In addition, the reconstructed result is further refined through the orthogonality constraints between the surface and its normal, which fully exploits the detailed information of estimated normal maps. Experiments on the Adobe and MonoPerfCap datasets show that TCR achieves the state-of-the-art performance. Furthermore, TCR is more robust and can maintain the integrity and accuracy of the reconstructed human body even with extreme poses and partial occlusions.

1. Introduction

Reconstructing 3D clothed human surface from still images or video is an important research topic in computer vision. It plays a significant role in many applications, such as holographic stereo communication, virtual reality technology and match broadcasting. Unlike human pose and shape estimation (HPS) [1], which reconstructs only naked body, the 3D clothed human reconstruction [2] requires fine-grained clothing details.

Currently, several representations are widely used for human reconstruction, but they all suffer from their own problems. *Parametric models*, e.g. SMPL [5], can represent reasonable human shape and pose with prior knowledge. However, these models can only represent the naked body without clothing details. Although coupling the parametric models with vertex displacements [24,25] can reconstruct some small details, large deformations such as long skirts cannot be reconstructed due to the limited topology of parametric meshes. *Neural implicit functions*[27,28] utilize neural networks such as multi-layer perceptrons (MLPs) to fit continuous occupancy functions or signed distance functions in 3D space. Owing to the continuity of the implicit function and the strong fitting ability of the neural network, neural implicit functions have theoretically infinite resolution and can represent arbitrary shapes. However, the lack of human body prior makes it difficult to maintain the integrity (unbroken body structure) and accuracy (no non-human parts) of human models. *Combined representations* adopt parametric models to provide the prior knowledge of human body in implicit functions. These methods [34] typically use a geometric encoder to integrate the 3D features of parametric models into the implicit neural network. Nevertheless, similar situations to neural implicit functions still occur in some difficult poses and partial occlusions when the parametric models are mis-estimated.

In view of the above problems, we propose a method for **T**emporally **C**onsistent **R**econstruction (TCR) of 3D clothed human surface with warp field. As the general shape of a person does not change largely over time, the temporally consistent shape information from previous frames can facilitate subsequent reconstructions. Therefore, we use a canonical model (usually the reconstructed model of the first frame) to hold the shape information of each frame. The space where the canonical model

* Corresponding author.
E-mail address: zhaoxu@sjtu.edu.cn (X. Zhao).

https://doi.org/10.1016/j.imavis.2023.104782
Received 6 April 2023; Received in revised form 7 July 2023; Accepted 24 July 2023
Available online 28 July 2023

resides is called *canonical space*, where the pose of all models is identical to the canonical model. And the space where the reconstruction result resides is called *observed space*. To bridge the gap between the two spaces, a *warp field*, formed by a set of sparse transformation nodes, is estimated to warp the single-frame human models from the observed space to the canonical space. The temporally consistent shape information is then preserved by fusing these warped models with the canonical model. As shown in Fig. 1, the integrity and accuracy of the canonical model is gradually improved through the fusion of temporally consistent shape information.

The pipeline of TCR is shown in Fig. 2, where three stages are involved. 1) In the *single frame prediction* stage, the human model of each frame is predicted by an implicit reconstruction method [35] based on normal maps and parametric models. To enhance the details of the back normal map (invisible surface), the front normal map (visible surface) is fed into the back normal prediction network. Compared to the original image feature, the normal map feature is more direct for the prediction of the normal network. 2) In the *temporally consistent reconstruction* stage, the predicted human model of each frame is first transformed into the canonical space by a warp field, and then used to update the canonical model through a probabilistic fusion strategy. Finally, the updated canonical model is inversely warped into the observed space as a preliminary result. 3) In the *normal map carving* stage, the preliminary result is carved based on the orthogonality relations between the surface and its normal, which makes full use of the predicted normal maps and recovers the clothing details eroded by the previous stage.

Experiments show that our method can not only maintain the integrity and accuracy of human models but also recover fine-grained details. We evaluate TCR on both indoor and outdoor videos. Meanwhile, we also show its robustness in challenging poses and partial occlusions. Furthermore, compared to previous methods [27,28,34,35], our method achieves the state-of-the-art performance in the field of human reconstruction. In summary, this paper has three main contributions:

- We present TCR to reconstruct an integral and accurate clothed human model by utilizing temporally consistent information even with challenging poses and partial occlusions.
- We improve the normal prediction network for the invisible human surface by exploiting the normal map of the visible surface, which predicts more details of the back normal map.

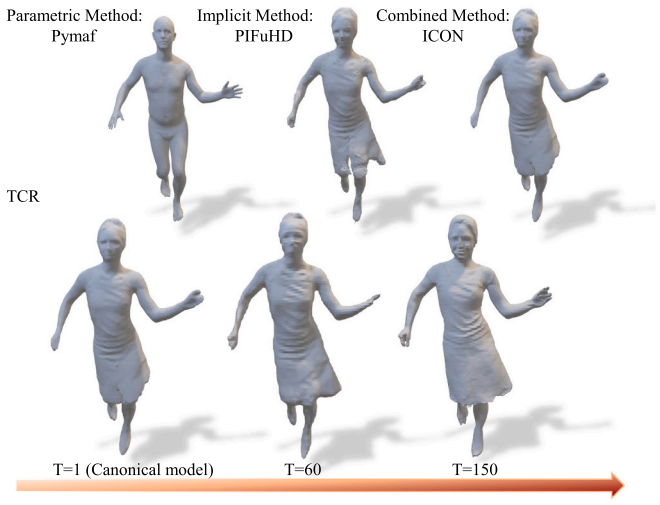

Fig. 1. Unlike existing single-frame methods such as parametric method [37], implicit method [28], and combined method [35], TCR gradually refines a canonical model by fusing the temporally consistent shape information of each frame (from left to right).

- We design a post-optimization algorithm that employs the predicted normal maps to refine the details of human models through orthogonality constraints.

The rest of this paper is organized as follows. In Section 2, we summarize the relevant work on 3D human reconstruction. Section 3 introduces the pipeline and details of our method. Section 4 presents some experiments to validate the effectiveness of our method and show relevant results. Finally, we conclude this paper and discuss future work in Section 5.

2. Related work

The field of 3D human reconstruction has been booming in recent years. Various representations for human models have been used by related methods, which can be divided into explicit representations (such as voxels [11] and triangle meshes [12]), and implicit representations (such as signed distance functions [38] and occupancy functions [27]). Currently, the explicit triangle meshes and implicit neural functions are two of the most popular representations.

2.1. Explicit mesh representation

In recent years, statistical body models [3–8] have been proposed that use a set of pose and shape parameters to control model deformation. These models are learned from 3D body scans, with the prior of human shape and pose. SCAPE [3] is a notable early study on deformable human model. It decouples the representation of human body into pose-dependent and individual shape-dependent triangle deformations. SMPL [5] is a vertex-based linear model for human body, which is compatible with existing rendering engines. The above models simply ignore hand joints and facial expressions which have a significant impact on human communication. SMPL-X [8] is an extension of SMPL by combining the MANO [10] hand model and the FLAME [9] face model. SMPL(-X) is the most widely used model because of its parameter specification and compatibility with many industrial applications.

