RobustFusion: Human Volumetric Capture with Data-driven Visual Cues using a RGBD Camera

Zhuo Su[1*], Lan Xu[1,2*], Zerong Zheng[1], Tao Yu[1], Yebin Liu[1], and Lu Fang[1†]

[1] Tsinghua University
[2] ShanghaiTech University

Abstract. High-quality and complete 4D reconstruction of human activities is critical for immersive VR/AR experience, but it suffers from inherent self-scanning constraint and consequent fragile tracking under the monocular setting. In this paper, inspired by the huge potential of learning-based human modeling, we propose RobustFusion, a robust human performance capture system combined with various data-driven visual cues using a single RGBD camera. To break the orchestrated self-scanning constraint, we propose a data-driven model completion scheme to generate a complete and fine-detailed initial model using only the front-view input. To enable robust tracking, we embrace both the initial model and the various visual cues into a novel performance capture scheme with hybrid motion optimization and semantic volumetric fusion, which can successfully capture challenging human motions under the monocular setting without pre-scanned detailed template and owns the reinitialization ability to recover from tracking failures and the disappear-reoccur scenarios. Extensive experiments demonstrate the robustness of our approach to achieve high-quality 4D reconstruction for challenging human motions, liberating the cumbersome self-scanning constraint.

Keywords: Dynamic Reconstruction; Volumetric Capture; Robust; RGBD camera

1 Introduction

With the recent popularity of virtual and augmented reality (VR and AR) to present information in an innovative and immersive way, the 4D (3D spatial plus 1D time) content generation evolves as a cutting-edge yet bottleneck technique. Reconstructing the 4D models of challenging human activities conveniently for better VR/AR experience has recently attracted substantive attention of both the computer vision and computer graphics communities.

Early solutions [53, 34, 27, 28, 52] requires pre-scanned templates or two to four orders of magnitude more time than is available for daily usages such as immersive tele-presence. Recent volumetric approaches have eliminated the reliance of a pre-scanned template model and led to a profound progress in terms of both effectiveness and efficiency, by leveraging the RGBD sensors and high-end

* Equal Contribution. † Corresponding Author.

GPUs. The high-end solutions [7, 10, 9, 24, 61] rely on multi-view studio setup to achieve high-fidelity reconstruction but are expensive and difficult to be deployed, leading to the high restriction of the wide applications for daily usage. Besides, a number of approaches [35, 22, 45, 60, 16, 65–67] adopt the most common single RGBD camera setup with a temporal fusion pipeline to achieve complete reconstruction. However, these single-view approaches suffer from careful and orchestrated motions, especially when the performer needs to turn around carefully to obtain complete reconstruction. When the captured model is incomplete, the non-rigid tracking in those newly fused regions is fragile, leading to inferior results and impractical usage for VR/AR applications. On the other hand, the learning-based techniques have achieved significant progress recently for human attribute prediction using only the RGB input. This overcomes inherent constraint of existing monocular volumetric capture approaches, since such data-driven visual cues encode various prior information of human models such as motion [6, 32, 25] or geometry [42, 2, 68]. However, researchers did not explore these solutions to strengthen the volumetric performance capture.

In this paper, we attack the above challenges and propose *RobustFusion* – the first human volumetric capture system combined with various data-driven visual cues using only a single RGBD sensor, which does not require a pre-scanned template and outperforms existing state-of-the-art approaches significantly. Our novel pipeline not only eliminates the tedious self-scanning constraint but also captures challenging human motions robustly with the re-initialization ability to handle the severe tracking failures or disappear-reoccur scenarios, whilst still maintaining light-weight computation and monocular setup.

To maintain the fast running performance for the wide daily usages, we utilize those light-weight data-driven visual cues including implicit occupancy representation, human pose, shape and body part parsing. Combining such light-weight data-driven priors with the non-rigid fusion pipeline to achieve more robust and superior human volumetric capture is non-trivial. More specifically, to eliminate the inherent orchestrated self-scanning constraint of single-view capture, we first combine the data-driven implicit occupancy representation and the volumetric fusion within a completion optimization pipeline to generate an initial complete human model with fine geometric details. Such complete model is utilized to initialize both the performance capture parameters and the associated human priors. To enable robust tracking, based on both the initial complete model and the various visual cues, a novel performance capture scheme is proposed to combine the non-rigid tracking pipeline with human pose, shape and parsing priors, through a hybrid and flip-flop motion optimization and an effective semantic volumetric fusion strategy. Our hybrid optimization handles challenging fast human motions and recovers from tracking failures and the disappear-reoccur scenarios, while the volumetric fusion strategy estimates semantic motion tracking behavior to achieve robust and precise geometry update and avoid deteriorated fusion model caused by challenging fast motion and self-occulusion. To summarize, the main contributions of RobustFusion include:

- We propose a robust human volumetric capture method, which is the first to embrace various data-driven visual cues under the monocular setting without pre-scanned template, achieving significant superiority to state-of-the-arts.
- To eliminate the tedious self-scanning constraint, we propose a novel optimization pipeline to combine data-driven occupancy representation with volumetric fusion only using the front-view input.
- We propose an effective robust performance capture scheme with human pose, shape and parsing priors, which can handle challenging human motions with reinitialization ability.

2 Related Work

Human Performance Capture. Marker-based performance capture systems are widely used [55, 59, 56] but they are costly and quite intrusive to wear the marker suites. Thus, markerless performance capture [5, 1, 54] technologies have been widely investigated. The multi-view markerless approaches require studio-setup with a controlled imaging environment [11, 48, 29, 23, 7, 24, 15], while recent work [40, 38, 44] even demonstrates robust out-of-studio capture but synchronizing and calibrating multi-camera systems are still cumbersome. Some recent work only relies on a light-weight single-view setup [64, 19, 63] and even enables hand-held capture [20, 57, 37, 58] or drone-based capture [62, 60] for more practical application of performance capture. However, these methods require pre-scanned template model or can only reconstruct naked human model.

Recently, free-form dynamic reconstruction methods combine the volumetric fusion [8] and the nonrigid tracking [49, 27, 71, 17]. The high-end solutions [10, 9, 61] rely on multi-view studio to achieve high-fidelity reconstruction but are difficult to be deployed for daily usage, while some work [35, 22, 18, 45–47] adopt the most common single RGBD camera setting. Yu *et al.* [65, 66, 51] constrain the motion to be articulated to increase tracking robustness, while HybridFusion [67] utilizes extra IMU sensors for more reliable reconstruction. Xu *et al.* [60] further model the mutual gains between capture view selection and reconstruction. Besides, some recent work [36, 31] combine the neural rendering techniques to provide more visually pleasant results. However, these methods still suffer from careful and orchestrated motions, especially for a tedious self-scanning process where the performer need to turn around carefully to obtain complete reconstruction. Comparably, our approach is more robust for capturing challenging motions with reinitialization ability, and eliminates the self-scanning constraint.

