Lightweight Multi-person Total Motion Capture Using Sparse Multi-view Cameras

Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu*, Yebin Liu*
Department of Automation and BNRist, Tsinghua University

Abstract

Multi-person total motion capture is extremely challenging when it comes to handle severe occlusions, different reconstruction granularities from body to face and hands, drastically changing observation scales and fast body movements. To overcome these challenges above, we contribute a lightweight total motion capture system for multi-person interactive scenarios using only sparse multi-view cameras. By contributing a novel hand and face bootstrapping algorithm, our method is capable of efficient localization and accurate association of the hands and faces even on severe occluded occasions. We leverage both pose regression and keypoints detection methods and further propose a unified two-stage parametric fitting method for achieving pixel-aligned accuracy. Moreover, for extremely self-occluded poses and close interactions, a novel feedback mechanism is proposed to propagate the pixel-aligned reconstructions into the next frame for more accurate association. Overall, we propose the first light-weight total capture system and achieves fast, robust and accurate multi-person total motion capture performance. The results and experiments show that our method achieves more accurate results than existing methods under sparse-view setups.

1. Introduction

Marker-less motion capture, due to its great potentials for behaviour understanding, sports analysis, human animation, video editing and virtual reality, has been a popular research topic in computer vision and graphics for decades. Within this research field, total motion capture, pioneered by [22] using an extremely dense-view setup (hundreds of cameras), shows impressive results of simultaneous capture of multi-person total interactive behaviours including facial expressions, body and hand poses, and has aroused widespread interest in computer vision community. However, this work [22] suffers from expensive and sophisticated hardware setup and low run-time efficiency.

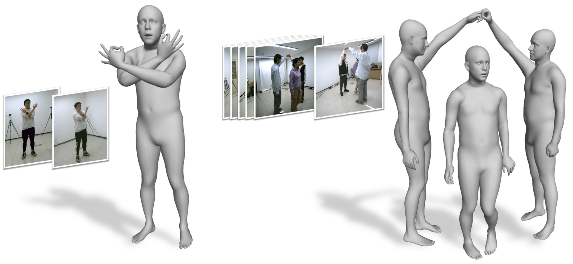

Figure 1. Our lightweight total capture system produces expressive human models with sparse multi-view cameras.

Recently, to reduce the capture complexity, more and more researches try to perform total motion capture from only a single image or video [41, 55, 13, 45, 35, 64]. By either optimizing a parametric models like SMPL-X [41] and Adam [22] ([55]) or regressing the model parameters directly from the input images [13], these methods even achieve real-time total motion capture performance for single-person [45, 64]. However, it remains difficult for the monocular methods to handle severe occlusions and challenging poses under multi-person interactive scenarios.

To guarantee both lightweight setups and robust performance, we propose the first lightweight total capture system using only sparse multi-view cameras. However, extending the existing monocular total capture methods to sparse-view multi-person total capture is not trivial. Although the incorporation of multi-view observations may resolve the depth ambiguities for monocular methods, the severe occlusions caused by complex-poses and multi-person interactions will significantly deteriorate the performance for current total capture methods. Specifically, the main challenges include: i) hand/face association across multiple views under drastically changing observation scales and unstable detection results, ii) pixel-aligned fitting between the reconstructed 3D model and the input images, and iii) robust and accurate body association under severe occlusions even for close interactions. To resolve all the challenges above, we propose, as far as we know, the first method to achieve fast, robust and accurate multi-person total motion capture using only light-weight sparse-view cameras.

First of all, compared with the relatively fixed body part scales and satisfactory occlusion-free view point in monocular single-person total capture cases [41, 55, 13, 45, 64],

* Corresponding Author

sparse multi-view setups suffer from hand/face fragments on account of severe occlusions, blurs on hands and even fingers by fast limb movements, and varying hand/face scales across different cameras. Moreover, it remains challenging to associate hands correctly when different hands are located very closely on an image. To resolve these challenges, we propose a novel hand and face bootstrapping algorithm to extract accurate body part features effectively from the sparse and multi-scale images for accurate association. Benefiting from the recent progress in multi-person skeleton pose capture [62], the skeleton-level results are utilized to guide the following object-detection network for more robust and accurate detection. Moreover, we introduce the cross-modality consistency and cross-scale consistency to filter unexpected detection results of fragments caused by occlusions or improper view points.

Secondly, using only the pose-regression methods or the key-point detection methods cannot yet guarantee accurate parametric model fitting. Firstly, pose-regression methods [13, 45, 64] are able to reconstruct decent hand gestures in self-occlusion cases, but these one-shot methods cannot guarantee pixel level alignment with 2D joint positions on the image. On the other hand, keypoint-detection methods [41, 55] are capable of providing pixel-aligned geometric features for visible joints, but may need heavy post-processing optimizations, which is quite sensitive to the initialization and usually fails due to self-occlusion. To fully leverage the advantages of both categories and avoid their drawbacks, we propose a new unified two-stage parametric fitting method, in which we leverage the pose-regression result as the initial value to accelerate the convergence for parametric model fitting based on the detected keypoints, and finally achieves pixel-aligned fitting accuracy without losing the efficiency.

Last but not least, for extremely complex poses and close interactions, even 4D association [62] may fail in the body association step, which is an inherent and natural limitation for sparse multi-view setups. To this end, we propose a feedback mechanism in which the reconstructed pixel-aligned human parametric models in the previous frame are propagated into the current frame for enhancing soft visibility information and finally achieve accurate association result. Benefiting from this novel feedback mechanism, our method is able to capture accurate human behaviours even under scenarios with severe occlusions and close interactions.