Some parametric methods [12–15] learn parameters of statistical body models end-to-end and others [16–20] use intermediate features such as contour to enhance the supervision of learning. In addition, some methods [21–23] also consider temporal consistency and smoothness to boost the accuracy of parameter estimation. Although these methods can accurately estimate poses and naked human shapes, they are difficult to be applied to visual enhancement applications due to the lack of clothing details. To address this, some methods [24–26] add displacements to the vertices of the parametric model to represent the details. Alldieck et al. [25] estimates details in the UV-space through normal and displacement maps which are applied to SMPL model for rendering. Octopus [26] predicts pose-invariant shape and adds an offset parameter to the template model in T-pose space to represent details. Nonetheless, some large deformations such as long skirt are still not well reconstructed. Parametric methods can maintain the integrity of the reconstructed human body in any difficult case owing to strong prior constraints. In this paper, to constrain the implicit human reconstruction, we also use a parametric model to provide prior knowledge of the human body.

2.2. Implicit neural representation

Unlike explicit meshes, which can only represent limited details, implicit neural representations can be adapted to arbitrary topological shape. A 3D human model is typically represented by an implicit isosurface, whose the explicit surface is extracted by a preset level set. PIFu [27] applies pixel-aligned features and depth values as input to an implicit function represented by multi-layer perceptrons (MLPs). Saito et al. argue that pixel-aligned image features are more conducive to the function retaining local details than global features. PIFuHD [28]

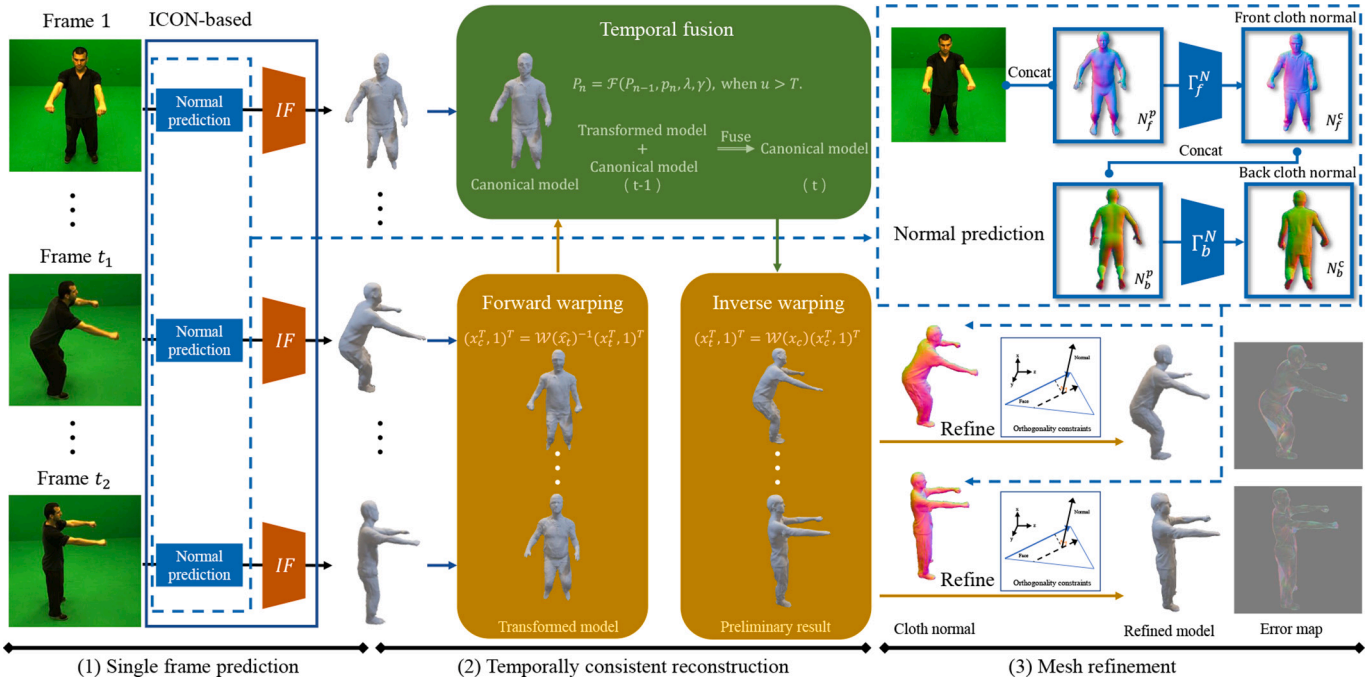

Fig. 2. The pipeline of TCR includes three stages: 1) single frame prediction, 2) temporally consistent reconstruction, and 3) mesh refinement. Firstly, the human body of each frame is predicted by an ICON-based [35] neural implicit function, where the normal prediction network is re-designed. Secondly, the predicted human model is transformed into the canonical space by a warp field. And the warped model of the first frame is selected as a canonical model. The canonical model can be updated through probabilistic fusion. Then a preliminary intact result can be obtained by inverse warping. Finally, the predicted normal maps are utilized to refine the result. The refined model is the final output of TCR.

introduces a two-level network architecture releasing the limitations of image resolution. The coarse level is identical to PIFu and observes the lower resolution images. The fine level uses higher resolution images as input and focuses on detailed geometry.

Thanks to the representation ability of implicit functions, above methods can represent clothing details of the human body. Nevertheless, these methods are prone to missing limbs and excess parts due to the absence of prior constraints. Since the pixel-aligned features cannot represent adequately spatial information in some views, human reconstruction are easily subject to depth ambiguity. Hence, some methods [31–33] employ parametric model to extract spatial features. Zheng et al. [34] propose to employ a parametric model to regularize the implicit function. They extract a voxel-aligned feature from the voxelized parametric model besides a pixel-aligned feature from the image. Xiu et al. [35] use SMPL body to guide detailed clothed-human surface normal prediction and visibility-aware implicit surface inference. Parametric models limit shape recovery away from the body, such as loose clothing. ECON [36] recovers 2.5D front and back detailed surfaces from normal maps using normal integration. The parametric model is used only as a guide to assist with front and back surface recovery and shape completion. These methods are highly dependent on the accuracy of the model parameter estimation. In the case of difficult poses or occlusions, the inaccurate estimation of the parametric model will still result in broken model.

To address this problem, we propose a method, called TCR, which uses temporally consistent shape information to ensure the integrity and accuracy of the reconstructed human model. Unlike the above methods, which only observe from a single frame, TCR combines the previous reconstruction in the same video to boost the quality of the current model. Following some previous work [38–40], we employs a canonical model in canonical space to hold the information of invariant human shape in each frame. As an ingenious method, ICON [35] takes full advantage of the representation ability of implicit functions and the human shape prior of parametric models. TCR enhances ICON's normal prediction network by replacing the original image features with features from the front normal map to enhance the reconstruction of back details. In addition, we use the normal maps to optimize the details of the reconstruction results.