Data-driven Human Modeling Early human modeling techniques [43, 12] are related to the discriminative approaches of performance capture, which take advantage of data driven machine learning strategies to convert the capture problem into a regression or pose classification problem. With the advent of deep neural networks, recent approaches obtain various attributes of human successfully from only the RGB input, which encodes rich prior information of human models. Some recent work [6, 41, 32, 25, 26] learns the skeletal pose and even human shape prior by using human parametric models [3, 30]. Various approaches [69, 50, 39,

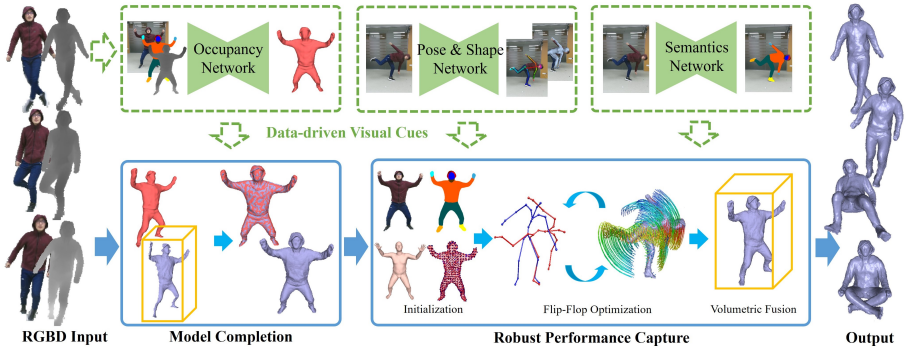

Fig. 1. The pipeline of RobustFusion. Assuming monocular RGBD input with various data-driven human visual priors, our approach consists of a model completion stage (Sec. 4) and a robust performance capture stage (Sec. 5) to generate live 4D results.

2, 68] propose to predict human geometry from a single RGB image by utilizing parametric human model as a basic estimation. Several work [21, 33, 42] further reveals the effectiveness of learning the implicit occupancy directly for detailed geometry modeling. However, such predicted geometry lacks fine details for face region and clothes wrinkle, which is important for immersive human modeling. Besides, researchers [14, 13, 70] propose to fetch the semantic information of human model. Several works [64, 19, 63] leverage learnable pose detections [6, 32] to improve the accuracy of human motion capture, but these methods rely on pre-scanned template models. However, even though these visual attributes yield huge potential for human performance modeling, researchers pay less attention surprisingly to explicitly combine such various data-driven visual priors with the existing volumetric performance capture pipeline. In contrast, we explore to build a robust volumetric capture algorithm on top of these visual priors and achieve significant superiority to previous capture methods.

3 Overview

RobustFusion marries volumetric capture to various data-driven human visual cues, which not only eliminates the tedious self-scanning constraint but also captures challenging human motions robustly handling the severe tracking failures or disappear-reoccur scenarios. As illustrated in Fig. 1, our approach takes a RGBD video from Kinect v2 as input and generates 4D meshes, achieving considerably more robust results than previous methods. Similar to [66, 61], we utilize TSDF [8] volume and ED model [49] for representation.

Model Completion. Only using the front-view input, we propose to combine the data-driven implicit occupancy network with the non-rigid fusion to eliminate the orchestrated self-scanning constraint of monocular capture. A novel completion optimization scheme is adopted to generate a high-quality watertight human model with fine geometric details.

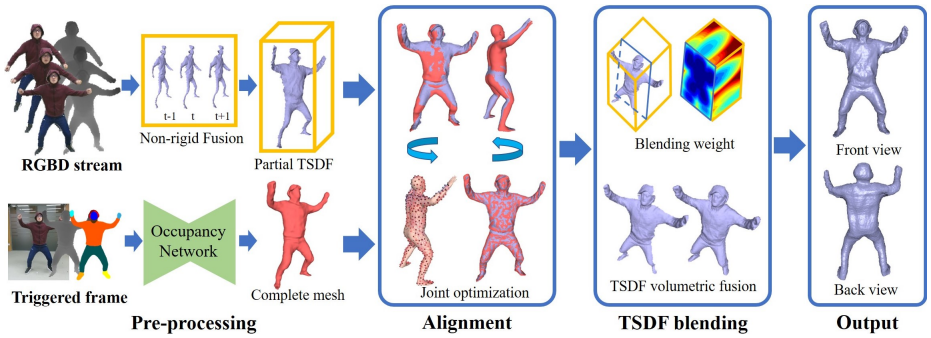

Fig. 2. Model completion pipeline. Assuming the front-view RGBD input, both a partial TSDF volume and a complete mesh are generated, followed by the alignment and blending operations to obtain a complete human model with fine geometry details.

Motion Initialization. Before the tracking stage, we further utilize the watertight mesh to initialize both the human motions and the visual priors. A hybrid motion representation based on the mesh is adopted, while various human pose and parsing priors based on the front-view input are associated to the mesh.
Robust Tracking. The core of our pipeline is to solve the hybrid motion parameters from the canonical frame to current camera view. We propose a robust tracking scheme which utilizes the reliable visual priors to optimize both the skeletal and non-rigid motions in an iterative flip-flop manner. Our scheme can handle challenging motions with the reinitialization ability.
Volumetric Fusion. After estimating the motions, we fuse the depth stream into a canonical TSDF volume to provide temporal coherent results. Based on various visual priors, our adaptive strategy models semantic tracking behavior to avoid deteriorated fusion caused by challenging motions. Finally, dynamic atlas [61] is adopted to obtain 4D textured reconstruction results.

