POSEFusion: Pose-guided Selective Fusion for Single-view Human Volumetric Capture

Zhe Li[1], Tao Yu[1], Zerong Zheng[1], Kaiwen Guo[2], Yebin Liu[1]
[1]Department of Automation, Tsinghua University, China [2]Google, Switzerland

Abstract

*We propose **PO**se-guided **SE**lective **Fusion** (POSEFusion), a single-view human volumetric capture method that leverages tracking-based methods and tracking-free inference to achieve high-fidelity and dynamic 3D reconstruction. By contributing a novel reconstruction framework which contains pose-guided keyframe selection and robust implicit surface fusion, our method fully utilizes the advantages of both tracking-based methods and tracking-free inference methods, and finally enables the high-fidelity reconstruction of dynamic surface details even in the invisible regions. We formulate the keyframe selection as a dynamic programming problem to guarantee the temporal continuity of the reconstructed sequence. Moreover, the novel robust implicit surface fusion involves an adaptive blending weight to preserve high-fidelity surface details and an automatic collision handling method to deal with the potential self-collisions. Overall, our method enables high-fidelity and dynamic capture in both visible and invisible regions from a single RGBD camera, and the results and experiments show that our method outperforms state-of-the-art methods.*

1. Introduction

Human volumetric capture, due to their potential value in holographic communication, online education, games and the movie industry has been a popular topic in computer vision and graphics for decades. Multi-view camera array methods [4, 11, 28, 5, 29, 49, 32, 9, 37, 21] can achieve high-fidelity human volumetric capture using multiple RGB or depth sensors but suffer from sophisticated equipment or run-time inefficiency, which limits their application deployment. In contrast, single-view human volumetric capture [23, 59, 12, 34, 52, 14, 38, 39, 15, 25, 44] has attracted more and more attention for its convenient setup.

Current methods for single-view human volumetric capture can be roughly classified into two categories: tracking-based methods and tracking-free ones. Tracking-based methods utilize a pre-scanned template [23, 12, 14, 15] or continuously fused mesh [34, 52, 44] as the reference model, and solve or infer the deformation of the reference model parameterized by embedded skeletons [50, 3, 47], node graph [23, 34], parametric body models (e.g., SMPL [30]) [57], or a combination of them [51, 52, 14, 15]. In these tracking-based methods, previous observations are integrated into the current frame after calculating the deformations across frames, thus plausible geometric details are preserved in the invisible regions. In addition, the reconstructed models are temporally continuous thanks to the frame-by-frame tracking. However, none of their deformation representations, neither skeletons nor node graph, is able to describe topological changes or track extremely large non-rigid deformations (Fig. 7(c)), which is an inherent drawback of the tracking-based methods.

Figure 1. High-fidelity and dynamic results reconstructed using our method.

On the other end of the spectrum, tracking-free methods [46, 33, 10, 42, 58, 1, 56, 38] mainly focus on geometric and/or texture recovery from a single RGB(D) image. By learning from a large amount of 3D human data, these methods demonstrate promising human reconstruction with high-fidelity details in visible regions and plausible shape in invisible areas [38, 39]. As the reconstruction for the current frame is independent from the previous frames, these methods can easily handle topological changes. However, their results may deteriorate in the cases of challenging human poses and/or severe self-occlusions. Besides, the reconstructions in the invisible regions are usually oversmoothed for the lack of observations (Fig. 7(d)). What's worse, tracking-free methods are incapable of generating temporally continuous results when applied on video inputs.

By reviewing the advantages and the drawbacks of these two types of methods, it is easy to notice that tracking-based methods and tracking-free inference are naturally complementary as shown in Tab. 1. A straightforward way is to

combine both branches by integrating the inferred models of all the other frames in the monocular RGBD sequence into the current frame to recover the invisible regions. The benefits of such a pipeline are: a) topological changes and large deformations can be accurately reconstructed using tracking-free inference directly, b) the invisible surfaces can be faithfully recovered by integrating the other frames into current frame, and finally c) temporal continuity is guaranteed by tracking the whole sequence frame-by-frame. However, the afore-mentioned pipeline still has limitations. Specifically, if we fuse all the other frames indiscriminately, we can only generate static surfaces with all the dynamic changing details averaged together. Moreover, it remains difficult for such a pipeline to handle the artifacts caused by self-collisions. To this end, we further propose POse-guided SElective Fusion (POSEFusion), a novel pipeline that contains pose-guided keyframe selection and adaptive implicit surface fusion. In this pipeline, we only integrate the keyframes selected by our proposed pose-guided metric, which takes into account both visibility complementarity and pose similarity.

Our key observations are: a) keyframes with similar poses to the current frame enable the recovery of physically plausible dynamic details, b) keyframes with complementary viewpoints to the current frame avoid to oversmooth the visible regions, and c) the adaptive fusion considers depth observations and visibility promotes to preserve the surface details and resolves collision artifacts. Based on these observations, the limitations of the simple pipeline are successfully overcome.

Specifically, we start with SMPL [30] tracking for all the frames given a monocular RGBD sequence as input. We then utilize the SMPL model as a robust and effective proxy and propose a novel criterion to quantify pose similarity and visibility complementarity. Based on this criterion, we select appropriate keyframes for each frame. Note that per-frame keyframe selection cannot guarantee the temporal continuity of invisible details; therefore, we further formulate the selection as a dynamic programming of min-cost path to reconstruct dynamic and temporally continuous invisible details. In implicit surface fusion, we propose an adaptive blending weight which considers depth and visibility information to avoid oversmooth fusion and preserve the observed details. Finally, we propose an automatic collision handling scheme to deal with possible self-collisions while maintaining adjacent details.

