Shape-aware Multi-Person Pose Estimation from Multi-View Images

Zijian Dong[1] Jie Song[1] Xu Chen[1,2] Chen Guo[1] Otmar Hilliges[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen

Abstract

In this paper we contribute a simple yet effective approach for estimating 3D poses of multiple people from multi-view images. Our proposed coarse-to-fine pipeline first aggregates noisy 2D observations from multiple camera views into 3D space and then associates them into individual instances based on a confidence-aware majority voting technique. The final pose estimates are attained from a novel optimization scheme which links high-confidence multi-view 2D observations and 3D joint candidates. Moreover, a statistical parametric body model such as SMPL is leveraged as a regularizing prior for these 3D joint candidates. Specifically, both 3D poses and SMPL parameters are optimized jointly in an alternating fashion. Here the parametric models help in correcting implausible 3D pose estimates and filling in missing joint detections while updated 3D poses in turn guide obtaining better SMPL estimations. By linking 2D and 3D observations, our method is both accurate and generalizes to different data sources because it better decouples the final 3D pose from the inter-person constellation and is more robust to noisy 2D detections. We systematically evaluate our method on public datasets and achieve state-of-the-art performance. The code and video will be available on the project page: https://ait.ethz.ch/projects/2021/multi-human-pose/.

1. Introduction

Markerless human motion capture is one of the fundamental problems in computer vision. In recent years much progress has been made in estimating the configuration of the human body in 2D [5, 16, 34, 37, 50] and 3D [4, 25, 31, 32, 52] from a single RGB image as input. However, if we consider settings in which multiple people are depicted and in particular if these people are interacting with each other at close range, we can expect a multitude of difficulties due to the heavy and complicated occlusions and depth ambiguities. To robustly estimate the poses of such groups, multi-camera setups are indispensable to provide additional observations from different views which can resolve occlusion and provide stereo cues for 3D estimation.

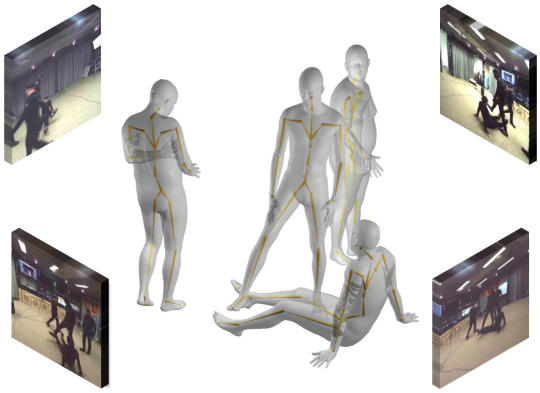

Figure 1. **Shape-aware multi-person pose estimation:** We propose a novel pipeline for robust recovery of 3D poses and shapes of multiple people from a few camera views. A formulation that links 2D and 3D observations and that is regularized via a parametric body model is robust to noisy and missing 2D detections. Articulated poses can even be recovered under heavy occlusion.

Due to the real-world importance of this problem, several recent approaches have attempted to predict the poses of multiple people, observed from multiple cameras [6, 10, 18, 45, 46, 54]. Such methods can loosely be categorized into two groups. The first group formulates the problem as a cross-view matching and association problem [6, 10, 54]. For example, Zhang et al. [54] introduce an optimization formulation that attempts to jointly solve the per-view parsing and cross-view matching problems as an instance of the multicut problem. The formulation is based on an association graph that links joints within and across multiple views. While the formulation is elegant, in practice it requires traversing the dense, cyclic association graph which results in an NP-hard problem. To attain a computationally tractable approach, the authors revert to a greedy heuristic which is sensitive to noisy 2D joint detections and imperfect visual features which limits the accuracy of the method.

Other methods, such as Tu et al. [46] combine the features from individual camera views into a 3D voxel space. This volume is then segmented into sub-volumes by a learned person detector. The final 3D human pose configurations are regressed from these sub-volumes. Because the

pipeline can be trained end-to-end, high accuracy can be achieved if the training and test distributions are similar. However, owing to the reliance on the volumetric feature representation – which encapsulates the joint distribution of different individuals, their location in 3D space, the camera setup and even the 2D joint detections – learning such a representation requires a vast amount of data. In the absence of a large corpus of annotated multi-people, multi-view data, such methods face generalization issues and are sensitive to distribution shifts (Tab. 4).

Embracing this challenging problem, we propose a simple yet effective coarse-to-fine pipeline to estimate 3D multi-person poses from multi-view images. Our method combines concepts from both bottom-up and top-down methods. To avoid having to solve the association problem with partial local evidence, we aggregate initial 3D pose proposals in a 3D feature space. Our first insight is that, in pose estimation, the uncertainty associated with the 2D features (i.e., joint detections) can be trusted more than in many other computer vision domains due to semantics. Thus we forgo neural-network based classifiers and propose a simple confidence-aware majority voting technique to obtain initial 3D proposals. We experimentally show that it is more robust to distribution differences in terms of human poses, location of individuals and cameras in space and thus leads to better generalization behavior. This coarse 3D localization step is followed by a refinement step to correct poses and fill in missing joints via an optimization scheme that leverages multi-view constraints directly, where high confidence 2D observations are available, and regularizes the 3D pose via a parametric body model.