4 Model Completion

To eliminate the orchestrated self-scanning constraint and the consequent fragile tracking of monocular capture, we propose a model completion scheme using only the front-view RGBD input. As illustrated in Fig. 2, our completion scheme combines the data-driven implicit representation and the non-rigid fusion to obtain a complete human model with fine geometry details.
Pre-processing. To generate high-fidelity geometry details, we utilize the traditional ED-based non-rigid alignment method [35, 60] to fuse the depth stream into live partial TSDF volume. Once the average accumulated TSDF weight in the front-view voxels reaches a threshold (32 in our setting), a data-driven occupancy regression network is triggered to generate a watertight mesh from only the triggered RGBD frame. To this end, we pre-train the PIFu [42] network using 1820 scans from Twindom, which learns a pixel-aligned implicit function by

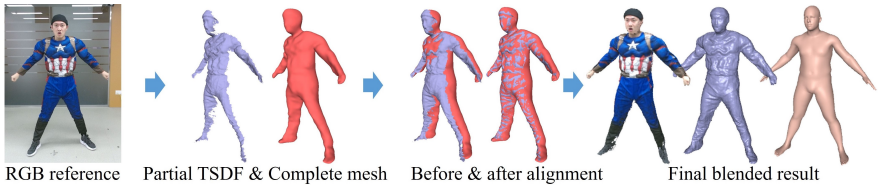

Fig. 3. The results of our model completion pipeline.

combining an image encoder and an implicit occupancy regression. To improve the scale and pose consistency, both the depth image and the human parsing image are added to the input of the image encoder.

Alignment. Note that the unique human prior can serve as a reliable reference to eliminate the misalignment between the partial TSDF and the complete mesh caused by their different input modalities. Thus, we adopt the double-layer motion representation [66, 61], which combines the ED model and the linear human body model SMPL [30]. For any 3D vertex $\mathbf{v}_c$, let $\tilde{\mathbf{v}}_c = ED(\mathbf{v}_c; G)$ denote the warped position after ED motion, where G is the non-rigid motion field. As for the SMPL model [30], the body model $\bar{\mathbf{T}}$ deforms into the morphed model $T(\boldsymbol{\beta}, \boldsymbol{\theta})$ with the shape parameters $\boldsymbol{\beta}$ and pose parameters $\boldsymbol{\theta}$. For any vertex $\bar{\mathbf{v}} \in \bar{\mathbf{T}}$, let $W(T(\bar{\mathbf{v}}; \boldsymbol{\beta}, \boldsymbol{\theta}); \boldsymbol{\beta}, \boldsymbol{\theta})$ denote the corresponding posed 3D position. Please refer to [66, 61] for details about the motion representation.

To align the partial TSDF and the complete mesh, we jointly optimize the unique human shape $\boldsymbol{\beta}_0$ and skeleton pose $\boldsymbol{\theta}_0$, as well as the ED non-rigid motion field G_0 from the TSDF volume to the complete mesh as follows:

$$\boldsymbol{E}_{\text{comp}}(G_0, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \lambda_{\text{vd}}\boldsymbol{E}_{\text{vdata}} + \lambda_{\text{md}}\boldsymbol{E}_{\text{mdata}} + \lambda_{\text{bind}}\boldsymbol{E}_{\text{bind}} + \lambda_{\text{prior}}\boldsymbol{E}_{\text{prior}}. \quad (1)$$

Here the volumetric data term $\boldsymbol{E}_{\text{vdata}}$ measures the misalignment error between the SMPL model and the reconstructed geometry in the partial TSDF volume:

$$\boldsymbol{E}_{\text{vdata}}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \sum_{\bar{\mathbf{v}} \in \bar{\mathbf{T}}} \psi(\mathbf{D}(W(T(\bar{\mathbf{v}}; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0); \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)), \quad (2)$$

where $\mathbf{D}(\cdot)$ takes a point in the canonical volume and returns the bilinear interpolated TSDF, and $\psi(\cdot)$ is the robust Geman-McClure penalty function.

The mutual data term E_{mdata} further measures the fitting from both the TSDF volume and the SMPL model to the complete mesh, which is formulated as the sum of point-to-plane distances:

$$\boldsymbol{E}_{\text{mdata}} = \sum_{(\bar{\mathbf{v}}, \mathbf{u}) \in \mathcal{C}} \psi(\mathbf{n}_{\mathbf{u}}^T(W(T(\bar{\mathbf{v}}; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)) - \mathbf{u})) + \sum_{(\tilde{\mathbf{v}}_c, \mathbf{u}) \in \mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T(\tilde{\mathbf{v}}_c - \mathbf{u})), \quad (3)$$

where $\mathcal{C}$ and $\mathcal{P}$ are the correspondence pair sets found via closest searching; $\mathbf{u}$ is a corresponding 3D vertex on the complete mesh. Besides, the pose prior term $\boldsymbol{E}_{\text{prior}}$ from [4] penalizes the unnatural poses while the binding term $\boldsymbol{E}_{\text{bind}}$ from [66] constrains both the non-rigid and skeletal motions to be consistent. We

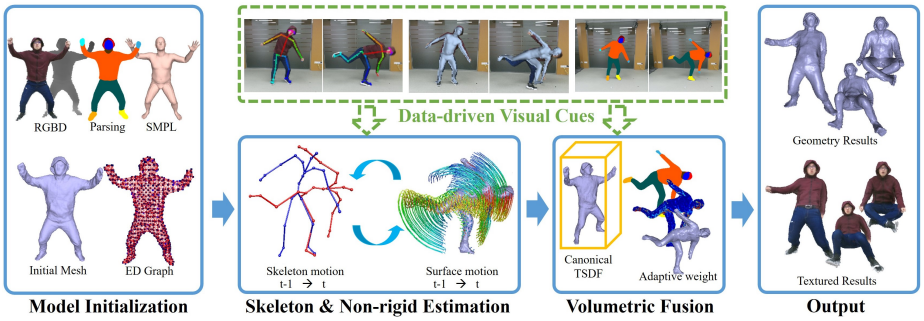

Fig. 4. The pipeline of our robust performance capture scheme. We first initialize both the motions and visual priors. Then, both skeletal and non-rigid motions are optimized with the associated visual priors. Finally, an adaptive volumetric fusion scheme is adopted to generated 4D textured results.

solve the resulting energy $\boldsymbol{E}_{\text{comp}}$ under the Iterative Closest Point (ICP) framework, where the non-linear least squares problem is solved using Levenberg-Marquardt (LM) method with a custom designed Preconditioned Conjugate Gradient (PCG) solver on GPU [18, 10].