In summary, this paper proposes the following technical contributions:

- A new human volumetric capture pipeline that leverages tracking-based methods and tracking-free inference, and achieves high-fidelity and dynamic reconstruction in both visible and invisible regions from a single RGBD camera (Sec. 3.2).
- A new pose-guided keyframe selection scheme that considers both pose similarity and visibility complementarity and enables detailed and pose-guided reconstruction in the invisible regions (Sec. 4.2).
- A robust implicit surface fusion scheme that involves an adaptive blending weight conditioned by depth observations and visibility, and an automatic collision handling method which considers an adjacent no-collision model into the fusion procedure to maintain the adjacent details while eliminating collision artifacts (Sec. 4.3).

Building on these novel techniques, POSEFusion is the first single-view approach that is able to capture high-fidelity and dynamic details in both visible and invisible regions. Given a monocular RGBD sequence as input, our method is able to produce compelling human reconstruction results with complete, dynamic, temporally continuous, and high-fidelity details. The experimental results prove that our method outperforms state-of-the-art methods.

2. Related Work

2.1. Tracking-based Human Reconstruction

Some works in tracking-based methods utilize a pre-scanned person-specific model as a template, and deform it to fit with depth input of each frame. Especially, for human reconstruction, Gall *et al.* [11] and Liu *et al.* [29] modeled body motion by skeletons embedded in the template. Besides skeletal motion, embedded deformation graph [45] is an alternative parameterization method for non-rigid reconstruction. Li *et al.* [23] solved the warp field modeled by [45] and reconstructed detailed 3D geometric sequences from a single-view depth stream. Zollhöfer *et al.* [59] enabled real-time performance for general non-rigid tracking based on the parallelism of GPU. Guo *et al.* [12] introduced a L_0-based regularizer to implicitly constrain articulated motion. LiveCap [14] utilized a person-specific template and achieved real-time monocular performance capture. On another branch, volumetric fusion methods replaced the pre-scanned template with a continuously fused model for online incremental reconstruction. The pioneering work KinecFusion [17] reconstructed a rigid scene incrementally by using a commercial RGBD camera. The following work [35, 8, 20] focused on memory cost, geometry, and texture quality for rigid scene reconstruction, respectively. DynamicFusion [34] extended [17] and introduced a dense non-rigid warp field for real-time non-rigid reconstruction. The following work [16, 40, 13, 51, 41, 22, 52, 54, 44] incorporated different cues for more robust and accurate reconstruction. SimulCap [53] combined cloth simulation into the fusion pipeline but the quality of invisible details suffered from a simple cloth simulator. LiveCap [14] and DeepCap [15] respectively solved or regressed the skeleton and non-rigid motion of a person-specific template from a monocular RGB video. TexMesh [57] deformed SMPL [30] to generate a parametric coarse mesh and generate high-quality but static texture from a single-view RGBD video. However, because of the requirement of deforming a reference model,

Methods		Topological Change	Natural Deformation	Details in Invisible Regions	Temporal Continuity	Texture
Tracking-based	DoubleFusion [52]	✗	✗	✓(Static)	✓	None
	RobustFusion [44]	✗	✗	✓(Static)	✓	✓(Low-quality)
	TexMesh [57]	✗	✗	✓(Low-quality)	✓	✓(Static)
Tracking-free	PIFu [38]/PIFuHD [39]	✓	✓	✗	✗	✓(Low-quality)
	Ours	✓	✓	✓ (Dynamic, High-quality)	✓	✓ (Dynamic, High-quality)

Table 1. Comparison of our method with other state-of-the-art works when applying a monocular RGBD video as input. Our method inherits all the advantages of tracking-based and tracking-free methods while avoiding their drawbacks. Moreover, our method can reconstruct dynamic pose-guided geometric details in both visible and invisible regions.

all these methods cannot handle topological change and reconstruct extremely non-rigid deformations.

2.2. Tracking-free Human Inference

Recently, more and more works focused on single RGB(D) image reconstruction because of the rise of deep learning. [36, 18, 19, 27, 48] regressed the pose and shape parameters of a human parametric model (e.g., SMPL [30]) from a single image. Moreover, to address the challenge of general clothed human reconstruction, recent work tackled this problem by multi-view silhouettes [33], depth maps [10, 42], template deformation [58, 1], volumetric reconstruction [46, 56] and implicit function [38]. DeepHuman [56] conditioned single image reconstruction on the parametric SMPL model [30] to address the problem of challenging poses. PaMIR [55] combined implicit function [38] with convoluted SMPL feature for more robust and accurate inference. PIFuHD [39] extended PIFu [38] to a coarse-to-fine framework and demonstrated detailed geometric results learned from high-resolution single-view RGB images. Li *et al.* [25] accelerated PIFu [38] to achieve monocular real-time human performance capture. But all these methods focus on single frame reconstruction, but ignore temporal continuity and lack details in the invisible region.

3. Overview

3.1. Preliminaries

We firstly introduce the parametric body model [30] and the occupancy inference network adopted in this paper.

Parametric Body Model The parametric body model SMPL [30] is a function that maps the pose parameters $\theta \in \mathbb{R}^{75}$ and shape parameters $\beta \in \mathbb{R}^{10}$ to a human mesh with $N = 6890$ vertices:

$$\begin{aligned} T(\beta,\theta) &= \bar{\mathbf{T}} + B_s(\beta) + B_p(\theta), \\ M(\beta,\theta) &= W(T(\beta,\theta), J(\beta), \theta, \mathbf{W}), \end{aligned} \tag{1}$$

where $W(\cdot)$ is a skinning function that takes T-pose model $T(\theta,\beta)$, pose parameters θ, joint positions $J(\beta)$ and skinning weights $\mathbf{W}$ as input, and returns the posed model $M(\beta,\theta)$, and $T(\theta,\beta)$ is an individual model in T-pose with shape and pose based offsets ($B_s(\beta)$ and $B_p(\theta)$).