More precisely, the first part of our pipeline consists of triangulating the 3D coordinates of all pairs of 2D detections with the same part label. This is followed by a confidence-aware majority voting technique to cluster the proposals. The technique is based on the insight that if a joint has been seen and predicted accurately (i.e., with high confidence) in several views, then there will be a dense cluster of 3D candidates for that joint and low confidence, isolated candidates can be discarded. Furthermore, we leverage the observation that certain joints, for example, the hip, are detected more reliably than end-effectors and can be used as a heuristic to decide the number and location of individual humans. While simple, experiments (Tab. 1) show that our approach outperforms the SOTA learning-based method [46] and matching-based method [54] in terms of the detection performance.

The second part of our pipeline refines the initial 3D estimates based on a novel 2D-3D objective (Eq.(5)). In our formulation, we optimize the 3D joint locations directly by minimizing the 2D re-projection error if the corresponding 2D joint detections are of high confidence (Eq.(1)). To regularize the fitting procedure and to attain complete and kinematically plausible poses we leverage SMPL for low-confidence 3D candidates (Eq.(3)). Importantly, the SMPL parameters are aligned directly to the updated 3D observations (current state of the 3D joint locations). For this we use the learned per-parameter gradient method (Eq.(6)). This approach is fundamentally different from most existing approaches [4] that fit SMPL parameters directly to 2D observations. We experimentally show that the triangulated 3D joints are more accurate than the PCA-based SMPL skeleton – if they stem from confident 2D observations (Fig. 4). Both initial 3D pose proposal and SMPL parameters are optimized in an alternating manner (Alg. 1). This is motivated by the insight that a good estimate of 3D poses helps in fitting SMPL, while better SMPL estimates make 3D poses more robust. Finally, detailed experiments are performed to demonstrate that both components improve the robustness and accuracy of the pose estimation task. In summary, our main contributions are:

- A coarse-to-fine confidence-aware pipeline to aggregate noisy 2D observations from all camera views into 3D space and associate them into individual instances.
- A novel refinement pipeline which optimizes 3D poses and their corresponding SMPL models in an alternating fashion. The parametric models help in regularizing low-confidence 3D poses while updated 3D poses in turn guide the SMPL parameter estimation.
- Our method is general since we only leverage off-the-shelf 2D pose detector and body pose priors distilled from motion capture datasets. SOTA performance is achieved on public datasets.

2. Related Work

A vast amount of work on single-person 3D pose estimation from monocular [8, 9, 11, 26, 36, 40, 40, 47] and multi-view [17, 20, 39] images exists. Since we study the setting of multi-person pose estimation from multiple views [1, 2, 10, 12, 23, 24, 28, 54], the focus of this literature review is on multi-person pose estimation.

Multi-Person 2D Pose Estimation

A natural approach to multi-person 2D pose estimation is to detect people first and then estimate the body pose independently. Pishchulin et al. [38] employs a pictorial structure model to locate the person and subsequently estimate the pose. More recent top-down approaches also follow a similar strategy but instead use CNN-based person detectors and pose estimation models [14, 16, 27, 49, 51].

In contrast, bottom-up approaches [5, 7, 33, 35, 37, 43] begin with localizing identity-free body part proposals and associates them into individual instances. The seminal work by Pishchulin et al. [37] proposed a framework that jointly

labels part candidates and also associates them into individual people. More recently, Cao et al. [5] introduced a representation of pairwise scores via the so-called Part Affinity Fields (PAF). The authors demonstrated that PAFs are able to provide effective features for the part association that a simple greedy bipartite parse can be directly applied achieving state-of-the-art results.

Multi-Person 3D Pose Estimation
When only one camera is available, the problem is underdetermined since many 3D poses may correspond to the same 2D pose. Leveraging the learning-based method, 3D poses can be recovered by lifting detected 2D poses [41, 42, 55], or directly regressing 3D poses [3, 13, 48, 53], or by fitting parametric human body models [21, 53]. However, the reconstruction accuracy of these methods is limited due to the depth ambiguities and strong occlusions when multiple humans are close to each other.

Most closely related to ours are methods that leverage multi-view images. A straightforward approach for this problem is to find correspondences across views, either leveraging high-level features such as human instances, or low-level features such as joints. Early work implicitly solves this matching and parsing problem by leveraging 3D pictorial structure models, in which nodes represent 3D locations of body joints and edges encode pairwise relations between them [2]. However, such methods are computationally expensive due to the large state space in 3D. Joo et al. [23] rely on local features from dense multi-view images to vote for possible 3D joint positions, which can be seen as an implicit form of matching.

The method proposed by Dong et al. [10] first performs per-view person parsing, followed by a cross-view person matching via a convex optimization method constrained by cycle consistency. In [54], the authors formulate parsing, matching, and tracking in a unified graph optimization framework to simultaneously address 4D information. In contrast to these matching-based methods, a recent work [46] directly localizes all people and estimates their corresponding 3D poses in 3D voxel space. Due to the reliance on the 3D feature representation as input for subsequent learning-based steps, this method faces challenges in generalization with different configurations of people, poses and cameras. In our work, we propose a simple yet effective pipeline that triangulates joint candidates and associates them into individual instances via a simple confidence-aware voting scheme. A 2D-3D optimization technique, optimized via learned gradient descent, produces highly accurate 3D pose estimates. We show that this pipeline outperforms both matching-based approaches and end-to-end learning methods.

3. Method

Fig. 2 provides an overview of our proposed approach, which contains two stages: 3D human proposal generation and shape-aware 3D pose optimization.