TSDF Blending. After the alignment, we blend both the partial volume and the complete mesh seamlessly in the TSDF domain. For any 3D voxel $\mathbf{v}$, $\tilde{\mathbf{v}}$ denotes its warped position after applying the ED motion field; $\mathbf{N}(\mathbf{v})$ denotes the number of non-empty neighboring voxels of $\mathbf{v}$ in the partial volume which indicates the reliability of the fused geometry; $\mathbf{D}(\mathbf{v})$ and $\mathbf{W}(\mathbf{v})$ denote its TSDF value and accumulated weight, respectively. Then, to enable smooth blending, we calculate the corresponding projective SDF value $\mathbf{d}(\mathbf{v})$ and the updating weight $\mathbf{w}(\mathbf{v})$ as follows:

$$\mathbf{d}(\mathbf{v}) = (\mathbf{u} - \tilde{\mathbf{v}})\mathbf{sgn}(\mathbf{n}_{\mathbf{u}}^{T}(\mathbf{u} - \tilde{\mathbf{v}})), w(\mathbf{v}) = 1/(1 + \mathbf{N}(\mathbf{v})). \tag{4}$$

Here, recall that $\mathbf{u}$ is the corresponding 3D vertex of $\tilde{\mathbf{v}}$ on the complete mesh and $\mathbf{n}_{\mathbf{u}}$ is its normal; $\mathbf{sgn}(\cdot)$ is the sign function to distinguish positive and negative SDF. The voxel is further updated by the following blending operation:

$$\mathbf{D}(\mathbf{v}) \leftarrow \frac{\mathbf{D}(\mathbf{v})\mathbf{W}(\mathbf{v}) + \mathbf{d}(\mathbf{v})w(\mathbf{v})}{\mathbf{W}(\mathbf{v}) + w(\mathbf{v})}, \mathbf{W}(\mathbf{v}) \leftarrow \mathbf{W}(\mathbf{v}) + w(\mathbf{v}). \tag{5}$$

Finally, as illustrated in Fig. 3, marching cubes algorithm is adopted to obtain a complete and watertight human model with fine geometry details, which further enables robust motion initialization and tracking in Sec. 5.

5 Robust Performance Capture

As illustrated in Fig. 4, a novel performance capture scheme is proposed to track challenging human motions robustly with re-initialization ability with the aid of reliable data-driven visual cues.

Initialization. Note that the final complete model from Sec. 4 provides a reliable initialization for both the human motion and the utilized visual priors. To this end, before the tracking stage, we first re-sample the sparse ED nodes $\{\mathbf{x}_i\}$ on the mesh to form a non-rigid motion field, denoted as G. Besides, we rig the mesh with the output pose parameters $\boldsymbol{\theta}_0$ from its embedded SMPL model in Sec. 4 and transfer the SMPL skinning weights to the ED nodes $\{\mathbf{x}_i\}$. Then, for any 3D point $\mathbf{v}_c$ in the capture volume, let $\tilde{\mathbf{v}}_c$ and $\hat{\mathbf{v}}_c$ denote the warped positions after the embedded deformation and skeletal motion, respectively. Note that the skinning weights of $\mathbf{v}_c$ for the skeletal motion are given by the weighted average of the skinning weights of its knn-nodes. Please refer to [66, 30, 61] for more detail about the motion formulation. To initialize the pose prior, we apply OpenPose [6] on the RGBD image to obtain the 2D and lifted 3D joint positions, denoted as $\mathbf{P}_l^{2D}$ and $\mathbf{P}_l^{3D}$, respectively, with a detection confidence $\mathbf{C}_l$.Then, we find the closest vertex from the watertight mesh to $\mathbf{P}_l^{3D}$, denoted as $\mathbf{J}_l$, which is the associated marker position for the l-th joint. To utilize the semantic visual prior, we apply the light-weight human parsing method [70] to the triggered RGB image to obtain a human parsing image L. Then, we project each ED node $\mathbf{x}_i$ into L to obtain its initial semantic label $\mathbf{l}_i$. After the initialization, inspired by [64, 19], we propose to optimize the motion parameters G and $\boldsymbol{\theta}$ in an iterative flip-flop manner, so as to fully utilize the rich motion prior information of the visual cues to capture challenging motions.

Skeletal Pose Estimation. During each ICP iteration, we first optimize the skeletal pose $\boldsymbol{\theta}$ of the watertight mesh, which is formulated as follows:

$$\boldsymbol{E}_{\text{smot}}(\boldsymbol{\theta}) = \lambda_{\text{sd}}\boldsymbol{E}_{\text{sdata}} + \lambda_{\text{pose}}\boldsymbol{E}_{\text{pose}} + \lambda_{\text{prior}}\boldsymbol{E}_{\text{prior}} + \lambda_{\text{temp}}\boldsymbol{E}_{\text{temp}}. \tag{6}$$

Here, the dense data term $\boldsymbol{E}_{\text{sdata}}$ measures the point-to-plane misalignment error between the warped geometry in the TSDF volume and the depth input:

$$\boldsymbol{E}_{\text{sdata}} = \sum_{(\mathbf{v}_c,\mathbf{u})\in\mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T(\hat{\mathbf{v}}_c - \mathbf{u})), \tag{7}$$

where $\mathcal{P}$ is the corresponding set found via a projective searching; $\mathbf{u}$ is a sampled point on the depth map while $\mathbf{v}_c$ is the closet vertex on the fused surface. The pose term $\boldsymbol{E}_{\text{pose}}$ encourages the skeleton to match the detections obtained by CNN from the RGB image, including the 2D position $\mathbf{P}_l^{2D}$, lifted 3D position $\mathbf{P}_l^{3D}$ and the pose parameters $\boldsymbol{\theta}_d$ from OpenPose [6] and HMR [25]:

$$\boldsymbol{E}_{\text{pose}} = \psi(\Phi^T(\boldsymbol{\theta} - \boldsymbol{\theta}_d)) + \sum_{l=1}^{N_J} \phi(l)(\|\pi(\hat{\mathbf{J}}_l) - \mathbf{P}_l^{2D}\|_2^2 + \|\hat{\mathbf{J}}_l - \mathbf{P}_l^{3D}\|_2^2), \tag{8}$$

where $\psi(\cdot)$ is the robust Geman-McClure penalty function; $\hat{\mathbf{J}}_l$ is the warped associated 3D position and $\pi(\cdot)$ is the projection operator. The indicator $\phi(l)$ equals to 1 if the confidence $\mathbf{C}_l$ for the l-th joint is larger than 0.5, while Φ is the vectorized representation of $\{\phi(l)\}$. The prior term $\boldsymbol{E}_{\text{prior}}$ from [4] penalizes the unnatural poses, while the temporal term $\boldsymbol{E}_{\text{temp}}$ encourages coherent deformations by constraining the skeletal motion to be consistent to the previous ED

motion:

$$\boldsymbol{E}_{\text{temp}} = \sum_{\mathbf{x}_i} \|\hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_i\|_2^2, \tag{9}$$

where $\tilde{\mathbf{x}}_i$ is the warped ED node using non-rigid motion from previous iteration.

