RobustFusion: Robust Volumetric Performance Reconstruction Under Human-Object Interactions From Monocular RGBD Stream

Zhuo Su, Lan Xu, Dawei Zhong, Zhong Li, Fan Deng, Shuxue Quan, and Lu Fang, *Senior Member, IEEE*

Abstract—High-quality 4D reconstruction of human performance with complex interactions to various objects is essential in real-world scenarios, which enables numerous immersive VR/AR applications. However, recent advances still fail to provide reliable performance reconstruction, suffering from challenging interaction patterns and severe occlusions, especially for the monocular setting. To fill this gap, in this paper, we propose RobustFusion, a robust volumetric performance reconstruction system for human-object interaction scenarios using only a single RGBD sensor, which combines various data-driven visual and interaction cues to handle the complex interaction patterns and severe occlusions. We propose a semantic-aware scene decoupling scheme to model the occlusions explicitly, with a segmentation refinement and robust object tracking to prevent disentanglement uncertainty and maintain temporal consistency. We further introduce a robust performance capture scheme with the aid of various data-driven cues, which not only enables re-initialization ability, but also models the complex human-object interaction patterns in a data-driven manner. To this end, we introduce a spatial relation prior to prevent implausible intersections, as well as data-driven interaction cues to maintain natural motions, especially for those regions under severe human-object occlusions. We also adopt an adaptive fusion scheme for temporally coherent human-object reconstruction with occlusion analysis and human parsing cue. Extensive experiments demonstrate the effectiveness of our approach to achieve high-quality 4D human performance reconstruction under complex human-object interactions whilst still maintaining the lightweight monocular setting.

Index Terms—4D reconstruction, human-object interaction, performance reconstruction, RGBD camera, robust

1 Introduction

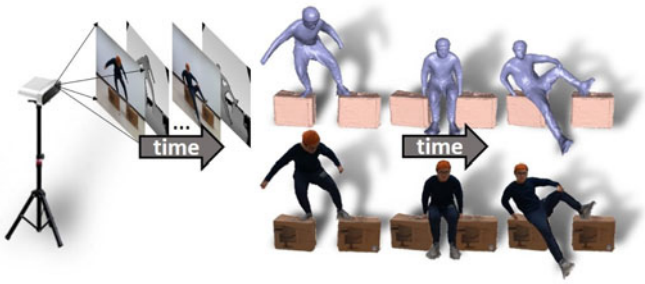
The rise of virtual reality and augmented reality (VR and AR) to present information in an innovative and immersive way has increased the demand for human-centric 4D (3D spatial plus 1D temporal) content generation, with various applications from entertainment to commerce, from gaming to education, from military to art. Further, reconstructing the 4D models of human activities under human-object interactions both robustly and conveniently remains unsolved, which suffers from challenging interaction patterns and severe occlusions. It evolves as a cutting-edge yet bottleneck technique and has recently attracted substantive attention of both the computer vision and computer graphics communities.

Early model-based methods [1], [2], [3], [4], [5] suffer from pre-scanned templates or inefficient run-time performance, which are unacceptable for daily interactive application. Recent volumetric approaches have eliminated the reliance on the templates and increased both the effectiveness and efficiency with modern GPUs.

The high-end solutions [6], [7], [8], [9], [10] achieve realistic human-object reconstruction using multi-view studio setup which provides sufficient view observation to solve the challenging interaction and occlusion ambiguity. However, their complex and expensive multi-view studio setting leads to the high restriction of the daily applications. Differently, the monocular volumetric approaches adopt the handiest commercial RGBD camera and temporal fusion pipeline. Early general solutions [11], [12], [13], [14] handle general dynamic scenes without disentangling human and objects, suffering from careful and orchestrated motions. Recent solutions [15], [16], [17] embed the human parametric models

- *Zhuo Su and Lu Fang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100190, China, and also with the Tsinghua-Berkeley Shenzhen Institute (TBSI), Guangdong 518055, China. E-mail: suzhuo13@gmail.com, fanglu@tsinghua.edu.cn.*
- *Lan Xu is with the School of Information Science and Technology, Shanghaitech University, Shanghai 201210, China. E-mail: lxuan@connect.ust.hk.*
- *Dawei Zhong is with the Department of Electronic Engineering, Tsinghua University, Beijing 100190, China, also with the Tsinghua-Berkeley Shenzhen Institute (TBSI), Guangdong 518055, China, and also with Zhejiang future technology institute (Jiaxing), Jiaxing, Zhejiang 314000, China. E-mail: zdw19@mails.tsinghua.edu.cn.*
- *Zhong Li, Fan Deng, and Shuxue Quan are with OPPO US Research Center, Palo Alto, CA 94301 USA. E-mail: {zhong.li, fand}@oppo.com, quanshuxue@outlook.com.*

Manuscript received 7 May 2021; revised 10 May 2022; accepted 13 October 2022. Date of publication 19 October 2022; date of current version 3 April 2023.
This work was supported in part by the Natural Science Foundation of China (NSFC) under Grants 62125106, 61860206003, and 62088102, in part by the Ministry of Science and Technology of China under Grant 2021ZD0109901, in part by the Shenzhen Science and Technology Research and Development Funds under Grant JCYJ20180507183706645, in part by the Provincial Key R&D Program of Zhejiang under Grant 2021C01016.
(Corresponding author: Lu Fang.)
Recommended for acceptance by L. Agapito.
This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2022.3215746, provided by the authors.
Digital Object Identifier no. 10.1109/TPAMI.2022.3215746



(a) RGBD stream (Input) (b) 4D reconstruction (Output)

Fig. 1. Illustration of the system and results of our RobustFusion.

like SMPL [18] into the fusion pipeline to handle more complex motions. Within this category, our conference version RobustFusion [19] (denoted as RobustFusion(Conf.)) further enables more robust monocular capture using various data-driven visual cues such as motion [20], [21], geometry [22], [23] or semantic segmentation [24]. It gets rid of the self-scanning constraint for monocular capture with re-initialization ability, where the captured performer does not need to turn around carefully to obtain complete reconstruction. However, these monocular approaches with human priors neglect to model the mutual influence between human and object, leading to limited reconstruction under the challenging interaction scenarios. On the other hand, various researchers [25], [26], [27], [28], [29], [30], [31], [32] reconstruct the 4D relations between humans and the objects or the environments. However, they only recover the naked human bodies or heavily rely on specific pre-scanned object and scene templates to optimize the spatial arrangement. Researchers pay little attention to strengthening the template-less volumetric performance capture by utilizing the rich human-object interaction priors, especially for the monocular setting.

In this paper, we attack the above challenges and propose *RobustFusion*, a robust human-object volumetric performance capture system combined with various data-driven visual and interaction cues using only a single RGBD sensor (with optional multi-view setup). As illustrated in Fig. 1, our approach solves the challenging ambiguity and severe occlusions under complex human-object interactions, achieving robust volumetric performance reconstruction, which outperforms the baselines favorably without using any pre-scanned templates.

Combining data-driven cues for robust volumetric reconstruction under challenging human-object interactions is non-trivial, let alone maintaining the lightweight property and fast running performance under the monocular setting. To encode the interaction pattern and alleviate the occlusion ambiguity, our key idea is to utilize the data-driven interaction cues for human motions prior under occlusions, as well as the rich visual cues including scene semantic segmentation, body part parsing estimation, implicit occupancy learning, and human pose and shape detection. More specifically, we first embrace the scene semantic cue for scene decoupling to model the challenging occlusions explicitly for human-object interactions. To prevent the disentanglement uncertainty, we refine the human-object segmentation through robust object tracking in an iterative manner, which utilizes previous reconstruction results for temporal consistency. We also adopt a human initialization in the first frame similar to RobustFusion(Conf.), which utilizes the SMPL model [18], human parsing and implicit occupancy learning to generate a complete and fine-detailed initial model and non-rigid motion for the human. Such initialization eliminates the tedious self-scanning constraint for more robust human-object performance capture. Then, we propose a robust human performance capture scheme with the aid of various data-driven cues. Besides the original strategy with visual priors including the human pose, shape, and parsing to enable re-initialization ability similar to RobustFusion (Conf.), we further model the interaction patterns for the challenging human-object occlusions in a data-driven manner. To this end, we introduce a novel spatial relation prior to prevent physically implausible intersections, as well as the interaction poses prior based on Gaussian Mixture Model (GMM) and the temporal interaction prior based on LSTM predictor to maintain natural motions, especially for those regions under severe occlusions. Finally, we adopt an adaptive fusion scheme to obtain temporally coherent reconstruction results. With both the human-object occlusion analysis and human parsing cue, the fusion weights are adaptively adjusted to avoid deteriorated fusion caused by tracking errors and occlusions. To summarize, our main contributions include:

- We propose a robust volumetric performance reconstruction approach for challenging human-object interaction scenarios using only a single RGBD camera, which embraces data-driven visual and interaction cues to achieve significant superiority to existing state-of-the-art methods.
- We introduce a novel scene decoupling scheme under the volumetric capture framework for explicit disentanglement of human-object interactions, with the aid of robust object tracking and semantic refinement.
- We propose a novel and robust human-object performance capture scheme with various data-driven interaction cues, which can handle challenging human motions with complex interaction patterns and severe occlusions.

2 Related Work

Human Volumetric Capture. In recent years, free-form dynamic reconstruction methods combine the volumetric fusion [33] and the embedded deformation [34]. The multi-view solutions [7], [8], [10] are difficult to be deployed for daily usage. In contrast, [11] utilizes only one common single RGBD camera and achieves real-time dynamic reconstruction. Yu et al. [16], [35] further take human articulated skeleton prior into account to increase tracking robustness, while HybridFusion [36] utilizes extra IMU sensors for more reliable motion tracking and Xu et al. [37] model the mutual gains between capture view selection and reconstruction. Besides, POSEFusion [17] combines both implicit inference network and temporal volumetric fusion in a key-frame selection scheme and can capture more dynamic details in invisible regions, and some methods [38], [39] combine the neural rendering techniques. Above methods still suffer from careful and orchestrated motions, especially

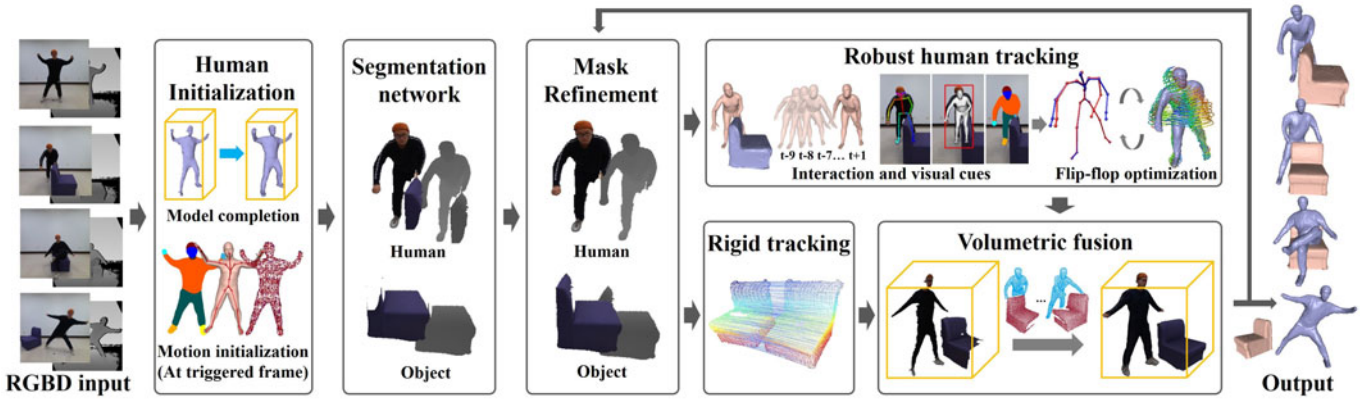


Fig. 2. The pipeline of RobustFusion. Assuming monocular RGBD input, our approach consists of a human initialization stage only at the triggered frame (Section 4.2), a scene decoupling stage that includes mask refinement and object tracking (Section 4.3), a robust human tracking stage(Section 4.4) and volumertic fusion stage (Section 4.5) to generate live 4D results.

for a tedious self-scanning process where the performer needs to turn around carefully to obtain complete reconstruction, and RobustFusion(Conf.) [19] liberates this constraint by introducing implicit occupancy method [22]. However, these methods either cannot handle modeling human-object interactions (e.g., [10], [16], [17], [19], [36]) or cannot robustly handle the fast human non-rigid motions (e.g., [2], [13]). Comparably, our approach is more robust for capturing challenging motions under human-object interaction scenarios with the re-initialization ability and enables the simultaneous reconstruction of both human and object without the self-scanning constraint.

