Pacific Graphics 2018
H. Fu, A. Ghosh, and J. Kopf
(Guest Editors)

*Volume 37 (2018), Number 7*

# Shape and Pose Estimation for Closely Interacting Persons Using Multi-view Images

*Kun Li*[1], Nianhong Jiao[1], Yebin Liu[2], Yangang Wang[3] and Jingyu Yang[4†]

[1]School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
[2]Department of Automation, Tsinghua University, Beijing 100084, China
[3]School of Automation, Southeast University, Nanjing 210096, China
[4]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

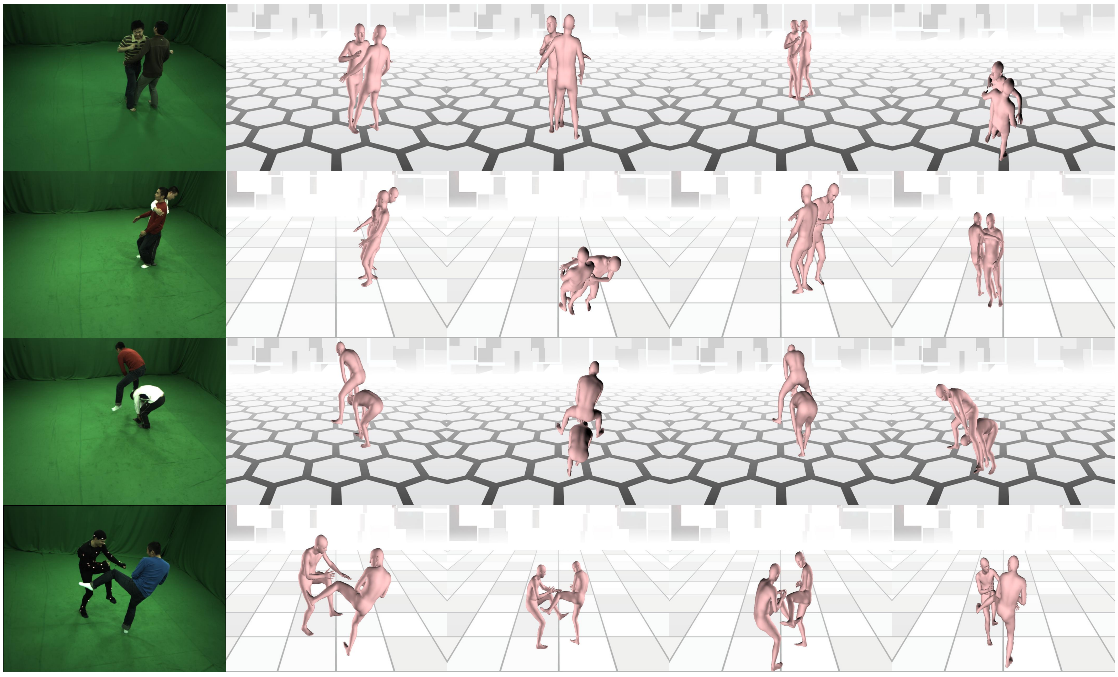

**Figure 1:** *3D shape reconstruction for multiple persons with close interactions. The results are presented from 4 different views as illustrated for each sequence. This paper proposes a fully-automatic 3D pose and shape estimation method for closely interacting persons.*

**Abstract**

*Multi-person pose and shape estimation is very challenging, especially when the persons have close interactions. Existing methods only work well when people are well spaced out in the captured images. However, close interaction among people is very common in real life, which is more challenge due to complex articulation, frequent occlusion and inherent ambiguities. We present a fully-automatic markerless motion capture method to simultaneously estimate 3D poses and shapes of closely interacting people from multi-view sequences. We first predict the 2D joints for each person in an image, and then design a spatio-temporal tracker for multi-person pose tracking based on multi-view videos. Finally, we estimate 3D poses and shapes of all the persons with multi-view constraints using a skinned multi-person linear model (SMPL). Experimental results demonstrate that our method achieves fast but accurate pose and shape estimation results for multi-person close interaction cases. Compared with existing methods, our method does not need pre-segmentation for each person and manual intervention, which greatly reduces the complexity of the system including time complexity and system processing complexity.*

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [Computer Graphics]: Scene Analysis—Shape, Motion



† Corresponding author: yjy@tju.edu.cn.

## 1. Introduction

Markerless human motion capture has been a popular and challenging topic in computer vision and computer graphics. Its main task is to recover a temporally coherent representation of dynamic 3D shape by tracking the motion of a moving object from videos. Motion capture for a single person has made tremendous advance for the last decade [AST*08, GSA*09, VBM08, LDX11, WLT*18]. However, these methods require carefully designed camera settings or controlled studios, and rely on good segmentation. In the case of multiple persons, direct application of existing methods for the single person will fail to generate satisfying results due to the difficulties of multi-person segmentation and pose estimation. Although some methods [CBI10, MKGH17, WSVT13] can handle the multi-person situation, the captured scene is limited to very simple interaction without inter-occlusion, such as face-to-face playing a ball. However, the interaction among multiple persons in real life is usually very close, *e.g.*, a hug, a double dance or a fight, *etc.*, which is also common in games and movies. Therefore, reconstructing the shapes and poses of closely interacting people is crucial for practical application.

To our best knowledge, no existing methods can fully-automatically and simultaneously estimate 3D shapes and poses of closely interacting people. More importantly, there are tremendous ambiguities where commonly-used features like color, edges, or keypoints cannot be individually assigned to each person. When people interact closely, problems become more complicated and more challenging due to occlusion, truncation and inherent ambiguity. Liu *et al.* [LGS*13] propose a markerless motion capture method for closely interacting persons using multi-view image segmentation, but this method need a laser scan to capture a template mesh and manual intervention to rig a skeleton. Moreover, the computational complexity of this method is very high, and its result heavily relies on the segmentation.

In this paper, we propose a new markerless motion capture method to achieve automatic 3D shape and pose estimation for closely interacting persons from multi-view videos. We first utilize RMPE [FXTL17] to estimate 2D joints of each person in a single image and track the same person using spatio-temporal tracking. Then, we employ a popular statistical body shape model, SMPL [LMR*15], as an implicit representation to estimate 3D poses and shapes of all the persons with multi-view constraints. Experimental results show that our method achieves appealing results with much less computation time and without manual intervention, *e.g.*, in Figure 1. Our method does not need the segmentation for multi-view videos, and has no requirement for the capture system, *e.g.*, lighting and background. Therefore, our method has more flexibility and the computation time is much less than the state-of-the-art method.

The main contributions of this paper are summarized into the following three aspects:

- **Fully-automatic 3D pose and shape estimation for close interacting persons**. Our method do not need manual intervention, template mesh scan and segmentation for each person. It has more flexibility and less computation time (about $1min$ per person per time instance without GPU acceleration).
- **Multi-person spatio-temporal pose tracking**. We indicate the same person in the multi-view videos by considering spatio-temporal correspondence. The tracking strategy uses both bounding box and pose information. This drastically reduces the ambiguities.
- **Multi-view 3D pose and shape estimation**. We optimize the 3D poses and shapes of multi-persons from 2D joints estimated by RMPE in videos with multi-view constraints. This is robust to close interaction and occlusion of multi-person.