3. Method

The three main stages of TCR are shown in Fig. 2: a) single frame prediction, b) temporally consistent reconstruction, and c) mesh refinement. We first predict a coarse model for a single frame using the implicit network (Section 3.1). In the stage of temporally consistent reconstruction (Section 3.2), we transform the coarse model into the canonical space, and fuse the transformed model with the preset canonical model to obtain the updated one. A preliminary intact result can then be obtained by inverse warping. In the mesh refinement phase (Section 3.3), we utilize normal maps to refine the preliminary result and obtain a refined mesh.

3.1. Single frame model

The single-frame model framework in our approach is based on ICON [35], whose framework contains two stages: a) parametric model based normal prediction and b) implicit 3D reconstruction. In the stage of normal prediction, parametric models are first predicted using PyMAF [37] or other human pose and shape (HPS) regressors. Then the front/back normal map $N^p = \{N_f^p, N_b^p\}$ of parametric model is obtained naturally. Finally, the clothed-body normal maps $N^c = \{N_f^c, N_b^c\}$ are predicted by two normal networks $\Gamma^N = \{\Gamma_f^N, \Gamma_b^N\}$ with SMPL-body normal maps N^p and original image I as input:

$$\Gamma^N(N^p, I) \rightarrow N^c. \tag{1}$$

The quality of the clothed-body normal map estimation directly affects the details of reconstruction results. As can be seen from Fig. 5 the prediction of normal maps on the back tends to be smooth and lacks details. We argue that features extracted from normal maps are more

direct than those extracted from images. Therefore, we use front normal map instead of original image in the prediction of back normal map:

$$\begin{aligned}&\Gamma_f^N(N_f^p, I)\rightarrow N_f^c,\\&\Gamma_b^N(N_b^p, N_f^c)\rightarrow N_b^c.\end{aligned} \tag{2}$$

In the stage of implicit 3D reconstruction, ICON employs a Multi-Layer Perceptron (MLP) as implicit neural network to estimate volumetric occupancy field. The local features F_x of the implicit function input are as follows:

$$F_x = \{F_d(x), F_n^p(x), F_n^c(x)\} \tag{3}$$

where F_d is the distance from the query point x to the surface of the SMPL mesh. F_n^p and $F_n^c$is a normal vector extracted from N^p and N^c respectively. Whether the normal map is N_f or N_b depends on the visibility of the point P. We choose N_f if visible and N_b otherwise.

Meanwhile, like ICON, a feedback loop is designed to alternately optimize the SMPL model and refine the normal map. The normal network can obtain human prior from SMPL model to help prediction of clothed-body normal maps. In turn, the SMPL model is optimized to achieve pixel-aligned fits by punishing the difference between the SMPL-body normal map N^p and the clothed-body normal map N^c.

3.2. Temporally consistent reconstruction

The overall body shape of the same person in different frames does not change with posture. To preserve the information of human shape for subsequent model reconstruction, we build a canonical model in canonical space and fuse the reconstruction results of each frame into the canonical model. Usually we set the single-frame result of first frame as the original canonical model. The reconstruction result is obtained by warping the canonical model back into observed space. As shown in the Fig. 3, the process involves three steps: forward warping, temporal fusion and inverse warping.

Forward and inverse warping. Since calculating the deformation of each voxel is very computative, we first construct a deformation node graph to drive the deformation of the entire model. The node of graph is attached to vertex of canonical model. We construct the node graph with sampling radius r and resample the node graph after each frame fusion. Refer to [38], the node graph at time t is defined as a set of n nodes:

$$\mathcal{G}_t = \{dn_l, dn_w, dn_{se3}\}_t, \tag{4}$$

where dn_l is the position of node in the canonical model, dn_w determines the effect of this node on the surrounding vertices and dn_{se3} is the associated transformation. A warp field is constructed by smooth interpolation through a k-nearest node average in the canonical space. The warp field is used to transform the human model between the canonical space and the observed space. In the warp field, the warp function is defined for each vertex using dual-quaternion blending $DQB(\cdot)$:

$$\mathscr{W}(v^i) = SE3(DQB(v^i)), \tag{5}$$

where $SE3(\cdot)$ converts from quaternions back to an SE(3) transformation matrix and v is vertex of the canonical model. $DQB(v)$ is the weighted average over dual quaternion transformations of the k-nearest nodes to the vertex v:

$$DQB(v^i) = \frac{\sum_1^k w_k(v^i) q_{ki}}{\| \sum_1^k w_k(v^i) q_{ki} \|}, \tag{6}$$

where $q \in \mathbb{R}^8$ is unit dual quaternion. The influence of a node on a vertex rest with the position of node dn_l and influencing factors dn_w:

$$w_k(v^i) = exp(-\|dn_l^k - v^i\|^2 \Big/ (2(dn_w^k)^2)). \tag{7}$$

We optimally solve for the parameters of warp function that transforms the canonical model M_c to the pose of single-frame model M_t at time t via following energy function:

$$E(M_t, M_c, \xi) = E_{data}(M_t, M_c) + E_{reg}(M_c, \xi), \tag{8}$$

where ξ is the set of edges connecting nodes. The energy function contains a data term E_{data} and a regularization term E_{reg}. We compute the ICP cost of two models in the data term to make the transformed model as consistent as possible with single-frame model. We compute under the Tukey loss function Ψ_{data}:

$$E_{data}(M_t, M_c) = \sum_{v \in M_t} \Psi_{data}(\hat{v} - v_c), \tag{9}$$

where v_c is the closest point to vertex v on the single-frame model. We employ warp function to transform vertex into observed space $\hat{v} = \mathscr{W}(v)$. In order to maintain the topological shape and smoothness of the transformed model, we use a regularization term to punish the non-smooth transformation. The transformation of the edges connecting the nodes should be rigid, so the regularization term is as follows:

$$E_{reg}(M_c, \xi) = \sum_{i=0}^{n} \sum_{j \in \xi(i)} \Psi_{reg}(q_i^{-1} dn_l^j - q_j^{-1} dn_l^j), \tag{10}$$