In the first stage, we generate 3D joint candidates by triangulating 2D human pose estimates from different views. Then a confidence-aware voting-based technique is applied to cluster joint candidates from noisy observations and determine human instances. To generate a pose proposal for each human instance, a 3D bounding box is placed around its hip and is projected back to the images. The image observations surrounding the body's limbs are used to filter joint candidates from closely interacting people.

In the second stage, an energy formulation which includes a multi-view re-projection term $E_{2d}(X)$ and a 3D body model fitting term $E_{shape}(X, \Theta)$ is introduced, to refine the initial pose X_0. Both 3D poses X and SMPL parameters Θ are optimized jointly in an alternating manner. For each iteration, the gradient updating network first takes current 3D poses X and SMPL estimation Θ as input to guide updating SMPL prediction. Then the current 3D poses X are optimized by minimizing the multi-view re-projection error when they stem from confident 2D observations and the updated SMPL prediction is leveraged for regularizing low-confidence or missing 3D joint candidates. After a small number of iterations, our method can generate complete and accurate 3D human poses.

3.1. 3D Human Proposal Generation

One of the main challenges for multi-person pose estimation from multiple views is to associate 2D poses from different views with consistent identities. Prior matching-based work [54], is sensitive to imperfect 2D detections due to its local heuristic, and purely learning-based methods [18, 46] are prone to overfitting. In contrast, we propose an effective approach to generate initial 3D pose proposals based on a confidence-aware voting technique, operating in the global 3D space of joint candidates that have been triangulated from pairs of 2D noisy detections.

3D Joint Candidates Reconstruction.
To reconstruct 3D joint candidates, we first run an off-the-shelf 2D human pose detector [5] on each input image to generate 2D joint detections (Fig. 2, (a)). Then pairs of joints with the same label from different views are triangulated into 3D joint candidates (Fig. 2, (b)). We use standard linear algebraic triangulation [15], solving the linear system defined on the homogeneous 3D coordinate vector $\tilde{y_j} : A_j \tilde{y_j} = 0$, where $A_j \in \mathbb{R}^{(2C,4)}$ is a matrix composed of the components from the projection matrices and 2D poses. In our case, we perform the triangulation from each pair of 2D poses and set C to 2.