Our contributions can be concluded as:

- A new hand and face bootstrapping method that involves the body-level skeleton guidance for more accurate body part localization and self-validated consistency scores to filter out the noise of fragmented detection results by unexpected view points or occlusion observations (Sec. 4).
- A new unified two-stage parametric fitting method that fully utilizes both pose-regression and keypoint-detection methods to produce accurate pixel-aligned 3D human models with expressive motion (Sec. 5).
- A new feedback mechanism that propagates the accurate reconstruction into the next frame to further improve the association accuracy especially on the severe occluded occasions (Sec. 6).

2. Related Work

2.1. Total Motion Capture

Total motion capture methods, which aim at marker-less multi-scale human behaviour capture (including body motion, facial expressions and hand gestures), have shown great potentials in human 4D reconstruction and high-fidelity neural rendering [42, 49, 28, 63]. As the pioneering method of total motion capture, [22] achieved promising human behaviours capture results under the setup of hundreds of cameras, however, this method relies on the expensive and sophisticated hardware and is therefore hard for applications. On the other end of the spectrum, to achieve lightweight and convenient capture, many works [41, 55, 13, 45, 35, 64] focused on total capture from a monocular setup. Monocular total capture [55] and SMPLify-X [41] optimized parametric human models (SMPL-X [41] and Adam [22]) to fit with the 2D detected keypoints. Choutas *et al.* [13] directly regressed the parameters of SMPL-X [41] from a single RGB image and refined the captured results of head and hands subsequently. Pose2Pose [35] combined global and local image features for more accurate prediction. FrankMocap [45] regressed parameters of hand and body poses separately and finally integrated two parts into a unified whole body output. Zhou *et al.* [64] exploited the motion relationship between body and hands to design the network and achieved real-time monocular capture. Overall, although current monocular methods could achieve plausible human total capture performance, they still suffer from depth ambiguity and occlusions.

2.2. Skeleton-based Pose Reconstruction

Single-view 2D and 3D pose estimation methods [54, 43, 19, 16, 9, 12, 26, 58, 1, 34, 23, 39, 33, 60] have been widely explored in recent years, however, they suffer from severe occlusions and ambiguity and cannot produce high-confidence results. To alleviate the occlusion and produce more accurate reconstruction, many works aimed to reconstruct human poses from multi-view input. On the first branch of this direction, some approaches [17, 50, 32, 57, 31, 27, 25, 40] preformed temporal skeleton-based tracking for each frame, but these methods suffer from imperfect initialization and accumulated errors. On another branch, cross-view matching methods associated correspondences (e.g., human instances and keypoints) from different viewpoints and finally reconstructed 3D pose for each performer. Some works utilized 3DPS models to solve 3D joint positions implicitly by skeletal constraints [3, 4] or body part detection [15]. Joo *et al.* [21] utilized 2D detection from dense multiple views to vote for possible 3D joint positions. Dong *et al.* [14] proposed a multi-way matching algorithm to guarantee cycle consistency across all the views. Zhang *et al.* [62] jointly formulated the temporal tracking and cross-view matching as a 4D association graph and achieved real-time performance. Tu *et al.* [51] proposed to directly operate in the 3D space while avoiding incorrect decisions in each viewpoint. Lin *et al.* [30] presented a plane-sweep-based approach to perform multi-view multi-person 3D pose estimation without the explicit cross-view matching. Even though these methods are able to capture 3D human poses using skeletons, they cannot reconstruct full body behaviours, i.e., facial expressions, hand motions, and body surfaces.

2.3. 3D Hand Reconstruction

3D hand reconstruction is an essential sub-problem in total capture. Many works [47, 8, 48, 56, 65, 20, 37] that focused on 3D hand pose estimation from a single RGB image have been proposed. Recently, more and more works aimed at recovery of a 3D hand mesh [18, 24, 10] or directly regressing the pose and shape parameters of a parametric hand model (MANO [44]) [2, 6, 61, 66, 11]. However, these methods only focused on single hand reconstruction while ignoring the interaction between hands. Moon *et al.* [36] proposed InterHand2.6M, a large-scale two hand interaction dataset. Several researchers have explored the problem of pose estimation under two hand interacting situation [38, 52, 29, 59], but the problem of hand pose estimation under multi-person interacting scenario with more hands involved is still unsolved.

3. Overview

3.1. Main Pipeline

As shown in Fig. 2, given multiple synchronous and calibrated RGB videos as input, our pipeline works in a frame-by-frame manner, and outputs a series of parametric human models naturally combining body posture, hand gesture and facial expressions by the following steps:

1. **4D Body Association** (Sec. 3.2): Given multi-view input, we associate the 2D keypoints and triangulate 3D body skeletons using 4D association [62].
2. **Hand and Face Bootstrapping** (Sec. 4): With the body skeletons, we perform hand and face bootstrapping to extract their 2D bounding boxes efficiently and also associate them to different subjects among different views.
3. **Two-stage Parametric Fitting** (Sec. 5): Then we fit parametric human model SMPL-X [41] to these posture, gesture and expression features in a two-stage manner to achieve efficient and accurate pixel-level alignment.
4. **Feedback Mechanism** (Sec. 6): Finally, the tracked human models are propagated into the 4D association step of the next frame to further improve the association accuracy especially on severe occluded occasions.

3.2. 4D Body Association

As a building block of our method, the 4D association [62] contributes a real-time multi-person skeleton tracking framework with sparse multi-view video inputs. By taking the tracked 3D joints from the previous frame and the detected 2D key-points in current frames as graph nodes $\mathcal{D}_j$, 4D association algorithm introduces a series of connecting edges: single-view parsing edges $\mathcal{E}_P$, cross-view matching edges $\mathcal{E}_V$ and temporal tracking edges $\mathcal{E}_T$, and finally formulate a unified association graph $\mathcal{G}_{4D}$ for optimizing the multi-view body association problem effectively.