Non-rigid Estimation. To capture realistic non-rigid deformation, on top of the pose estimation result, we solve the surface tracking energy as follows:

$$\boldsymbol{E}_{\text{emot}}(G) = \lambda_{\text{ed}}\boldsymbol{E}_{\text{edata}} + \lambda_{\text{reg}}\boldsymbol{E}_{\text{reg}} + \lambda_{\text{temp}}\boldsymbol{E}_{\text{temp}}. \tag{10}$$

Here the dense data term E_{edata} jointly measures the dense point-to-plane misalignment and the sparse landmark-based projected error:

$$\boldsymbol{E}_{\text{edata}} = \sum_{(\mathbf{v}_c,\mathbf{u})\in\mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T(\tilde{\mathbf{v}}_c - \mathbf{u})) + \sum_{l=1}^{N_J} \phi(l)\|\pi(\tilde{\mathbf{J}}_l) - \mathbf{P}_l^{2D}\|_2^2, \tag{11}$$

where $\tilde{\mathbf{J}}_l$ is the warped associated 3D joint of the l-th joint in the fused surface. The regularity term $\boldsymbol{E}_{\text{reg}}$ from [66] produces locally as-rigid-as-possible (ARAP) motions to prevent over-fitting to depth inputs. Besides, the $\hat{\mathbf{x}}_i$ after the skeletal motion in the temporal term $\boldsymbol{E}_{\text{temp}}$ is fixed during current optimization.

Both the pose and non-rigid optimizations in Eqn. 6 and Eqn. 10 are solved using LM method with the same PCG solver on GPU [18, 10]. Once the confidence $\mathbf{C}_l$ reaches 0.9 and the projective error $\|\pi(\tilde{\mathbf{J}}_l) - \mathbf{P}_l^{2D}\|_2^2$ is larger than 5.0 for the l-th joint, the associated 3D position $\mathbf{J}_l$ on the fused surface is updated via the same closest searching strategy of the initialization stage. When there is no human detected in the image, our whole pipeline will be suspended until the number of detected joints reaches a threshold (10 in our setting).

Volumetric Fusion. To temporally update the geometric details, after above optimization, we fuse the depth into the TSDF volume and discard the voxels which are collided or warped into invalid input to achieve robust geometry update. To avoid deteriorated fusion caused by challenging motion or reinitialization, an effective adaptive fusion strategy is proposed to model semantic motion tracking behavior. To this end, we apply the human parsing method [70] to current RGB image to obtain a human parsing image L. For each ED node $\mathbf{x}_i$, recall that $\mathbf{l}_i$ is its associated semantic label during initialization while $L(\pi(\tilde{\mathbf{x}}_i)$ is current corresponding projected label. Then, for any voxel $\mathbf{v}$, we formulate its updating weight $\mathbf{w}(\mathbf{v})$ as follows:

$$\mathbf{w}(\mathbf{v}) = \exp(\frac{-\|\boldsymbol{\Phi}^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}_d)\|_2^2}{2\pi}) \sum_{i\in\mathcal{N}(v_c)} \frac{\varphi(\mathbf{l}_i, L(\pi(\tilde{\mathbf{x}}_i)))}{|\mathcal{N}(v_c)|}, \tag{12}$$

where $\boldsymbol{\theta}^*$ is the optimized pose; $\mathcal{N}(v_c)$ is the collection of the knn-nodes of $\mathbf{v}$; $\varphi(\cdot,\cdot)$ denote an indicator which equals to 1 only if the two input labels are the same. Note that such robust weighting strategy measures the tracking performance based on human pose and semantic priors. Then, $\mathbf{w}(\mathbf{v})$ is set to be zero if it's less than a truncated threshold (0.2 in our setting), so as to control the minimal integration and further avoid deteriorated fusion of severe tracking failures. Finally, the voxel is updated using Eqn. 5 and the dynamic atlas scheme [61] is adopted to obtain 4D textured reconstruction.

Fig. 5. 4D reconstructed results of the proposed RobustFusion system.

6 Experiment

In this section, we evaluate our RobustFusion system on a variety of challenging scenarios. We run our experiments on a PC with a NVIDIA GeForce GTX TITAN X GPU and an Intel Core i7-7700K CPU. Our optimized GPU code takes 15 s for the model completion. The following robust performance capture pipeline runs at 123 ms per frame, where the visual priors collecting takes 90 ms, the flip-flop optimization takes around 18 ms with 4 ICP iteration, and 15 ms for all the remaining computations. In all experiments, we use the following empirically determined parameters: $\lambda_{vd} = 1.0$, $\lambda_{md} = 2.0$, $\lambda_{bind} = 1.0$, $\lambda_{prior} = 0.01$, $\lambda_{sdata} = 4.0$, $\lambda_{pose} = 2.0$, $\lambda_{temp} = 1.0$, $\lambda_{edata} = 4.0$ and $\lambda_{reg} = 5.0$. Fig. 5 demonstrates the results of RobustFusion, where both the challenging motion and the fine-detailed geometry are faithfully captured.

6.1 Comparison

We compare our RobustFusion against the state-of-the-art methods DoubleFusion [66] and UnstructuredFusion [61]. For fair comparison, we modify [61] into the monocular setting by removing their online calibration stage. As shown in Fig. 6, our approach achieves significantly better tracking results especially for challenging self-occluded and fast motions. Then, we utilize the sequence from [61] with available ground truth captured via the OptiTrack system to compute

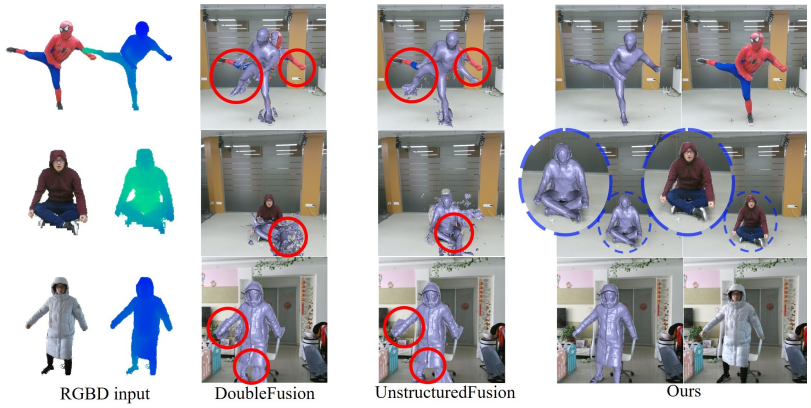

Fig. 6. Qualitative comparison. Our results overlay better with the RGB images.