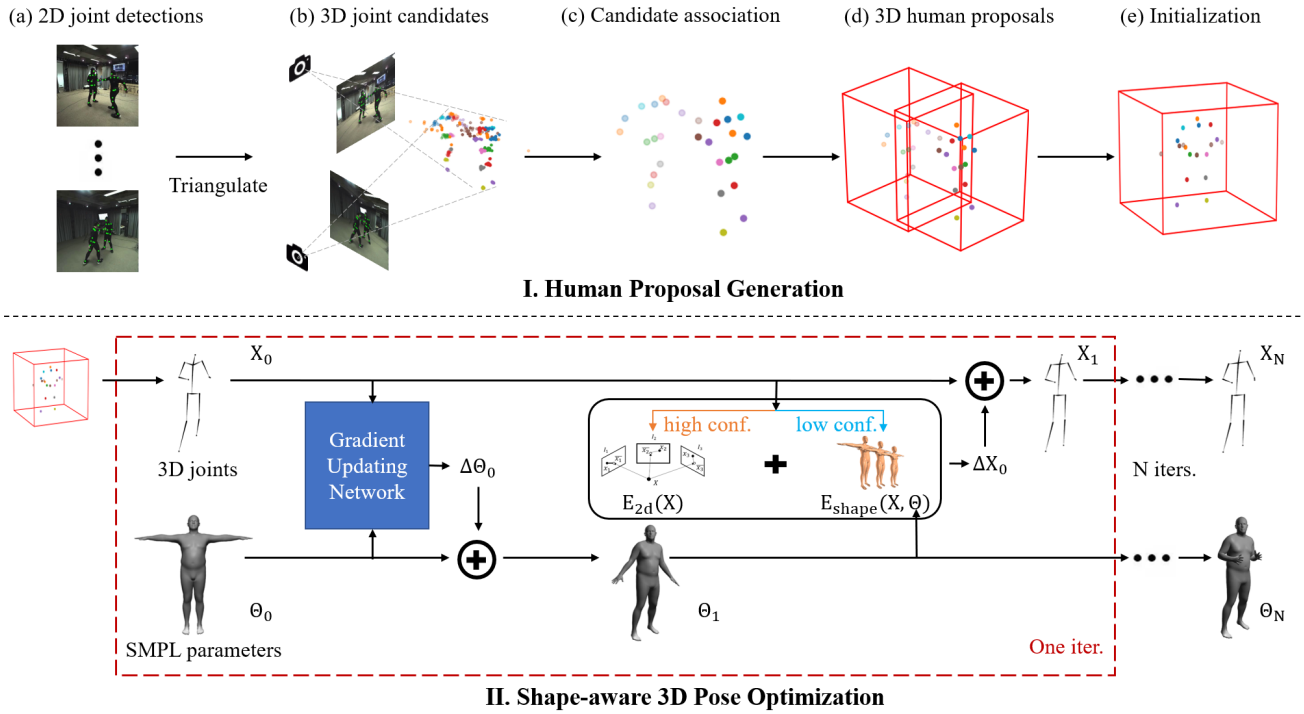

Figure 2. **Pipeline structure.** Stage I: (a): We apply a 2D human pose estimation method [5] to obtain 2D joint candidates. (b): 2D candidate pairs with the same part label are triangulated into 3D space to produce 3D joint candidates. (c): A confidence-aware voting-based algorithm is used for clustering joint candidates from partial observations. (d): The position of human instances can be detected based on a reliable joint. (e): For each 3D human proposal, we project it back into the image space and leverage the part affinity field feature (PAF [5]) to filter the joint candidates from closely interacting people and obtain initial 3D pose proposals. Stage II: We refine the initial 3D poses X_0 by optimizing a 2D-3D objective. Both 3D poses X and SMPL parameters Θ are optimized alternatively. For each iteration, the 3D joint locations X are optimized by a 2D re-projection error when the corresponding 2D joint detections are of high confidence. To obtain kinematically plausible poses, we leverage updated SMPL estimation for regularizing the low-confidence 3D joint candidates. The SMPL parameters Θ are encouraged to align to the updated 3D poses in each iteration via a learned gradient updating network. After a small number of iteration, our method can generate complete and accurate 3D human poses and output SMPL parameters.

Candidates Association.
The next step is to associate triangulated 3D joint candidates into individual instances. Our insight for the association is simple: since we triangulate pairs of joint detections, joints that are visible in several views produce dense clusters of 3D candidates. Based on this observation, we propose an efficient and effective voting-based algorithm.

For the set C_i of all the 3D joint candidates with part label i, we initialize an empty set S_i and update it iteratively. In each iteration, we first find the point p_k with the highest confidence in C_i. Next, a subset $s_k \in C_i$ containing all the neighboring 3D candidates around p_k with a distance less than threshold ρ is selected. We add s_k to S_i and remove s_k from C_i. We repeat the above until C_i becomes the empty set. Since outliers usually stem from falsely associated 2D detections or wrong detections in one specific view, there will only be few neighboring candidates around them. We thus eliminate clusters with less than three points. For the remaining clusters, we use their center to represent its position in 3D.

Human Proposal Generation.
The filtered 3D joint candidates need to be associated with individual instances. We experimentally find that the hip joint is one of the most reliable parts and can be leveraged to robustly decide the number of instances and also the location of each instance. Hence, we simply place a 3D bounding box with fixed size and orientation using the hip candidates as anchors. Furthermore, we keep the anchors whose corresponding 3D bounding box contains more than 90% of body parts and whose average confidences are larger than an empirically derived threshold. These bounding boxes may still contain joints of other closely interacting people. In order to distinguish them, we project the 3D bounding box back to the image space and use the part affinity fields [5] to determine which 3D joints belong to other person instances.

3.2. Shape-aware 3D Pose Optimization

The initial pose proposals do not yet adhere to kinematic constraints and may have missing joints due to imperfect 2D joint detections. We refine these initial poses X via both multi-view re-projection evidence $E_{2d}(X)$ and a parametric body model prior $E_{shape}(X, \Theta)$. The 3D poses X and SMPL parameters Θ are optimized alternatively. The re-projection term aligns 3D joints X with the 2D observations for high-confidence joint detections. Whereas missing or low-confidence joints are determined by leveraging the updated SMPL estimation to regularize the 3D pose, leading to complete and kinematically plausible 3D pose estimates. For SMPL parameters Θ, they are optimized to align to the current 3D joints X via a learned gradient updating network. This refinement finally leads to complete and kinematically plausible 3D pose estimates after a small number of iterations. The alternating process is shown in Fig. 2, II.

Objectives.

Given an initial 3D pose proposal X and its corresponding 2D observations x_{ij}, where $i \in (1 \dots N)$ stands for the joint label and $j \in (1 \dots K)$ represents the view indices, we want to refine the 3D pose by leveraging multi-view constraints where the 2D observations have high confidence:

$$E_{2D}(X) = \sum_{i=1}^{N} \sum_{j=1}^{K} w_{ij} \delta_{ij} d_{2D}(\Pi_j X_i, x_{ij}) \tag{1}$$

Here $E_{2D}(X)$ denotes the re-projection error between the 2D joint projections into each view and the detected 2D joints. Π_j is the projection matrix of view j. w_{ij} is the confidence of detected joint i in the view j and δ_{ij} is an indicator function denoting if joint i in view j is discarded:

$$\delta_{ij} = \begin{cases} 1, & d_{2D}(\Pi_j X_i, x_{ij}) < \rho_{2D} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where ρ_{2D} is a threshold for selecting inliers.

To complement E_{2D} we leverage a body shape term E_{shape} to regularize the 3D pose via a parametric body model when 2D detections are missing or have low associated confidence. For this purpose, we use the SMPL model[29]. It is a differentiable function that outputs a triangulated mesh $M(\theta, \beta)$ taking the pose parameters $\theta \in \mathbb{R}^{23\times 3}$ and the shape parameters $\beta \in \mathbb{R}^{10}$ as input. The 3D body joints can be obtained by a linear regressor W taking the mesh as input ($\bar{X}(\theta, \beta) = W(M(\theta, \beta))$). We jointly optimize predicted 3D poses X and SMPL parameters θ, β:

$$E_{shape}(X, \theta, \beta) = \sum_{i=1}^{N} \delta(w_i) d_{3D}(X_i, \bar{X}_i(\theta, \beta)) \tag{3}$$

where w_i is the average confidence of detected 2D joint i across views. $\delta(w_i)$ is an indicator function denoting whether the initial 3D joints i has high enough confidence:

$$\delta(w_i) = \begin{cases} 1, & w_i < \rho_{3D} \\ 0, & otherwise \end{cases} \tag{4}$$