4. Hand and Face Bootstrapping

We introduce a hand and face bootstrapping method to (i) extract local body part regions of interest (RoI) and detection from full-body inputs and (ii) eliminate incorrectly associated matches using the proposed non-maximum suppression (NMS) method. Body-level semantic features, hand pose regressions and keypoint detections are integrated into our pipeline. Note that the proposed bootstrapping methods for hand and face are quite similar, but the interactive hand behaviour is much more frequent under practical multi-person scenarios. So in this section, we mainly introduce the hand bootstrapping method which is more representative, and the method for face is similar.

Specifically, given sparse multi-view image inputs at frame t, we firstly leverage the 4D association algorithm (Sec. 3.2) to get the associated 2D body keypoints in each view and the triangulated 3D body skeletons. Secondly, we indicate preliminary screened RoIs $\{RoI^c_\alpha\}$ through body skeleton semantic information, then a lightweight object-detection network is utilized for further localizing tight and reliable RoIs $\{RoI^c_\beta\}$ in these initiatory screened areas to boost the key point detection and parametric regression performance of hands. However, there may exist several RoI^c_β corresponding to different hands in a single RoI^c_α due to close interactions and bad view directions as shown in Fig. 4 (b). This will lead to severe ambiguity in the later hand association step. To eliminate these ambiguous RoIs, we propose a double-check non-maximum suppression (NMS) method to guarantee both **cross-modality** (between key point detection and parametric regression of hands) and **cross-scale** (between body reconstruction and hand reconstruction) consistency. Next, we will introduce the 2D hand localization and association in detail.

4.1. 2D Hand Localization

We conduct 2D hand localization in a coarse-to-fine manner: first generating initial bounding box for each hand according to the reconstructed 3D body skeleton and semantic information, and then refine the initial bounding

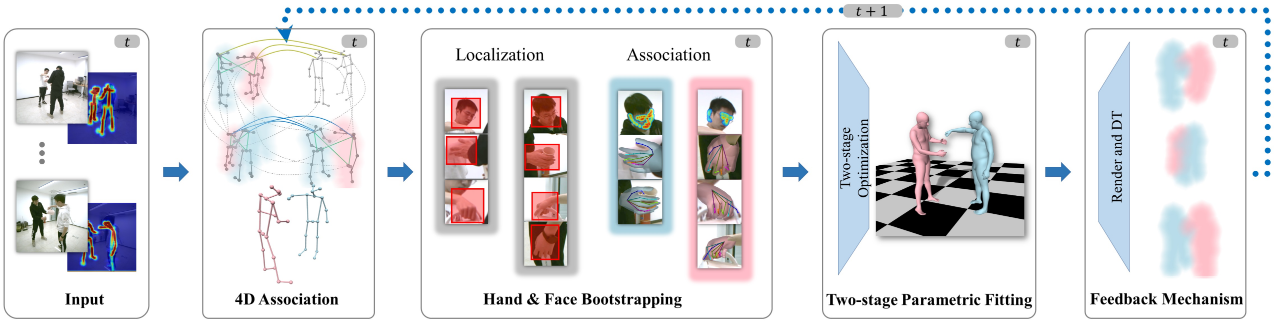

Figure 2. Method overview. Initially, we take multi-view RGB sequences and body estimation results as our inputs. Skeletons of each individuals are constructed by 4D association(Sect. 3.2). After that, we utilize our limb bootstrapping framework to localize(Sect. 4.1) and associate(Sect. 4.2) body part. After that, we optimize parametric SMPL-X models from all these outputs(Sect. 5). Finally, our feedback mechanism(Sect. 6) is introduced to boost the body association performance in next frame with the reconstructed human model.

boxes using the iterative hand detector [53]. Note that this strategy helps us filter out the inconsequential areas efficiently at the coarse level, thus reducing unnecessary computation and accelerating the hand localization process.

For generating the initial bounding box for a hand, we leverage the reconstructed 3D body skeleton to interpolate the center of hand and construct a 3D bounding sphere with constant radius to handle size variations of hands on 2D images caused by perspective projection. We then generate the initial 2D bounding box according to the projected center and radius of the 3D bounding ball in each view. Specifically, we estimate $\{RoI^c_{k,\alpha}\}$ of person k in each view c under the guidance of the reconstructed body skeletons:

$$\begin{aligned} o_p^c = P_c(O_p), \quad r_p^c = \frac{f_c \cdot R}{d_c(O_p)}, \\ \{RoI^c_{k,\alpha}\} = \{Rect(o_p^c, r_p^c) | z_p^c = 1, p = 1, 2, ..., P\}, \end{aligned} \tag{1}$$

where O_p and R is sphere center and radius. O_p could be simply extrapolated from the 3D position of wrist and elbow, and R is a constant parameter defined in terms of realistic physical scale. o_p^c and r_p^c are the projected circle center and radius in view c, respectively. f_c is the focal length of camera c, $P_c(\cdot)$ is perspective projection function, and $d_c(\cdot)$ is the distance to camera's image plane. An indicated variable z_p^c is introduced for whether the wrist joint of person p has been assigned a 2D keypoint detection in view c.

Since current hand keypoints detection and regression methods still rely on tight and accurate bounding boxes for achieving good performance, we further refine the initial bounding boxes using the iterative hand detector [53]. As shown in Fig. 3, we utilize that one-pass hand detection network to further extract more precise RoIs $\{RoI^c_{k,\beta}\}$. We demonstrate that our two-step localization method outperforms the light-weight detector (e.g., 100DOH [46] used in FrankMocap [45]) on both the speed and accuracy (Fig. 11).