Object-Related Reconstruction. As for object reconstruction, the method [40] utilizes structure-from-motion to recover complete 3D models from RGB images. [41] introduces color information to point cloud registration and gets more robust results. Apart from rigid objects, the reconstruction of non-rigid objects explored in work [11] mentioned above. Besides, [14] reconstruct dynamic objects and static indoor environment at the same time. Note that the dynamic motions that [11] and [14] capture are very limited. In addition to the traditional methods, recovering the 3D object shape using deep neural networks has attracted increasing attention [42], [43] in the past few years. Moreover, [44] propose a 4D human-object interaction model to detect human-object geometric relation and the interaction events. Recently, the work [27] learns the spatial arrangements of humans and objects with pose estimation in a 3D scene from a single RGB image. The methods [28], [30] try to generate the plausible human model(s) in existing 3D scenes. However, they are limited to the naked human body or the pre-scanned models.

Occlusion-Aware Tracking Methods. Human-object interaction scenarios always contain a lot of occlusions, which introduces the challenges for our dynamic reconstruction. There are several works dealing with the occlusion problem in human pose estimation. [26] utilize saliency masks as visibility information to handle the occlusions. [45] solves the occlusions by making heatmaps contain richer semantic information. [46] employs estimated 2D confidence heatmaps of keypoints and an optical-flow consistency constraint to filter out the unreliable estimations of occluded keypoints. [47] introduces a soft attention mechanism that learns to predict body-part-guided attention masks. Differently, our purpose is to keep natural motions and alleviate the impacts of object-human occlusions in a temporal RGBD sequence, with explicit human and object models. Therefore, we choose to model the interaction patterns for the challenging human-object occlusions in a data-driven manner. The use of GMM from [48] provides us an idea of utilizing a data-driven pose distribution cue for human-object interaction scenarios. [49] also utilizes the Gaussian mixture alignment strategy to track hands and objects. Besides, inspired by [50] which introduces an auto-regressive sequence modeling for 3D motion prediction, we also take pose prediction into account. Here, we choose the LSTM architecture.

Data-Driven Visual Cues. Recently, data-driven techniques have attracted more and more interest due to the rise of deep learning and RGB-based human modeling approaches bloom. First, the methods [20], [51] estimate human 2D or 3D skeletal pose, and human parametric models [18], [52] with human pose and shape parameters provide a good sparse representation for human models, based on which some recent work [21], [48], [53], [54], [55] also learn the human pose and shape from a single RGB image or video. Second, there are many approaches that directly estimate the human geometry from RGB images, such as the parametric representation [56], [57], implicit representation [22], [58], [59] and volume representation [23]. However, such predicted geometry lacks fine details and surfers inaccurate pose, which is important for immersive human modeling. Besides, some performance capture methods [60], [61] based on RGB videos leverage the above learnable pose detection [20], [51] or its own pose regression network to improve the accuracy of human motion capture, but these methods have to rely on pre-scanned template models. As for scene visual cues, scene segmentation methods [62], [63] fetch the semantic information of the whole scene including the people and objects in it, and human parsing methods [24], [64] also propose to fetch the semantic information of the human model. These data-driven methods yield colossal potential for human performance reconstruction. We explore building a robust human and object volumetric capture algorithm on top of these priors and then achieve significant superiority to previous methods.

3 Overview

RobustFusion achieves both human and objects volumetric capture under a unified framework in a model-specific way, which can perform human-object interaction reconstructions and maintain robust ability to handle challenging human motions. As illustrated in Fig. 2, our approach takes an RGBD video from Kinect v2 (or Kinect Azure) as input and generates 4D meshes, achieving more robust results than previous methods considerably. In our volumetric capture framework, we utilize TSDF [33] volume for geometry reconstruction, just like in [10], [16]. A brief introduction of our technical components is provided as follows:

Human Initialization. First, for model initialization, we follow [19] to generate a high-quality watertight human model with fine geometric details at the beginning, in which we combine the implicit occupancy regression network with the traditional non-rigid fusion pipeline using only the front-view RGBD input. Second, we further utilize the complete model to initialize both the human motions and the visual priors before the tracking stage. We adopt a hybrid motion representation [10], [16], [19], including the newly sampled ED-graph and embedded SMPL [18]. Besides, various human pose and parsing priors based on the front-view input are associated with the initialized model.

Scene Decoupling. To reconstruct the dynamic scene, we first apply a semantic segmentation network and background separation to obtain the foreground masks, including both human and objects. The segmented masks are too coarse to be applied for tracking. Thus, with the help of the reconstructed results, we can project 3D models to current 2D image and use a iterative strategy to refine the masks.

Object Tracking. After scene decoupling, we track the rigid motions of the objects by solving an optimization problem under the Iterative Closest Point (ICP) framework by taking account of color consistency, geometry consistency, and spatial relationship between the human and objects. The correct decoupling results provided by mask refinement enable accurate object tracking, and the correct tracked object models provide a good reference for scene decoupling in turn.

Robust Human Tracking. The core of our pipeline is to solve the hybrid motion parameters from the canonical frame to the current camera view. We propose a robust human tracking scheme which utilizes reliable interaction and visual data-driven priors to optimize both the skeletal and surface motions in an iterative flip-flop manner. Observed that human poses have particular patterns in the interaction with objects, we train a GMM model and LSTM predictor to exploit the spatial and temporal prior information in the optimization. Moreover, our scheme can handle challenging motions with the re-initialization ability.

Object-Aware Reconstruction. We fuse the masked depth stream into the canonical TSDF volume after motion tracing to provide temporal-coherent results for the human and objects separately. The human model is fused based on the non-rigid motion field, and the object model is fused based on the estimated rigid transformation. Based on various visual priors and object-aware occlusion ratios, we adaptively adjust the fusion weight to avoid deteriorated fusion caused by tracking errors and occlusions. Finally, dynamic atlas [10] and per-vertex color fusion are adopted to obtain 4D textured reconstruction results.

4 Technical Details

4.1 Problem Representation

Motion tracking is a core problem in our performance capture system. To robustly estimate both human and object motions, we decouple and track them separately with the data-driven cues. This subsection briefly overviews these motion representations and defines the mathematical notations in our tracking framework.

The motion of rigid objects is formulated by the rigid transformations $\mathbf{T} = \{T_i, i \in N\}$ in $\mathbf{SE}(3)$ space, where N is the number of objects. As for human motions, we adopt the efficient and robust double-layer surface representation for motion representation [16], which combines the embedded deformation (ED) and the linear human model SMPL [18]. Since we can get a complete human model after model initialization (Section 4.2), we modify the SMPL-sampled ED-graph by the ED-graph sampled on the complete model. We utilize SMPL to represent our skeleton motions. SMPL is a linear body model with $N = 6890$ vertices and $K = 24$ joints. Before posing, 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 $\boldsymbol{\theta}$ as $T(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\mathbf{T}} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta})$, where $B_s(\boldsymbol{\beta})$ and $B_p(\boldsymbol{\theta})$ are the shape blendshapes and pose blendshapes respectively. $T(\bar{\mathbf{v}}; \boldsymbol{\beta}, \boldsymbol{\theta})$ denotes the morphed 3D position for any vertex $\bar{\mathbf{v}} \in \bar{\mathbf{T}}$. The posed SMPL is further formulated as the blend skinning function: $W(T(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathcal{W})$, in terms of the body $T(\boldsymbol{\beta}, \boldsymbol{\theta})$, pose parameters $\boldsymbol{\theta}$, joint locations $J(\boldsymbol{\beta})$ and the skinning weights $\mathcal{W}$. Specifically, for any 3D vertex $\mathbf{v}_c$, the linear blend skinning (LBS) operation with the SMPL skeleton motions is formulated as $\hat{\mathbf{v}}_c = \mathbf{G}(\mathbf{v}_c, \boldsymbol{\theta})\mathbf{v}_c$, where $\mathbf{G}(\mathbf{v}_c, \boldsymbol{\theta}) = \sum_{i \in \mathcal{B}} w_{i,v_c} \mathbf{G}_i$ is the posed rigid transformation of $\mathbf{v}_c$, $\mathcal{B}$ is index set of bones, $\mathbf{G}_i = \prod_{k \in \mathcal{K}_i} \exp(\theta_k \hat{\xi}_k)$ is the rigid transformation of i-th bone referencing the parent bones whose indices are $\mathcal{K}_i$ in the backward kinematic chain, $\exp(\theta_k \hat{\xi}_k)$ is the exponential map of the twist associated with k-th bone, and w_{i,v_c} is the skinning weight associated with i-th bone and point $\mathbf{v}_c$. If $\mathbf{v}_c$ is on SMPL model, w_{i,v_c} is pre-defined in $\mathcal{W}$. If $\mathbf{v}_c$ is on the fused surface, w_{i,v_c} is given by the weighted average of its knn-nodes.

Non-rigid motions of the human is represented by a embedded deformation node-graph $G = \{\mathbf{dq}_j, \mathbf{x}_j\}$, consisting of the dual quaternions $\{\mathbf{dq}_j\}$ and the corresponding ED nodes $\{\mathbf{x}_j\}$. $SE3(\mathbf{dq}_j)$ denotes the rigid transformation in $\mathbf{SE}(3)$ space. Then for any 3D vertex $\mathbf{v}_c$ in the canonical volume, the ED warping operation is formulated as follows:

$$\tilde{\mathbf{v}}_c = ED(\mathbf{v}_c; G) = SE3\Big(\sum_{i \in \mathcal{N}(v_c)} w(\mathbf{x}_i, \mathbf{v}_c)\mathbf{dq}_i\Big)\mathbf{v}_c, \tag{1}$$

where $\mathcal{N}(v_c)$ is a set of node neighbors of $\mathbf{v}_c$, and $w(\mathbf{x}_i, \mathbf{v}_c) = \exp(-\|\mathbf{v}_c - \mathbf{x}_i\|_2^2/(2r_k^2))$ is the influence weight of the i-th node $\mathbf{x}_i$ to $\mathbf{v}_c$. The influence radius r_k is set as 0.075 m for all the ED nodes. Similarly, $\tilde{\mathbf{n}}_{v_c} = ED(\mathbf{n}_{v_c}; G)$ denotes the warped normal of $\mathbf{v}_c$ using the ED motion field G.

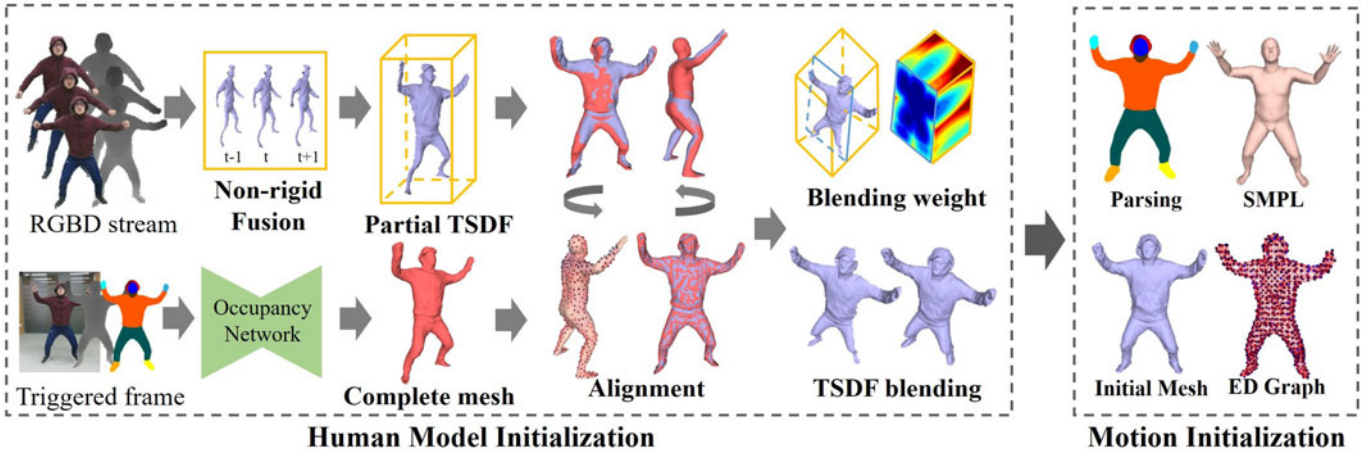


Fig. 3. Human model and motion initialization pipeline. Assuming the front-view RGBD input, both a partial TSDF volume and a complete mesh are generated, followed by the alignment and blending operations to obtain a complete human model with fine geometry details, based on which motion is initialized by re-sampling the ED-node-graph and semantic binding.