The final energy is the weighted sum of Eq. (1) and (3):

$$E(X, \theta, \beta) = w_{2D} E_{2D}(X) + w_{shape} E_{shape}(X, \theta, \beta). \tag{5}$$

Algorithm 1 - Alternative Optimization

$X_0 \leftarrow$ initial 3D pose proposal
$\Theta_0 \leftarrow \{\theta_0, \beta_0\} \leftarrow 0$
for $n = 0, ..., N-1$ **do**
  $\Theta^{(0)} \leftarrow \Theta_n$
  for $m = 0, ..., M-1$ **do**
    $\mathcal{L}_{shape}(\Theta^{(m)}) \leftarrow E_{shape}(X_n, \Theta^{(m)})$
    $\Delta\Theta^{(m)} \leftarrow \mathcal{N}_w(\frac{\partial \mathcal{L}_{shape}(\Theta^{(m)})}{\partial \Theta^{(m)}}, \Theta^{(m)}, X_n)$
    $\Theta^{(m+1)} \leftarrow \Theta^{(m)} + \Delta\Theta^{(m)}$
  end for
  $\Theta_{n+1} \leftarrow \Theta^{(M)}$
  $\mathcal{L}(X_n) \leftarrow w_{2d} E_{2d}(X_n) + w_{shape} E_{shape}(X_n, \Theta_{n+1})$
  $X_{n+1} = X_n + \lambda \frac{\partial \mathcal{L}(X_n)}{\partial X_n}$
end for

Alternating Optimization.

To optimize Eq. (5) we employ a custom gradient descent strategy. We start by fixing X to its initial value and optimize θ, β. Since fitting SMPL parameters to 3D observations is a non-convex and highly non-linear problem, this can be slow and error-prone with traditional methods such as that by Bogo et al. [4]. We take inspiration from a recent 2D-3D lifting approach [44] that solves the fitting of 3D human body by leveraging a neural network to predict the parameter update rule. We adopt a similar concept for fitting the human model to the 3D candidates. To accelerate fitting of $\Theta = \{\theta, \beta\}$ we replace the standard gradient descent rule by a learned per-parameter update:

$$\Theta^{(m+1)} = \Theta^{(m)} + \mathcal{N}_w(\frac{\partial E_{shape}}{\partial \Theta^{(m)}}, \Theta^{(m)}, X) \tag{6}$$

where $\mathcal{N}_w$ is a deep network parameterized by a set of weights w, $\frac{\partial E_{shape}}{\partial \Theta^{(m)}}$ is the partial derivative wrt Θ and X is the fitting target. $\mathcal{N}_w$ is trained with samples of poses and shapes from AMASS [30]. Please refer to [44] and supplementary for more details of the training process.

Once Θ is optimized, we keep it fixed and optimize X via standard gradient descent. We optimize X, Θ in an alternating fashion until convergence. The overall routine is detailed as pseudo-code in Alg. 1, where n is the iteration index for the overall optimization routine while m is for the body fitting process.

4. Experiments

4.1. Test Datasets

We conduct experiments on two standard datasets for multi-view multi-person 3D human pose estimation, which consist of challenging scenarios including interactions between individuals with heavy occlusions.

Shelf [1] contains 3200 frames from 5 cameras. In terms of evaluation setting and evaluation metric, we follow prior work [1, 46, 54] and test on a separate test set and report the percentage of correctly estimated parts (PCP@0.5) to measure performance.

Association Dataset [54] is one of the most challenging public datasets for this task. It contains 3 sequences with 2-4 closely interacting people observed from 6 cameras. Following [54], we use all the sequences for test and report the precision, recall and F1-score as evaluation metrics. A joint is considered correct if its Euclidean distance to the ground truth annotation is less than 0.2m.

4.2. Training Data Comparison

For training, our method *only* uses AMASS, a collection of 3D human bodies with varying poses [30]. For clarity, we would like to emphasize that our primary goal is to boost the robustness and effectiveness to new, entirely unseen humans and poses. Note that, learning-based methods [18, 46] generally require direct 3D supervision via annotated pairs of multi-people, multi-view data.

4.3. Ablation Study

To validate the effectiveness of our method, we conduct a detailed analysis on both the initial 3D pose proposal and the shape-aware pose refinement. All the experiments are conducted on the ***Association*** dataset. When comparing with the learning-based method [46], we deploy its model trained on the CMU Panoptic Dataset [22], which is the largest training dataset for this task.

4.3.1 Evaluation of 3D Human Proposal Generation

To evaluate the effectiveness of our human proposal generation technique, we compare our method with two SOTA methods [46, 54]. Following the metric of [54], one human proposal is valid if the error of its hip joint is less than 0.2m. As shown in Tab. 1, our method achieves significantly better performance compared to [46, 54]. The learning-based method [46] faces generalization issues to unseen human poses and motion. The bottom-up method [54] suffers from low recall since the greedy algorithm is sensitive to noisy 2D joint detections. More ablation study on the proposal generation can be found in the supplementary.

Method	Precision(%)	Recall(%)	F1-score(%)
VoxelPose [46]	68.8	77.3	72.8
Zhang *et al.* [54]	99.6	51.2	67.6
Ours	98.8	**94.2**	**96.4**

Table 1. **Evaluation of human proposal generation on the Association Dataset.** A human proposal is valid if the error of its hip joint is less than 0.2m. Ours achieves better performance compared to other SOTA ones especially for recall.