Occupancy Inference Network The occupancy value $\varphi(\mathbf{x})$ of a 3D point $\mathbf{x}$ is an occupancy probability of the point inside the 3D object, and the mapping function $\varphi : \mathbb{R}^3 \rightarrow \mathbb{R}$ is an implicit function. Recently, Saito *et al.* [38] conditioned the implicit function on an image-encoded feature and proposed the pixel-aligned implicit function (PIFu):

$$\varphi(\mathbf{x}; I) = f(G_I(\mathbf{x}_{2\mathrm{D}}), \mathbf{x}_z), \tag{2}$$

where $\mathbf{x}_{2\mathrm{D}} = \pi(\mathbf{x})$ is the 2D projection of $\mathbf{x}$, I is the conditional image, G_I is a feature map of I encoded by a deep image encoder, $G_I(\mathbf{x}_{2\mathrm{D}})$ represents the sampled feature vector of $\mathbf{x}_{2\mathrm{D}}$ on G_I, $\mathbf{x}_z$ is the depth value of $\mathbf{x}$, and $f(\cdot)$ is a mapping function represented by multi-layer perceptrons (MLP). Based on this pixel-level representation, PIFu can reconstruct high-fidelity details in visible regions of the object from the conditional image I. We improve on the original PIFu network [38] by incorporating the feature encoded from the depth input to remove the depth ambiguity.

3.2. Main Pipeline

Our goal is to reconstruct temporally continuous human models with high-quality dynamic geometric details and texture from a single-view RGBD video. At first, to construct the body motion among frames, we track the SMPL model [30] using the whole depth sequence. For the current frame, we allocate a volume which contains the current SMPL model. We suppose that the true body surface in the current frame is near the current SMPL model, so we select valid voxels (points) around SMPL without processing redundant invalid points. Our main idea is to warp these valid points to each keyframe and fetch the corresponding occupancy values, and finally fuse a complete model with high-fidelity details. Then our system performs the following 3 steps sequentially as shown in Fig. 2.

1. **Pose-guided Keyframe Selection** (Sec. 4.2): To enable dynamic and high-fidelity reconstruction in the invisible region, we propose a pose-guided keyframe selection scheme that considers both visibility complementarity and pose similarity. We calculate the two metrics relative to the current frame using the tracked SMPL models, then select keyframes which not only contain the invisible regions of the current frame but are also as similar as the current SMPL pose. The pose-guided keyframe selection is further formulated as a dynamic programming to guarantee high-fidelity, dynamic and temporally continuous details in both visible and invisible regions.
2. **Implicit Surface Fusion** (Sec. 4.3): After the keyframe selection, the selected valid points are warped to each

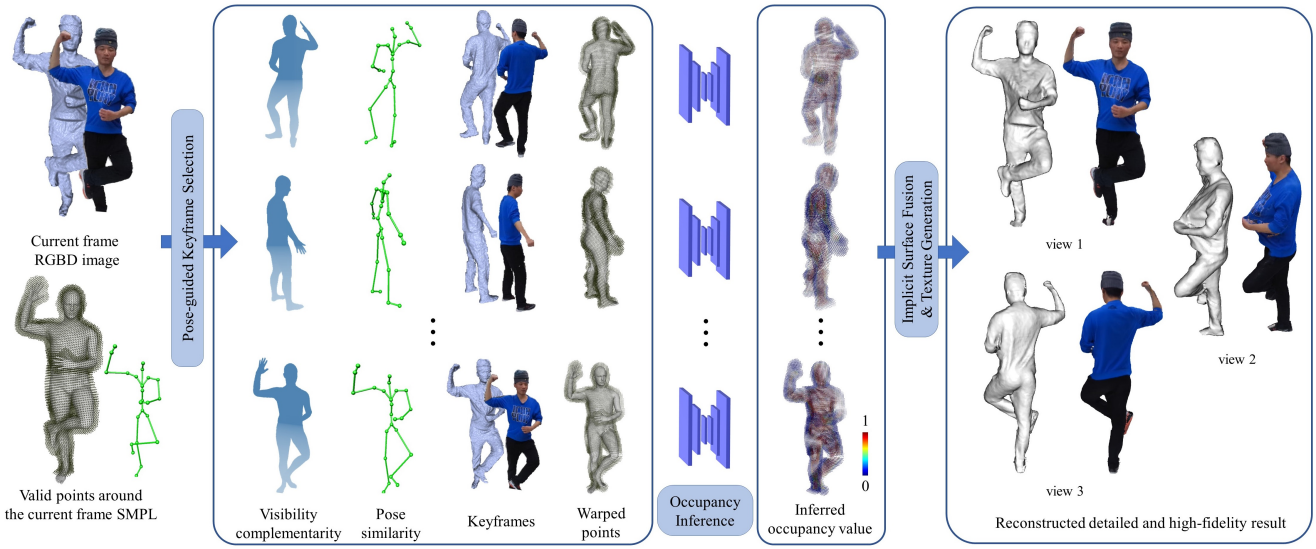


Figure 2. Reconstruction pipeline. Firstly we perform the pose-guided keyframe selection scheme to select appropriate keyframes by the visibility complementarity and pose similarity. Valid points around the current SMPL are then deformed to each keyframe by SMPL motion, and fed into a neural network with the corresponding RGBD image. The neural network infers occupancy values of each keyframe, and then we integrate all the inferred values to generate a complete model with high-fidelity and dynamic details. Finally, a high-resolution texture map is generated by projecting the reconstructed model to each keyframe RGB image and inpainted by a neural network.

keyframe by SMPL motion, and then fed into a neural network to infer occupancy values which indicate the surface location contributed by this keyframe. However, the inferred values may be inaccurate. We therefore design an adaptive blending weight as the confidence and integrate occupancy values of each keyframe into the current frame to preserve high-fidelity surface details in both visible and invisible regions and guarantee smooth transition on the fusion boundaries. Moreover, if the collision occurs among different body parts, we perform the collision handling to eliminate the collision artifacts while maintaining the adjacent geometric details.

3. **Texture Generation** (Sec 4.4): Finally, a high-resolution texture map is generated from all the keyframes and inpainted by a neural network.