4.2 Human Initialization

Due to complex non-rigid human motions, the good initial human model and motion is critical for us to worry-free focus on human-object interactions. Fortunately, [19] provides us an available robust human performance capture baseline as initialization.

Model Initialization. To eliminate the orchestrated self-scanning constraint and the consequent fragile tracking of monocular capture, we propose a model initialization scheme using only the front-view RGBD input. As illustrated in Fig. 3, to generate high-fidelity geometry details, we first utilize the traditional ED-based non-rigid fusion method [11], [37] 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), an RGBD-version PIFu network from previous work [19] is triggered to generate a watertight mesh, in which this frame is called triggered frame. Then, to align the partial TSDF and the complete mesh, we jointly optimize the unique human shape $\boldsymbol{\beta}_0$ and pose $\boldsymbol{\theta}_0$, as well as the ED motion field G_0 from the TSDF volume to the complete mesh as follows:

$$\boldsymbol{E}_{\text{comp}}(G_0, \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \lambda_{\text{vd}}\boldsymbol{E}_{\text{vdata}} + \lambda_{\text{md}}\boldsymbol{E}_{\text{mdata}} + \lambda_{\text{bind}}\boldsymbol{E}_{\text{bind}} + \lambda_{\text{prior}}\boldsymbol{E}_{\text{prior}}. \tag{2}$$

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

$$\boldsymbol{E}_{\text{vdata}}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \sum_{\bar{\mathbf{v}} \in \mathbf{T}} \psi(\mathbf{D}(W(T(\bar{\mathbf{v}}; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0); \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)), \tag{3}$$

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:

$$\boldsymbol{E}_{\text{mdata}} = \sum_{(\bar{\mathbf{v}}, \mathbf{u}) \in \mathcal{C}} \psi(\mathbf{n}_{\mathbf{u}}^T(W(T(\bar{\mathbf{v}}; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)) - \mathbf{u})) + \sum_{(\tilde{\mathbf{v}}_c, \mathbf{u}) \in \mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T(\tilde{\mathbf{v}}_c - \mathbf{u})), \tag{4}$$

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 and $\mathbf{n}_{\mathbf{u}}$ is its normal. Note that the pose prior term $\boldsymbol{E}_{\text{prior}}$ from [48] penalizes the unnatural poses while the binding term $\boldsymbol{E}_{\text{bind}}$ from [16] constrains both the non-rigid and skeletal motions to be consistent. We solve the resulting energy $\boldsymbol{E}_{\text{comp}}$ under the 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 [7], [65]. Finally, to seamlessly blend both the partial volume and the complete mesh in the TSDF domain, we update the voxel as follows:

$$\begin{aligned} \mathbf{D}(\mathbf{v}) &\leftarrow \frac{\mathbf{D}(\mathbf{v})\mathbf{W}(\mathbf{v}) + \mathbf{d}(\mathbf{v})w(\mathbf{v})}{\mathbf{W}(\mathbf{v}) + w(\mathbf{v})}, \\ \mathbf{W}(\mathbf{v}) &\leftarrow min(\mathbf{W}(\mathbf{v}) + w(\mathbf{v}), w_{max}), \end{aligned} \tag{5}$$

where $\mathbf{D}(\mathbf{v})$ and $\mathbf{W}(\mathbf{v})$ denote its TSDF value and accumulated weight, respectively, and w_{max} is set as 32 to prevent over-smoothness of geometry during volumetric fusion in Section 4.5 and the corresponding projective SDF value $\mathbf{d}(\mathbf{v})$ and the updating weight $\mathbf{w}(\mathbf{v})$ are as follows:

$$\mathbf{d}(\mathbf{v}) = (\mathbf{u} - \tilde{\mathbf{v}})\mathbf{sgn}(\mathbf{n}_{\mathbf{u}}^T(\mathbf{u} - \tilde{\mathbf{v}})), w(\mathbf{v}) = 1/(1 + \mathbf{N}(\mathbf{v})), \tag{6}$$

Here, 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, and $\mathbf{sgn}(\cdot)$ is the sign function to distinguish positive and negative SDF.

Motion Initialization. The complete model after model initialization provides a reliable initialization for both the human motion and the utilized visual priors. As described in Section 4.1, 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, and then we rig the mesh with the pose $\boldsymbol{\theta}_0$ from its embedded SMPL model in model initialization and transfer the SMPL skinning weights to the ED nodes $\{\mathbf{x}_i\}$. 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. To initialize the pose prior, we apply OpenPose [20] 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 [24] 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 $\mathbf{l}_i$.

4.3 Scene Decoupling and Object Tracking

Accurate scene decoupling is the premise of robust motion capture that makes full use of the object-specific priors. Otherwise, the wrong segmentation reduces tracking accuracy, and segmentation noise will be fused in the models. However, the semantic segmentation network unavoidably has noise in human-object junction and occlusion, which can not be handled only by the input data. Therefore, we take advantage of our reconstructed human and object models to iteratively refine the segmentation masks to prevent disentanglement uncertainty and maintain temporal consistency. As illustrated in Fig. 2, the proposed mask refinement based on initial semantic segmentation provides accurate segmentation of both human and object. Then we can decouple the dynamic scene between rigid object motions and non-rigid human motions to track and reconstruct them. In this subsection, we summarize the mask refinement and the object tracking as follows.

Algorithm 1. Mask Refinement

Input: M_o, M_h
Output: M_o^r, M_h^r
1: $M_o^p = \pi(R_o)$
2: $M_h^p = M_h$
3: **for** $i = 0, i < 3, i++$ **do**
4: $M_o^r = M_o + M_o^p$
5: $M_o^r = M_o^r - (dep(M_h^p) < dep(M_o^r))$
6: $T_o = ICP(M_o^r)$, $M_o^p = \pi(T_o * R_o)$
7: $M_h^r = M_h - (dep(M_o^p) < dep(M_h))$
8: $M_h^p = \pi(track(M_h^r))$
9: **end for**
11: **return** M_o^r, M_h^r

Mask Refinement. We utilize a semantic segmentation method [62] and the human segmentation from Kinect SDK to get the human masks. For objects, in addition to extract object masks by utilizing background subtraction, we can also optionally obtain the labeled object masks from the segmentation network. For operation efficiency, we execute the segmentation network every five frames. At the same time, Kinect SDK provides the human mask for the entire sequence, and the object masks are provided by background separation for the remaining frames. With the human and object masks, we use the masked point cloud for motion tracking and reconstruction. However, wrong segmentation often occurs in the place where people and objects connect. Directly using the coarse segmentation mask leads the unstable tracking and erroneous reconstruction. Therefore, we propose a refinement strategy to obtain the accurate object and human masks, illustrated in Algorithm 1. Besides, we demonstrate the results of the mask refinement pipeline in Fig. 4.

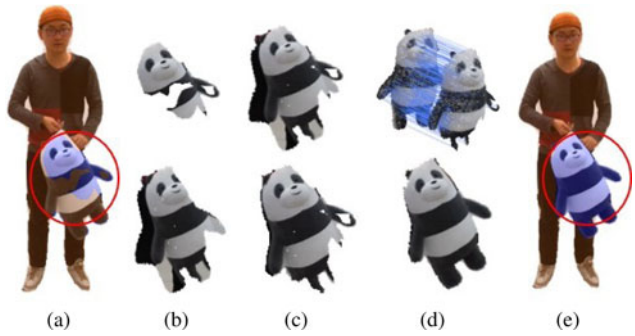


Fig. 4. The results of the proposed mask refinement. (a,e) are the segmentation mask before and after refinement. We sum the original object mask (top of (b)) and projection mask of the previous frame (bottom of (b)) to get the combined mask (top of (c)), then subtract the projection of the human body and remove noise based on point cloud continuity to get a denoised mask (bottom of (c)). Based on the denoised mask, we optimize Eq.(7) to get the transformation of the object between two adjacent frames (top of (d)) then use the estimated transformation to get the updated projection mask (bottom of (d)). The detailed iteration procedure is explained in Algorithm. 1.

Given the coarse object mask M_o and human mask M_h provided by the segmentation network and Kinect SDK, we aim to get the refined object mask M_o^r and human mask M_h^r. Here, R_o is the reconstructed object model of the previous frame. Function $\pi(\cdot)$ projects the reconstructed model to get the projected mask. M_o^p and M_h^p is the projected object and human mask based on the reconstructed models. Due to temporal continuity, the current object mask is similar to the previous frame. The current object mask for tracking is refined as $M_o^r = M_o + M_o^p$. Then we remove the human occlusion by comparing the depth (Algorithm 1 Line.5), where function $dep(\cdot)$ returns the depth value for the mask. Based on the refined object mask M_o^r, the transformation between current frame and previous frame T_o is solved by optimization $ICP(\cdot)$ and the projected object mask M_o^p is updated. Then we get the refined human mask M_h^r by comparing the depth (Algorithm 1 Line.7). Function $track(\cdot)$ returns the update human model based on the refined mask M_h^r. Then the projected human mask is updated by the tracked human model (Algorithm 1 Line.8). Moreover, we also utilize an iteration framework to raise the refinement accuracy. With the mask refinement, we successfully obtain the correct masks.

Object Tracking. To robustly track the objects, we optimize the rigid motions ($\mathbf{T} = \{T_i\}, i \in N$) of the corresponding object point clouds under ICP iteration framework as follows:

$$\boldsymbol{E}_{\text{object}}(\mathbf{T}) = \lambda_{color}\boldsymbol{E}_{\text{color}} + \lambda_{geo}\boldsymbol{E}_{\text{geo}} + \lambda_{sp\_o}\boldsymbol{E}_{\text{sp\_o}}. \tag{7}$$

The color term $\boldsymbol{E}_{\text{color}}$ is achieved by the colored point cloud registration [41], which encourages the color consistency as follows:

$$\boldsymbol{E}_{\text{color}} = \sum_{i \in N} \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{R}} (C_p(f(T_i\mathbf{q})) - C(\mathbf{q}))^2, \tag{8}$$

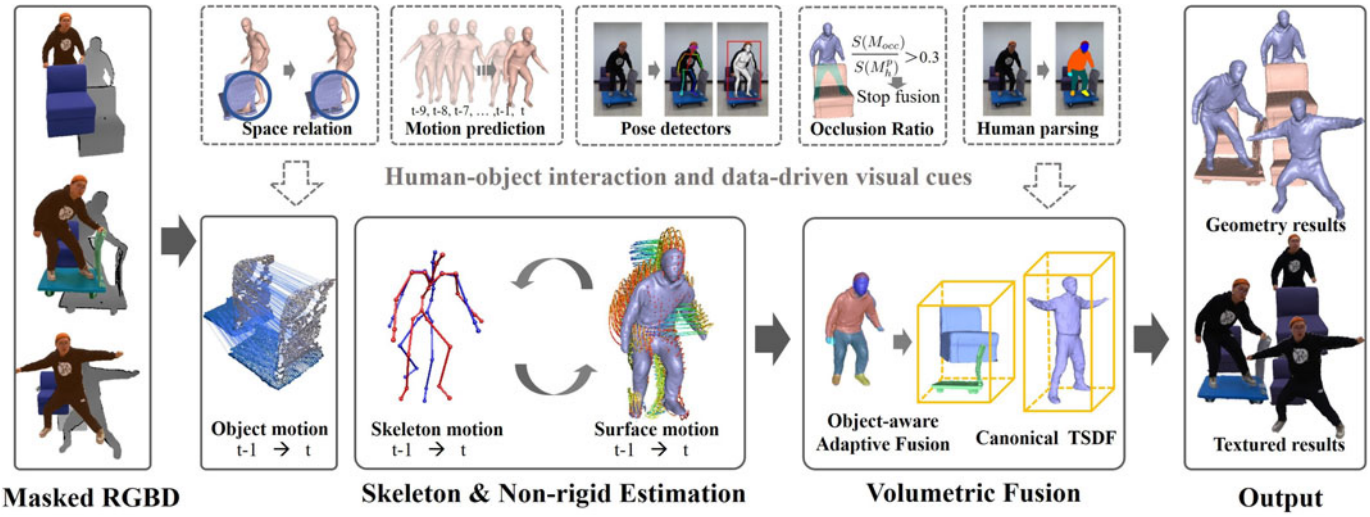


Fig. 5. The pipeline of our robust performance capture scheme. Assuming the masked RGBD input, We track the object based on the space relation cue. Then, both skeletal and non-rigid motions are optimized with the associated human-object interaction and data-driven visual cues. Finally, an object-aware adaptive volumetric fusion scheme is adopted to generated 4D models.