4.3.2 Evaluation of Shape-aware Pose Optimization

In this section, we validate the effectiveness of our shape-aware optimization. As seen in the last two rows in Tab. 4, the initial pose proposal can be dramatically improved under all metrics by leveraging multi-view constraints and regularization with the parametric body model. There are two main reasons for this improvement. First, the full-body constraint of SMPL can infill missing joints and corrects implausible poses (see orange and purple circles in Fig. 3). Furthermore, the weighted re-projection error contributes to refine the joint predictions, shown in blue circles in Fig. 3.

Human Body Constraint. Note that the SMPL model is used as a full-body constraint to help infill missing joints and also to correct non-kinematic poses. We conduct the experiment to fit SMPL model to the initial 3D proposal and draw the MPJPEs of the joints before and after fitting in Fig. 4. For joints with associated low average confidence, regularizing the 3D pose via a parametric body model can boost performance. For joints with higher confidence (confidence$>$0.25), the effect of full-body constraint is not obvious. This is caused by a slight difference in the skeletal configurations between [5] and [29], which causes some systematic error. Therefore we only use the multi-view constraints to optimize high-confidence joints as described in the method section.

Comparison with SMPL-only. We compare our method with approaches that optimize only SMPL parameters to align the joint re-projections with 2D observations [4, 19]. For a fair comparison, we apply a multi-view variant of SMPLify. The comparison on the ***Association*** dataset is summarized in Tab. 2. Our method outperforms the SMPL-only baseline by a significant margin. This is in part due to the inherent approximation error stemming from differences in the skeletal configurations between [5] and [29] in part due to fitting errors. This experimental evidence is an important motivation for the design of our 3D pose fitting formulation. Furthermore, leveraging learned gradient descent for optimization leads to significant performance increases both in terms of runtime (20x speed-up, 0.1s vs 2s), convergence rate (14 vs 100 iters) and precision (90.1% vs 78.3%).

4.4. Quantitative Results

We compare our method with SOTA methods quantitatively on the ***Shelf*** and ***Association*** datasets. The compar-

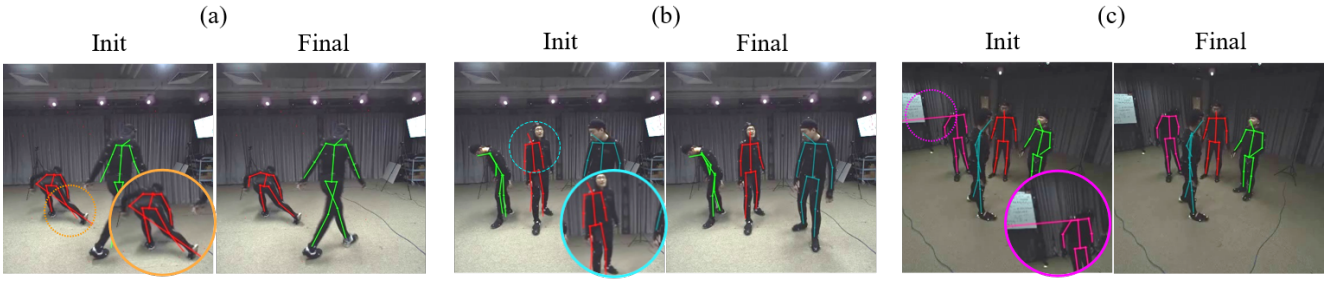

Figure 3. **Comparison between our initial poses and final poses after optimization.** (a) The orange circle is set for missing joints (note that we set the missing joints to the origin point) (b) The blue circle is set for incorrect joint predictions;(c) The purple circle is set for abnormal human poses. After the shape-aware optimization, our method generates more complete and accurate 3D human poses.

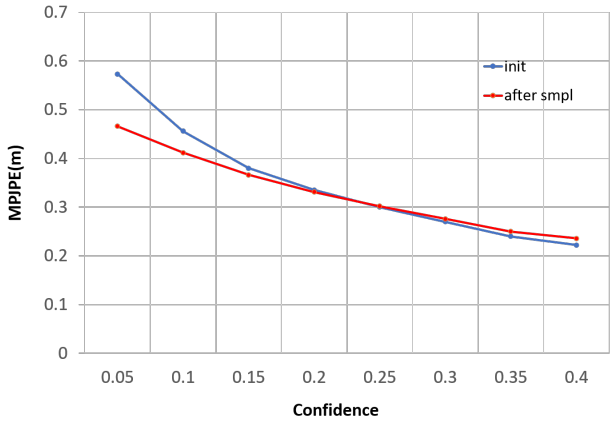

Figure 4. **Comparison of MPJPE (m) between our initial proposals (init) and the predicted poses after regularizing with SMPL (after smpl).** The horizontal axis represents the associated confidence of the 3D joints from the initial proposal. When the confidence is under 0.25, our algorithm achieves better performance after regularizing the 3D pose via a parametric body model.

Method	Precision(%)	Recall(%)	F1-score(%)
Multi-view SMPLify [4]	78.3	77.4	77.8
Ours	**90.1**	**89.0**	**89.2**

Table 2. **Comparison between our method and the traditional method [4]** Our method outperforms the SMPL-only baseline for both precision and recall.