## 2. Related Work

### 2.1. Multi-Person 2D Pose Estimation

In recent years, multi-person pose estimation from an image is gaining increasing popularity because of the high demand for practical applications. However, multi-person pose estimation is challenging due to occlusion, specificity of individual postures and unpredictable interactions between different people. Existing work can be mainly divided into two categories: bottom-up approaches and top-down approaches.

**Bottom-Up Approaches** Bottom-up approaches [CSWS17, PIT*16, IPA*16, NHD17] first directly predict all the 2D joints and then assemble them into a complete skeleton for each person. DeepCut [PIT*16] detects all the body parts at first, and then label and assemble these parts via integer linear programming. DeeperCut [IPA*16] improves DeepCut [PIT*16] using a stronger part detector based on ResNet [HZRS16] and a better incremental optimization strategy proposed by Insafutdinov *et al.* [IPA*16]. OpenPose [CSWS17] adopts Part Affinity Fields (PAFs) to associate body parts with individuals and assemble detected keypoints into different poses of persons.

**Top-Down Approaches** Top-down approaches [PZK*17, HGT17, HGDG17, FXTL17, CWP*17] separate the multi-person pose estimation into a two-stage pipeline, *i.e.*, detecting and cropping each person from the image and then applying single person pose estimator for each individual in the cropped patch. Papandreou *et al.* [PZK*17] use the heatmaps with offsets to estimate the position of keypoints. Mask-RCNN [HGDG17] first obtains human bounding boxes and then predicts the keypoints from the cropped feature map of the corresponding human bounding box. Some recent work [FXTL17, CWP*17] combine different human detectors and single person pose estimators to obtain better performance. Currently, top-down approaches have achieved the state-of-the-art performance in almost all benchmark datasets, *e.g.*, MSCOCO [LMB*14] and MPII [APGS14a].

### 2.2. Multi-Person Tracking

Multi-person tracking is a traditional topic studied intensively in computer vision. Although great progress has been made, challenges remain in the cases of false position detection, long-term occlusions and camera motion, especially tracking multi-person under crowded scenes. Recent work mainly focused on tracking-by-detection pipeline. Some work operates on online linking people detection over time [KLCR15, Cho15], and the other work groups the detections into tracklets and then merges them into tracks [WWCW17]. Kim *et al.* [KLCR15] use a generic CNN (Convolution Neural Network) to represent person appearance by estimating a target-specific appearance model online. Tang *et al.* [TAAS16] propose a pairwise feature based on local image patch matching that is similar to [Cho15]. Some trackers depend on data association methods such as greedy or Hungarian Algorithm [XAS15, BGO*16], and consider tracking problem as a maximum weight bipartite matching issue. The nodes of this bipartite graph are the human bounding boxes in two adjacent frames. However, pose information is not taken into account as a major factor in the crowded images. Xiu *et al.* [XLW*18] proposed an effective pose tracker based on pose flows and re-design two kinds of ORB [RRKB12]-based similarity criteria. We utilize these two kinds of criteria and extend our tracker to multi-view cases for tracking closely interacting multi-people.

### 2.3. 3D Pose and Shape Estimation

Markerless 3D pose and shape estimation for human bodies has been a longstanding and challenging research topic in computer vision and computer graphics. Most previous methods [AARS13, DR05, DWL*16, GRBS10, RKS13, YGUU11, SIHB12] ignore 3D human shape and only focus on pose. They assume no explicit human body model and directly infer 3D pose from 2D image features. The stochastic gradient-based method [YGUU11] has very good optimization performance by using a Gaussian Process Latent Variable Model (GPLVM). Zhou *et al.* [ZZL*16] create a sparse prior over human pose that captures how these poses appear from multi-views, and demonstrate that the resulting optimization problem is easier to solve. Meanwhile, plenty of methods based on deep learning achieve accurate pose estimation results [DWL*16, MN17, PZS17, TRLF16, TGHC16]. Rhodin *et al.* [RSK*18] propose a deep network to predict 3D pose for actions by using multiple views. Alternatives [ICS14, LZC15, MRC*16, PZDD17, TKS*16, ZSZ*16] directly regress from a single image to the 3D pose, but lead to temporally incoherent reconstructions. Some work can estimate 3D body shape from images. However, good silhouettes are often assumed to be available [Bla08, HAR*10] and manual initialization is required [AST*08, PF03, WVT12, VBM08].

Recently, Xu *et al.* [XCZ*17] present a general 3D performance capture of a person from monocular video, but a template mesh and the corresponding skeleton are needed at first and manual intervention is required. Yin *et al.* [YHH*18] propose a data-driven method to generate closely interacting 3D pose-pairs from 2D video annotations based on Markov Chain Monte Carlo (MCMC) sampling. Kanazawa *et al.* [KBJM18]introduce an end-to-end framework to reconstruct a full 3D mesh of a human body from a single image. However, their method can only use paired 2D data with labels, instead of ground-truth 3D data which is hard to acquire. Bogo *et al.* [BKL*16] automatically and simultaneously estimate 3D pose and convincing shape of a person from a single unconstrained image, which do not require any user intervention or complex optimization techniques. They utilize 2D joints estimated by a 2D joint detector, *e.g.*, DeepCut [PIT*16] or CPM [WRKS16] to fit the projection of 3D SMPL [LMR*15] joints, and infer human shape and pose parameters. Huang *et al.* [HBC*17] extend that work [BKL*16] to multi-view case by utilizing silhouettes and temporal coherence similar to the method in [RRD*16].

However, all the above methods only focus on a single person or multi-person without close interactions. Liu *et al.* [LGS*13] propose a multi-person motion capture method to solve the close interaction problem using multi-view image segmentation, but this method need a laser scan to capture a template mesh and manual intervention to rig a skeleton. Moreover, its computational complexity is very high, and its result heavily relies on the segmentation. Ye *et al.* [YLH*12] use three hand-held Kinect cameras with depth videos to reconstruct human skeletal poses, deforming surface geometries and camera poses, but this method also need a scanned template mesh and manual rigging. In this paper, we adopt the generative human body model SMPL [LMR*15] to reduce the computational complexity, which is a skinned vertex-based model and can accurately represent a wide variety of body shapes in natural human poses. We estimate 2D joints of each person and track the same person using spatio-temporal tracking, which is robust to the close interaction cases. Then, we estimate 3D poses and shapes of all the persons with multi-view constraints.

## 3. Method

In this section, we present the details of the proposed method. As shown in Fig. 2, We first employ RMPE [FXTL17] to predict 2D joints of each person from each image of every camera for each frame. Then, multi-person tracking via spatio-temporal optimization is used to better exploit the temporal correlation between frames and spatial correlation among multi-views. Finally, we fit a statistical 3D body model to the 2D joints of each person by multi-view optimization, and obtain the estimated 3D poses and temporally consistent 3D shapes.