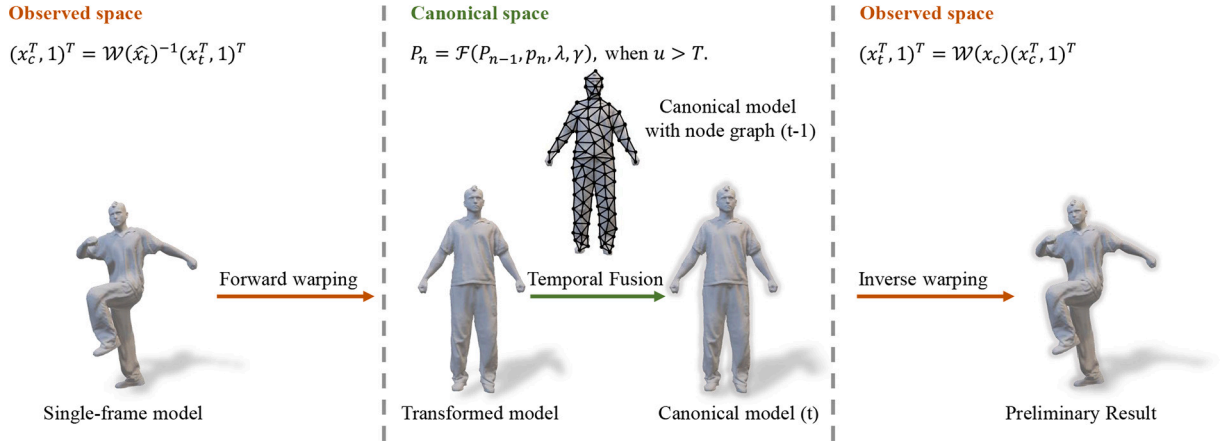

Fig. 3. Details of the temporally consistent reconstruction stage. Firstly, the single-frame model is transformed into the canonical space by a warp field. Then, the transformed model is fused with the canonical model at $(t-1)$. Finally, the updated canonical model at (t) is inversely warped into the observed space to obtain a preliminary result.

where $\xi(i)$ are all the nodes connected to node i and Ψ_{reg} is Huber penalty function. After solving for the parameters, the point x_c in canonical space can be transformed into observed space by warp function $(x_t^T, 1)^T = \mathscr{W}(x_c)(x_c^T, 1)^T$. Naturally, the point x_t in observed space can be converted into canonical space using the inverse of the warp function $(x_c^T, 1)^T = \mathscr{W}(\widehat{x}_t)(x_t^T, 1)^T$. The $\widehat{x}_t$ is the closest point to x_t on the transformed canonical model.

Temporal fusion. After transforming the model at time t into the pose of the canonical model through the warp field, we obtain the canonical model at time t by probability fusion between the transformed model and the canonical model at time $t-1$. We use an array $\mathscr{A} = \{P_n\}$ to store the occupancy probability of truncated volume. The value of $\mathscr{A}$ is set as 0 when corresponding voxel is non-occupied in canonical model. The transformed model is fused to canonical model by updating the occupation probability. For the original occupied voxels (OV) and the non-occupied voxels (NOV), we adopt different fusion strategies to ensure that the correct shape information is integrated:

$$\begin{aligned} P_n &= \mathscr{F}(P_{n-1}, p_n, \lambda, \gamma) \\ &= \begin{cases} \lambda P_{n-1} + (1-\lambda)p_n & OV \\ \gamma p_n & NOV \end{cases}, when\, T > \mu. \end{aligned} \tag{11}$$

The p_n is occupancy probability from the implicit function of the current frame. The parameter λ, γ controls the influence of the new occupancy probability on the original. The value of them determines whether the non-human voxel is excluded and whether the new occupied voxel is added or not. However, direct fusion will result in poor model detail due to alignment errors, especially in the original visible region. We argue that the prediction of the occupancy probability of visible voxels is more accurate than the invisible ones. Therefore, we set a confidence parameter μ to control the change of voxel occupancy state:

$$\mu = 1 - |tanh(\theta \cdot VSDF(x))|, \tag{12}$$

where $VSDF(x)$ is the value of signed distance function from the visible surface generated by SMPL model and θ is a hyper-parameter. When μ is less than the set threshold T, the occupancy probability of voxel x is not updated. This fusion strategy allows us to integrate the shape information of the model from each frame into the canonical model.

3.3. Mesh refinement

In the stage of temporally consistent reconstruction, we reconstruct a preliminary model with intact body structure. The model cannot adequately show the details of the normal maps. Therefore, we added a post-optimization step to improve the details of the reconstruction model using the clothed-body normal maps. By the orthogonality relationship between tangent vectors and normal vectors of the reconstruction model surface, the depth of the surface vertex in the front view is updated in a traversal way. The normal vector to pixel i in the clothed-body normal map is $\mathscr{N} = (\mathscr{N}_{ix}, \mathscr{N}_{iy}, \mathscr{N}_{iz})$ and the position of the corresponding vertex on the model surface is (X_i, Y_i, Z_i). The model surface vertex corresponding to the adjacent pixel point j is (X_j, Y_j, Z_j). From the orthogonality relation we can obtain the following equality:

$$((X_i, Y_i, Z_i) - (X_j, Y_j, Z_j)) \cdot \mathscr{N} = 0. \tag{13}$$

Then the updated depth value can be obtained:

$$Z_i = (\mathscr{N}_{ix}(X_j - X_i) + \mathscr{N}_{iy}(Y_j - Y_i) + \mathscr{N}_{iz}Z_j) / \mathscr{N}_{iz}. \tag{14}$$

With this equation, the preliminary model is refined pixel by pixel from top to bottom and left to right. However, there are wrong adjustment when the adjacent pixels do not correspond to the same body part. Therefore, we set a threshold μ to filter out these pixels. We ignore these pixels during the mesh refinement phase when the following two situations occur: 1) The difference between the depth values of vertex corresponding to adjacent pixels is greater than the threshold value μ. 2) The pixel of normal maps cannot be projected onto the surface of the model.

4. Experiments

In this section, we present a quantitative and qualitative evaluation of our method and compare TCR with state-of-the-art human reconstruction methods on two public benchmark datasets. All experiments are performed on an Intel Xeon Gold 6226R 2.9 GHz CPU and an NVIDIA GeForce RTX 3090 GPU.

4.1. Datasets

Training data. As training data we use THuman2.0 [41], which contains 500 realistic human models captured by a dense DSLR rig and provides corresponding parameters for SMPL(-X) [5,8]. All 3D models have corresponding texture maps, so the image can be generated from 36 different perspectives by model rendering. Meanwhile, all the methods [27,28,34,35] are retrained on this dataset for a fair comparison.

Testing data. To demonstrate the effectiveness of the proposed method, we conduct some experiments on Adobe [43] and MonoPerfCap [42]. The Adobe dataset is an indoor dataset with eight views, containing three subjects and ten action sequences. The videos were recorded in rooms covered with green cloth. To complement this, we use two sequences with accurate surface reconstruction from the Monopercap dataset as an evaluation on outdoor videos. The ground truth surfaces of these two sequences are derived from the work of [44,45]. They are reconstructed from multi-view images.