4. Method

4.1. Initialization

Given a single-view depth stream $\{D_1, D_2, ..., D_n\}$, firstly we solve the pose and shape parameters of SMPL [30] to track each frame following the skeleton tracking of DoubleFusion [52]. After that, we can obtain a SMPL sequence with pose parameters $\{\theta_1, \theta_2, ..., \theta_n\}$ corresponding to the depth stream. For the current frame, we allocate a 3D volume which contains the SMPL model. Then valid voxels (points) are selected by the distance of each voxel to the current SMPL being less than a threshold (8cm in our experiments). What we need to do next is to solve the occupancy values of these valid points.

4.2. Pose-guided Keyframe Selection

Consider the i-th frame as the current frame, the proposed pose-guided keyframe selection chooses appropriate frames by both pose similarity and visibility complementarity from other frames $F = \{1, 2, ..., i-1, i+1, ..., n\}$. The parametric SMPL model [30] across the whole sequence contributes to quantify visibility complementarity and pose similarity. More importantly, the keyframe selection should guarantee the temporal continuity of the selected keyframes between adjacent frames. In the keyframe selection, our goal is to select K ($K = 4$ in our experiments) keyframes from F by our proposed pose and visibility metrics, and in each iteration, we select one keyframe.

Pose Similarity Based on the parametric SMPL model, we can formulate the pose similarity energy between the j-th frame and the current frame conveniently as:

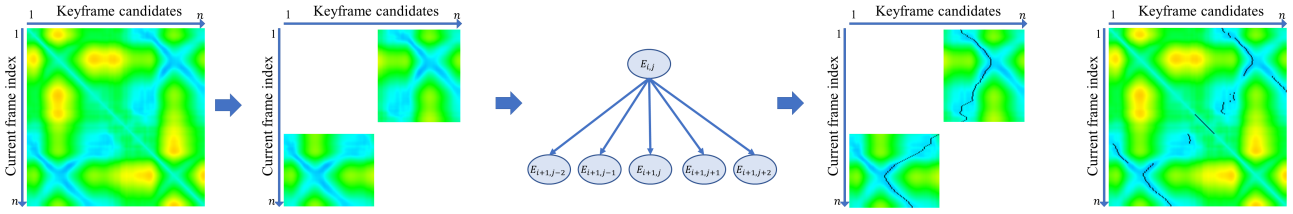
$$E_{\text{pose}}(i, j) = \sum_{k \in \mathcal{J}} w_k |\theta_i^k - \theta_j^k|^2, \tag{3}$$

where $\mathcal{J}$ is the joint index set except the global rotation and translation, and w_k is the influence weight of the k-th joint to the keyframe selection. The pose similarity constrains that the body pose in the selected keyframe is similar to that in the current frame.

Visibility Complementarity With the topology-consistent SMPL model across the whole sequence, we can set a visible flag for each vertex of SMPL in the j-th frame:

$$f_j^k = \begin{cases} 1, & \mathbf{v}_k \text{ is visible in the } j\text{-th frame} \\ 0, & \mathbf{v}_k \text{ is not visible in the } j\text{-th frame} \end{cases}, \tag{4}$$

where $\mathbf{v}_k$ is the k-th vertex of SMPL. So we can define a visible vector $\mathbf{f}_j = [f_j^1, f_j^2, \cdots, f_j^N]^\top$ which encodes the



(a) Energy matrix E (b) Divided blocks by visibility energy (c) Directed graph (d) Selected keyframe trail by the joint keyframe selection (e) Selected keyframe trail by the per-frame keyframe selection

Figure 3. Illustration of the two solutions to the keyframe selection. (a)(b)(c)(d) We formulate the keyframe selection as a dynamic programming (DP) problem to select temporally continuous keyframes, (e) the keyframe trail by the per-frame keyframe selection.

visible region of human body for each frame. The selected keyframe set of the i-th frame in previous iterations is denoted as $\mathcal{K}_i$, and the visibility $\mathbf{F}_{\mathcal{K}_i}$ over $\mathcal{K}_i$ is defined as:

$$\mathbf{F}_{\mathcal{K}_i} = \vee_{k \in \mathcal{K}_i} \mathbf{f}_k, \tag{5}$$

where $\vee$ is the element-wise logical OR operation. Before the first iteration, we initialize $\mathcal{K}_i = \{i\}$. In each iteration, the visibility complementarity energy is defined as:

$$E_{\text{visibility}}(\mathcal{K}_i, j) = \frac{\|\neg \mathbf{F}_{\mathcal{K}_i} \wedge \mathbf{f}_j\|_0}{\|\neg \mathbf{F}_{\mathcal{K}_i}\|_0}, \tag{6}$$

where $\neg$ and $\wedge$ are element-wise logical NOT and AND operations respectively, and Eq. 6 represents the proportion of "new" visible vertices that are visible in the j-th frame but not in $\mathcal{K}_i$ to all the invisible vertices in $\mathcal{K}_i$.

Joint Keyframe Selection In each iteration, we construct an energy matrix $E \in \mathbb{R}^{n \times n}$ (n is the frame number), and the (i, j)-th element of E is defined as

$$E_{i,j} = E_{\text{pose}}(i, j) - \lambda_{\text{visibility}} E_{\text{visibility}}(\mathcal{K}_i, j), \tag{7}$$

where $\lambda_{\text{visibility}}$ is a term weight. The energy matrix E encodes the visibility complementarity and pose similarity between the current frame and each keyframe candidate as shown in Fig. 3(a). We consider the row and column indices of E as the current frame index and keyframe candidates, respectively. We define the selected keyframe trail $\mathcal{T}$ in each iteration as $\mathcal{T} = \{t_1, t_2, ..., t_n\}$, where t_i is the selected keyframe of the i-th frame. A potential solution of the keyframe selection is to select the minimal element of each row, i.e., $t_i = \arg\min_j E_{i,j}$. However, the drawback of the per-frame selection is the temporal discontinuity of selected keyframes between the adjacent frames (Fig. 3(e)).