### 3.1. Multi-Person 2D Pose Estimation

We utilize RMPE [FXTL17] to predict 2D joints of each person and obtain the corresponding confidence scores $\{c_i\}_{1\leq i\leq J}$ where $J$ denotes the number of joints. RMPE adopts Faster R-CNN [RHGS17] as human detector and Pyramid Network [YLO*17] as single person pose estimator, respectively. We run RMPE on each frame of each camera. We find that this configuration has a great performance on inferring the keypoints of closely interacting persons, even in the presence of inaccurate human bounding boxes which is due to close interaction among multiple people.

### 3.2. Spatio-Temporal Tracking

In Section 3.1, we predict all the 2D joints of per person, but we do not know the everyone's order in a single image, *i.e.*, the multi-person pose estimation result in an image is unordered for each person and we need employ the tracking method to indicate the same person in each frame of each camera. One way is to use temporal tracking for each sequence, but this may fail for serious occlusion which is common for close interaction. Therefore, we propose a spatio-temporal tracker to better label each person in sequences. Specifically, we first unify the order of the characters in the starting

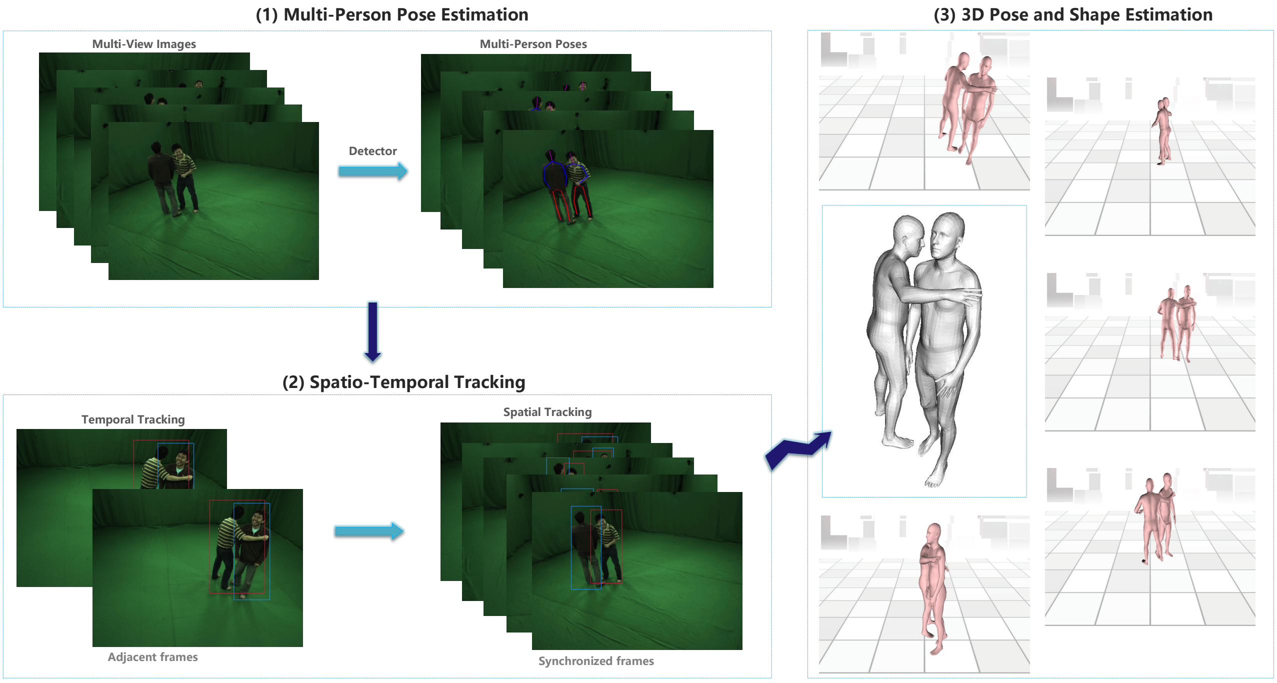


**Figure 2:** *Overview of our pipeline: 1) Multi-Person Pose Estimation. 2) Spatio-Temporal Tracker for Multi-Person Pose Tracking based on Multi-View Videos. 3) 3D Pose and Shape Estimation.*

frame of each video using a spatial criterion, and then use temporal tracking and spatial tracking alternately. Moreover, we take pose information into account to improve the accuracy of tracking. In addition, we do not use the segmentation, which is time-consuming to apply graph-cut model in spatial and temporal domains.

**Temporal Tracking** Here, we use important temporal information to infer the similarity of two poses in two adjacent frames. Using Hungarian algorithm to match the closest pose in the next frame is an effective method. We first perform frame-by-frame pose estimation on a sequence, and adopt the inter-frame pose distance defined in [XLW*18]:

$$P_d(P_1,P_2) = \sum_i \frac{n_i}{m_i}, \tag{1}$$

where $P_1$ and $P_2$ are the poses of two consecutive frames. Denote $p_1^i$ and $p_2^i$ as the $i^{th}$ keypoints of pose $P_1$ and $P_2$, respectively. Bounding boxes surrounding $p_1^i$ and $p_2^i$ are denoted as $B_1^i$ and $B_2^i$. According to the standard PCK [APGS14b], the size of box is 10% person bounding box size. We evaluate the similarity of $B_1^i$ and $B_2^i$ by the ORB matching [RRKB12] percentage $\frac{n_i}{m_i}$, where ORB matching is a very fast binary descriptor based on BRIEF (Binary Robust IndependentElementary Features) similar to SIFT, $m_i$ is the feature point extracted from $B_1^i$ and $n_i$ is the matching point in $B_2^i$.

Except the bounding boxes of pose information as a crucial factor, the bounding box of full body is also indispensable, which includes some feature points that pose cannot perceive. Therefore, given the detected bounding boxes $B_1$ and $B_2$ between frames, we define $BU = |B_1 \cup B_2|$ as the total feature points in $B_1$ and $B_2$, and $BI = |B_1 \cap B_2|$ as the matching feature points between $B_1$ and $B_2$. The similarity of $B_1$ and $B_2$ is defined as

$$B_o(B_1,B_2) = BI/BU. \tag{2}$$

We combine Eq. (1) and Eq. (2) to track the same person in two adjacent frames. The final metric function is defined as

$$T(P_1,P_2,B_1,B_2) = P_d(P_1,P_2) + B_o(B_1,B_2). \tag{3}$$

Note that, if we lose the pose in the current frame, we will add it from the previous frame, and we introduce a function to penalize the confidence score $c_i$ of the $i^{th}$ keypoint in that 2D pose:

$$C(c_i) = c_i \times mean(\sum_i c_i). \tag{4}$$

Figure 3 show a comparison result of without/with using the proposed penalty function in Eq. (4).