4.2. Evaluation

Metrics. We use three common metrics to evaluate the proposed method. Chamfer distance (**Chamfer**) measures the similarity between the sampled point clouds of the reconstruction meshes and the ground truth. For each point in the point cloud, this metric finds the nearest point in another point cloud, and sums the square of the distance. Point-to-surface Euclidean distance (**P2S**) is the average Euclidean distance from the vertices on the reconstruction mesh to the ground-truth surface. The previous two metrics can only reflect the quality of the general shape, not the quality of local details. Normal reprojection error (**L2-norm**) indicates the fineness of local details by calculating the L2 error between the normal map of the reconstruction meshes and the ground truth.

TCR vs. SOTA. TCR makes full use of temporal information and normal maps to boost the quality of the clothed human reconstruction. Table 1 shows the performance of our method on Adobe and MonoPerfcap datasets. It is also compared with several well-known single-frame methods, such as PIFu [27] and ECON [36]. We notice that TCR outperforms the state-of-the-art methods on all three metrics. Furthermore, we evaluate the performance of our method qualitatively and compare it with these methods in Fig. 4. We can see that compared to ICON and TCR, the reconstruction results of PIFuHD [28] are often broken and often contain non-human parts, due to the lack of prior constraints. Meanwhile, thanks to the temporal information and the detailed information from normal maps, TCR is superior to the other two methods in terms of details and completeness.

Ablation experiments. Meanwhile, to explore the effect of Step II (Section 3.2) and Step III (Section 3.3) on the overall performance, two sets of ablation experiments are performed. In the first experiment, we only refine the single-frame result with normal maps (w/o II). In the second experiment, we use only the temporally consistent information to obtain the intact human model (w/o III). It can be seen from the experimental results in Table 1 that all of them have a significant effect

Table 1
Quantitative evaluation (cm) of our method and previous work using three metrics: point-to-surface Euclidean distance, chamfer distance and normal re-projection error. Besides, the performance without Step II or Step III is also shown below.

	Adobe			MonoPerfCap		
	P2S ↓	Chamfer ↓	L2-norm ↓	P2S ↓	Chamfer ↓	L2-norm ↓
PIFu [27]	3.856	3.256	0.199	3.596	2.956	0.177
PIFuHD [28]	3.689	3.102	0.196	3.355	2.785	0.175
PaMIR [34]	3.125	2.896	0.181	2.865	2.262	0.139
ICON [35]	2.572	2.389	0.178	1.977	1.747	0.132
ECON [36]	2.316	2.351	0.152	1.842	1.627	0.123
TCR(w/o II)	2.549	2.377	0.162	1.935	1.706	0.125
TCR(w/o III)	2.359	2.268	0.155	1.756	1.638	0.121
TCR	**2.261**	**2.178**	**0.127**	**1.717**	**1.598**	**0.108**

on the performance improvement of the whole method. The Step II ensures the integrity of the reconstruction mannequin and the Step III refines the details of the model.

Normal prediction. In Section 3.1, in order to enhance the back detail prediction of normal network, we replace the original features I with front normal map features N_f^c. In Table 2 and Fig. 5, we compare the performance of the normal networks before and after replacing image features with front normal map features. In Table 2, the metric is the L2 error between the clothed-body normal map and the ground truth generated by model rendering. From Table 2 we can see that the error is reduced by using the front normal map. As shown in Fig. 5, the network can learn more details of the back more directly from the front normal maps.

Temporal Fusion. As discussed in Section 3.2, the parameter λ, γ controls the influence of the new occupancy probability on the original and directly determines the quality of the canonical model after fusion. In Table 3, we show the influence of different values of the fusion parameters on the overall performance of our method. We preset several sets of parameter values for λ, γ and fix one parameter when another changes. As shown in Fig. 7, the tendency of performance to change with parameter values is approximated by a parabola with an upward opening and the parameters $\lambda = 0.70, \gamma = 0.80$ are optimal. When λ is too large, the dynamic shape of the current frame cannot be fused into the model. While λ is too small, useful shape information cannot be completely retained. Similarly, if γ is too large, some non-human parts will be included in the canonical model. In Fig. 8, the reconstruction results on in-the-wild videos from the MonoPerfCap dataset [42] are shown in canonical space. With the fusion of the shape information from the video frame, the integrity and accuracy of the mannequin is progressively improved. The improvement in the quality of our reconstruction results depends on the new shape information constantly provided in the video frame.

Robustness to hard cases. With the help of parametric models, while some implicit methods such as ICON [35] are able to keep the mannequin intact in most cases, the model is still prone to the loss of limbs when the parameters of parametric models is misestimated. Fig. 6 shows the performance of our method and ICON on some examples of difficult

Table 2
Comparison of normal map estimation results between image features (I) and front normal map features (NF). The metric is the L2 error of the clothed-body normal map.