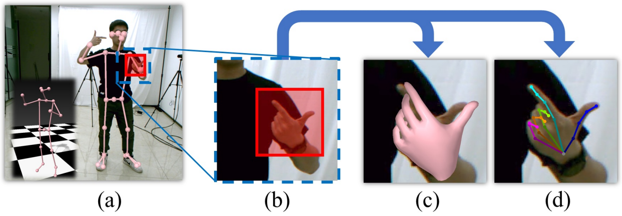

Figure 3. Illustration of hand localization and detection. (a) Reconstructed body skeletons using [62] in advance which will guide us to focus on key areas $\{RoI^c_{k,\alpha}\}$ (blue dotted line rectangle). (b) After that, a light-weight network is utilized for regressing more precise and tight bounding boxes $\{RoI^c_{k,\beta}\}$(red solid line rectangle). Then we clipped $\{RoI^c_{k,\beta}\}$ from the full-body image and then feed them to both the pose-regression network and the keypoint-detection network. (c) The regressed hand gesture which is decent but not pixel-aligned accurate enough. (d) 2D Keypoints are accurate but suffer from depth ambiguity.

4.2. Hand Association

Since 2D body joints have been associated and 3D body skeletons of different subjects have been triangulated in previous steps, in this section, we mainly focusing on how to assign correct bounding boxes of hands to the 3D wrist joints in each view.

We leverage classical non-maximum suppression (NMS) [5] algorithm but proposed two novel consistency scores to effectively filter out ambiguous RoIs. Specifically, cross-modality consistency score ζ^c_k and cross-scale consistency score ξ^c_k are proposed to judge which match will be finally retained. In practice, hands usually come close or even overlapped in some side views, and interactions among individuals will lead to more ambiguities. Specifically, considering the case that $RoI^c_{k_1,\alpha} \cap RoI^c_{k_2,\alpha} \neq \varnothing, k_1 \neq k_2$, as illustrated in Fig. 4, $RoI^c_{k_1,\beta_2}$ (dark blue) and $RoI^c_{k_2,\beta}$ (red) share the same sub-region, which results in association ambiguities. Inspired by traditional NMS algorithms in object-detection pipeline, we come up with a self-validated association algorithm to filter out redundant RoI^c_β and re-

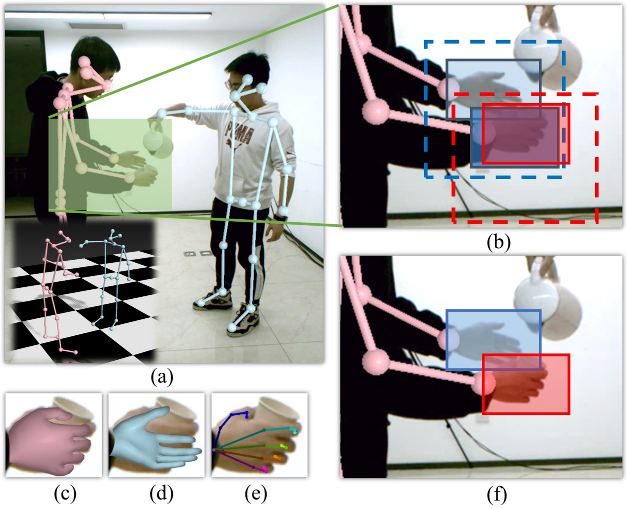

Figure 4. Illustration of hand association algorithm. (a) Body skeletons obtained by [62]. (b) Association ambiguity may happen when $RoI^c_{k_1,\alpha} \cap RoI^c_{k_2,\alpha} \neq \varnothing, k_1 \neq k_2$. Blue and red dotted line rectangles are $RoI^c_{k_1,\alpha}$ and $RoI^c_{k_2,\alpha}$, respectively. Then 3 tight bounding boxes $RoI^c_{k_1,\beta_1}$ (light blue), $RoI^c_{k_1,\beta_2}$ (dark blue) and $RoI^c_{k_2,\beta}$ (red) are further extracted from the two initial rectangles. We can observe that the right hand lies in the overlapping area of $RoI^c_{k_1,\alpha}$ and $RoI^c_{k_2,\alpha}$, leading to redundant proposals and confusing partition. (c) and (d) show that pose-regression network is specific for one chirality input. (e) is the result of heatmap-based detector which is invariant with chirality. (f) shows that after our double-check NMS procedure, the correct distributed $RoI^c_{k_1,\beta_1}$ are retained and the false one $RoI^c_{k_1,\beta_2}$ is discarded.

tain the correct match. Firstly, for each view c, we locate all redundant RoIs by calculating Intersection of Union (IoU) from total individuals' hands proposals. Secondly, we calculate **cross-modality consistency** score ζ^c_k and **cross-scale consistency** score ξ^c_k for each RoI proposal.

The first metric, cross-modality consistency score ζ, is used to penalize the inconsistency between different detection modalities. As shown in Fig. 4 (c), (d) and (e), the heatmap-based feature is invariant with respect to flipping translation, but pose-regression network needs right chirality assurance for achieving reasonable results. This divergence can help us to distinguish the left or right association ambiguities. Denote $J_h = 21$ as the hand joint number, $S^h_{regr} \in \mathbb{R}^{2\times J_h}$ as 2D hand joint positions from pose-regression network, $S^h_{dect} \in \mathbb{R}^{2\times J_h}$ as the output of keypoint-detection network, and w, h as the size of RoI_β. Then ζ is formulated as

$$\zeta = \frac{1}{J_h}\sum_{j=1}^{J_h} max(0, 1 - \frac{2\|S^h_{regr,j} - S^h_{dect,j}\|_2}{\sqrt{w^2+h^2}}). \tag{2}$$

On the other hand, the second metric, cross-scale consistency score ξ, is formulated to punish unreasonable wrist misalignment between the local hand estimator and global body estimator. Denote $S^b_{wrt} \in \mathbb{R}^2$ as the associated 2D wrist position from full-body detections, S^h_{dect,j_w} as wrist joint position by local hand keypoints detector. Finally, ξ is defined as

$$\xi = max(0, 1 - \frac{2\|S^b_{wrt} - S^h_{dect,j_w}\|_2}{\sqrt{w^2+h^2}}). \tag{3}$$

Finally, we sum up these two scores as confidence measurements to apply the NMS algorithm to reserve the one with the highest score. We demonstrate that our double-check NMS method helps to improve the association accuracy in confusing situations.