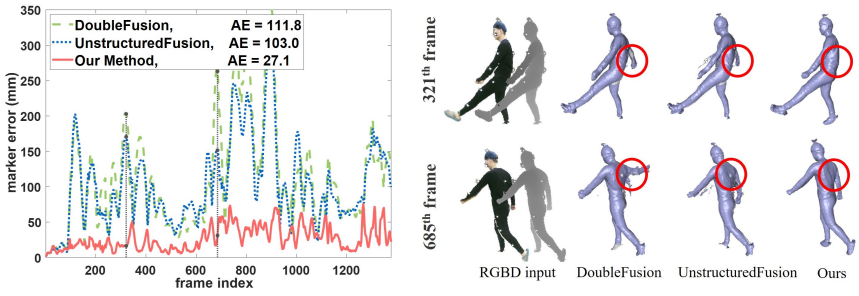

Fig. 7. Quantitative comparison against UnstructuredFusion [61] and DoubleFusion [66]. Left: the error curves. Right: the reconstruction results.

the per-frame mean error of all the markers as well as the average mean error of the whole sequence (AE). As illustrated in Fig. 7, our approach achieves the highest tracking accuracy with the aid of various visual priors.

We further compare against HybridFusion [67], which uses extra IMU sensors. We utilize the challenging sequence with ground truth from [67] and remove their orchestrated self-scanning process before the tracking stage (the first 514 frames). As shown in Fig. 8, our approach achieves significantly better result than DoubleFusion and even comparable performance than HybridFusion only using the RGBD input. Note that HybridFusion still relies on the self-scanning stage for sensor calibration and suffers from missing geometry caused by the body-worn IMUs, while our approach eliminates such tedious self-scanning and achieves more complete reconstruction.

6.2 Evaluation

Model Completion. As shown in Fig. 9 (a), without model completion, only partial initial geometry with SMPL-based ED-graph leads to inferior tracking

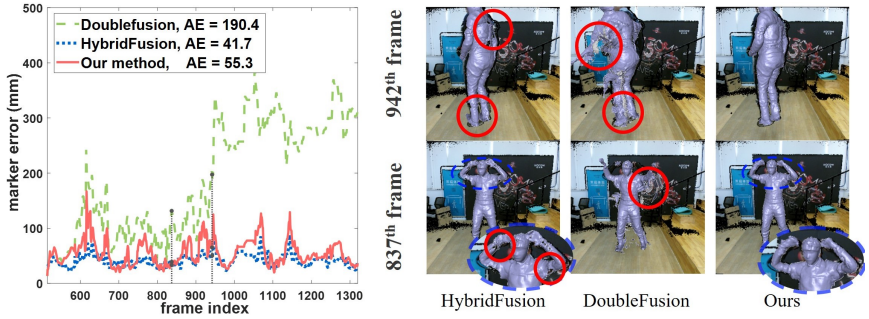

Fig. 8. Quantitative comparison against HybridFusion [67] and DoubleFusion [66]. Left: the error curves. Right: the reconstruction results.

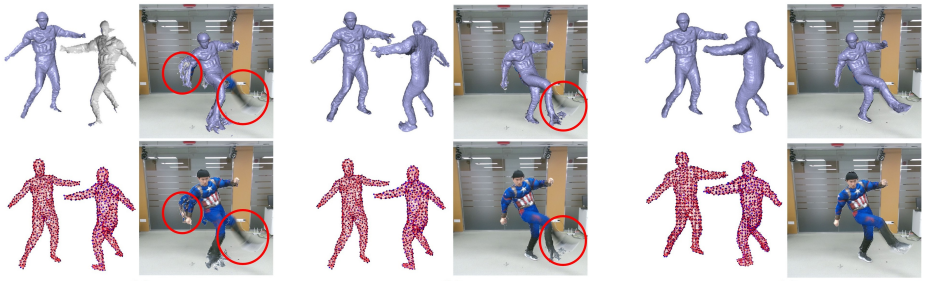

Fig. 9. Evaluation of model completion. (a-c) are the results without completion, with completion only using skeleton optimization and SMPL-based ED-graph, with our full model completion, respectively. Each triple includes the output mesh, corresponding ED-graph and the overlaid tracking geometry/texture results for a following frame.

results. The result in Fig. 9 (b) is still imperfect because only the skeletal pose is optimized during completion optimization and only SMPL-based ED-graph is adopted for motion tracking. In contrast, our approach with model completion in Fig. 9 (c) successfully obtains a watertight and fine-detailed human mesh to enable both robust motion initialization and tracking.

Robust Tracking. We further evaluate our robust performance capture scheme. In Fig. 10, we compare to the results using traditional tracking pipeline [66, 61] without data-driven visual priors in two scenarios where fast motion or disappear-reoccurred case happens. Note that our variation without visual priors suffers from severe accumulated error and even totally tracking lost, while our approach achieves superior tracking results for these challenging cases. Furthermore, we compare to the baseline which jointly optimizes skeletal and non-rigid motions without our flip-flop strategy. As shown in Fig 11, our approach with flip-flop strategy makes full use of the visual priors, achieving more robust and visually pleasant reconstruction especially for the challenging human motions.

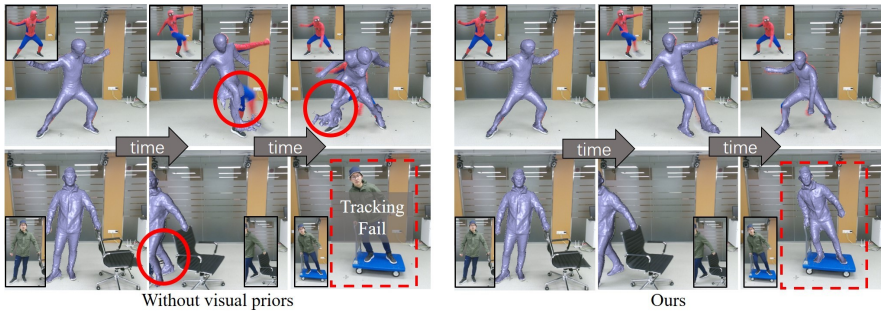

Fig. 10. Evaluation of robust tracking. Our approach with various human visual priors achieves superior results for challenging motions and has reinitialization ability.