	L-NF ↓	L-NB ↓
I + SMPL	0.155	0.213
NF + SMPL	**0.151**	**0.192**

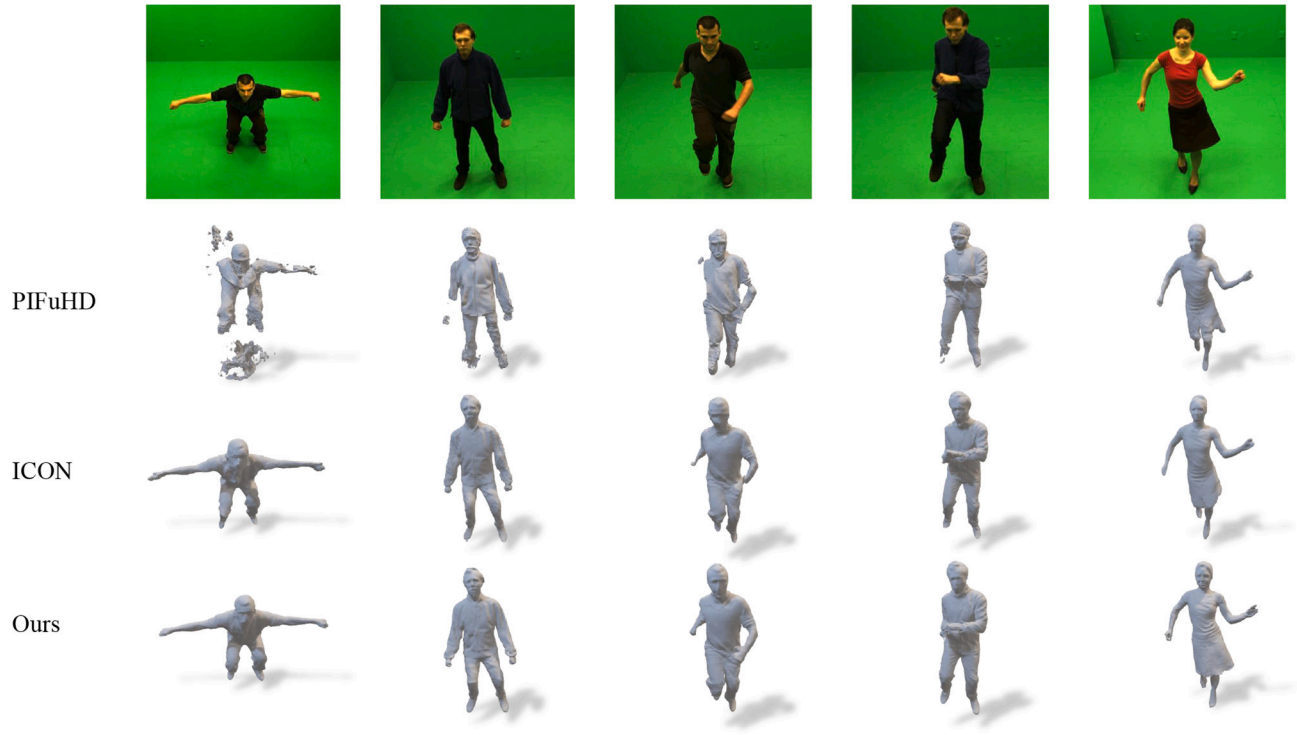

Fig. 4. Qualitative comparison of our method with the previous methods, PIFuHD [28] and ICON [35]. The images are selected from the Adobe dataset [43]. PIFuHD tends to produce broken structures due to the lack of a prior knowledge for human body. Meanwhile, TCR is superior to the other two methods in terms of the integrity and accuracy of the reconstruction results.

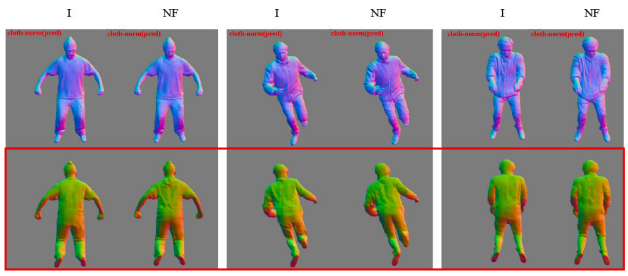

Fig. 5. Comparison on the predicted normal maps with different features, i.e. the input image (I) or the front normal map (NF), fed into the back normal map prediction network.

poses and partial occlusions. These difficult examples are taken from the Adobe dataset. As can be seen from the circled part of the reconstruction model, compared to ICON, TCR can preserve the integrity of human body by using temporally consistent information in these hard cases.

Limitations. TCR has two main limitations: (1) As shown in Table 4, the proposed method has a longer run time compared to other methods, such as the latest ECON [36], due to the optimization process of solving the warp field. The time is the average run time of each method on the Adobe dataset [43], excluding rendering. (2) The accuracy of our method requires video input, and the Step II of TCR cannot play a role for a single frame image. TCR relies on the information provided by human motion in the video, and reconstruction accuracy is limited when the person in the video is stationary or has little movement.

5. Conclusions and discussion

In this paper we present a method, called TCR, for maintaining the integrity and accuracy of human models by fusing the temporally consistent information. TCR includes three main stages: single frame prediction, temporally consistent reconstruction and mesh refinement. Considering that the general shape of a person is changeless over time, we utilize the shape information from previous frames to reconstruct the current human model. Unlike previous single-frame based methods, we fuse the shape information of the previous reconstruction results into a canonical model. In addition, we employ the normal maps to refine the details of human models.

We evaluate the performance of TCR qualitatively and quantitatively on the Adobe and MonoPerfCap datasets. The results show that our method outperforms the state of the art. Experiments on difficult cases such as extreme poses and partial occlusions show that TCR can maintain the integrity and accuracy of human models even when the pose estimation is inaccurate. For practical applications, TCR is very time-consuming due to the optimization process. The issue of the reconstruction speed needs to be further studied in the future. Meanwhile, a large-scale dataset with high quality videos and ground-truth human body meshes is desired to promote the development of the field in temporal human surface reconstruction.

Table 3
Performance comparison of the fusion strategies with different parameters λ, γ. The best set of parameters are highlighted.

	P2S ↓	Chamfer ↓	L2-norm ↓
$\lambda = 0.70, \gamma = 0.70$	2.566	2.346	0.151
$\lambda = 0.70, \gamma = 0.75$	2.410	2.221	0.140
$\lambda = 0.70, \gamma = 0.90$	2.398	2.215	0.135
$\lambda = 0.70, \gamma = 0.85$	2.366	2.188	0.134
$\lambda = 0.70, \gamma = 0.80$	**2.261**	**2.178**	**0.127**
$\lambda = 0.65, \gamma = 0.80$	2.411	2.228	0.136
$\lambda = 0.60, \gamma = 0.80$	2.585	2.366	0.155
$\lambda = 0.75, \gamma = 0.80$	2.365	2.196	0.132
$\lambda = 0.80, \gamma = 0.80$	2.465	2.257	0.146

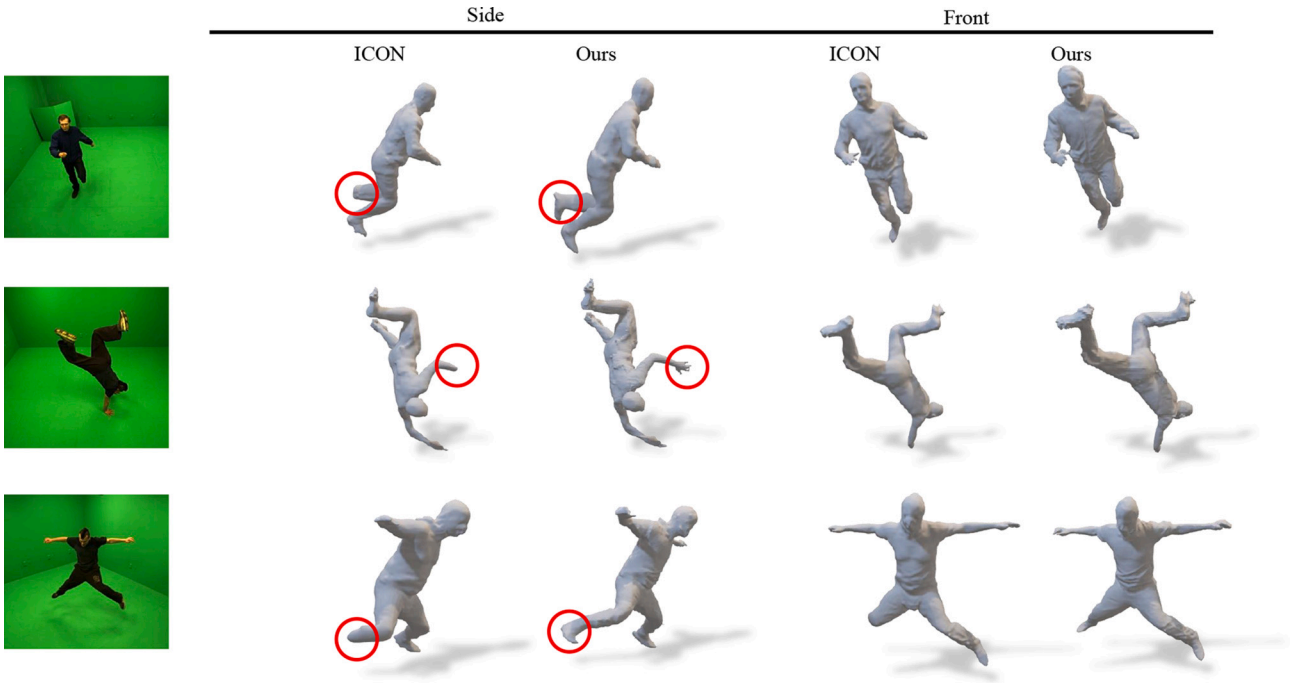

Fig. 6. Qualitative comparison on examples with difficult poses and partial occlusions. Compared to ICON [35], which tends to lose limbs due to the mis-estimation of SMPL [5] model, TCR maintains the integrity and accuracy of the reconstruction.