5. Two-stage Parametric Fitting

We observe that previous methods usually utilize either parametric pose regression [45] or heatmap-based keypoints [41] for total motion capture. However, they do have their own limitations. On one hand, although pose-regression networks can produce plausible results even under occlusions, they can not guarantee accurate 2D alignment with the input image. On the other hand, heatmap-based networks provide accurate 2D detections for visible joints, but they are still suffer from depth ambiguities and are susceptible to local minima during optimization. In this paper, we unify them together in a two-stage parametric fitting scheme, which contains local initialization and total optimization, to boost the total motion capture performance.

Local Initialization To accelerate convergence and prevent optimization deviation, it is essential to initialize the motion of each body part to a reasonable status. Specifically, for hand pose initialization, we pick a decent initial value from the semantic pose-regression gestures according to the hand association score ζ (Eqn. 2) and ξ (Eqn. 3). Besides, for body/head pose initialization, we solve the SMPL-X body pose by minimizing the following energy function directly to guarantee more accurate initialization:

$$E_{body} = \lambda_{b3d}E_{b3d} + \lambda_{pri}E_{pri} + \lambda_\beta E_\beta \tag{4}$$

Here, E_{b3d} is the distance from parametric model's joints to the corresponding reconstructed 3D body skeletons. E_{pri} and E_β are used to regularize natural pose and shape optimization as in SMPLify-X [41].

Total Optimization In this stage, we leverage the accurate 2D hands keypoints and faces landmarkers to further optimize the initial SMPL-X model for accurate total capture:

$$\begin{aligned} E_{total} &= E_{data} + E_{reg}, \\ E_{data} &= \lambda_{b3d}E_{b3d} + \lambda_{h2d}E_{h2d} + \lambda_{f2d}E_{f2d}, \\ E_{reg} &= \lambda_{pri}E_{pri} + \lambda_\beta E_\beta + \lambda_{\theta,h}E_{\theta,h} + \lambda_\varepsilon E_\varepsilon, \end{aligned} \tag{5}$$

where E_{h2d} and E_{f2d} are 2D data terms to minimize the distances between the 2D projections of the SMPL-X joints

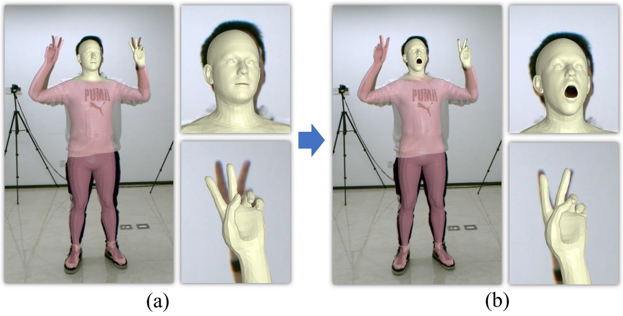

Figure 5. Illustration of two-stage parametric fitting. (a) Stage 1: we solve the body posture as well as arm kinematics and assign the gesture of pose regression with the highest association score ζ and ξ. (b) Stage 2: we then perform total optimization to achieve accurate total motion capture.

and the detected 2D keypoints in all the valid viewpoints. $E_{\theta,h}$ and E_ε are L-2 norm to keep the optimized gestures and expressions within reasonable ranges. Note that we can additionally leverage the consistency scores ζ and ξ (Eqn. 2 and 3) to balance the detection results in different views, so $E_{h2d} = \Sigma_c \frac{\zeta^c+\xi^c}{2} \cdot e^c_{h2d}$, where c is the view index.

6. Feedback Mechanism

Last but not least, for severe occlusions and close interactions, we put forward a feedback mechanism to boost the tracking performance of the association algorithm in return. On one hand, detailed limb detector contributes to extremity reconstruction with higher precision, which are leveraged to refine the body skeleton results. On the other hand, we re-render the human model to each view for the next frame to extend the tracking edges $\mathcal{E}_T$ of $\mathcal{G}_{4D}$ with additional visibility information. As shown in Fig. 6, we obtain the initial segmentation by rendering the optimized parametric models back to input images. Meanwhile, in order to enhance the robustness with body movements, we implement distance transformation to smooth the boundary of the rendered mask.

For a given 2D keypoint detection candidate c, we use the same denotation $z^k(c)$ in [62] to refer the possibility to connect that candidate to person k. Benefiting from our feedback module, the tracking edges in 4D association [62] (Sec. 3.2) are extended with visibility priors. We define the enhanced tracking edges $\widehat{z}^k(c)$ as:

$$\widehat{z}^k(c) = \frac{\tau^k(c)}{\sum_{i=1}^{K} \tau^i(c)} z^k(c), \tag{6}$$

where $\tau^i(c) \in [0,1]$ means the continuous occupancy of person i. As shown in Fig. 6 (d), $\tau^i(c)$ is negative interrelated to the distance to its binary mask, $\tau^k(c) = 1$ refers to the fully contained situation. As a result, our feed-back mechanism enhances skeleton tracking performance and reduces jittering in projection coincidence cases.

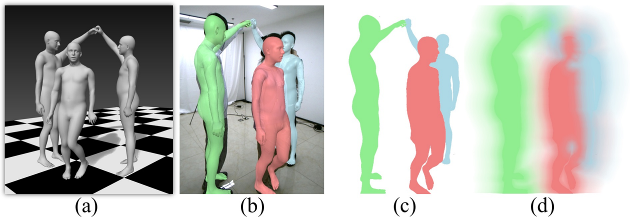

Figure 6. Illustration of our feedback mechanism. (a) and (b) are the aligned parametric models. (c) Segmented results by rendering. (d) is the softened mask generated by distance transformation to enhance the robustness of association during fast motion.