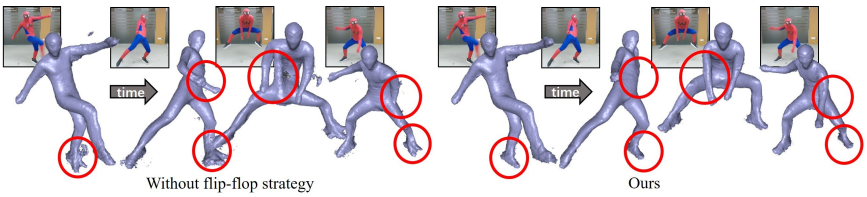

Fig. 11. Evaluation of robust tracking. Our approach with the flip-flop optimization strategy achieves more visually pleasant 4D geometry for challenging motions.

Adaptive Fusion. To evaluate our adaptive fusion scheme based on the pose and semantic cues, we compare to the variation of our pipeline using traditional volumetric fusion [35, 61, 67]. As shown in Fig. 12, the results without adaptive fusion suffer from severe accumulated error, especially for those regions with high-speed motions. In contrast, our adaptive fusion successfully models semantic tracking behavior and avoids deteriorated fusion.

For further analysis of the individual components of RobustFusion, we utilize the sequence with ground truth from [61]. We compute the per-frame mean error for the three variation of our approach without model completion, prior-based robust tracking and adaptive fusion, respectively. Fig 13 shows our full pipeline consistently outperforms the three baselines, yielding the lowest AE. This not only highlights the contribution of each algorithmic component but also illustrates that our approach can robustly capture human motion details.

7 Discussion

Limitation. First, we cannot handle surface splitting topology changes like clothes removal, which we plan to address by incorporating the key-volume update technique [10]. Our method is also restricted to human reconstruction, without modeling human-object interactions. This could be alleviated in the future by combining the static object reconstruction methods into current framework.

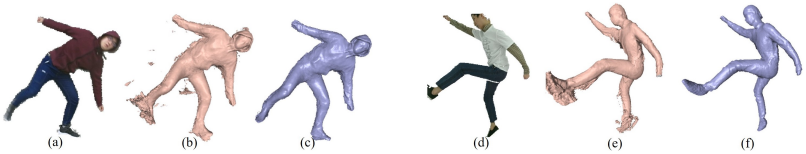

Fig. 12. Evaluation of the adaptive fusion. (a, d) Reference color images. (b, e) The results without adaptive fusion. (c, f) The results with adaptive fusion.

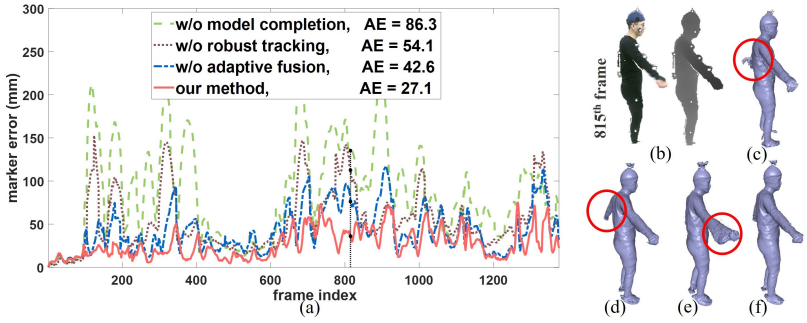

Fig. 13. Quantitative evaluation. (a) Numerical error curves. (b) RGBD input. (c)-(e) The results of three baselines without model completion, robust tracking and adaptive fusion, respectively. (f) The reconstruction results of our full pipeline.

As is common for learning methods, the utilized visual cue regressions fail for extreme poses not seen in training, such as severe and extensive (self-)occlusion. Fortunately, our approach is able to instantly recover robustly with our reinitialization ability as soon as the occluded parts become visible again. Our current pipeline turns to utilize the data-driven cues in an optimization framework. It's an promising direction to jointly model both visual cues and volumetric capture in an end-to-end learning-based framework.

Conclusions. We have presented a superior approach for robust volumetric human capture combined with data-driven visual cues. Our completion optimization alleviates the orchestrated self-scanning constraints for monocular capture, while our robust capture scheme enables to capture challenging human motions and reinitialize from the tracking failures and disappear-reoccur scenarios. Our experimental results demonstrate the robustness of RobustFusion for compelling performance capture in various challenging scenarios, which compares favorably to the state-of-the-arts. We believe that it is a significant step to enable robust and light-weight human volumetric capture, with many potential applications in VR/AR, gaming, entertainment and immersive telepresence.

Acknowledgement. This work is supported in part by Natural Science Foundation of China under contract No. 61722209 and 6181001011, in part by Shenzhen Science and Technology Research and Development Funds (JCYJ20180507183706645).

References

1. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video **27**(3), 98:1–10 (2008)
2. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers. p. 408–416. SIGGRAPH '05, Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1186822.1073207, https://doi.org/10.1145/1186822.1073207
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 561–578. Springer International Publishing, Cham (2016)
5. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: Computer Vision and Pattern Recognition (CVPR) (1998). https://doi.org/10.1109/CVPR.1998.698581
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Computer Vision and Pattern Recognition (CVPR) (2017)
7. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics (TOG) **34**(4), 69 (2015)
8. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. pp. 303–312. SIGGRAPH '96, ACM, New York, NY, USA (1996). https://doi.org/10.1145/237170.237269, http://doi.acm.org/10.1145/237170.237269
9. Dou, M., Davidson, P., Fanello, S.R., Khamis, S., Kowdle, A., Rhemann, C., Tankovich, V., Izadi, S.: Motion2fusion: Real-time volumetric performance capture. ACM Trans. Graph. **36**(6), 246:1–246:16 (Nov 2017)
10. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., Izadi, S.: Fusion4D: Real-time Performance Capture of Challenging Scenes. In: ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques (2016)
11. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. International Journal of Computer Vision (IJCV) **87**(1–2), 75–92 (2010)
12. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera (2010)
13. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: The European Conference on Computer Vision (ECCV) (September 2018)
14. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
15. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., et al.: The relightables: Volumetric

performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6) (Nov 2019)
16. Guo, K., Taylor, J., Fanello, S., Tagliasacchi, A., Dou, M., Davidson, P., Kowdle, A., Izadi, S.: Twinfusion: High framerate non-rigid fusion through fast correspondence tracking. In: International Conference on 3D Vision (3DV). pp. 596–605 (2018)
17. Guo, K., Xu, F., Wang, Y., Liu, Y., Dai, Q.: Robust Non-Rigid Motion Tracking and Surface Reconstruction Using L0 Regularization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3083–3091 (2015)
18. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, albedo and motion reconstruction using a single rgbd camera. ACM Transactions on Graphics (TOG) (2017)
19. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) **38**(2), 14:1–14:17 (2019)
20. Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. In: Computer Vision and Pattern Recognition (CVPR). pp. 224–231 (2009)
21. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., Li, H.: Deep volumetric video from very sparse multi-view performance capture. In: The European Conference on Computer Vision (ECCV) (September 2018)
22. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: VolumeDeform: Real-time Volumetric Non-rigid Reconstruction (October 2016)
23. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic Studio: A Massively Multiview System for Social Motion Capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015)
24. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
25. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
26. Kovalenko, O., Golyanik, V., Malik, J., Elhayek, A., Stricker, D.: Structure from Articulated Motion: Accurate and Stable Monocular 3D Reconstruction without Training Data. Sensors **19**(20) (2019)
27. Li, H., Adams, B., Guibas, L.J., Pauly, M.: Robust single-view geometry and motion reconstruction **28**(5), 175 (2009)
28. Li, H., Luo, L., Vlasic, D., Peers, P., Popović, J., Pauly, M., Rusinkiewicz, S.: Temporally coherent completion of dynamic shapes. In: ACM Trans. Graph. vol. 31 (Feb 2012)
29. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multiview image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(11), 2720–2735 (2013)
30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (Oct 2015)
31. Martin-Brualla, R., Pandey, R., Yang, S., Pidlypenskyi, P., Taylor, J., Valentin, J., Khamis, S., Davidson, P., Tkach, A., Lincoln, P., et al.: Lookingood: Enhancing performance capture with real-time neural re-rendering. ACM Trans. Graph. **37**(6) (Dec 2018)

32. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4) (2017)
33. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
34. Mitra NJ, Flöry S, O.M.G.N.G.L.P.H.: Dynamic geometry registration. In: In Symposium on geometry processing (2007)
35. Newcombe, R.A., Fox, D., Seitz, S.M.: DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time (June 2015)
36. Pandey, R., Tkach, A., Yang, S., Pidlypenskyi, P., Taylor, J., Martin-Brualla, R., Tagliasacchi, A., Papandreou, G., Davidson, P., Keskin, C., Izadi, S., Fanello, S.: Volumetric capture of humans with a single rgbd camera via semi-parametric learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (Jun 2019), http://smpl-x.is.tue.mpg.de
38. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Harvesting multiple views for marker-less 3d human pose annotations. In: Computer Vision and Pattern Recognition (CVPR) (2017)
39. Pumarola, A., Sanchez-Riera, J., Choi, G.P.T., Sanfeliu, A., Moreno-Noguer, F.: 3dpeople: Modeling the geometry of dressed humans. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
40. Robertini, N., Casas, D., Rhodin, H., Seidel, H.P., Theobalt, C.: Model-based outdoor performance capture. In: International Conference on 3D Vision (3DV) (2016), http://gvv.mpi-inf.mpg.de/projects/OutdoorPerfcap/
41. Rogez, G., Schmid, C.: Mocap Guided Data Augmentation for 3D Pose Estimation in the Wild. In: Neural Information Processing Systems (NIPS) (2016)
42. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
43. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time Human Pose Recognition in Parts from Single Depth Images (2011)
44. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Computer Vision and Pattern Recognition (CVPR) (2017)
45. Slavcheva, M., Baust, M., Cremers, D., Ilic, S.: KillingFusion: Non-rigid 3D Reconstruction without Correspondences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
46. Slavcheva, M., Baust, M., Ilic, S.: SobolevFusion: 3D Reconstruction of Scenes Undergoing Free Non-rigid Motion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
47. Slavcheva, M., Baust, M., Ilic, S.: Variational Level Set Evolution for Non-rigid 3D Reconstruction from a Single Depth Camera. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2020)
48. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of Gaussians body model. In: International Conference on Computer Vision (ICCV) (2011)

49. Sumner, R.W., Schmid, J., Pauly, M.: Embedded deformation for shape manipulation. ACM Transactions on Graphics (TOG) **26**(3), 80 (2007)
50. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
51. Tao Yu, Jianhui Zhao, Y.H.Y.L.Y.L.: Towards robust and accurate single-view fast human motion capture. In: IEEE Access (2019)
52. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 103–110 (2012)
53. Tevs, A., B.A.W.M.I.I.B.M.K.J..S.H.P.: Animation cartography—intrinsic reconstruction of shape and motion. In: ACM Transactions on Graphics (TOG) (2012)
54. Theobalt, C., de Aguiar, E., Stoll, C., Seidel, H.P., Thrun, S.: Performance capture from multi-view video. In: Image and Geometry Processing for 3-D Cinematography, pp. 127–149. Springer (2010)
55. Vicon Motion Systems. https://www.vicon.com/ (2019)
56. Vlasic, D., Adelsberger, R., Vannucci, G., Barnwell, J., Gross, M., Matusik, W., Popović, J.: Practical Motion Capture in Everyday Surroundings **26**(3) (2007)
57. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera **32**(6) (2013)
58. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
59. Xsens Technologies B.V. https://www.xsens.com/ (2019)
60. Xu, L., Cheng, W., Guo, K., Han, L., Liu, Y., Fang, L.: Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera. IEEE Transactions on Visualization and Computer Graphics pp. 1–1 (2019)
61. Xu, L., Su, Z., Han, L., Yu, T., Liu, Y., FANG, L.: Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercialrgbd cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2019)
62. Xu, L., Liu, Y., Cheng, W., Guo, K., Zhou, G., Dai, Q., Fang, L.: Flycap: Markerless motion capture using multiple autonomous flying cameras. IEEE Transactions on Visualization and Computer Graphics **24**(8), 2284–2297 (Aug 2018)
63. Xu, L., Xu, W., Golyanik, V., Habermann, M., Fang, L., Theobalt, C.: EventCap: Monocular 3D Capture of High-Speed Human Motions using an Event Camera. arXiv e-prints (2019)
64. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM Transactions on Graphics (TOG) **37**(2), 27:1–27:15 (2018)
65. Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y.: Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In: The IEEE International Conference on Computer Vision (ICCV). ACM (October 2017)
66. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
67. Zheng, Z., Yu, T., Li, H., Guo, K., Dai, Q., Fang, L., Liu, Y.: Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In: European Conference on Computer Vision (ECCV) (Sept 2018)

68. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
69. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
70. Zhu, T., Oved, D.: Bodypix github repository. https://github.com/tensorflow/tfjs-models/tree/master/body-pix (2019)
71. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al.: Real-time Non-rigid Reconstruction using an RGB-D Camera. ACM Transactions on Graphics (TOG) **33**(4), 156 (2014)