where N is the number of objects, $\mathcal{R}$ is the correspondence pair sets found via closest searching and $\mathbf{p}$, $\mathbf{q}$ are the closest points of frame t and frame $t-1$. Function $C(\cdot)$ returns color of the point $\mathbf{q}$ while $C_p(\cdot)$ is a pre-computed function continuously defined on the tangent plane of $\mathbf{p}$ and $f(\cdot)$ is the projection function that projects a 3D point to the tangent plane. The geometry term $\boldsymbol{E}_{\text{geo}}$ encourages the geometry consistency as follows:

$$\boldsymbol{E}_{\text{geo}} = \sum_{i \in N} \sum_{(\mathbf{p},\mathbf{q}) \in \mathcal{R}} (\mathbf{n}_{\mathbf{p}}^{T}(\mathbf{p} - T_i\mathbf{q}))^2, \tag{9}$$

where $\mathbf{n_p}$ is the normal of the point $\mathbf{p}$. To generate a healthy spatial relation without implausible interpenetration between human and objects, we introduce an interpenetration term $E_{\text{sp\_o}}$ as follows:

$$\begin{aligned}\boldsymbol{E}_{\text{sp\_o}} = &\sum_{i \in N} \sum_{\mathbf{p} \in O_i} \psi(\tau\mathbf{D}(T_i p)) + \\ &\sum_{i,j(i \neq j) \in N} \sum_{\mathbf{p} \in O_i} \psi(\tau\mathbf{D}_{oj}(T_i p)).\end{aligned} \tag{10}$$

where O_i is the i-th object, p is point of O_i, $\mathbf{D}(\cdot)$ is the same as in Eq. (3), $\mathbf{D}_{oj}(\cdot)$ takes a point in the live TSDF volume of j-th object and returns the bilinear interpolated TSDF value, and $\psi(\cdot)$ is the robust Geman-McClure penalty function, and τ is the indicator function which equals to 1 only if the visited TSDF value is positive (inside the corresponding volume).

4.4 Robust Human Tracking

As illustrated in Fig. 5, we propose a novel performance capture scheme to track challenging human motions under complex human-object interaction scenarios robustly, in which we introduce a spatial relation prior to prevent implausible interactions, data-driven interaction cues to maintain natural motions, especially for those regions under severe human-object occlusions, as well as the human pose, shape and parsing priors to enable re-initialization ability. We first optimize the motion fields described in Section 4.1 including both the skeletal pose and surface-sampled ED node-graph in a flip-flop iteration manner.

Skeleton Tracking. During each ICP iteration, we first optimize the skeletal pose $\boldsymbol{\theta}$ of the human model, which is formulated as follows:

$$\begin{aligned}\boldsymbol{E}_{\text{smot}}(\boldsymbol{\theta}) = \lambda_{\text{sd}}\boldsymbol{E}_{\text{sdata}} + \lambda_{\text{pose}}\boldsymbol{E}_{\text{pose}} + \lambda_{\text{prior}}\boldsymbol{E}_{\text{prior}} + \\ \lambda_{\text{temp}}\boldsymbol{E}_{\text{temp}} + \lambda_{\text{inter}}\boldsymbol{E}_{\text{inter}}.\end{aligned} \tag{11}$$

Here, since human motions have particular patterns in the interactions with objects, we introduce a human-object interaction term $\boldsymbol{E}_{\text{inter}}$ to Eq. (11), which includes the spacial relation prior, data-driven interaction pose prior, and motion prediction prior to keep natural motions and alleviate the impacts of severe object-human occlusions, formulated as follows:

$$\boldsymbol{E}_{\text{inter}} = \lambda_{\text{gmm}}\boldsymbol{E}_{\text{gmm}} + \lambda_{\text{lstm}}\boldsymbol{E}_{\text{lstm}} + \lambda_{\text{sp\_h1}}\boldsymbol{E}_{\text{sp\_h}}, \tag{12}$$

where $\boldsymbol{E}_{\text{gmm}}$, $\boldsymbol{E}_{\text{lstm}}$ and $\boldsymbol{E}_{\text{sp\_h}}$ are energies of interaction pose prior term, motion prediction prior term and interpenetration term respectively. The interaction pose and motion prior terms come from the experiential data-driven cues, representing the single-frame and temporal priors. At the same time, the interpenetration term represents the spatial prior of human-object interaction. We collect approximately 200000 human-object interaction temporal poses with SMPL parameters using RGBD systems [7], [10], in which we avoid object occlusion or imitate human-object interactions in the target camera view. The interaction pose prior term resembles the prior term $\boldsymbol{E}_{\text{prior}}$ from [48]. It is based on a GMM (16 Gaussians) fitted to the above data and formulated as follow:

$$E_{\text{gmm}} = -log\Big(\sum_{j} w_j N(\boldsymbol{\theta}; \mu_j, \delta_j)\Big), \tag{13}$$

where w_j, μ_j and δ_j are the mixture weight, the mean, and the variance of jth Gaussian model, respectively. Moreover,

we also train an LSTM predictor using the above data to predict the current pose in terms of the poses of the previous nine frames and formulate the motion prediction prior term $\boldsymbol{E}_{\text{lstm}}$ as follows:

$$\boldsymbol{E}_{\text{lstm}} = \psi(\boldsymbol{\theta} - L(\boldsymbol{\theta}_{t-9}, \boldsymbol{\theta}_{t-8}, \ldots, \boldsymbol{\theta}_{t-1})), \tag{14}$$

where $\psi(\cdot)$ is the robust Geman-McClure penalty function; $\boldsymbol{\theta}_i, (i = t-9, t-8, \ldots, t-1)$ are the skeleton poses of the previous 9 frames; $L(\cdot)$ is the LSTM prediction fuction. The interpenetration term $\boldsymbol{E}_{\text{sp\_h}}$ prevents unphysical intersections from human to objects in space dimension:

$$\boldsymbol{E}_{\text{sp\_h}} = \sum_{\mathbf{v}\in\mathbf{T}} \psi(\tau \mathbf{D}_o(W(T(\mathbf{v}; \boldsymbol{\beta}, \boldsymbol{\theta}); \boldsymbol{\theta})), \tag{15}$$

where $\mathbf{D}_o(\cdot)$ takes a point in the live TSDF volume of the object and returns the bilinear interpolated TSDF value, and $\psi(\cdot)$ is the robust Geman-McClure penalty function, τ is the indicator function which equals to 1 only if the TSDF value for the object of the vertex $\mathbf{v}$ on SMPL is positive (inside the object volume).

Besides, we also introduce the pose term $\boldsymbol{E}_{\text{pose}}$ in [19] and update its pose detectors to better encourage 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 [20] and EFT [53] , respectively:

$$\boldsymbol{E}_{\text{pose}} = \psi(\Phi^T(\boldsymbol{\theta} - \boldsymbol{\theta}_d)) + \sum_{l=1}^{N_J} \phi(l)(\|\pi(\hat{\mathbf{J}}_l) - \mathbf{P}_l^{2D}\|_2^2 + \|\hat{\mathbf{J}}_l - \mathbf{P}_l^{3D}\|_2^2), \tag{16}$$

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 lth joint is larger than 0.5, while Φ is the vectorized representation of $\{\phi(l)\}$. Finally, among the other terms, $\boldsymbol{E}_{\text{sdata}}$ measures the point-to-plane misalignment error between the warped geometry in the TSDF volume and the depth input:

$$\boldsymbol{E}_{\text{sdata}} = \sum_{(\mathbf{v}_c, \mathbf{u})\in\mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T(\hat{\mathbf{v}}_c - \mathbf{u})), \tag{17}$$

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 temporal term $\boldsymbol{E}_{\text{temp}}$ encourages coherent deformations by constraining the skeletal motion to be consistent with the previous ED motion:

$$\boldsymbol{E}_{\text{temp}} = \sum_{\mathbf{x}_i} \|\hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_i\|_2^2, \tag{18}$$

where $\tilde{\mathbf{x}}_i$ is the warped ED node using non-rigid motion from previous iteration; and the prior term $\boldsymbol{E}_{\text{prior}}$ from [48] penalizes the unnatural poses.

Surface Tracking. To capture realistic non-rigid deformation defined by ED-node graph G, on top of the skeleton tracking result, we solve the surface tracking energy as follows:

$$\boldsymbol{E}_{\text{emot}}(G) = \lambda_{\text{ed}}\boldsymbol{E}_{\text{edata}} + \lambda_{\text{sp\_h2}}\boldsymbol{E}_{\text{sp\_h}} + \lambda_{\text{reg}}\boldsymbol{E}_{\text{reg}} + \lambda_{\text{temp}}\boldsymbol{E}_{\text{temp}}. \tag{19}$$

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

$$\boldsymbol{E}_{\text{edata}} = \sum_{(\mathbf{v}_c, \mathbf{u})\in\mathcal{P}} \psi(\mathbf{n}_{\mathbf{u}}^T(\tilde{\mathbf{v}}_c - \mathbf{u})) + \sum_{l=1}^{N_J} \phi(l)\|\pi(\tilde{\mathbf{J}}_l) - \mathbf{P}_l^{2D}\|_2^2, \tag{20}$$

where $\tilde{\mathbf{J}}_l$ is the warped associated 3D joint of the lth joint in the fused surface. The interpenetration term $\boldsymbol{E}_{\text{sp\_h2}}$ is as follows:

$$\boldsymbol{E}_{\text{sp\_h2}} = \sum_{\mathbf{v}\in\mathbf{T}} \psi(\tau \mathbf{D}_o(\tilde{\mathbf{v}})), \tag{21}$$

note that the interpenetration term is associated with a smaller weight : $\text{sp\_h2} = \text{sp\_h1}/10$. The regularity term $\boldsymbol{E}_{\text{reg}}$ from [16] 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 $\boldsymbol{E}_{\text{temp}}$ as formulated above is fixed during current optimization.

Both the pose and non-rigid optimizations in Eqs. (11) and (19) are solved using LM method with the same PCG solver on GPU [7], [65]. 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 lth 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).

4.5 Object-Aware Reconstruction

After the above optimization, we separately fuse the masked depth into the respective canonical TSDF volume of the human and objects with occlusion analysis and human semantic cue to temporally update the geometric details. Note that each voxel in canonical space is updated using Eq. (5), while updating weight $\mathbf{w}(\mathbf{v})$ is different between human and objects.

For human reconstruction, we first discard the voxels which are collided or warped into invalid input. Then, to avoid deteriorated fusion caused by challenging motion, an effective adaptive fusion strategy as shown in Fig. 5 is proposed to model semantic motion tracking behavior. To this end, we apply the human parsing method [24] to the 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:

$$\mathbf{w}(\mathbf{v}) = \exp\left(\frac{-\|\Phi^T(\boldsymbol{\theta}^* - \boldsymbol{\theta}_d)\|_2^2}{2\pi}\right) \sum_{i\in\mathcal{N}(v_c)} \frac{\varphi(\mathbf{l}_i, L(\pi(\tilde{\mathbf{x}}_i)))}{|\mathcal{N}(v_c)|}, \tag{22}$$

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 a robust weighting strategy measures the tracking performance based on the human pose and semantic priors. Then, $\mathbf{w}(\mathbf{v})$ is set to be zero if it is less than a truncated threshold (0.2 in our setting), to control the minimal integration and further avoid deteriorated fusion of severe tracking failures. $M_{occ} = (dep(M_h^p) > dep(M_o^p))$ is the mask of human occluded by object. Function $S(\cdot)$ returns the pixel number of mask. When the severe object-human occlusion occurs as $S(M_{occ})/S(M_h^p)$ is bigger than a threshold (0.3 in our setting), we also set $\mathbf{w}(\mathbf{v})$ to zero in case deteriorated fusion caused by false segmentation results and tracking errors caused by occlusions. As for object reconstruction, R_{in} is the root mean square error (RMSE) of all inlier correspondences in the ICP framework. TSDF fusion is performed every five frames based on Eq. (5) only if R_{in} is less than a certain value (0.003 in our setting), in which the updating weight $\mathbf{w}(\mathbf{v})$ is formulated as $\mathbf{w}(\mathbf{v}) = \frac{0.0048}{R_{in}+0.0024}$. Again, similar to human volumetric fusion, once the human occludes an object, we stop the TSDF fusion for the whole object. Finally, the dynamic atlas scheme [10] and per-vertex color fusion are adopted to obtain 4D textured reconstruction results for human and objects, respectively.