7. Results

In Fig. 7, we demonstrate example results by our system. With the sparse multi-view setup, our method produces expressive human parametric models under multi-person interactive scenarios.

7.1. Implementation Details

Our light-weight total capture system are implemented with 6 synchronized RGB cameras (resolution 2048 × 2048) on a single PC (i5-6600K CPU, NVIDIA RTX 3090 GPU). We use Openpose [9] as our body pose estimator, SRHandNet [53] as our hand instance detector and keypoint detector. We leverage the hand pose-regression network of Frankmocap [45] gesture regressor. FaceAlignment [7] are used for face keypoints extraction. Besides, we accelerate all neural network inference by implementing half-precision arithmetic on NVIDIA TensorRT platform. The CNN performance is shown in Table. 1. Besides, our body association backbone takes nearly 10ms to recover human skeletons, the limb localization and association method is fast enough to be neglected. Our parametric fitting workflow takes 150 ms for stage one and 350 ms for stage two (20 Gauss-Newton iterations and parallel for each person). On the whole, our system run-time depends on the captured individual number and view number. Empirically, our pipeline runs about 1 fps for 2 person with 6 views, and the processing speed slows down to 0.3 fps for 7 persons with 8 views. As for hyper-parameters, sphere radius R in hand localization is set to $0.15m$, and association NMS threshold is 0.5. In two-stage parametric fitting, we set $\lambda_{b3d} = 10$, $\lambda_{h2d} = 0.0001$, $\lambda_{f2d} = 0.0003$, $\lambda_{pri} = \lambda_{\theta,h} = 0.01$, and $\lambda_\beta = \lambda_\varepsilon = 0.01$.

Network	Input	Batchsize	Speed(FPS)
Openpose [9]	368 × 368	6	43.1
FaceAlignment [7]	256 × 256	4	109.5
SRHandNet [53]	256 × 256	8	50.0
HandHMR [45]	224 × 224	8	202.1

Table 1. Inference speed of the CNN networks used in our system.

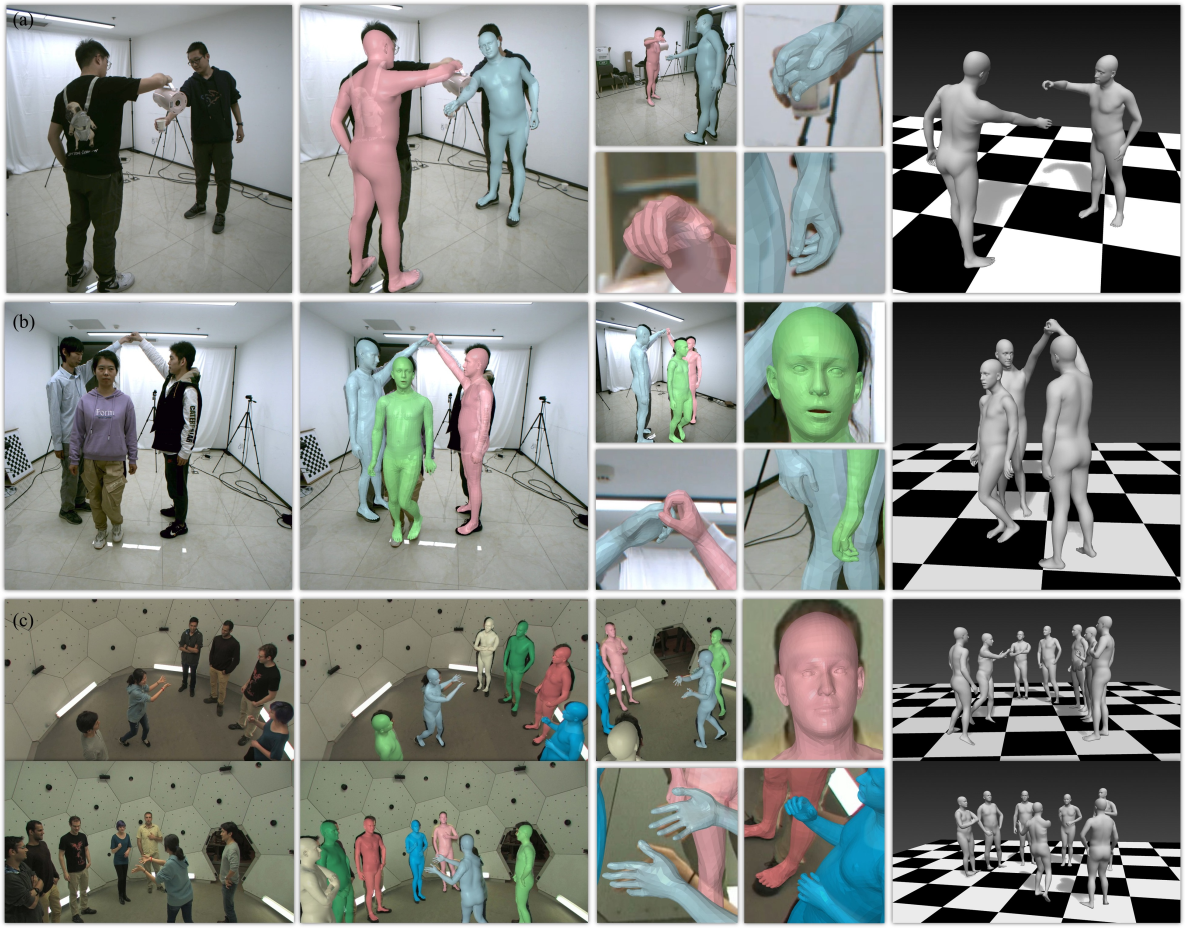

Figure 7. Results by our system. From the left to right are input reference images, parametric model alignment, facial and hand alignment and 3D visualization from a novel view, respectively. (a) Results of the hand-object-interaction case from our captured data using 6 views, (b) results of a multi-person-interaction scenario using 6 views, (c) results on CMU dataset [22] using 8 views.