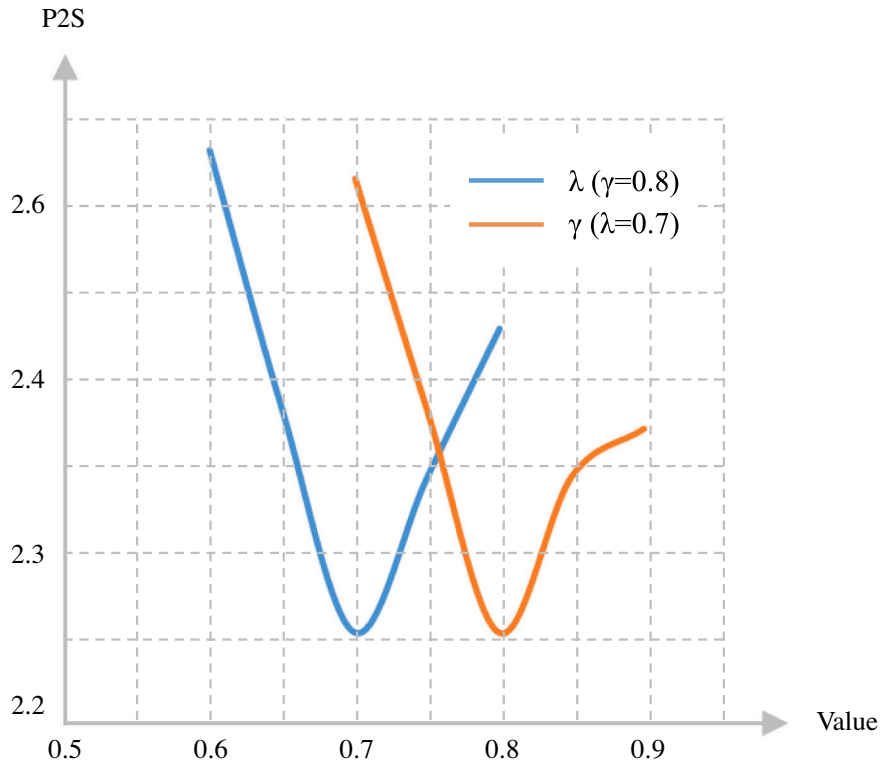

Fig. 7. The influence of fusion parameters λ and γ on the final performance. The ordinate represents the P2S metric on the Adobe dataset.

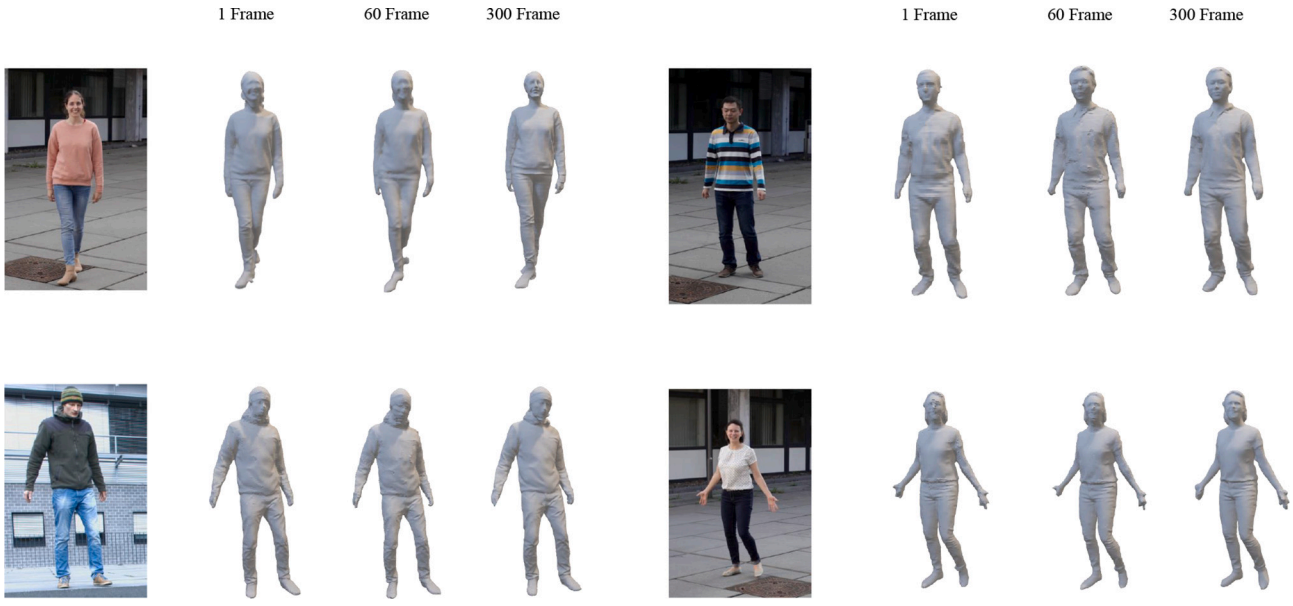

Fig. 8. The performance of TCR on in-the-wild videos from the MonoPerfCap dataset [42]. We show all the reconstruction results of video frames in canonical space (fixed pose) for comparison. The quality of the human model is improved by the fusion of shape information from the video frames.

Statement

We have taken account of the referee comments in preparing the revision. We added evaluation experiments according to reviewer #1 and highlighted the limitations of this paper. Meanwhile, following reviewer #2's suggestion, we have better summarised the abstract and adapted the citation format. We thank all reviewers and editors for their careful review and detailed feedback.

CRediT authorship contribution statement

Yong Deng: Conceptualization, Methodology, Software, Investigation, Writing-original-draft, Visualization. **Baoxing Li:** Validation, Writing-review-editing. **Yehui Yang:** Validation, Writing-review-

Table 4
The average run time of the proposed method and the early method for mesh reconstruction on the specified testbed. The time is the average run time on the Adobe dataset [43], excluding rendering.