5 Experimental Results

Fig. 6 demonstrates the results of RobustFusion, where both the challenging motions with human-object interactions and the fine geometry and texture details are faithfully captured. Our approach can even faithfully reconstruct the interaction scenarios with multiple performers and various objects (see the last row of Fig. 6). Please also kindly refer to the supplemental video for the sequential 4D reconstruction results.

5.1 Performance Runtime and Experimental Setting

We run our experiments on a PC with an NVIDIA GeForce GTX TITAN Xp GPU and an Intel Core i7-7700 K CPU. Our human initialization takes 15 s, and the following robust performance capture pipeline runs at an average of 135 ms per frame, where the visual priors collecting takes 97 ms, the robust human-object tracking takes around 21 ms with 4 ICP iteration and 17 ms on average for all the remaining computations. Note that the semantic segmentation network and volumetric fusion for objects are executed every five frames.

As for optimization parameters in all experiments, we use the following empirically determined parameters: $\lambda_{vd} = 1.0$, $\lambda_{md} = 2.0$, $\lambda_{bind} = 1.0$, $\lambda_{prior} = 0.01$, $\lambda_{color} = 0.1$, $\lambda_{geo} = 0.9$, $\lambda_{sp\_o} = 1.0$, $\lambda_{sd} = 4.0$, $\lambda_{pose} = 2.0$, $\lambda_{temp} = 1.0$, $\lambda_{inter} = 1.0$, $\lambda_{gmm} = 0.02$, $\lambda_{lstm} = 0.1$, $\lambda_{sp\_h1} = 2.0$, $\lambda_{sp\_h2} = 0.2$, $\lambda_{ed} = 4.0$ and $\lambda_{reg} = 5.0$. For the ED model, we use the four nearest node neighbors for ED warping and the eight nearest node neighbors to construct the ED graph. For the TSDF voxel, the size is set as 4 mm in each dimension. As for experimental data, we capture 16 human-object interaction sequences using an Azure Kinect RGBD sensor and also borrow one human-only sequence from [36], with 48000 frames, 7 performers and 7 objects.

5.2 Comparison to the State-of-The-Arts

For throughout comparison, we compare our RobustFusion against the state-of-the-art methods in this subsection both qualitatively and quantitatively, including DoubleFusion [16], UnstructuredFusion [10], HybridFusion [36], RobustFusion (Conf.) [19] and POSEFusion [17].

Qualitative Comparison. These state-of-the-art methods are restricted to human reconstruction without modeling human-object interactions, and UnstructuredFusion [10] is a multi-view method. For a fair comparison of dynamic reconstruction at the scenes with objects, we test the above state-of-the-art methods on the same refined segmentation results of the human in our setting and modify UnstructuredFusion [10] into the monocular setting by removing their online calibration stage.

The qualitative comparison of our approach against DoubleFusion [16], single-view UnstructuredFusion [10] and RobustFusion(Conf.) [19] is as shown in Fig. 7. Both DoubleFusion [16] and UnstructuredFusion [10] suffer from the fast human motions and the severe occlusions due to the human-object interactions. Moreover, without a complete model due to the lack of orchestrated self-circling motions, they tend to integrate erroneous surfaces at the newly fused region. With the aid of various visual priors, RobustFusion(Conf.) [19] is more robust to the fast motions but still suffers from severe occlusions, leading to wrong tracking results in the limb regions. In contrast, benefit from our robust human tracking scheme based on data-driven interaction and visual cues, our approach achieves significantly more robust tracking results, especially for challenging occluded and fast motions. Besides, we compare against the latest volumetric method POSEFusion [17], which combines implicit inference network with a key-frame selection strategy to capture details in invisible regions. As shown in Fig. 8, our approach can achieve more accurate human tracking and visually pleasant reconstruction results with the aid of human-object interaction cues. Nevertheless, our approach can faithfully reconstruct both the humans and objects in the interaction scenarios, which is unseen in the previous monocular fusion approaches.

Quantitative Comparison. For quantitative comparison, we first utilize the average projective numerical metric. Specifically, we render the reconstructed result to a depth map in the camera view and compute its MAE (Mean Absolute Error) by taking the depth input as the reference only in the intersection between the rendered surface and the human depth. Note that even without ground truth reconstruction, this MAE metric encodes the reconstruction error for the non-rigid motion capture process of each method, providing a reliable quantitative comparison. We only compute MAE in the human regions for a fair comparison since previous methods cannot reconstruct objects. Table 1 demonstrates the MAE of different sequences and average value across all sequences in our experiments, in which our method leads to considerably less error. Moreover, Fig. 9 demonstrates that our method achieves high-quality reconstruction results with less accumulated artifacts, using the corresponding sequence "Human-object interactions with a sofa" in Table 1. Note that our MAE for this sequence is 6.60 mm, compared favorably with 17.24 mm for the

Fig. 6. 4D human and object reconstructed results of the proposed RobustFusion system, and the interacted objects include a sofa, a cart, two cartons, a piece of luggage, a chair, and a toy.

reconstructed results provided by POSEFusion [17]. These quantitative comparisons above reveal the effectiveness of our method for more robust and accurate human motion tracking and reconstruction.

To illustrate our robustness for human-specific motions, we further compare against HybridFusion [36], which uses extra body-worn IMU sensors. We utilize the challenging sequence with ground truth from [36] and remove their

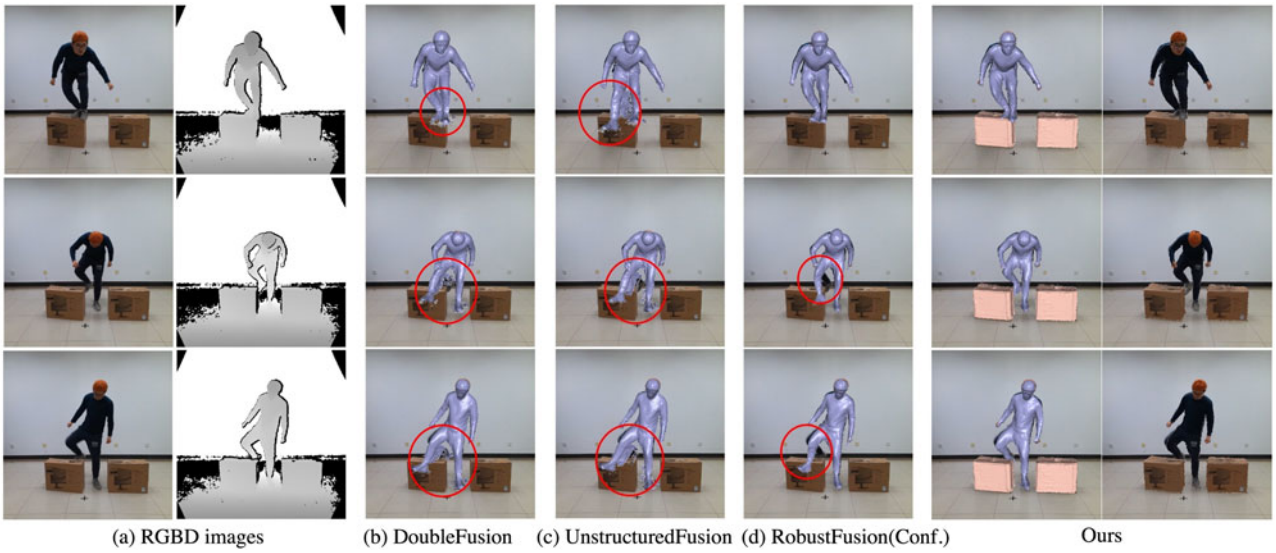


Fig. 7. Qualitative comparison. (b-d) are the geometry results of State-of-the-arts. (e) are the geometry/texture results of our method.

orchestrated self-scanning process for our methods. Even though this sequence does not include human-object interaction scenarios, such an experiment further illustrates that our approach with data-driven interaction cues can handle challenging human-only motions. The quantitative comparison in terms of the per-frame error in Fig. 10e and the average errors among the whole sequence in Table 2 demonstrate that both our approach and our preliminary version [19] achieve a significantly better result than DoubleFusion and even comparable performance against HybridFusion. Note that HybridFusion still relies on the self-scanning stage for sensor calibration and suffers from missing geometry caused by the body-worn IMUs as shown in Fig. 10a, while our approach eliminates such tedious self-scanning and achieves complete and plausible reconstruction results.

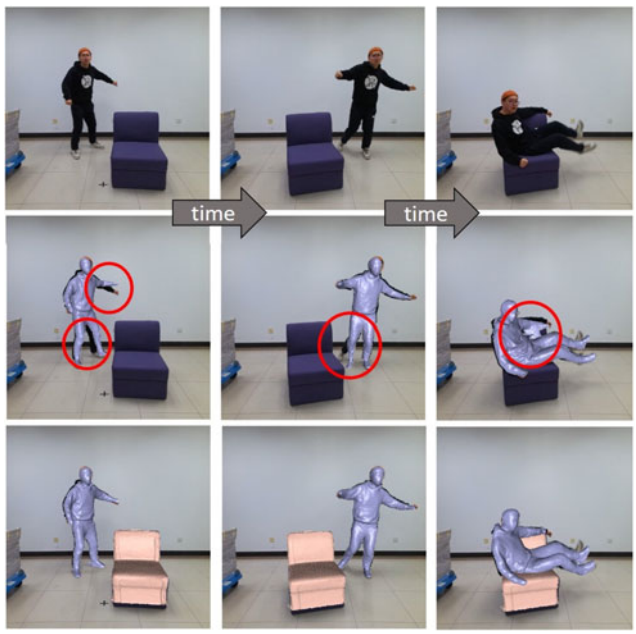


Fig. 8. Qualitative comparison. The first row is the reference RGB images. The second to third row are the geometry results of POSEFusion [17] and our method, respectively.

5.3 Ablation Study

In this subsection, we evaluate each technical contribution of our RobustFusion separately. Specifically, we evaluate the human initialization, mask refinement, object tracking, robust human tracking, and object-aware adaptive fusion, respectively. Moreover, we also evaluate our extension capability by experiments in multi-person and multi-camera scenarios.

Human Initialization. For completeness of evaluation, we first evaluate the human initialization scheme on a sequence without a carefully designed self-scanning process organized as model completion and initialization in performance capture stages in [19]. As shown in Fig. 11, without model initialization, only partial initial geometry with

TABLE 1
Average Projective Numerical Errors (mm) of Our Captured Sequences for the Concerned Methods: DoubleFusion [16], UnstructuredFusion [10], RobustFusion(Conf.) [19] and Our Methods, Where the Corresponding Sequences Can Refer to the Supplementary Video

Human-object Interactions	[16]	[10]	[19]	Ours
with luggage & chair	29.66	28.34	22.32	17.74
with two cartons	16.56	11.38	8.37	7.61
dragging things	9.11	8.56	6.45	4.96
rotating a chair	17.45	15.10	10.34	8.61
with a luggage	18.14	13.24	9.42	7.26
with luggage & carton (1)	14.35	10.57	8.51	7.87
with luggage & carton (2)	18.48	13.82	10.72	8.50
with a cart (girl)	17.75	12.47	9.49	8.87
with cart & carton (1)	17.83	13.27	10.34	8.80
with cart & carton (2)	12.93	7.91	5.48	5.05
with a sofa	11.64	7.66	7.11	6.60
with a cart (boy)	23.14	18.96	15.55	15.04
with backpack & toy	34.34	32.12	28.34	23.37
average across above	**18.57**	**14.88**	**11.73**	**10.02**

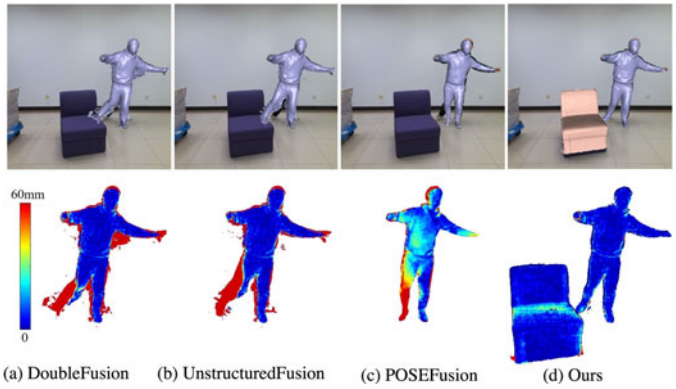


Fig. 9. Qualitative comparison. The first row indicate the geometry results and the color-coded maps indicate the projective errors.