To guarantee the temporal continuity of details in invisible regions, we jointly select keyframes for the whole sequence, and formulate this procedure as a dynamic programming (DP) problem as illustrated in Fig. 3(b, c, d). To avoid the continuous keyframe trial to cross the diagonal of E, we firstly utilize visibility energy to divide several blocks and select keyframes within each block independently as shown in Fig. 3(b). For two adjacent blocks, we maintain a FIFO keyframe queue for smooth transition. Within each block, we constrain that the selected frames between two adjacent frames should be temporally continuous with each other, i.e., $|t_{i+1} - t_i| \leq \tau$, where τ is a half window size, and $\tau = 2$ in Fig. 3(c). Then we connect $E_{i,j}$ with $\{E_{i+1,j-\tau}, ..., E_{i+1,j+\tau}\}$ respectively to construct a directed graph (Fig. 3(c)), and our goal is to find a trail from the first row to the last row with the minimal energy sum for a global optimal solution, which is naturally a dynamic programming problem of minimum cost path. Based on this formulation, we can obtain a temporally continuous keyframe trail as illustrated in Fig. 3(d).

After each iteration, we update the selected keyframe set of each frame and the visibility over it:

$$\mathcal{K}_i \leftarrow \mathcal{K}_i \cup \{t_i\},\ \mathbf{F}_{\mathcal{K}_i} \leftarrow \mathbf{F}_{\mathcal{K}_i} \vee \mathbf{f}_{t_i},\ i = 1, ..., n. \tag{8}$$

In the next iteration, we construct the energy matrix E using the updated $\{\mathcal{K}_i\}_{i=1}^n$ and $\{\mathbf{F}_{\mathcal{K}_i}\}_{i=1}^n$ to search another continuous keyframe trail to cover other invisible regions.

Based on the proposed pose-guided keyframe selection, our method can reconstruct dynamic and high-fidelity details in the invisible regions as shown in Fig. 4.

4.3. Implicit Surface Fusion

Occupancy Inference After the keyframe selection, we have a keyframe set $\mathcal{K}_i$ for the current i-th frame. For each keyframe $k \in \mathcal{K}_i$, we firstly deform the valid points from the current frame to the k-th frame by the SMPL motion, then feed them to the occupancy inference network with the corresponding RGBD image, and finally obtain the occupancy values contributed by this keyframe.

Adaptive Blending Weight The inferred occupancy values may be inaccurate in invisible regions especially for self-occluded input or challenging poses. So directly averaging inferred occupancy values provided by all the keyframes just like in DynamicFusion [34] is improper. Our observation is that thanks to the depth information, PIFu can provide quite precise inference near the depth point clouds and in the visible region. We therefore design an adaptive blending weight according to visibility and the distance between each valid point and the depth point clouds, and the weight is formulated as:

$$w(\mathbf{x}; D_k) = \begin{cases} 1 & , p(\mathbf{x}; D_k) < \tau \\ e^{-\sigma(p(\mathbf{x}; D_k) - \tau)} & , p(\mathbf{x}; D_k) \geq \tau \end{cases}, \tag{9}$$
$$p(\mathbf{x}; D_k) = \mathbf{x}_z - D_k(\pi(\mathbf{x})),$$

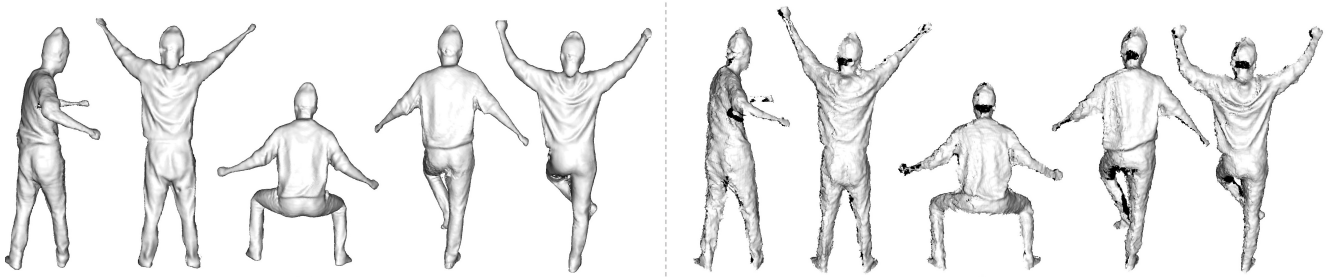


Figure 4. Comparison of reconstructed invisible details (a) and the ground truth (b) captured by multiple calibrated Kinects. These dynamic invisible details are similar to the physical ones due to the pose guidance.

where $p(\mathbf{x}; D_k)$ is the projected signed distance function (PSDF) [7] that takes a 3D point $\mathbf{x}$ as input and returns the difference between the depth value of $\mathbf{x}$ and the sampled depth value at the projected location $\pi(\mathbf{x})$ on the depth image D_k, σ is a factor to control the descending speed of the blending weight along the projection direction, and τ is a threshold to define high-confident regions. Then we integrate the inferred occupancy values from each keyframe using Eq. 9, and finally extract a model with high-fidelity geometric details using Marching Cubes [31].

Collision Handling Though the pipeline has been carefully designed to achieve dynamic and high-fidelity reconstruction, it still suffers from the collision problem. The self-collision in the live frame has been studied in [9, 13], however, the collision in the reference frame (i.e., the current frame in our method) is urgent to be resolved. For example, consider that the left arm of the performer collides with the torso as shown in Fig. 5(a), it is confusing to decide which body part to drive these collided points around this region. Due to the tracking error and the difference between SMPL and the real clothed human body, these points may be warped to incorrect positions and fetch wrong occupancy values, so that a crack occurs in the collided region as shown in the red ellipse of Fig. 5(a). A potential solution is to follow some tracking-based methods [52, 54, 44] in which they maintain a no-collision model under A-pose or T-pose, and deform the no-collision model to another frame. However, the continuously maintained model is being oversmoothed and loses the adjacent details due to the continuous fusion. Our observation is that since the body motion is continuous, a no-collision reconstructed model $\mathbf{M}'$ exists near the collided current frame. So we can deform this adjacent and no-collision model to the current frame, and voxelize it into an occupancy field $\mathbf{O}'$, and finally integrate $\mathbf{O}'$ into the implicit surface fusion. As shown in Fig. 5(b)(c), our collision handling scheme not only eliminates the collision artifacts but also maintains the geometric details in other regions, while fusing a continuous model loses the adjacent details.