**Spatial Tracking** The spatial criterion used for the starting frame is defined as

$$B_d(B_1,B_2) = \frac{\sum_i m_i^1}{\sum_j m_j^2}, \tag{5}$$

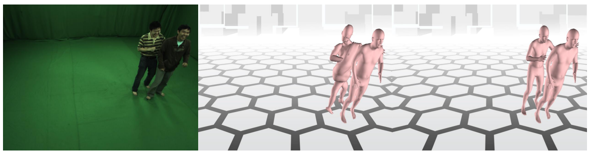

**Figure 3:** *Comparison result of shape estimation without using the propposed penalty function (Middle) and with the penalty function (Right).*

where $B_1$ and $B_2$ are the bounding boxes in two synchronized frames of different views. $m_j^2$ is the $j^{th}$ feature point detected from $B_2$, and $m_i^1$ is the $i^{th}$ feature point in $B_1$ that matches $m_j^2$. The similarity of $B_1$ and $B_2$ is evaluated by finding all matching points in $B_1$ from $B_2$, and we can identify and label the same person according to the similarity of a pair of bounding boxes. Note that we use DeepMatching [WRHS13] to robustly match the feature points between multi-view images, which involves a deep, multi-layer, convolutional architecture designed for matching images. Figure 4 shows some matching results using DeepMatching.

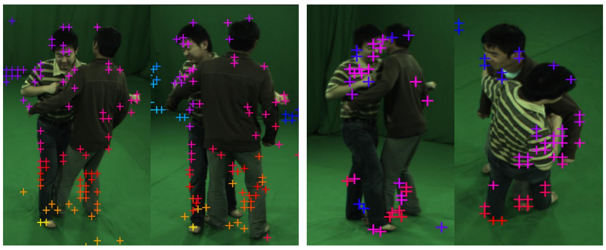

**Figure 4:** *Multi-view matching results of feature points using DeepMatching [WRHS13].*

For the frames after the starting frame, we use the interleaved spatio-temporal tracking. Specifically, we first track the same person in two adjacent frames of a video using temporal tracking, and then we track the same label of people in the synchronized frames of multi-view video sequences. The verification function for the person label $l$ of view $v$ is defined as

$$K_s(B_l^v) = \begin{cases} 1 & mean(\sum_{k=1,k\neq v}^{V} \frac{\sum_i m_j^k}{\sum_j m_i^v}) \geq \varepsilon \\ 0 & otherwise \end{cases}, \tag{6}$$

where $B_l^v$ is the bounding box of person $l$ in view $v$, $m_i^v$ is the feature point extracted from $B_l^v$, and $m_j^k$ is the matching point in the bounding box $B_l^k$ of person $l$ in view $k$. We set $\varepsilon = 0.3$ by cross-validation. The label $l$ in view $v$ is correct if $K_s(B_l^v) = 1$. If $K_s(B_l^v) = 0$, we will re-compute the labels in view $v$ by determining the most accurate label one by one. Specifically, we determine the most accurate label by computing the maximum of similarity mean:

$$H_s(B^v) = \max_p\{mean(\sum_{k=1,k\neq v}^{V} \frac{\sum_i m_j^k}{\sum_j m_i^v})|p \in N_p\}, \tag{7}$$

where $N_p$ is the number of people to be tracked. In order to find the most accurate label, we first calculate the sum of the scores of the matching similarities of the bounding boxes from other views in the remaining labels. Then, we find the most accurate label with the highest score. We repeat Eq. (7) to get the most appropriate label for each person. In this way, we can obtain the labels of poses in view $v$.

### 3.3. Multi-Person 3D Pose and Shape Estimation

Given the estimated 2D poses of different persons, we fit a skinned multi-person linear model (SMPL) [LMR*15] for each person by combining multi-view constraints.

**Model** The Skinned Multi-Person Linear model (SMPL) is a generative model that decomposes human body shape into identity-dependent shape and non-rigid pose-dependent shape. SMPL is defined as a function $M(\beta, \theta; \Phi)$, where $\beta$ is a vector of shape parameters containing 10 coefficients of a PCA shape space, $\theta$ is a vector of pose parameters using the axis-angle representation by a skeleton rig with $J = 23$ joints, $\Phi$ is a vector of the learned model parameters from a large number of 3D body meshes. The function outputs a triangulated surface with 6980 vertices. Please refer to [LMR*15] for more detailed meaning of all these parameters.

**Estimation** Using the single-view SMPLify [BKL*16] to fit multiple 3D human body models is impracticable and infeasible, because a lot of errors will occur due to occlusions especially for close interaction. If we use two or more camera views, many mistakes can be eliminated directly. Therefore, we estimate 3D pose and shape of each person using multi-view contraints. Specifically, we estimate the pose and shape parameters of the 3D body model at each time instance for each person. From the previous subsections, we obtain the 2D joints $J_{est}^v$ in the $v^{th}$ view together with confidence scores $\{c_i\}_{1\leq i\leq J}$ where $J$ denotes the number of joints. We minimize a robust weighted error function to fit a 3D body model by projecting joints of the model to multi-view images in a staged approach. Our energy function is defined as

$$E(\beta,\theta) = E_p(\beta,\theta) + \sum_{v=1}^{V} E_j(\beta,\theta; K_v, J_{est}^v), \tag{8}$$

where $E_p$ is the prior term, $E_j$ is the joint-based data term, $K_v$ are the camera parameters of the $v^{th}$ view. The prior term $E_p$ is defined as

$$E_p = \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta), \tag{9}$$

which contains a pose prior $E_\theta$ and a shape prior $E_\beta$ learnt from the CMU dataset [oCMU] and the SMPL body shape training set respectively, similar to SMPLify [BKL*16]. $\lambda_\theta$ and $\lambda_\beta$

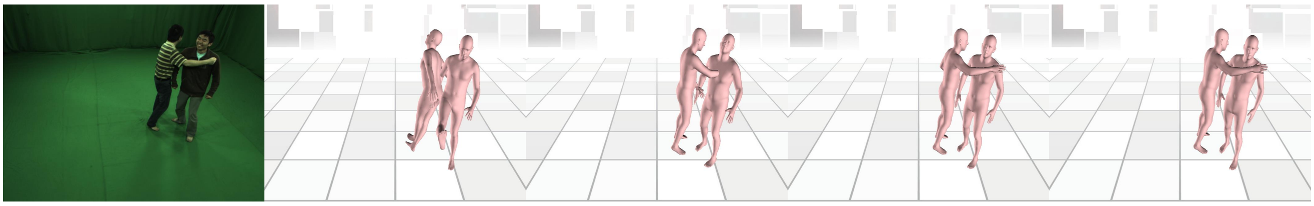

**Figure 5:** *Reconstruction results of using 1, 2, 4, 8 camera views (from left to right).*

are scalar weights, and are set to be $\{404, 404, 57.4, 4.78\}$ and $\{100, 50, 10, 5\}$ for four optimization stages, respectively. We remove $\lambda_\alpha E_\alpha(\theta)$ and $\lambda_{sp} E_{sp}(\theta;\beta)$ terms from SMPLify [BKL*16] because their contributions are no longer obvious.

The multi-view data term $E_j$ is defined as

$$E_j(\beta,\theta;K_v,J_{est}^v) = \sum_{\text{joint } i} c_i \rho_\sigma(\Pi_{K_v}(R_\theta(J_i(\beta))) - J_{est,i}^v), \quad (10)$$

where $J_i(\beta)$ is a function that predicts the $i^{th}$ skeleton joint location, $R_\theta$ is the global rigid transformation via pose $\theta$, $\Pi$ is the projection function, and $c_i$ is the confidence value of the $i^{th}$ joint. We use a robust Geman-McClure penalty function to help alleviate the impact of noise, which is defined as

$$\rho_\sigma(e) = \frac{e^2}{\sigma^2 + e^2}, \quad (11)$$

where $\sigma$ is a constant that is set to be 100 in our experiments, and $e$ is the residual error. We solve the optimization problem by using Powell's dogleg method [NW06], OpenDR [LB14] and Chumpy [Lop].