SMPL-based ED node-graph leads to inferior tracking and erroneous reconstruction results. This exactly explains the reason that DoubleFusion [16] and UnstructuredFusion [10] fail without careful self-scanning process. To evaluate our alignment during model initialization and demonstrate the superiority of our modified motion representation over original representation in previous methods [10], [16], the skeletal pose is optimized during alignment optimization, and only SMPL-based double-layer ED-graph is adopted for motion tracking, where the results are still imperfect. In contrast, our approach with both model and motion initialization successfully obtains a watertight and fine-detailed human mesh and enables more robust motion tracking.

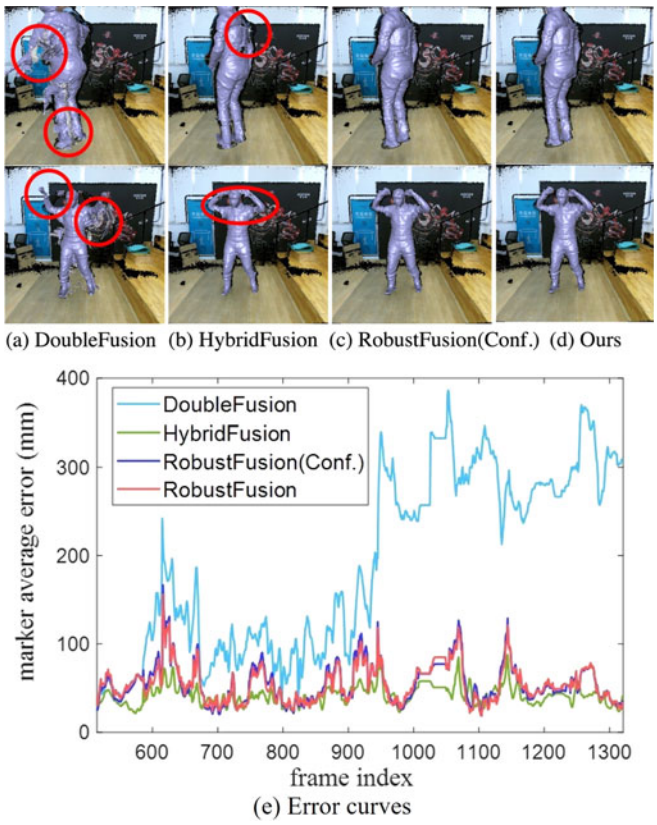


Fig. 10. Quantitative comparison. (a-d) are the reconstruction geometry results. (e) is the error curves.

TABLE 2
Average Errors on the Entire Sequence Compared to the Ground Truth Observation From the Vicon System, for These Three Methods: DoubleFusion [16], HybridFusion [36], RobustFusion(Conf.) [19] and Our Method, Respectively

	[16]	[36]	[19]	Ours
average error	0.1904 m	0.0417 m	0.0553 m	0.0546 m

Mask Refinement and Object Tracking. Here we evaluate the proposed mask refinement in Fig. 12. Since the toy is unlabeled in the network [62], we extract its masks by utilizing background subtracting. Therefore, the segmentation of objects is also dependent on human segmentation results. The original human segmentation mask in Fig. 12b is inaccurate, especially when human-object interaction occurs, leading to misaligned object masks due to the overlaying of the object on the performer. Then, directly using such coarse segmentation results leads to unstable tracking and erroneous object reconstructions as shown in Fig. 12c. In contrast, with our mask refinement scheme, the object is separated from the human segmentation result correctly (Fig. 12d). As a result, our approach achieves more robust human and object tracking results in Fig. 12e, which illustrates the effectiveness of our layer-wise strategy and mask refinement scheme.

Besides, Fig. 13 further demonstrates the robustness of our mask refinement for object mask segmentation and rigid tracking. Although sofa/chair is labeled in the network [62], it occasionally fails to extract the masks as shown in the second row of Fig. 13. With the refined masks and

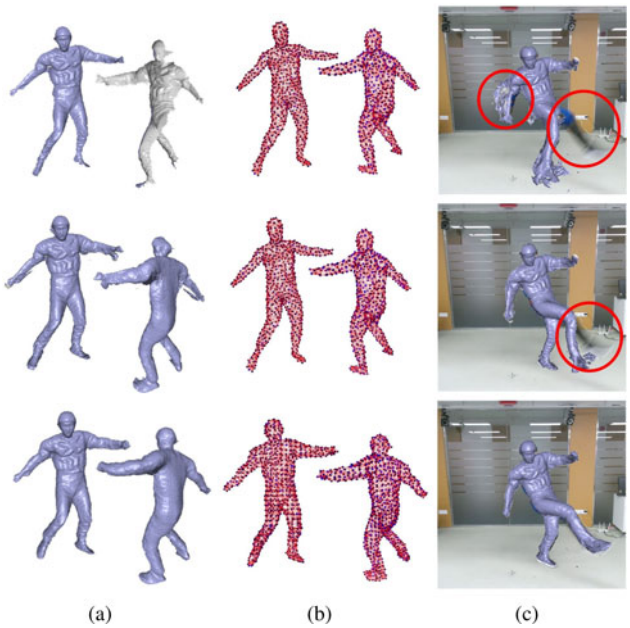


Fig. 11. Evaluation of human initialization. (a) is the 3D human model in two views. (b) is the ED node-graph that formulates the surface motions. (c) is the following tracking results that overlay on reference color image based on the corresponding 3D surface geometry and the ED node-graph.The results from the first row to the third are the results without human initialization, with initialization only using skeleton optimization and SMPL-based node-graph, and with our entire initialization process, respectively.

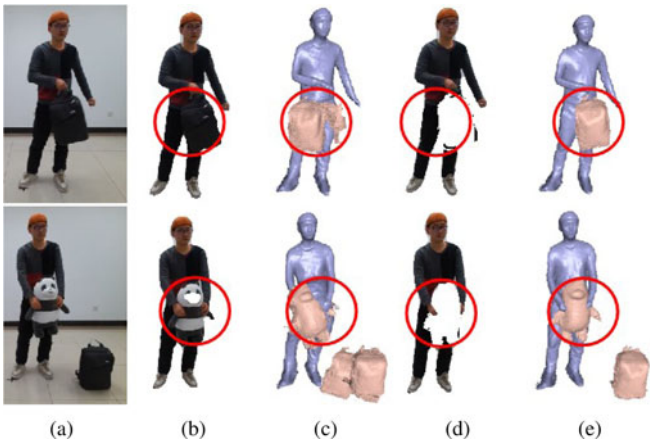


Fig. 12. Evaluation of the mask refinement. (a) Reference color images. (b) The human masks without refinement. (c) The reconstruction results without refinement. (d) The human masks with refinement. (e) The reconstruction results with refinement.

object tracking optimization in the fourth row of Fig. 13, we can achieve more accurate object tracking. The corresponding quantitative comparison in Fig. 14 further demonstrates that our method achieves the highest accuracy, where the MAE for the entire object sequence is 35.10 mm and

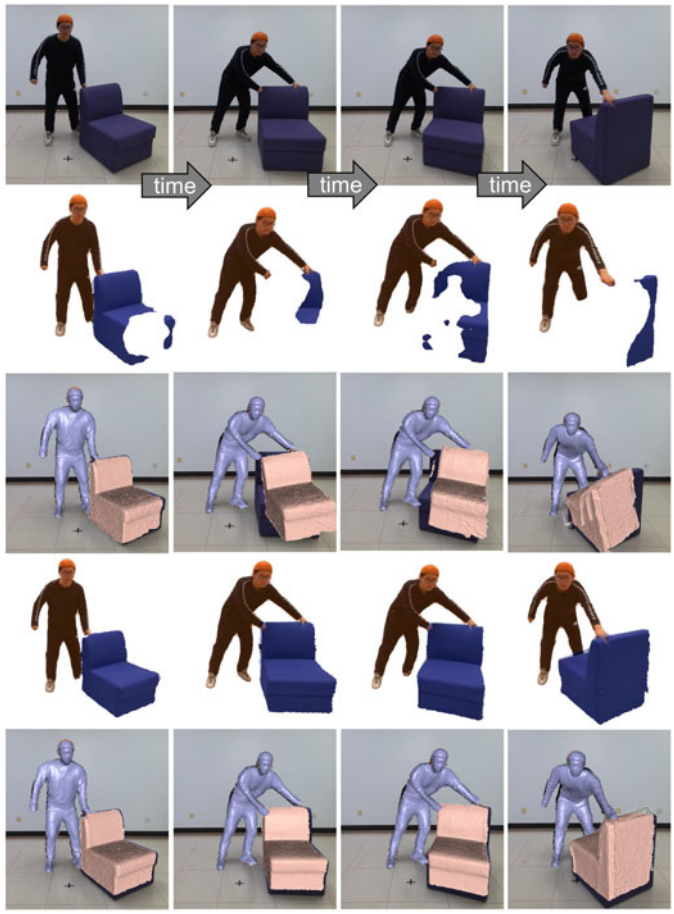


Fig. 13. Evaluation of the mask refinement (for objects). The first row demonstrates the input color images. The second and the fourth row are the original masks from the network [62] and our refined masks, respectively. The third and the last row overlaying on the color images are the tracking results using the object masks before and after mask refinement, respectively.

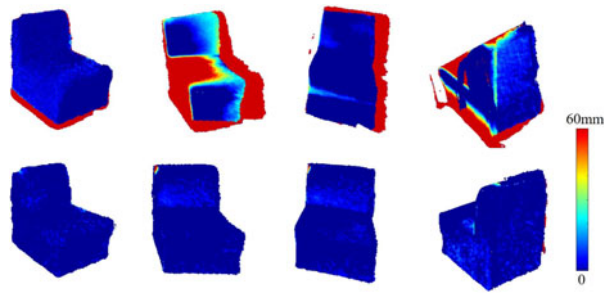


Fig. 14. Evaluation of the mask refinement (for objects). The color-coded maps indicate the projective error maps, in which the two rows are corresponding to the results in Fig. 13.

11.11 mm for our method w/o and with mask refinement, respectively.

Robust Human Tracking. Our robust human tracking is based on human-object spatial relation analysis using various data-driven cues. Here, we evaluate them one by one. First, as shown in Fig. 15, we evaluate our human-object spatial relation cue – the interpenetration term for human motion optimization. Note that we eliminate the interpenetration term by setting $\lambda_{sp\_h1} = 0$ and $\lambda_{sp\_h2} = 0$ in Fig. 15b, where the human model erroneously inserts into the cart. Differently, our full pipeline provides an essential spatial constraint for human motions estimation, especially in occlusion cases where no direct observation is available like the leg in Fig. 15. Benefit from our interpenetration term, we successfully avoid the interpenetration between human and cart models as demonstrated in Fig. 15c.