7.2. Comparison

Since our method is the first to enable lightweight total capture from sparse multi-view, we compared our method with SOTA single view method FrankMocap [45] in Fig. 11 and ground truth from Total Capture [22] in Fig. 9. What's more, MPJPE (Mean Per Joint Position Error) are provided for Total Capture dataset Tab. 2.

7.3. Evaluation: Hand Bootstrapping

We compare our hand bootstrapping method with SOTA monocular total capture method Frankmocap [45]. To ensure fairness as much as possible, we reduce our system to 2 close front view camera. Fig. 11 (a) shows the reconstruction failure of FrankMocap [45] caused by mixing up left and right hands to the same region proposal. Thanks to the proposed NMS method in hand association, our method can robustly reconstruct more accurate hands in Fig. 11 (b).

7.4. Evaluation: Two-stage Parametric Fitting

We conduct ablation study of two-stage fitting metric on CMU dataset [55] and demonstrate that our method makes different modality detectors benefit from each other. On the one hand, as shown in Fig. 8 (a)(d), we perform our two-stage fitting algorithm with only pose regression results, namely we leverage orthogonal projected joints from pose-regression network to take over heatmap-based 2D correspondences in stage two. Misalignment artifacts are shown in detail as pose-regression detector could not guarantee pixel-aligned accuracy. One the other hand, keypoints-detection-only results are shown in Fig. 8 (b)(e). Without pose regression network to initialize hand poses with reasonable gestures, the optimization is easy to fall into a local minimum.

7.5. Evaluation: Feedback Mechanism

We evaluate our feedback module in Shelf dataset [3]. As shown in Fig. 10(a), the left elbow the of salmon person is distributed to the background green person without feed-

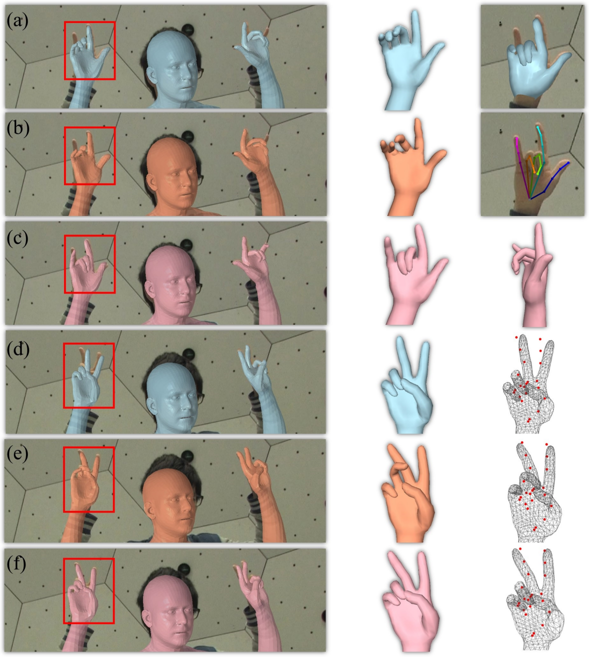

Figure 8. Qualitative evaluation of two-stage parametric fitting. (a) and (d) are the results with parametric-regression-only metric (blue). (b) and (e) are the results with keypoints-detection-only metric (salmon). (c) and (f) are the results with our two-stage fitting strategy combing both metrics (pink). Meanwhile, we visualize pose-regression network outputs and heatmap-based network outputs in right of (a) and (b), respectively. Red dots in right of (d) (e) (f) refer to the ground-truth 3D hand annotations.

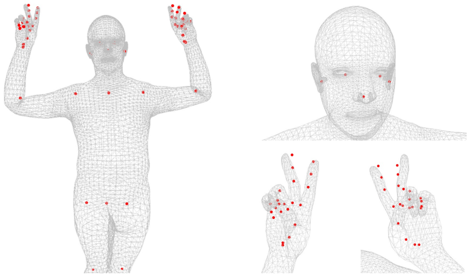

Figure 9. Comparison of our sparse-view method(8 view used) to ground truth from Total Capture Dataset[22]. The mesh refers to our reconstructed parametric model (SMPL-X), and the red keypoints are the ground truth from the Total Capture dataset.

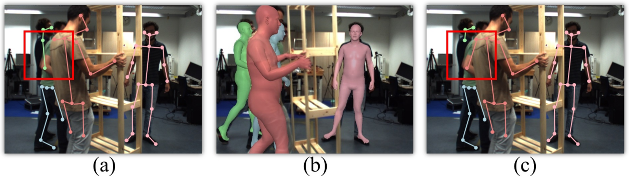

Figure 10. Qualitative evaluation of feedback mechanism. (a) shows the original association results of [62]. (b) are our reconstructed model of last frame. (c) shows that our feedback mechanism boosts body association performance with visibility prior.

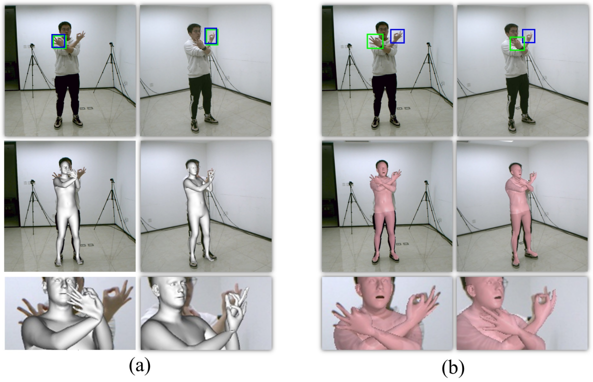

Figure 11. Qualitative evaluation of hand bootstrapping & comparison against SOTA monocular method, FrankMocap [45]. (a) Results of Frankmocap [45], only single ROI are extracted for each view, and left hand (blue rectangle) and right hand (green rectangle) have been distributed to the same ROI proposal. (b) Results of our method, all hands are extracted and associated correctly.