4.4. Texture Generation

Given the geometric model $\mathbf{M}_i$ in the current frame, we utilize a per-face tile [43] to represent the texture on each face of $\mathbf{M}_i$ for high-resolution texture and fast UV unwrapping. Then we deform $\mathbf{M}_i$ to each keyframe and project it to the RGB image to fetch RGB values and finally blend them together. For the invisible faces, we utlize a neural network [38] to infer their textures.

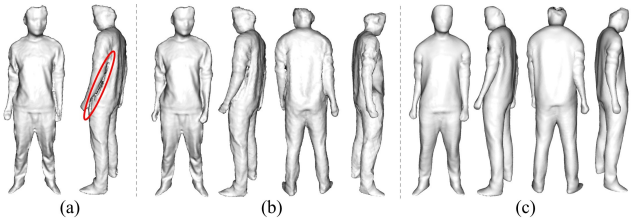


Figure 5. Illustration of the collision handling. (a) Reconstructed model without collision handling, (b) reconstructed model by fusing an adjacent no-collision model, (c) reconstructed model by fusing a continuously maintained template.

5. Results

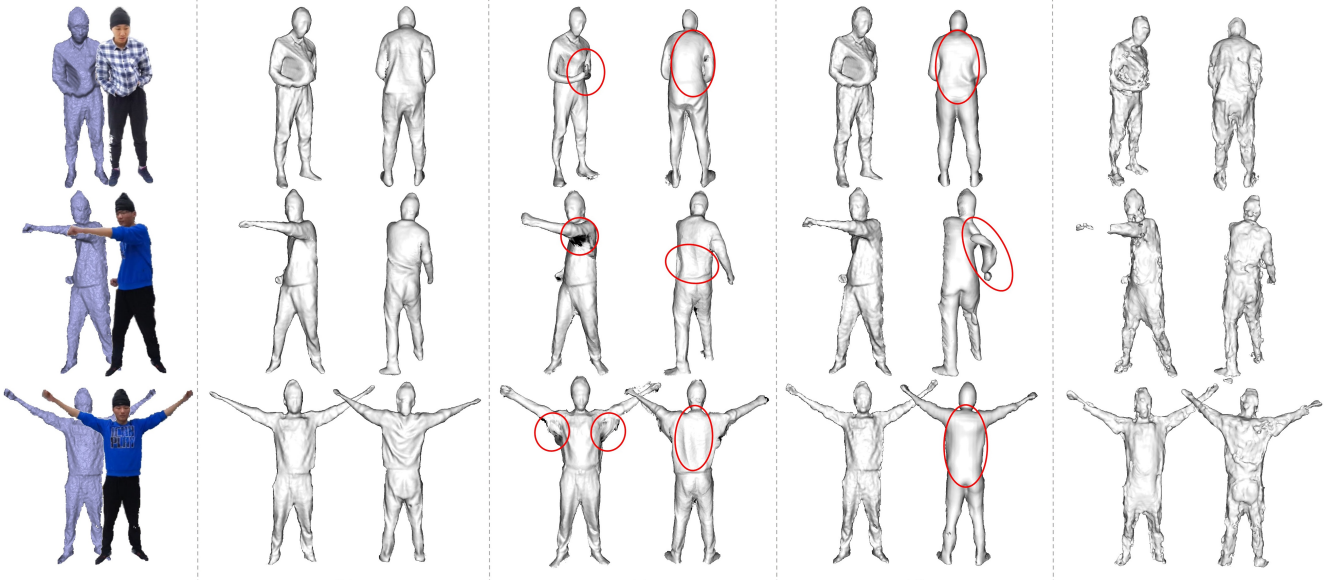
In this section, we firstly compare our method with current state-of-the-art works qualitatively and quantitatively. Then we evaluate our main contributions. Some results captured by our system are demonstrated in Fig. 6. Please refer to the supplemental material for the implementation details.

5.1. Comparison

Qualitative Comparison We compare the geometric reconstruction of our method with some representative and state-of-the-art tracking-based [52] and tracking-free [38] works as well as 3D human completion method [2] qualitatively using our captured data by a Kinect Azure in Fig. 7. And our method outperforms these methods on topological changes (top row of Fig. 7(c)), natural deformations (middle and bottom rows of Fig. 7(c)), dynamic pose-guided details (Fig. 7(c, d)) and invisible details (Fig. 7(d)). For a fair comparison, we retrain PIFu [38] with depth inputs and denote it as RGBD-PIFu, and the input of IP-Net [2] is a roughly complete depth point cloud merged from keyframes. However, due to the skeleton-level deformation error and depth noise, IP-Net fails to recover a complete detailed human model. We also compare our method with TexMesh [57] using their data in Fig. 8, and our method can reconstruct much more detailed geometry.

Quantitative Comparison We compare our method with

Figure 6. Results with dynamic and high-fidelity details reconstructed by our method. The bottom row is the input view, and the top row is another rendering view. Dynamic and high-fidelity details are reconstructed in both visible and invisible regions.



(a) (b) (c) (d) (e)

Figure 7. Qualitative comparison against other state-of-the-art methods. (a) RGBD images in the current frame, and results by our method (b), DoubleFusion [52] (c), RGBD-PIFu [38] (d) and IP-Net [2] (e).