Method	Anno	A1	A2	A3	Avg
VoxelPose [46]	Yes	99.3	94.1	97.6	97.0
Huang *et al.* [18]	Yes	98.8	96.2	97.2	97.3
Belagiannis *et al.* [1]	No	66.1	65.0	83.2	71.4
Belagiannis *et al.* [2]	No	75.3	69.7	87.6	77.5
Ershadi-Nasab *et al.* [12]	No	93.3	75.9	94.8	88.0
*Zhang *et al.* [54]	No	96.5	86.8	97.0	93.4
Dong *et al.* [10]	No	98.6	93.7	97.8	96.7(96.9)
Ours(final)	No	**99.1**	93.5	**98.1**	**96.9**

Table 3. **Evaluation on Shelf dataset.** Quantitative comparison on the Shelf dataset using a percentage of correct parts (PCP) metric. '*' denotes that the method discards temporal information from its original setting. 'A1'-'A3' correspond to the results of three actors, respectively. The averaged result is in column 'Avg'. The column 'Anno' indicates if the method relies on 3D supervision.

Method	Precision(%)	Recall(%)	F1-score(%)
VoxelPose [46]	55.1	66.5	60.3
*Zhang *et al.* [54]	97.1	48.8	65.0
†VoxelPose [46]	68.8	79.2	73.6
Dong *et al.* [10]	71.0	80.2	75.3
Ours(init)	83.7	82.8	83.4
Ours(final)	90.1	**89.0**	**89.2**

Table 4. **Evaluation on the Association Dataset.** '*' denotes that the method discards temporal information from its original setting. '†' means the method uses 3D bounding box ground truth.

ison on the ***Shelf*** is shown in Tab. 3. We achieve slightly better results compared to methods [1, 2, 10, 12, 54] which do not rely on 3D supervision and achieve comparable performance compared to learning-based methods [18, 46] which train the model based on this dataset. Since the test frames lack pose variations compared to the training set, this dataset is considered less challenging than the ***Association*** one, which also has a more strict evaluation metric.

The quantitative result on the ***Association*** dataset is shown in Tab. 4. Since this is a pure test set, for SOTA learning-based method [46], we directly deploy their trained model for test. As a result, our method outperforms this learning-based method largely, even when they use ground truth 3D bounding boxes. This shows that the learning-based method is prone to overfit to the training distribution. We also compare our algorithm with matching-based methods [10, 54]. Note that to have a fair comparison with [54], we compare with its static version only relying on images while it is originally a tracking method which leverages video information. We can see that these bottom-up based methods have relatively low recall due to the fact that they solve the global optimization with a greedy algorithm which is sensitive to missing 2D joint detections.

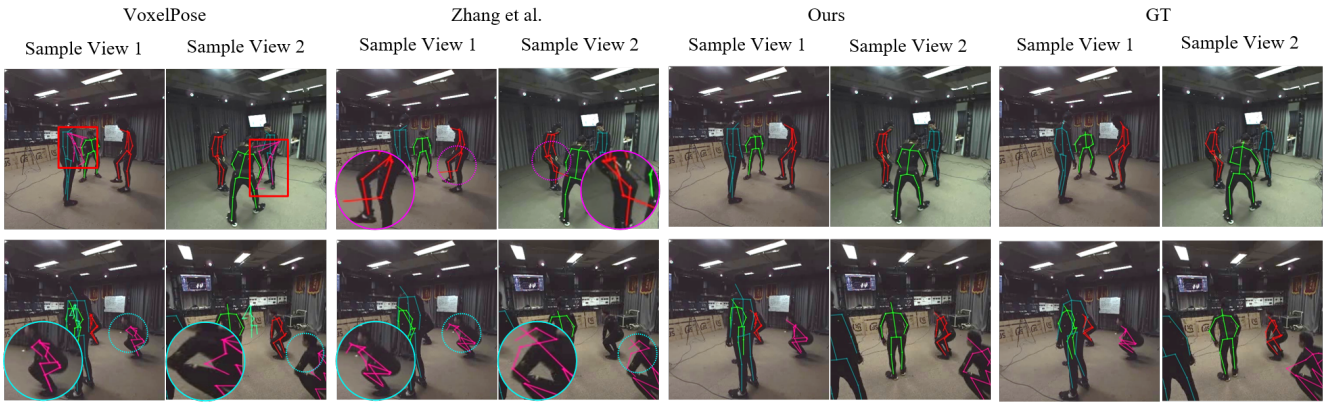

Figure 5. **Qualitative comparison with VoxelPose [46] and Zhang et al. [54] on the Association Dataset.** Different colors stand for different types of errors: red rectangles for extra actors; blue circles for incorrect joint positions; purple circles for abnormal human poses. Each row is an independent sample and the results are the projected 2D poses from the 3D predictions. Our method is more accurate compared to others especially in challenging scenarios with strong occlusions or highly articulated poses.