Our Shape and Pose Estimation for Close Interaction ALgorithm (SPECIAL) is summarized in Algorithm 1.

**Algorithm 1** Our SPECIAL algorithm

**Require:** Multi-view videos, $T \in \mathbb{N}, T \geq 1$.

```
for t = 1 to T do
    2D pose estimation using RMPE for multi-view images, and obtain poses {P_p} and the corresponding scores {c_i^p}.
    if t = 1 then
        Spatial tracking with multi-view images, and obtain the order(labels) of poses.
    else
        Temporal tracking using the (t-1)^th frame, and obtain the order(labels) of poses.
        Spatial tracking with multi-view images, and obtain the updated order(labels) of poses.
    end if
end for
for t = 1 to T do
    3D pose and shape estimation using multi-view constraints, and obtain the model M_t(β, θ; Φ).
end for
return the models {M_t(β, θ; Φ)}_{1≤t≤T}.
```

## 4. Experimental Results

In this section, we first evaluate the proposed method with ablation study in Section 4.2 on a public available multi-person interaction dataset (MHHI) [LSG*11] (Section 4.1), and then compare our method with the state-of-the-art methods qualitatively and quantitatively in Section 4.3. Finally, we give the detailed running times of our method in Section 4.4.

### 4.1. Dataset

We use MHHI dataset [LSG*11] to perform various ablation and comparison experiments. This dataset collects 7 different sequences consisting of 12 synchronized views with the image resolution of $1296 \times 972$, including multi-person markerless motion capture data and multi-person marker-based motion capture data that can be used for quantitative evaluation. Each sequence provides more than 200 frames with frame rates between 15fps and 60fps. In the marker-based data, one of the persons is attached with 38 markers and a commercial marker-based motion capture system *PhaseSpace*$^{TM}$ is used to capture his/her motion as ground truth. There are four challenging sequences available online (*Crash*, *Jump*, *Wrestle*, and *Fight*), with the frame rate of 45fps. The *Fight* sequence is a marker-based motion capture sequence, and is very challenging due to fast and complex motion. These sequences record a wide range of close interaction motions, which contain complex and extreme poses.

### 4.2. Ablation Study

In this section, we perform an ablation study to analyze the effect of different components of our approach.

#### 4.2.1. Multi-View

We investigate how the final reconstruction quality is affected by the number of camera views in Figure 5. From left to right shows the original captured image of the *Crash* dataset and the reconstruction results of using 1, 2, 4, 8 camera views, respectively. The cameras are sequentially selected according to their indices. As shown in Figure 5, the reconstruction result becomes better as the number of camera views increases. For the single view, close interaction causes the wrong estimation for the pose orientation. For the two views, the intersection of the final shapes occurs due to incorrect pose estimation. The reconstructed poses and shapes have already been good with four views, which demonstrates that our method can achieve accurate reconstruction for sparse camera settings. Table 1 gives the quantitative evaluation by comparing the position of markers and the corresponding reconstructed vertices on the *Fight* dataset. It can be observed that the estimation errors gradually decrease as the number of camera views increases. Multi-view provides more useful information than single-view, which helps eliminate inaccurate pose estimation and improve the accuracy of pose and shape estimation, especially for occlusion.

**Table 1:** *Quantitative evaluation for different number of cameras.*

| Number of views | 1 *view* | 2 *views* | 4 *views* | 8 *views* |
|---|---|---|---|---|
| Mean (mm) | 1549.88 | 242.27 | 58.42 | **48.57** |
| Std. | 2589.18 | 985.75 | 177.56 | **10.06** |

#### 4.2.2. Tracking

After 2D pose estimation for each person, it is essential to perform multi-person tracking before multi-person 3D pose and shape estimation. If only using temporal tracking, the pose may be lost or wrongly estimated due to occlusion. Hence, we propose to add spatial tracking to track the poses in spatio-temporal domain based on multi-view, not only between two adjacent frames. Figure 6 gives the comparison results of without and with spatial tracking on the *Wrestle* dataset. As shown in the figure, one of the persons disappears without using spatial tracking, while both poses and shapes of the persons are correctly estimated when the proposed spatio-temporal pose tracking is used.

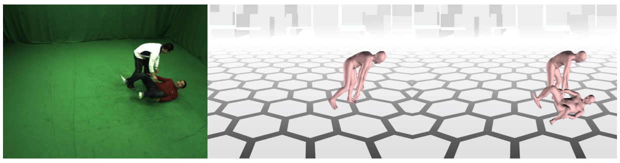

**Figure 6:** *Comparison results for an image of the Wrestle dataset without (Middle) and with (Right) spatial tracking.*

Figure 7 shows some results of without and with multi-person pose tracking on the *Crash* dataset. It can be seen that there are some very obvious mistakes, such as distorted shapes and wrong poses, without using pose tracking. When we track the poses in spatio-temporal domain, many of the above terrible artifacts are avoided. Quantitative evaluation is given in Table 2 by comparing the position of markers and the corresponding reconstructed vertices on the *Fight* dataset. It can be seen that the result with pose tracking has smaller error than that without pose tracking. Figure 8 shows more results on the four datasets by projecting the estimated shapes on the captured images. It can be seen that our estimated shapes basically coincide with the images without using the silhouettes.

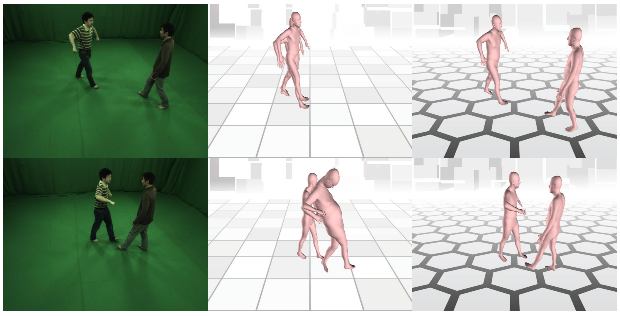

**Figure 7:** *Multi-person 3D reconstruction results on the Crash dataset (Left) of without (Middle) and with (Right) pose tracking.*

**Table 2:** *Quantitative evaluation of without tracking (N. track), with temporal tracking (T. track), and with spatio-temporal tracking (S. T. track).*

| Tracking | N. track | T. track | S. T. track |
|---|---|---|---|
| Mean (mm) | 214.79 | 74.95 | **43.30** |
| Std. | 361.32 | 45.46 | **9.45** |