Then, we evaluate the data-driven visual cues – the pose term in human motion optimization. Similar to the preliminary method [19], we compare to the variation of our pipeline without pose prior in two scenarios where fast motion or disappear-reoccurred case happens. The first row of Fig. 16 demonstrates that our variation without pose term

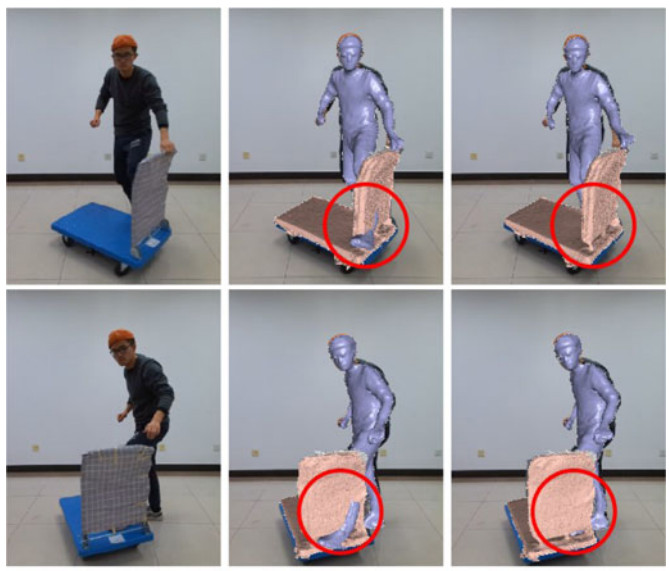
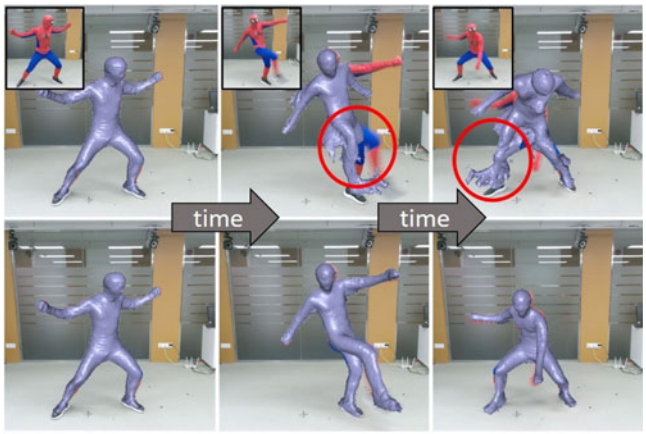
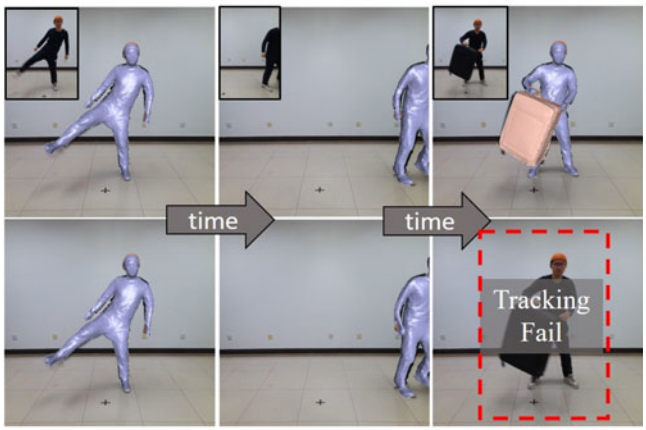


Fig. 15. Evaluation of the interpenetration term. (a) Reference color images. (b) The results without interpenetration term. (c) The results with interpenetration term.



(a) Fast motion scenarios

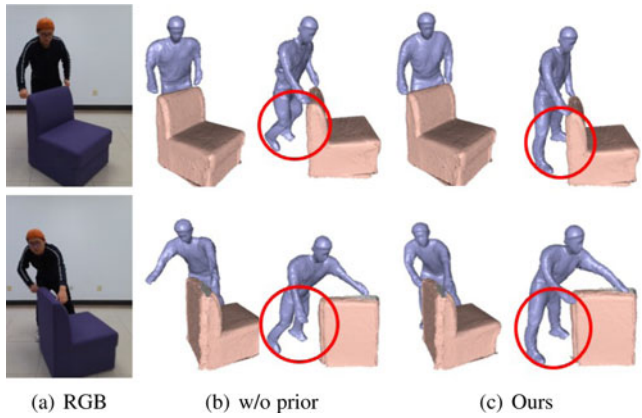


(b) Disappear and reoccur scenarios

Fig. 16. Evaluation of pose term. The top row of (a) and (b) are results of our method without pose term. The bottom row of (a) and (b) are results of our approach with pose term in optimization.

suffers from severe accumulated error, especially for the limb region with faster motions and more depth noise from the commercial sensor. Our full pipeline relieves this problem and help achieve superior tracking results for these challenging cases. Besides, as shown in the second row of Fig. 16, with the aid of pose detection and scene segmentation, the system can screen the disappearance and reappearance of the person with reconstructing object and achieve the recovery from the failing track.

Besides, the evaluation of empirical data-driven terms, including an interaction pose prior and a temporal motion prediction prior, is provided in Fig. 17. We eliminate the data-driven interaction terms by setting $\lambda_{lstm} = 0$ and $\lambda_{gmm} = 0$. Then, due to the severe occlusions between the performer and the sofa, this variation generates unnatural motion estimation for the occluded legs as shown in the different rendered views in Fig. 17b. With the aid of the empirical constraint, our approach can generate more plausible and reasonable results, as shown in the corresponding subfigures (c) of Fig. 17. In this comparison, the pose estimation of the unobserved region is up to the continuity of observed joints in the kinematic chain, similar to the unconstrained optimization that may lead to severe deviation from the real situation. Differently, the human skeletal pose estimation with an empirical constraint can be well deducted by the historical experience, including both interaction pose distribution



(a) RGB (b) w/o prior (c) Ours

Fig. 17. Evaluation of the prior constraints. (a) Reference color images. (b) The results without prior constraints (front view and side view respectively). (c) The results with prior constraints (front view and side view respectively).

prior based on GMM and temporal motion prediction based on LSTM.

Furthermore, we also evaluate the impact of the occlusion ratio on the tracking results. Here, the occlusion ratio is calculate as $S(M_{occ})/S(M_h^p)$ in Section 4.5. To simulate the different occlusion ratios, we expand the object occlusion masks manually through image expansion. Then, we evaluate the tracking error by calculating the mean square error (MSE) between the projection pixels of the reconstructed model and pixels of input images. As shown in Fig. 18, with the increase of occlusion ratios, the tracking accuracy decreases in a small range. The qualitative results are as demonstrated in Fig. 19, in which we can see that the severe occlusion leads to inaccurate but reasonable results due to the lack of information.

Object-Aware Adaptive Fusion. To evaluate our object-aware adaptive fusion scheme based on the occlusion relation and semantic errors, we compare our full volumetric fusion pipeline and the method variation without the object-aware adaptive fusion strategy. The comparison in Fig. 20a demonstrates that our full pipeline can effectively avoid the severe accumulated error for those regions with

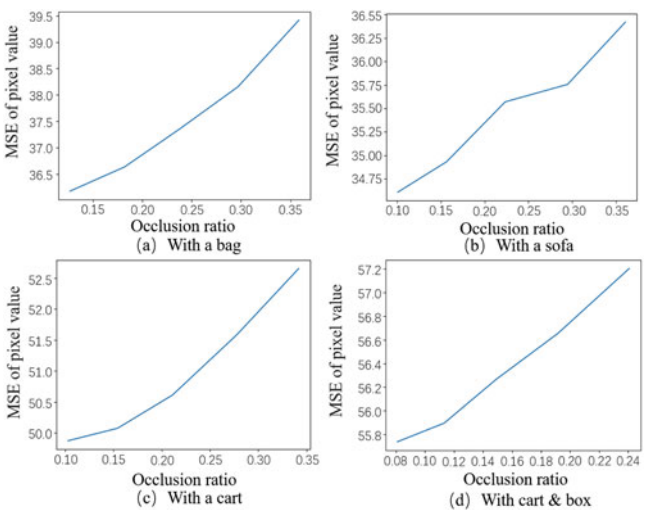


Fig. 18. Quantitative evaluation of occlusion ratios. The error curves for 4 sequences.

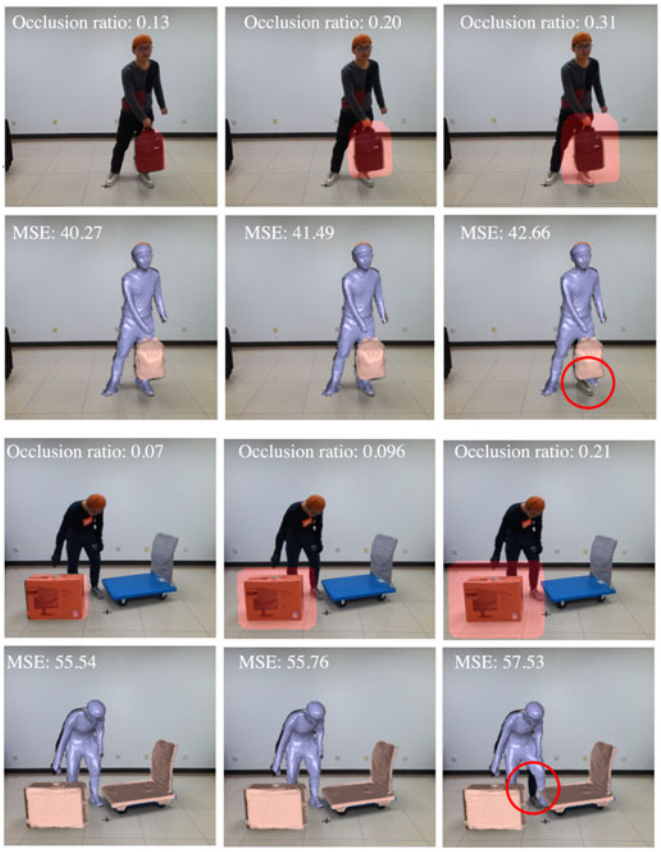


Fig. 19. Evaluation of occlusion ratios. The images listed above are the input image, and the red masks are the expanded mask of occlusion. The images below are the reconstruction results.

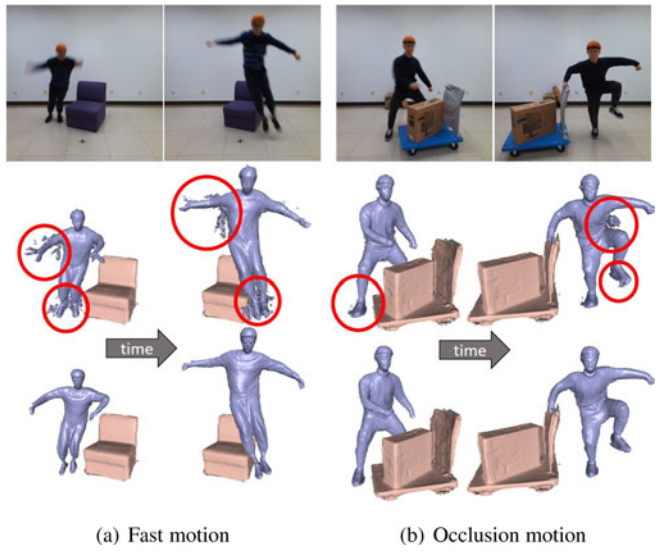


Fig. 20. Evaluation of the object-aware adaptive fusion. (a) A sequence with high-speed motions. (b) A sequence with occlusions. The first to the third row are reference color images, the geometry results without object-aware adaptive fusion, and the geometry results with object-aware adaptive fusion.

high-speed motions, such as jumping over the chair. Besides, as shown in the second row of Fig. 20b, the occlusion by object leads to erroneous surface generation (e.g., when the performer pulls the cart and walks behind the object). In contrast, the geometry results in the bottom row of Fig. 20b demonstrate that our object-aware adaptive fusion can successfully model occluded scenarios and avoid deteriorated fusion.

Expansion for Multi-Person Scenarios. Here we show our capability for extending to multi-person capture. As demonstrated in the last row of Fig. 6, with semantic segmentation labels of different subjects of the whole scene and our mask refinement process, we can also enable multi-person reconstruction. Specifically, we reconstruct one person and the interacted object firstly and then reconstruct another person by treating the first one as an object. Note that to capture such a larger scene with two persons, we deploy a two-camera system using the online calibration from [10]. We believe that it is an essential step for the reconstruction of more general dynamic scenes.

5.4 Limitation

As a trial for robust monocular volumetric performance capture under human-object interactions, we have demonstrated compelling 4D reconstruction results. Nonetheless, our approach is subject to some limitations.