DoubleFusion [52] and RGBD-PIFu [38] on the multi-view depth fitting error quantitatively. We utilize 4 calibrated Kinects to capture multi-view point clouds as the target, and evaluate the mean fitting error of each frame as shown in Fig. 9. It shows that our method reconstructs much more accurate results than DoubleFusion and RGBD-PIFu because in our method the invisible regions are similar to the physical ones thanks to the pose guidance and the visible regions are exactly same as the current observation.

5.2. Evaluation

Joint Keyframe Selection

– Comparison against Per-frame Keyframe Selection We demonstrate invisible details of several adjacent frames reconstructed with the joint keyframe selection and per-frame keyframe selection in Fig. 10, respectively. It shows the superiority of the joint keyframe selection on the temporal coherence compared with the per-frame keyframe selection.

– Comparison against Greedy Algorithm We compare the dynamic programming (DP) solution against the greedy algorithm[1] in Fig. 11. It shows that even though the greedy algorithm can obtain a temporally continuous keyframe trail, this algorithm may fall into a local minimum and some frames select keyframes with high energies in which the human poses are not similar to the current ones, so that the reconstructed invisible details are not physically plausible.

Ablation Study of Pose-guided Keyframe Selection We

[1]The greedy algorithm: considering the directed graph constructed in Fig. 3(c), given the keyframe t_i of the i-th frame, the keyframe t_{i+1} of the next frame is selected using $t_{i+1} = \arg\min_{j \in \{t_i-\tau, \cdots, t_i+\tau\}} E_{i,j}$.

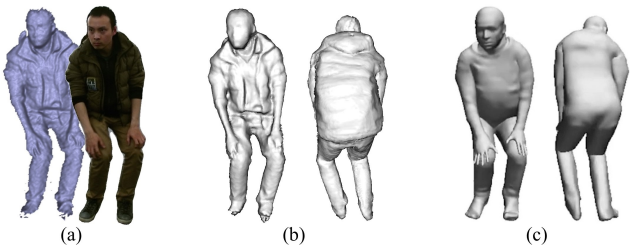


Figure 8. Comparison against TexMesh [57]. (a) RGBD image in the current frame, and results by our method (b) and TexMesh (c).

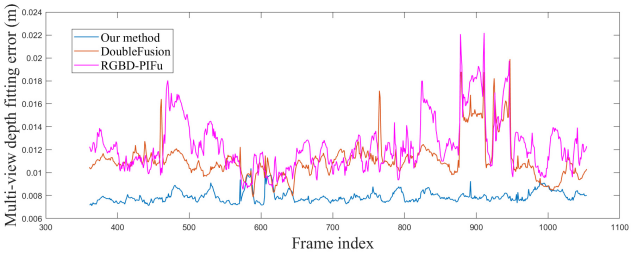


Figure 9. Quantitative comparison on the multi-view depth fitting error of our method, DoubleFusion [52] and RGBD-PIFu [38].

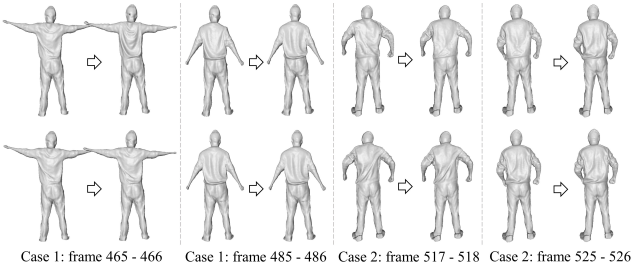


Figure 10. Comparison against per-frame keyframe selection. Invisible details of several adjacent frames by joint keyframe selection (bottom row) and per-frame selection (top row), respectively.

eliminate the pose or visibility energy, and visualize the keyframe trail in each situation in Fig. 12. **Pose Energy:** The red ellipse in Fig. 12(a) demonstrates that these frames select the same keyframe, so that the reconstructed invisible details in these frames are static. **Visibility Energy:** Fig. 12(b) demonstrates that without the visibility energy, the pose energy guides the selection scheme to choose the current frames, which is irrational. Fig. 12(c) shows that with both energies the keyframe selection provides a temporally continuous keyframe trail to generate dynamic pose-guided details in the invisible regions. For each situation (Fig.12(a, b, c)), we evaluate the multi-view depth fitting error quantitatively as shown in Fig. 13, and using both energies reconstructs more accurate geometry because the pose-guided invisible details are similar to the physical ones.

6. Discussion

Conclusion In this paper, we propose Pose-guided Selective Fusion (POSEFusion), the first method that can reconstruct high-fidelity and dynamic details of a performer even in the invisible regions from only a single RGBD camera. Based on the proposed pose-guided selective fusion framework, our method effectively combines the advantages of tracking-based methods and tracking-free inference methods while avoiding their drawbacks. As a result, our method outperforms the other state-of-the-art monocular capture methods.

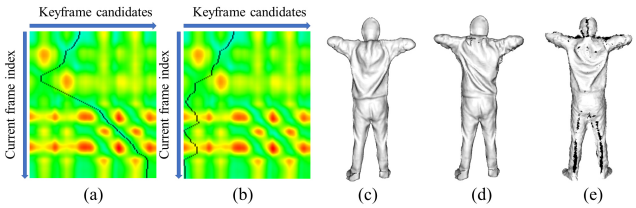


Figure 11. Comparison against the greedy algorithm. (a)(b) Keyframe trails obtained by DP and greedy algorithm, respectively, (c)(d) invisible details reconstructed by DP and greedy algorithm, respectively, (e) ground truth of the invisible regions.

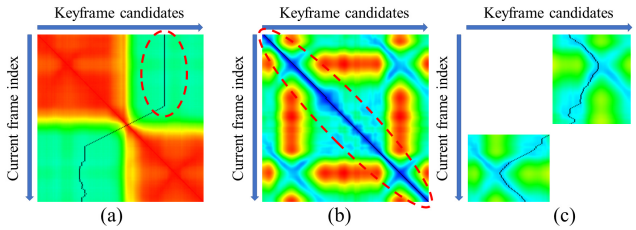


Figure 12. Qualitative ablation study of the pose-guided keyframe selection. (a) The keyframe trail without the pose energy, (b) the keyframe trail without the visibility energy, (c) the keyframe trail using the total energy within each divided block.