### 4.3. Comparisons

Very few works can achieve multi-person 3D pose and shape estimation for closely interacting persons. The only one is proposed by Liu *et al.* [LGS*13], which need a laser scan to capture a template mesh and manual intervention to rig a skeleton. Moreover, their results depend on the careful segmentation for each person. On the contrast, our method is fully-automatic, fast, and without manual intervention and segmentation. Figure 9 shows the comparison results with the method in [LGS*13]. It can be seen that our method achieves the same level of accuracy for pose and shape estimation as the method in [LGS*13], although lacking of some geometry details. For the first image, we have even more accurate pose estimation result for the right hand of the left person than the method in [LGS*13]. For the last image, we have also more accurate pose estimation for the head of the flying man than the method in [LGS*13]. Table 3 gives quantitative evaluation on the *Fight* dataset. For error measurement, we similarly calculate the average distance with standard deviation between the markers and the corresponding vertices of the reconstructed model across all 500 frames of the sequence, which is the same as the evaluation method in [LGS*13]. As shown in the table, our method outperforms the method in [LGS*13] on both the average error and the standard deviation. This demonstrates that our method achieves more accurate estimation for poses and shapes than the method in [LGS*13]. Although the method in [LGS*13] has more rich geometry details, the accuracy of pose estimation is lower than our method. Moreover, the reconstructed models of the method in [LGS*13] have

**Figure 8:** *Qualitative evaluation by projecting the reconstructed shapes on the original images.*

clenched fists due to using the laser scan. Our approach is conceptually simpler and more accurate without any manual intervention.

We also compare our method with the newest 3D pose and shape estimation method [KBJM18] in Fig. 10. It can be seen that our method achives more accurate estimation for multi-person poses and shapes. Table 3 shows the qutitative evaluation on the *Fight* dataset. For the method in [KBJM18], we calculate the average distance with standard deviation between the markers and the corresponding vertices of the reconstructed model across all 500 frames for each camera, and obtain the final mean and standard deviation by averaging the multi-view means and standard deviations. The qutitative result further proves the effectiveness of the proposed method.

Fig. 11 shows some failure examples using our method due to wrong estimation of 2D joints for occlusion and complex motion cases.

**Table 3:** *Quantitative comparison with different methods.*

| Method | [LGS*13] | [KBJM18] | Ours |
|---|---|---|---|
| Mean(mm) | 51.67 | 753.69 | **43.30** |
| Std. | 23.44 | 337.54 | **9.45** |

### 4.4. Running Times

All the experiments are run on a desktop with a 32-core Intel Xeon(R) ES-2620 v4 2.1-GHz CPU, two 16.0-GB RAMs, and two GPUs of NVIDIA GeForce GTX1080Ti. Note that our method does not use GPU acceleration except the 2D pose estimation part. The 2D pose estimation using RMPE [FXTL17] takes about $1.2s$ per frame, the temporal tracking takes about $0.15s$ per frame, the spatial tracking takes about $49.5s$ per frame due to the time-consuming DeepMatching, and the 3D pose and shape estimation using multi-view constraints takes about $2s$ per frame. The total computation time of a person for a time instance is about $67s$, while

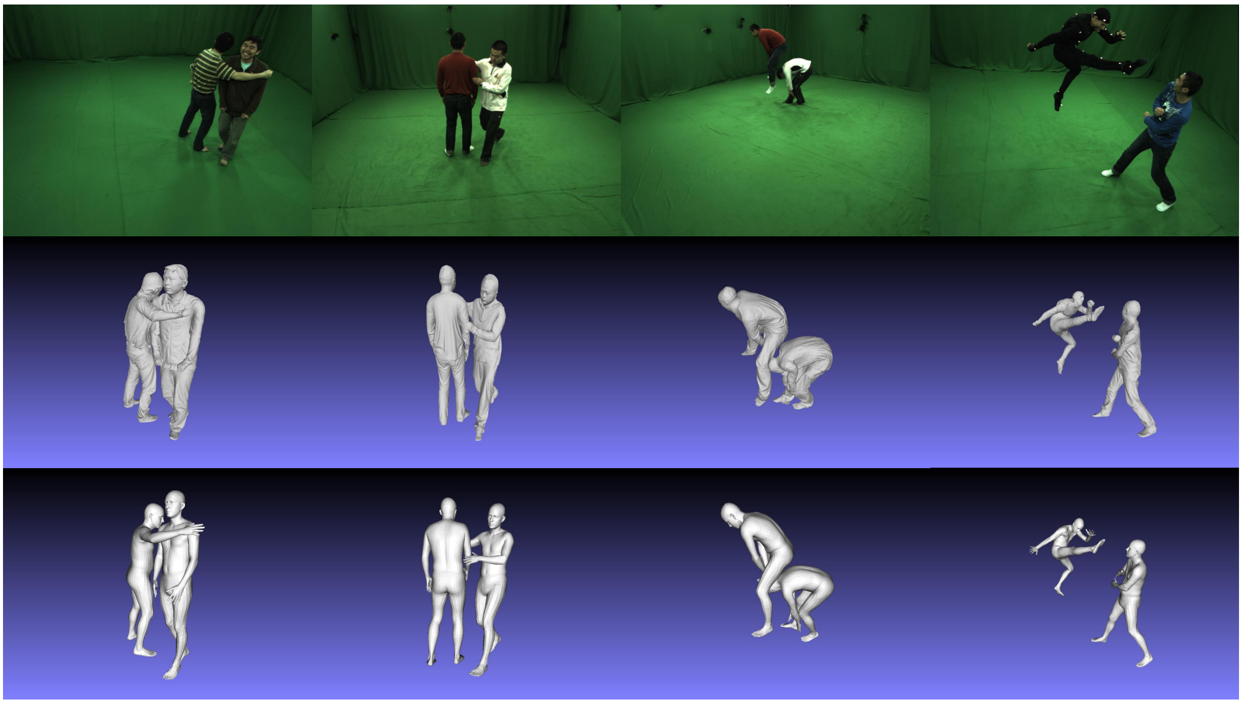

**Figure 9:** *Multi-person 3D reconstruction results for four images of the datasets (Top) by using the method in [LGS*13] (Middle) and our method (Bottom).*

the running time of the method in [LGS*13] is about 300$s$ except the time of scanning and manual rigging.

**Figure 10:** *Multi-person 3D reconstruction results for four images of the datasets by using the method in [KBJM18] (Top) and our method (Bottom).*

## 5. Conclusions

In this paper, we propose a new markerless multi-person motion capture method to estimate 3D shapes and poses for closely interacting persons from multi-view videos. To estimate more accurate and reliable 3D shape and pose, we design a novel tracking method based on spatio-temporal multi-view information, and combine a skinned multi-person linear model(SMPL) with multi-view constraints which enables our system robust to more complex scenarios. Experimental results show that our method achieves the same results with much less computational time and without manual intervention, compared with the state-of-the-art method.

In future work, we will try to achieve real-time 3D pose and shape estimation by using GPU acceleration, and obtain improved geometry by depth optimization. Besides, body language expression is a key content of human-human interactions, and hence we can estimate the meaning of human interaction by combining pose, shape and body parts, such as faces, hands and feet.


## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant 61571322, Grant 61771339, Grant 61522111 and Grant 61806054).