Similar to the previous methods [10], [16], [17], [19], our method cannot reconstruct the extremely fine details of the performer, such as the fingers, the subtle expression, and shaggy hair, due to the limited resolution and inherent noise of the depth input. It is promising to adopt data-driven techniques to further generate visually pleasant synthetic geometry details in those model-specific regions. Besides, the reconstruction of loose and wide cloth such as a long skirt with high-speed motions remains challenging since it is difficult to track such large free-form non-rigid deformation beyond human skeletal motions. It is also challenging for human initialization in Section 4.2. A better human model regression algorithm during model initialization or utilizing a pre-scanned detailed template will help remove this limitation. Furthermore, we cannot handle surface splitting topology changes like clothes removal, which we plan to address by incorporating the key-volume update technique [7]. As common for learning methods, the utilized scene semantic segmentation, human parsing, and pose estimation fail for extreme scenarios not seen in training, such as severe and extensive (self-)occlusions under extreme side-view observation. However, our mask refinement strategy turns to obtain accurate masks, and data-driven cues of motion prediction and pose prior help us to relieve the occlusion problem with re-initialization ability. As for more general interactions, our current system still cannot handle tiny objects which can be played with fingers or non-rigid objects like dolls or papers, which restricts the wide practical applications of our approach. The limitation of tiny objects is also due to the limited image resolution and quality of the available commercial RGBD sensors. We plan to combine those the task-specific approach such as [32], [66] for fine-grained interaction modeling and extend our method to non-rigid objects by modifying our node-graph sampling and updating strategy. Besides, we has tried to handle multi-person scenarios with inter-person interactions at a certain level. It is a promising and challenging direction to deal with more general inter-person interactions such as dancing, wrestling, and hugging, even using the same monocular RGBD input.

6 Conclusion

We have presented RobustFusion, a robust volumetric performance reconstruction approach for complex human-object interactions and challenging human motions using only a single RGBD sensor. It combines various data-driven visual and interaction cues for robust human-object 4D reconstruction whilst still maintaining light-weight computation and monocular setup. Our scene decoupling scheme with segmentation refinement and robust object tracking enables explicit human-object disentanglement and temporal-consistent modeling, while our human initialization gets rid of the tedious self-scanning constraint. Our robust human performance capture with various visual and interaction cues models complex interaction patterns in a data-driven manner and enables natural motion reconstruction under challenging human-object occlusions, with unique re-initialization ability. Our object-aware adaptive fusion scheme successfully avoids deteriorated fusion and obtains temporally coherent human-object reconstruction with the aid of occlusion analysis and human parsing cue. Extensive experimental results demonstrate the effectiveness and robustness of our approach for compelling performance capture in various challenging scenarios with human-object interactions. We believe that it is a critical step to enable robust and lightweight dynamic scene reconstruction under human-object interactions, with many potential applications in VR/AR, entertainment, human behavior analysis, and immersive telepresence.

Acknowledgments

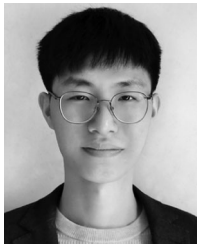
Zhuo Su, Lan Xu, and Dawei Zhong contributed equally to this work.

References

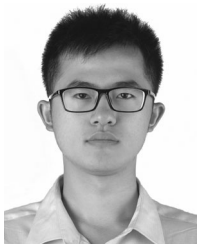
[1] A. Tevs et al., "Animation cartography–intrinsic reconstruction of shape and motion," *ACM Trans. Graph.*, vol. 31, no. 2, pp. 1–15, 2012.
[2] N. J. Mitra, S. Flöry, M. Ovsjanikov, N. Gelfand, L. J. Guibas, and H. Pottmann, "Dynamic geometry registration," in *Proc. Symp. Geometry Process.*, 2007, pp. 173–182.
[3] H. Li, B. Adams, L. J. Guibas, and M. Pauly, "Robust single-view geometry and motion reconstruction," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. no. 175.
[4] H. Li et al., "Temporally coherent completion of dynamic shapes," *ACM Trans. Graph.*, vol. 31, no. 1, Feb. 2012, Art. no. 2.
[5] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 103–110.
[6] A. Collet et al., "High-quality streamable free-viewpoint video," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 69.
[7] M. Dou et al., "Fusion4D: Real-time performance capture of challenging scenes," in *Proc. ACM SIGGRAPH Conf. Comput. Graph. Interactive Techn.*, 2016, Art. no. 114.
[8] M. Dou et al., "Motion2Fusion: Real-time volumetric performance capture," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 246:1–246:16, Nov. 2017.
[9] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3D deformation model for tracking faces, hands, and bodies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8320–8329.
[10] L. Xu, Z. Su, L. Han, T. Yu, Y. Liu, and L. FANG, "Unstructuredfusion: Realtime 4D geometry and texture reconstruction using commercialrgbd cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2508–2522, Oct. 2020.
[11] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 343–352.
[12] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger, "Volumedeform: Real-time volumetric non-rigid reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 362–379.
[13] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic, "KillingFusion: Non-rigid 3D reconstruction without correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5474–5483.
[14] H. Zhang and F. Xu, "Mixedfusion: Real-time reconstruction of an indoor scene with dynamic objects," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 12, pp. 3137–3146, 2018.
[15] T. Yu et al., "BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera," in *Proc. IEEE Int. Conf. Comput. Vis.*. ACM, 2017, pp. 910–919.
[16] T. Yu et al., "DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7287–7296.
[17] Z. Li, T. Yu, Z. Zheng, K. Guo, and Y. Liu, "Posefusion: Pose-guided selective fusion for single-view human volumetric capture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14162–14172.
[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
[19] Z. Su, L. Xu, Z. Zheng, T. Yu, Y. Liu, and L. Fang, "RobustFusion: Human volumetric capture with data-driven visual cues using a RGBD camera," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 246–264.
[20] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.
[21] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Regognit.*, 2018, pp. 7122–7131.
[22] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2304–2314.
[23] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3D human reconstruction from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7738–7748.
[24] T. Zhu and D. Oved, "Bodypix github repository," 2019. [Online]. Available: https://github.com/tensorflow/tfjs-models/tree/master/body-pix
[25] Y. Zhang, M. Hassan, H. Neumann, M. J. Black, and S. Tang, "Generating 3D people in scenes without people," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6194–6204.
[26] T. Zhang, B. Huang, and Y. Wang, "Object-occluded human shape and pose estimation from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7376–7385.
[27] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa, "Perceiving 3D human-object spatial arrangements from a single image in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 34–51.
[28] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang, "PLACE: Proximity learning of articulation and contact in 3D environments," in *Proc. IEEE Int. Conf. 3D Vis.*, 2020, pp. 642–651.
[29] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human poseitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4318–4329.
[30] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3D scenes by learning human-scene interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 703–14 713.
[31] P. Patel, C.-H. P. Huang, J. Tesch, D. T. Hoffmann, S. Tripathi, and M. J. Black, "AGORA: Avatars in geography optimized for regression analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 468–13 478.
[32] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 581–600.
[33] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, 1996, pp. 303–312, doi: 10.1145/237170.237269.
[34] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. no. 80.
[35] T. Yu, J. Zhao, Y. Huang, Y. Li, and Y. Liu, "Towards robust and accurate single-view fast human motion capture," *IEEE Access*, vol. 7, pp. 85 548–85 559, 2019.
[36] Z. Zheng et al., "HybridFusion: Real-time performance capture using a single depth sensor and sparse imus," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 389–406.

[37] L. Xu, W. Cheng, K. Guo, L. Han, Y. Liu, and L. Fang, "FlyFusion: Realtime dynamic scene reconstruction using a flying depth camera," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 1, pp. 68–82, Jan. 2021.
[38] R. Pandey et al., "Volumetric capture of humans with a single RGBD camera via semi-parametric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9701–9710.
[39] R. Martin-Brualla et al., "LookinGood: Enhancing performance capture with real-time neural re-rendering," *ACM Trans. Graph.*, vol. 37, no. 6, Dec. 2018, Art. no. 255.
[40] X. Huang, I. Walker, and S. Birchfield, "Occlusion-aware reconstruction and manipulation of 3D articulated objects," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2012, pp. 1365–1371.
[41] J. Park, Q.-Y. Zhou, and V. Koltun, "Colored point cloud registration revisited," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 143–152.
[42] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 165–174.
[43] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, "Pix2vox : Multi-scale context-aware 3D object reconstruction from single and multiple images," *Int. J. Comput. Vis.*, vol. 128, no. 12, pp. 2919–2935, 2020.
[44] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1165–1179, Jun. 2016.
[45] Y. Cheng, B. Yang, B. Wang, and R. T. Tan, "3D human pose estimation using spatio-temporal networks with explicit occlusion training," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10 631–10 638.
[46] Y. Cheng, B. Yang, B. Wang, W. Yan, and R. T. Tan, "Occlusion-aware networks for 3D human pose estimation in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 723–732.
[47] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3D human body estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11 127–11 137.
[48] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 561–578.
[49] S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, and C. Theobalt, "Real-time joint tracking of a hand manipulating an object from RGB-D input," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 294–310.
[50] J. Y. Zhang, P. Felsen, A. Kanazawa, and J. Malik, "Predicting 3D human dynamics from video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7114–7123.
[51] D. Mehta et al., "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 44.
[52] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: Shape completion and animation of people," in *Proc. ACM SIGGRAPH Papers Conf.*, 2005, pp. 408–416, doi: 10.1145/1186822.1073207.
[53] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation," 2020, *arXiv:2004.03686*.
[54] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5252–5262.
[55] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5607–5616.
[56] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2Shape: Detailed full human body geometry from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2293–2303.
[57] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1175–1186.
[58] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 81–90.
[59] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, "Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction," 2020, *arXiv:2007.03858*.
[60] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "LiveCap: Real-time human performance capture from monocular video," *ACM Trans. Graph.*, vol. 38, no. 2, pp. 14:1–14:17, 2019.
[61] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "DeepCap: Monocular human performance capture using weak supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5051–5062.
[62] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[63] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[64] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6757–6765.
[65] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu, "Real-time geometry, albedo and motion reconstruction using a single RGBD camera," *ACM Trans. Graph.*, vol. 36, 2017, Art. no. 44.
[66] K. Karunratanakul, J. Yang, Y. Zhang, M. Black, K. Muandet, and S. Tang, "Grasping field: Learning implicit representations for human grasps," in *Proc. Int. Conf. 3D Vis.*, 2020, pp. 333–344.

Zhuo Su received the BS degree from the Department of Automation, College of Information Science and Engineering, Northeastern University, Shenyang, China, in 2018, and the MS degree from the Automation Department, Tsinghua University, Beijing, China, in 2021. His current research interests include computer vison and graphics.



Lan Xu received the BE degree from Zhejiang University, in 2015, and the PhD degree from the Hong Kong University of Science and Technology, in 2020. He is currently an assistant professor with the School of Information Science and Technology, ShanghaiTech University. His research interests include computer vision and computer graphics and computational photography.



Dawei Zhong received the bachelor degree from Tongji University, in Jul. 2019. He studied Data Science and Information Technology with the Tsinghua-Berkeley Shenzhen Insititute(TBSI), Tsinghua University. His current research interest includes about 3D computer vision.

Zhong Li received the MSc degree in computer science from the University of Missouri, Columbia, MO, in 2015, and the PhD degree in computer science from the University of Delaware, Newark, DE, in 2019. He is currently a senior staff research scientist with InnoPeak Technology (OPPO US Research Center). His research interests include computational photography, computer graphics, and computer vision.

Fan Deng is now the director of perception Laboratory of OPPO US Research Center, mainly responsible for the research and development of CV technology, and supporting the products in AR, VR and robotics. He has nearly ten years of research and development experience with the fields of image processing, machine vision, camera algorithm, and Android system.

Shuxue Quan received the BS and MS degrees in optical engineering from the Beijing Institute of Technology, and the PhD degree in imaging science from the Rochester Institute of Technology. He is the senior director of research on perception and 3D vision with OPPO US Research Center where he leads teams developing algorithms of SLAM, 3D reconstruction, and human computer interaction. His team also builds software framework and applications to enable AR on mobile and wearable devices. Before OPPO, he worked at Qualcomm on computer vision, machine learning and mobile camera, besides a few other positions at Sony, Micron and Broadcom. He has been granted more than 30 international patents and published more than 20 journal or conference papers.

Lu Fang (Senior Member, IEEE) received the BE degree from the University of Science and Technology of China, in 2007, and the PhD degree from the Hong Kong Universsity of Science and Technology, in 2011. He is currently an associate professor with the Department of Electronic Engineering, Tsinghua University. Her research interests include computational imaging and visual intelligence. He serves as associate editor for Optica, *IEEE Transactions on Image Processing* and *IEEE Transactions on Multimedia*.