Type	Body	Head	LHand	RHand
MPJPE(mm)	33.4	21.7	22.6	19.3

Table 2. Quantitative evaluation on Total Capture dataset. We calculate the MPJPE of body&head joints on a video segment (750 frames) which involves both large variation movements and meticulous hand gestures (noted that hand annotations are few for challenge gestures) with 5 cameras for a comprehensive evaluation.

back. We show enhanced association results in Fig. 10(c) and prove that visibility informations provided by reconstructed human models help to eliminate that ambiguity.

Shelf	A1	A2	A3	Avg
w/o feedback	99.0	96.2	97.6	97.6
w/ feedback	99.5	97.0	97.8	98.1

Table 3. Ablation study of feedback mechanism on Shelf dataset. Numbers are percentage of correct parts(PCP).

8. Discussion

Conclusion In this paper, we propose, as far as we know, the first multi-person total motion capture framework with only a sparse multi-view setup. Based on the proposed hand and face bootstrapping, two-stage parametric fitting and feedback mechanism, our method can enable lighweight, fast, robust and accurate capture of the body pose, hand gesture and facial expression of each character even under the scenarios with severe occlusions and close interactions.

Limitation and Future Work We can mainly recover the facial expression by the jaw joint, and cannot reconstruct subtle facial expressions due to the low-resolution facial image input, which we leave for future research.

Acknowledgements This work is supported by the National Key Research and Development Program of China No.2018YFB2100500 and the NSFC No.62125107, No.61827805, No.62171255, and the Shuimu Tsinghua Scholarship.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, pages 1067–1076, 2019. 3

[3] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 2, 7

[4] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *TPAMI*, 2016. 2

[5] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms – improving object detection with one line of code. In *3DV*, 2017. 4

[6] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 3

[7] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 6

[8] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, pages 666–682, 2018. 3

[9] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 2, 6

[10] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *CVPR*, pages 13274–13283, 2021. 3

[11] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *CVPR*, pages 10451–10460, 2021. 3

[12] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. In *CVPR*, 2018. 2

[13] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 1, 2

[14] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *CVPR*, 2019. 3

[15] Sara Ershadi-Nasab, Erfan Noury, Shohreh Kasaei, and Esmaeil Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 2018. 3

[16] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 2

[17] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, pages 1746–1753. IEEE, 2009. 2

[18] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, pages 10833–10842, 2019. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[20] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 3

[21] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, 2017. 3

[22] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 1, 2, 7, 8

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[24] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 3

[25] Oh-Hun Kwon, Julian Tanke, and Juergen Gall. Recursive bayesian filtering for multiple human pose tracking from multiple cameras. In *ACCV*, 2020. 2

[26] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 2

[27] Kun Li, Nianhong Jiao, Yebin Liu, Yangang Wang, and Jingyu Yang. Shape and pose estimation for closely interacting persons using multi-view images. In *CGF*, 2018. 2

[28] Zhe Li, Tao Yu, Zerong Zheng, Kaiwen Guo, and Yebin Liu. Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In *CVPR*, 2021. 2

[29] Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3d pose estimation using monocular rgb. In *WACV*, pages 2373–2381, January 2021. 3

[30] Jiahao Lin and Gim Hee Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *CVPR*, pages 11886–11895, 2021. 3

[31] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *TPAMI*, 2013. 2

[32] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, 2011. 2

[33] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020. 2

[34] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 2

[35] Gyeongsik Moon and Kyoung Mu Lee. Pose2pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation. *arXiv preprint arXiv:2011.11534*, 2020. 1, 2

[36] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*. Springer, 2020. 3

[37] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 3

[38] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM TOG*, 38(4):1–13, 2019. 3

[39] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *ICCV*, 2019. 2

[40] Takuya Ohashi, Yosuke Ikegami, and Yoshihiko Nakamura. Synergetic reconstruction from 2d pose and 3d motion for wide-space multi-person video motion capture in the wild. *Image and Vision Computing*, 104:104028, 2020. 2

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 2, 3, 5

[42] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 2

[43] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 2

[44] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *TOG*, 36(6):1–17, 2017. 3

[45] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 1, 2, 4, 5, 6, 7, 8

[46] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, pages 9869–9878, 2020. 4

[47] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, pages 1145–1153, 2017. 3

[48] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98, 2018. 3

[49] Guoxing Sun, Xin Chen, Yizhang Chen, Anqi Pang, Pei Lin, Yuheng Jiang, Lan Xu, Jingya Wang, and Jingyi Yu. Neural free-viewpoint performance rendering under complex human-object interactions. In *ACM MM*, 2021. 2

[50] Graham W Taylor, Leonid Sigal, David J Fleet, and Geoffrey E Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010. 2

[51] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *ECCV*, pages 197–212, 2020. 3

[52] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *TOG*, 39(6):1–16, 2020. 3

[53] Yangang Wang, Baowen Zhang, and Cong Peng. Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization. *IEEE TIP*, 29(1):2977 – 2986, 2019. 4, 6

[54] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[55] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *CVPR*, 2019. 1, 2, 7

[56] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *ICCV*, pages 2335–2343, 2019. 3

[57] Angela Yao, Juergen Gall, Luc V Gool, and Raquel Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *NIPS*, 2011. 2

[58] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *NIPS*, 2018. 2

[59] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *ICCV*, 2021. 3

[60] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 2

[61] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 3

[62] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, June 2020. 2, 3, 4, 5, 6, 8

[63] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *ICCV*, 2021. 2

[64] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular realtime full body capture with inter-part correlations. In *CVPR*, 2021. 1, 2

[65] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 3

[66] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 3