## References

[AARS13] Amin S., Andriluka M., Rohrbach M., Schiele B.: Multi-view pictorial structures for 3D human pose estimation. In *Proc.*

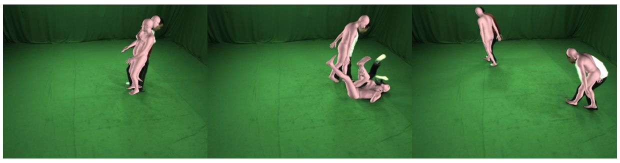

**Figure 11:** *Examples of failure cases.*

*British Machine Vision Conference* (2013), pp. 45.1–45.11. 3

[APGS14a] ANDRILUKA M., PISHCHULIN L., GEHLER P., SCHIELE B.: 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (June 2014). 2

[APGS14b] ANDRILUKA M., PISHCHULIN L., GEHLER P., SCHIELE B.: 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 3686–3693. 4

[AST*08] AGUIAR E. D., STOLL C., THEOBALT C., AHMED N., SEIDEL H. P., THRUN S.: Performance capture from sparse multi-view video. In *Proc. ACM SIGGRAPH* (2008), p. 98. 2, 3

[BGO*16] BEWLEY A., GE Z., OTT L., RAMOS F., UPCROFT B.: Simple online and realtime tracking. In *Proc. IEEE International Conference on Image Processing* (2016), pp. 3464–3468. 3

[BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P., ROMERO J., BLACK M. J.: Keep It SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. European Conference on Computer Vision* (2016), pp. 561–578. 3, 5, 6

[Bla08] BLACK M. J.: The naked truth: Estimating body shape under clothing. In *Proc. European Conference on Computer Vision* (2008), pp. 15–29. 3

[CBI10] CAGNIART C., BOYER E., ILIC S.: Free-form mesh tracking: A patch-based approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2010), pp. 1339–1346. 2

[Cho15] CHOI W.: Near-online multi-target tracking with aggregated local flow descriptor. In *Proc. IEEE International Conference on Computer Vision* (2015), pp. 3029–3037. 2, 3

[CSWS17] CAO Z., SIMON T., WEI S.-E., SHEIKH Y.: Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2017). 2

[CWP*17] CHEN Y., WANG Z., PENG Y., ZHANG Z., YU G., SUN J.: Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319* (2017). 2

[DR05] DEUTSCHER J., REID I.: Articulated body motion capture by stochastic search. *International Journal of Computer Vision 61*, 2 (2005), 185–205. 3

[DWL*16] DU Y., WONG Y., LIU Y., HAN F., GUI Y., WANG Z., KANKANHALLI M., GENG W.: Marker-less 3D human motion capture with monocular image sequence and height-maps. In *Proc. European Conference on Computer Vision* (2016), pp. 20–36. 3

[FXTL17] FANG H.-S., XIE S., TAI Y.-W., LU C.: RMPE: Regional multi-person pose estimation. In *Proc. IEEE International Conference on Computer Vision* (2017). 2, 3, 8

[GRBS10] GALL J., ROSENHAHN B., BROX T., SEIDEL H. P.: Optimization and filtering for human motion capture. *International Journal of Computer Vision 87*, 1-2 (2010), 75–92. 3

[GSA*09] GALL J., STOLL C., AGUIAR E. D., THEOBALT C., ROSENHAHN B., SEIDEL H. P.: Motion capture using joint skeleton tracking and surface estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 1746–1753. 2

[HAR*10] HASLER N., ACKERMANN H., ROSENHAHN B., THORMÄHLEN T., SEIDEL H.-P.: Multilinear pose and body shape estimation of dressed subjects from image sets. In *Proc. Computer Vision and Pattern Recognition* (2010), pp. 1823–1830. 3

[HBC*17] HUANG Y., BOGO F., CLASSNER C., KANAZAWA A., GEHLER P. V., AKHTER I., BLACK M. J.: Towards accurate markerless human shape and pose estimation over time. *arXiv preprint arXiv:1707.07548* (2017). 3

[HGDG17] HE K., GKIOXARI G., DOLLÁR P., GIRSHICK R.: Mask R-CNN. In *Proc. IEEE International Conference on Computer Vision* (2017), pp. 2980–2988. 2

[HGT17] HUANG S., GONG M., TAO D.: A coarse-fine network for keypoint localization. In *Proc. IEEE International Conference on Computer Vision* (2017), pp. 3047–3056. 2

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778. 2

[ICS14] IONESCU C., CARREIRA J., SMINCHISESCU C.: Iterated second-order label sensitive pooling for 3D human pose estimation. In *Proc. Computer Vision and Pattern Recognition* (2014), pp. 1661–1668. 3

[IPA*16] INSAFUTDINOV E., PISHCHULIN L., ANDRES B., ANDRILUKA M., SCHIELE B.: DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *Proc. European Conference on Computer Vision* (2016), pp. 34–50. 2

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *Proc. Computer Vision and Pattern Regognition* (2018). 3, 8, 9

[KLCR15] KIM C., LI F., CIPTADI A., REHG J. M.: Multiple hypothesis tracking revisited. In *Proc. IEEE International Conference on Computer Vision* (2015), pp. 4696–4704. 2, 3

[LB14] LOPER M. M., BLACK M. J.: OpenDR: An approximate differentiable renderer. In *Proc. European Conference on Computer Vision* (2014), pp. 154–169. 6

[LDX11] LI K., DAI Q., XU W.: Markerless shape and motion capture from multi-view video sequences. *IEEE Transactions on Circuits and Systems for Video Technology 21*, 3 (2011), 320–334. 2

[LGS*13] LIU Y., GALL J., STOLL C., DAI Q., SEIDEL H. P., THEOBALT C.: Markerless motion capture of multiple characters using multi-view image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*, 11 (2013), 2720–2735. 2, 3, 7, 8, 9

[LMB*14] LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P., ZITNICK C. L.: Microsoft COCO: Common objects in context. In *Proc. European Conference on Computer Vision* (2014), pp. 740–755. 2

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *Acm Transactions on Graphics 34*, 6 (2015), 248. 2, 3, 5

[Lop] LOPER M.: Chumpy. https://github.com/mattloper/chumpy. 6

[LSG*11] LIU Y., STOLL C., GALL J., SEIDEL H. P., THEOBALT C.: Markerless motion capture of interacting characters using multi-view image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2011), pp. 1249–1256. 6

[LZC15] LI S., ZHANG W., CHAN A. B.: Maximum-margin structured learning with deep networks for 3D human pose estimation. In *Proc. IEEE International Conference on Computer Vision* (2015), pp. 2848–2856. 3

[MKGH17] MUSTAFA A., KIM H., GUILLEMAUT J. Y., HILTON A.: General dynamic scene reconstruction from multiple view video. In *Proc. IEEE International Conference on Computer Vision* (2017), pp. 900–908. 2

[MN17] MORENO-NOGUER F.: 3D human pose estimation from a single image via distance matrix regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1561–1570. 3

[MRC*16] MEHTA D., RHODIN H., CASAS D., SOTNYCHENKO O., XU W., THEOBALT C.: Monocular 3D human pose estimation using transfer learning and improved CNN supervision. *arXiv preprint arXiv:1611.09813* (2016). 3