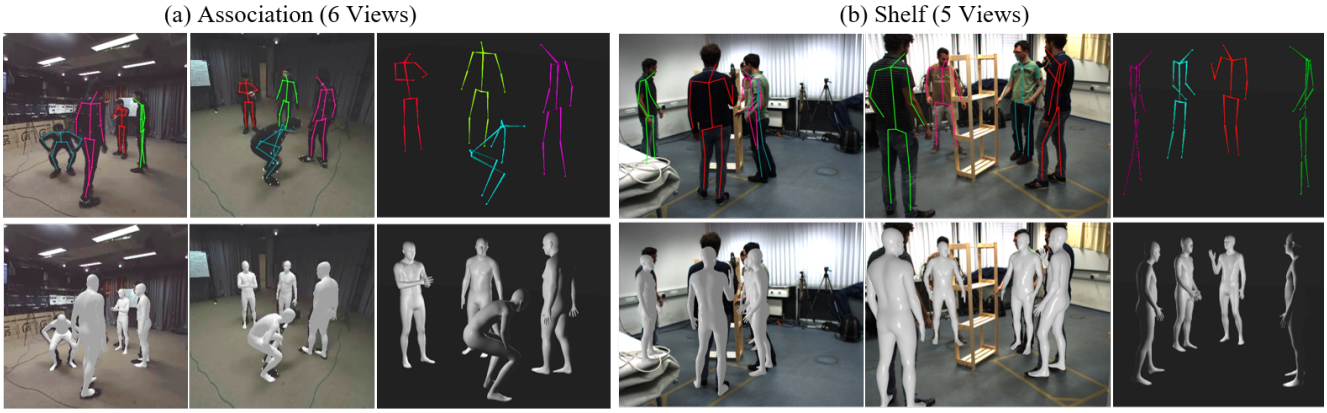

Figure 6. **Results of our method on (a) Association Dataset(6 views) and (b) Shelf (5 views).** The 2D reprojections of predicted 3D skeletons and SMPL mesh model are shown in two different views (left and middle); Skeletons and SMPL models of all the actors in 3D are demonstrated in the right column. More results can be found in the supplementary.

4.5. Qualitative Comparison

We show qualitative comparison in Fig. 5. Generally, our method is more accurate and robust compared to others especially in challenging scenarios with strong occlusions or when highly articulated poses are presented. Specifically, other methods tend to generate extra actors (red rectangles), abnormal (purple circles) or incorrect (blue circles) human poses. The reason is that it is difficult for the learning-based method [46] to generalize to unseen poses and motions. For the image-based version of [54], without temporal information, solving the 3D association graph is sensitive to noisy 2D joint detections. In Fig. 6, we demonstrate more results of our method on both the ***Association*** and ***Shelf*** datasets. The first row is the predicted 3D poses and their projections on two views. The second row shows the predicted SMPL models in 3D space and also their projections in 2D images.

5. Conclusion

In this paper, we propose an effective coarse-to-fine pipeline to estimate 3D multi-person poses from multi-view images. To avoid having to solve the association problem with local evidence, we aggregate initial 3D pose proposals in a 3D feature space and associate them into individual instances. This coarse 3D localization step is followed by a refinement step that corrects poses and fills in missing joints via an optimization routine leveraging multi-view constraints directly where high confidence 2D observations are available and regularizing the 3D pose via a parametric body model. We systematically evaluate our method on public datasets and SOTA performance is achieved.

Acknowledgements: Xu Chen was supported by the Max Planck ETH Center for Learning Systems.

References

[1] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, 2014.

[2] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):1929–1942, 2015.

[3] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6856–6865, 2020.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[6] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded scenes based on multi-view geometry. *arXiv preprint arXiv:2007.10986*, 2020.

[7] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020.

[8] Stefano Corazza, Lars Mündermann, Emiliano Gambaretto, Giancarlo Ferrigno, and Thomas P Andriacchi. Markerless motion capture through visual hull, articulated icp and subject specific model generation. *International journal of computer vision*, 87(1-2):156, 2010.

[9] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM SIGGRAPH 2008 papers*, pages 1–10. 2008.

[10] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019.

[11] Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3810–3818, 2015.

[12] Sara Ershadi-Nasab, Erfan Noury, Shohreh Kasaei, and Esmaeil Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018.

[13] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7213, 2020.

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

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[17] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformer for multi-view human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1036–1037, 2020.

[18] Congzhentao Huang, Shuai Jiang, Yang Li, Ziyue Zhang, Jason Traish, Chen Deng, Sam Ferguson, and Richard Yi Da Xu. End-to-end dynamic matching network for multi-view multi-person 3d pose estimation.

[19] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V Gehler, Javier Romero, Ijaz Akhter, and Michael J Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 international conference on 3D vision (3DV)*, pages 421–430. IEEE, 2017.

[20] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7718–7727, 2019.

[21] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020.

[22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[23] 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. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):190–204, 2017.

[24] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018.

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

[26] Roland Kehl and Luc Van Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding*, 104(2-3):190–209, 2006.

[27] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019.

[28] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2720–2735, 2013.

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[30] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5442–5451, 2019.

[31] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[32] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.

[33] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016.

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[35] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018.

[36] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[37] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. *arXiv preprint arXiv:1511.06645*, 2015.

[38] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012.

[39] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4342–4351, 2019.

[40] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6040–6049, 2020.

[41] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017.

[42] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1146–1161, 2019.

[43] Jie Song, Bjoern Andres, Michael J Black, Otmar Hilliges, and Siyu Tang. End-to-end learning for graph decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10093–10102, 2019.

[44] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 744–760. Springer, 2020.

[45] Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4220–4229, 2017.

[46] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. ECCV, 2020.

[47] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008.

[48] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *European Conference on Computer Vision*, pages 242–259. Springer, 2020.

[49] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[50] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016.

[51] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[52] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 2148–2157, 2018.

[53] 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. *Advances in Neural Information Processing Systems*, 31:8410–8419, 2018.

[54] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1324–1333, 2020.

[55] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *European Conference on Computer Vision*, pages 550–566. Springer, 2020.