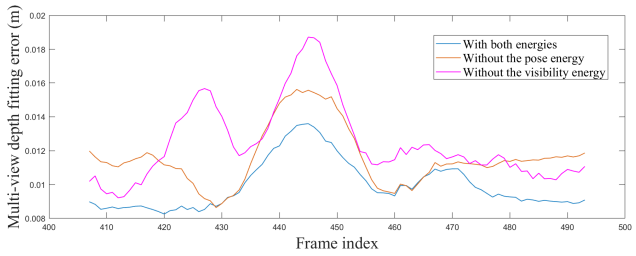


Figure 13. Quantitative ablation study of the pose-guided keyframe selection on the mean multi-view depth fitting error.

Limitation and Future Work It remains difficult for our method to handle very loose cloth (e.g., long skirt) because it is challenging to track a performer wearing loose cloth only using SMPL [30]. Replacing SMPL with a pre-scanned template (e.g., [24, 26]) may remove this limitation. Moreover, keyframe candidates in our keyframe selection are required to be sequential to guarantee temporal coherence, as for a non-sequential database, reorganizing these candidates by shape similarity [6] may resolve this problem. In addition, the reconstructed invisible details may not be exactly the same as the real ones, which we leave for future research.

Acknowledgement This paper is supported by the National Key Research and Development Program of China [2018YFB2100500], the NSFC No.61827805 and No.61861166002, and China Postdoctoral Science Foundation No.2020M670340.

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2293–2303, 2019. 1, 3

[2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 6, 7

[3] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2300–2308, 2015. 1

[4] Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. Markerless garment capture. *ACM Transactions on Graphics (TOG)*, 27(3):1–9, 2008. 1

[5] Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers. Combined region and motion-based 3d tracking of rigid and articulated objects. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):402–415, 2009. 1

[6] Chris Budd, Peng Huang, Martin Klaudiny, and Adrian Hilton. Global non-rigid alignment of surface sequences. *International Journal of Computer Vision*, 102(1-3):256–270, 2013. 8

[7] Brian Curless. *New methods for surface reconstruction from range images*. PhD thesis, Stanford University Stanford, CA, 1997. 6

[8] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics*, 36(4):1, 2017. 2

[9] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016. 1, 6

[10] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2232–2241, 2019. 1, 3

[11] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1753. IEEE, 2009. 1, 2

[12] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3083–3091, 2015. 1, 2

[13] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Transactions on Graphics*, 36(3):32:1–32:13, 2017. 2, 6

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

[15] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 1, 2

[16] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 362–379, Amsterdam, 2016. SPRINGER. 2

[17] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 2

[18] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 3

[20] Joo Ho Lee, Hyunho Ha, Yue Dong, Xin Tong, and Min H Kim. Texturefusion: High-quality texture acquisition for real-time rgb-d scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1272–1280, 2020. 2

[21] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3094–3103, 2017. 1

[22] Chao Li, Zheheng Zhang, and Xiaohu Guo. Articulatedfusion: Real-time reconstruction of motion, geometry and segmentation using a single depth camera. In *European Conference on Computer Vision (ECCV)*, pages 324–40, Munich, 2018. SPRINGER. 2

[23] Hao Li, Bart Adams, Leonidas J Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Transactions on Graphics*, 28(5):1–10, 2009. 1, 2

[24] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T Barron, and Gleb Gusev. 3d self-portraits. *ACM Transactions on Graphics (TOG)*, 32(6):1–9, 2013. 8

[25] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[26] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1344–1353, 2020. 8

[27] Junbang Liang and Ming C Lin. Shape-aware human pose and shape reconstruction using multi-view images. In *Pro-*

ceedings of the IEEE International Conference on Computer Vision, pages 4352–4362, 2019. 3

[28] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE transactions on visualization and computer graphics*, 16(3):407–418, 2009. 1

[29] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1256. IEEE, 2011. 1, 2

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 1, 2, 3, 4, 8

[31] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6

[32] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–908, 2015. 1

[33] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019. 1, 3

[34] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1, 2, 5

[35] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics*, 32(6):1–11, 2013. 2

[36] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 3

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

[38] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 3, 6, 7, 8

[39] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1, 3

[40] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1395, 2017. 2

[41] Miroslava Slavcheva, Maximilian Baust, and Slobodan Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2646–2655, Salt Lake City, June 2018. IEEE. 2

[42] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5330–5339, 2019. 1, 3

[43] Marc Soucy, Guy Godin, and Marc Rioux. A texture-mapping approach for the compression of colored 3d triangulations. *The Visual Computer*, 12(10):503–514, 1996. 6

[44] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. Robustfusion: human volumetric capture with data-driven visual cues using a rgbd camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 6

[45] Robert W. Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics*, 26(3), July 2007. 2

[46] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 1, 3

[47] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):1–15, 2018. 1

[48] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019. 3

[49] Genzhi Ye, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. Performance capture of interacting characters with handheld kinects. In *European Conference on Computer Vision*, pages 828–841. Springer, 2012. 1

[50] Mao Ye and Ruigang Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2352, 2014. 1

[51] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *IEEE International Conference on Computer Vision (ICCV)*, pages 910–919, Venice, 2017. IEEE. 1, 2

[52] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 7287–7296, Salt Lake City, June 2018. IEEE. 1, 2, 3, 4, 6, 7, 8

[53] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5499–5509. IEEE, 2019. 2

[54] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, pages 389–406, Munich, Sept 2018. SPRINGER. 2, 6

[55] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[56] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019. 1, 3

[57] Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 6, 8

[58] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4491–4500, 2019. 1, 3

[59] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 1, 2