[NHD17] NEWELL A., HUANG Z., DENG J.: Associative embedding: End-to-end learning for joint detection and grouping. In *Proc. Advances in Neural Information Processing Systems* (2017), pp. 2274–2284. 2

[NW06] NOCEDAL J., WRIGHT S. J.: *Numerical Optimization*, 2 ed. Springer Series in Operations Research and Financial Engineering, 2006. 6

[oCMU] OF CARNEGIE MELLON UNIVERSITY G. L.: CMU motion capture database. http://mocap.cs.cmu.edu/. 5

[PF03] PLNKERS R., FUA P.: Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 9 (2003), 1182–1187. 3

[PIT*16] PISHCHULIN L., INSAFUTDINOV E., TANG S., ANDRES B., ANDRILUKA M., GEHLER P. V., SCHIELE B.: DeepCut: Joint subset partition and labeling for multi person pose estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4929–4937. 2, 3

[PZDD17] PAVLAKOS G., ZHOU X., DERPANIS K. G., DANIILIDIS K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proc. Computer Vision and Pattern Recognition* (2017), pp. 1263–1272. 3

[PZK*17] PAPANDREOU G., ZHU T., KANAZAWA N., TOSHEV A., TOMPSON J., BREGLER C., MURPHY K.: Towards accurate multi-person pose estimation in the wild. *arXiv preprint arXiv:1701.01779 8* (2017). 2

[PZS17] POPA A.-I., ZANFIR M., SMINCHISESCU C.: Deep multitask architecture for integrated 2D and 3D human sensing. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2017). 3

[RHGS17] REN S., HE K., GIRSHICK R., SUN J.: Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 6 (2017), 1137–1149. 3

[RKS13] RAMAKRISHNA V., KANADE T., SHEIKH Y.: Reconstructing 3D human pose from 2D image landmarks. In *Proc. European Conference on Computer Vision* (2013), pp. 573–586. 3

[RRD*16] RHODIN H., ROBERTINI N., DAN C., RICHARDT C., SEIDEL H. P., THEOBALT C.: General automatic human shape and motion capture using volumetric contour cues. In *Proc. European Conference on Computer Vision* (2016), pp. 509–526. 3

[RRKB12] RUBLEE E., RABAUD V., KONOLIGE K., BRADSKI G.: ORB: An efficient alternative to SIFT or SURF. In *Proc. IEEE International Conference on Computer Vision* (2012), pp. 2564–2571. 3, 4

[RSK*18] RHODIN H., SPÖRRI J., KATIRCIOGLU I., CONSTANTIN V., MEYER F., MÜLLER E., SALZMANN M., FUA P.: Learning monocular 3D human pose estimation from multi-view images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2018). 3

[SIHB12] SIGAL L., ISARD M., HAUSSECKER H., BLACK M. J.: Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision 98*, 1 (2012), 15–48. 3

[TAAS16] TANG S., ANDRES B., ANDRILUKA M., SCHIELE B.: Multi-person tracking by multicut and deep matching. In *Proc. European Conference on Computer Vision* (2016), pp. 100–111. 3

[TGHC16] TRUMBLE M., GILBERT A., HILTON A., COLLOMOSSE J.: Deep convolutional networks for marker-less human pose estimation from multiple views. In *Proc. European Conference on Visual Media Production* (2016), p. 6. 3

[TKS*16] TEKIN B., KATIRCIOGLU I., SALZMANN M., LEPETIT V., FUA P.: Structured prediction of 3D human pose with deep neural networks. *arXiv preprint arXiv:1605.05180* (2016). 3

[TRLF16] TEKIN B., ROZANTSEV A., LEPETIT V., FUA P.: Direct prediction of 3D body poses from motion compensated sequences. In *Proc. Computer Vision and Pattern Recognition* (2016), pp. 991–1000. 3

[VBM08] VLASIC D., BARAN I., MATUSIK W.: Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics 27*, 3 (2008), 1–9. 2, 3

[WLT*18] WANG Y., LIU Y., TONG X., DAI Q., TAN P.: Outdoor markerless motion capture with sparse handheld video cameras. *IEEE Transactions on Visualization and Computer Graphics 24*, 5 (2018), 1856–1866. 2

[WRHS13] WEINZAEPFEL P., REVAUD J., HARCHAOUI Z., SCHMID C.: DeepFlow: Large displacement optical flow with deep matching. In *Proc. IEEE Intenational Conference on Computer Vision* (Sydney, Australia, Dec. 2013). 5

[WRKS16] WEI S.-E., RAMAKRISHNA V., KANADE T., SHEIKH Y.: Convolutional pose machines. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4724–4732. 3

[WSVT13] WU C., STOLL C., VALGAERTS L., THEOBALT C.: On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics 32*, 6 (2013), 161. 2

[WVT12] WU C., VARANASI K., THEOBALT C.: Full body performance capture under uncontrolled and varying illumination: a shading-based approach. In *Proc. European Conference on Computer Vision* (2012), pp. 757–770. 3

[WWCW17] WANG B., WANG G., CHAN K. L., WANG L.: Tracklet association by online target-specific metric learning and coherent dynamics estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 3 (2017), 589–602. 3

[XAS15] XIANG Y., ALAHI A., SAVARESE S.: Learning to track: Online multi-object tracking by decision making. In *Proc. IEEE International Conference on Computer Vision* (2015), pp. 4705–4713. 3

[XCZ*17] XU W., CHATTERJEE A., ZOLLHÖFER M., RHODIN H., MEHTA D., SEIDEL H.-P., THEOBALT C.: MonoPerfCap: Human performance capture from monocular video. *arXiv preprint arXiv:1708.02136* (2017). 3

[XLW*18] XIU Y., LI J., WANG H., FANG Y., LU C.: Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977* (2018). 3, 4

[YGUU11] YAO A., GALL J., URTASUN R., URTASUN R.: Learning probabilistic non-linear latent variable models for tracking complex activities. In *Proc. International Conference on Neural Information Processing Systems* (2011), pp. 1359–1367. 3

[YHH*18] YIN K., HUANG H., HO E. S. L., WANG H., KOMURA T., COHENOR D., ZHANG R.: A sampling approach to generating closely interacting 3D pose-pairs from 2D annotations. *IEEE Transactions on Visualization and Computer Graphics PP*, 99 (2018). 3

[YLH*12] YE G., LIU Y., HASLER N., JI X., DAI Q., THEOBALT C.: Performance capture of interacting characters with handheld kinects. In *Proc. European Conference on Computer Vision* (2012), pp. 828–841. 3

[YLO*17] YANG W., LI S., OUYANG W., LI H., WANG X.: Learning feature pyramids for human pose estimation. In *Proc. IEEE International Conference on Computer Vision* (2017), pp. 1290–1299. 3

[ZSZ*16] ZHOU X., SUN X., ZHANG W., LIANG S., WEI Y.: Deep kinematic pose regression. In *Proc. European Conference on Computer Vision* (2016), pp. 186–201. 3

[ZZL*16] ZHOU X., ZHU M., LEONARDOS S., DERPANIS K. G., DANIILIDIS K.: Sparseness meets deepness: 3D human pose estimation from monocular video. In *Proc. Computer Vision and Pattern Recognition* (2016), pp. 4966–4975. 3