Method	PIFu [27]	PIFuHD [28]	PaMIR [34]	ICON [35]	ECON [36]	TCR
Time	5.8s	23.8s	8.1s	24.9s	26.5s	47.6s

editing. **Xu Zhao:** Resources, Writing-review-editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Z.-Q. Cheng, Y. Chen, R.R. Martin, T. Wu, Z. Song, Parametric modeling of 3D human body shape—A survey, Comput. Graph. 71 (2018) 88–100.
[2] L. Chen, S. Peng, X. Zhou, Towards efficient and photorealistic 3d human reconstruction: a brief survey, Vis. Inform. 5 (2021) 11–19.
[3] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, J. Davis, SCAPE: shape completion and animation of people, ACM Trans. Graph. 24 (2005) 408–416.
[4] G. Pons-Moll, J. Romero, N. Mahmood, M.J. Black, Dyna: a model of dynamic human shape in motion, ACM Trans. Graph. 34 (2015) 1–14.
[5] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: a skinned multi-person linear model, ACM Trans. Graph. 34 (2015) 248:1–248:16.
[6] A.A.A. Osman, T. Bolkart, M.J. Black, STAR: Sparse Trained Articulated Human Body Regressor, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 598–613.
[7] H. Xu, E.G. Bazavan, A. Zanfir, W.T. Freeman, R. Sukthankar, C. Sminchisescu, GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020, pp. 6183–6192.
[8] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A. Osman, D. Tzionas, M.J. Black, Expressive Body Capture: 3D Hands, Face, and Body From a Single Image, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10967–10977.
[9] T. Li, T. Bolkart, M.J. Black, H. Li, J. Romero, Learning a model of facial shape and expression from 4D scans, ACM Trans. Graph. 36 (2017) 194:1–194:17.
[10] J. Romero, D. Tzionas, M.J. Black, Embodied hands: modeling and capturing hands and bodies together, ACM Trans. Graph. 36 (2017) 245:1–245:17.
[11] Z. Zheng, T. Yu, Y. Wei, Q. Dai, Y. Liu, DeepHuman: 3D Human Reconstruction From a Single Image, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7738–7748.
[12] A. Kanazawa, M.J. Black, D.W. Jacobs, J. Malik, End-to-End Recovery of Human Shape and Pose, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7122–7131.
[13] J. Liang, M. Lin, Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4351–4361.
[14] N. Kolotouros, G. Pavlakos, M.J. Black, K. Daniilidis, Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2252–2261.
[15] Z. Wan, Z. Li, M. Tian, J. Liu, S. Yi, H. Li, Encoder-Decoder With Multi-Level Attention for 3D Human Shape and Pose Estimation, in: ICCV, 2021, pp. 13033–13042.
[16] G. Pavlakos, L. Zhu, X. Zhou, K. Daniilidis, Learning to Estimate 3D Human Pose and Shape From a Single Color Image, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 459–468.
[17] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, B. Schiele, Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation, in: 2018 International Conference on 3D Vision (3DV), IEEE, Verona, 2018, pp. 484–494.
[18] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, B. Schiele, Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation, in: 2018 International Conference on 3D Vision (3DV), IEEE, Verona, 2018, pp. 484–494.
[19] G. Pavlakos, N. Kolotouros, K. Daniilidis, TexturePose: Supervising Human Mesh Estimation With Texture Consistency, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 803–812.
[20] G. Moon, K.M. Lee, I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 752–768.
[21] M. Kocabas, N. Athanasiou, M.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, 2020, pp. 5253–5263.
[22] H. Choi, G. Moon, J.Y. Chang, K.M. Lee, Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1964–1973.
[23] S. Guan, J. Xu, Y. Wang, B. Ni, X. Yang, Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10472–10481.
[24] H. Zhu, X. Zuo, S. Wang, X. Cao, R. Yang, Detailed Human Shape Estimation From a Single Image by Hierarchical Mesh Deformation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4486–4495.
[25] T. Alldieck, G. Pons-Moll, C. Theobalt, M. Magnor, Tex2Shape: Detailed Full Human Body Geometry From a Single Image, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 2293–2303.
[26] T. Alldieck, M. Magnor, B.L. Bhatnagar, C. Theobalt, G. Pons-Moll, Learning to Reconstruct People in Clothing From a Single RGB Camera, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 1175–1186.
[27] S. Saito, Z. Huang, R. Natsume, S. Morishima, H. Li, A. Kanazawa, PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2304–2314.
[28] S. Saito, T. Simon, J. Saragih, H. Joo, PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 81–90.
[31] Z. Huang, Y. Xu, C. Lassner, H. Li, T. Tung, ARCH: Animatable Reconstruction of Clothed Humans, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: pp. 3090–3099.
[32] T. He, Y. Xu, S. Saito, S. Soatto, T. Tung, ARCH++: Animation-Ready Clothed Human Reconstruction Revisited, in: ICCV, 2021, pp. 11046–11056.
[33] H. Kim, H. Nam, J. Kim, J. Park, S. Lee, LaplacianFusion: Detailed 3D Clothed-Human Body Reconstruction, ACM Trans. Graph. 41 (2022) 216:1–216:14.
[34] Z. Zheng, T. Yu, Y. Liu, Q. Dai, PaMIR: Parametric Model-Conditioned Implicit Representation for Image-based Human Reconstruction, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2021) 3170–3184.
[35] Xiu, Y., Yang, J., Tzionas, D., Black, M.J., 2022. ICON: Implicit Clothed humans Obtained from Normals, in: CVPR. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13286–13296.
[36] Xiu, Y., Yang, J., Cao, X., Tzionas, D., Black, M.J., 2023. ECON: Explicit Clothed humans Optimized via Normal integration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 512–523.
[37] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, Z. Sun, PyMAF: 3D Human Pose and Shape Regression With Pyramidal Mesh Alignment Feedback Loop, in: ICCV, 2021, pp. 11446–11456.
[38] R.A. Newcombe, D. Fox, S.M. Seitz, DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 343–352.
[39] Sida Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, H. Bao, Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies, in: ICCV, 2021, pp. 14314–14323.
[40] Z. Su, L. Xu, Z. Zheng, T. Yu, Y. Liu, L. Fang, Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer, 2020, pp. 246–264.
[41] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, Y. Liu, Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, 2021, pp. 5742–5752.
[42] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, C. Theobalt, Monoperfcap: Human performance capture from monocular video, ACM Trans. Graph. (ToG) 37 (2018) 1–15.
[43] D. Vlasic, I. Baran, W. Matusik, J. Popović, Articulated mesh animation from multi-view silhouettes, in: ACM SIGGRAPH 2008 Papers, 2008, pp. 1–9.
[44] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, C. Theobalt, Model-based outdoor performance capture, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 166–175.
[45] C. Wu, C. Stoll, L. Valgaerts, C. Theobalt, On-set performance capture of multiple actors with a stereo camera, ACM Trans. Graph. (TOG) 32 (2013) 1–11.