Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views

Junting Dong[1]* Wen Jiang[1] Qixing Huang[2] Hujun Bao[1] Xiaowei Zhou[1]†
[1]Zhejiang University [2]University of Texas at Austin

Abstract

This paper addresses the problem of 3D pose estimation for multiple people in a few calibrated camera views. The main challenge of this problem is to find the cross-view correspondences among noisy and incomplete 2D pose predictions. Most previous methods address this challenge by directly reasoning in 3D using a pictorial structure model, which is inefficient due to the huge state space. We propose a fast and robust approach to solve this problem. Our key idea is to use a multi-way matching algorithm to cluster the detected 2D poses in all views. Each resulting cluster encodes 2D poses of the same person across different views and consistent correspondences across the keypoints, from which the 3D pose of each person can be effectively inferred. The proposed convex optimization based multi-way matching algorithm is efficient and robust against missing and false detections, without knowing the number of people in the scene. Moreover, we propose to combine geometric and appearance cues for cross-view matching. The proposed approach achieves significant performance gains from the state-of-the-art (96.3% *vs.* 90.6% *and* 96.9% *vs.* 88% *on the Campus and Shelf datasets, respectively), while being efficient for real-time applications.*

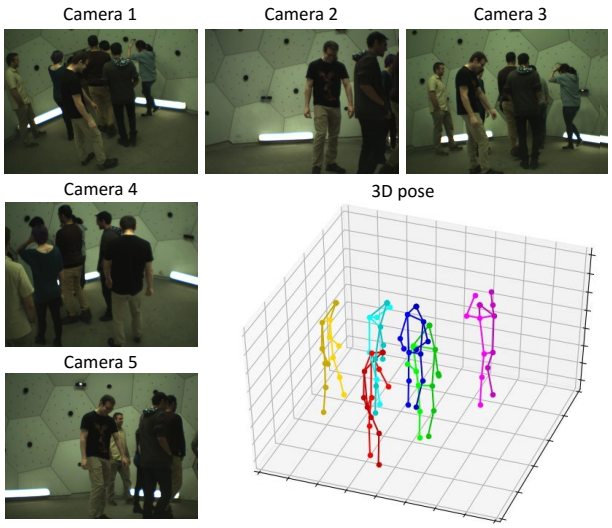

Figure 1: This work proposes a novel approach for fast and robust recovery of 3D poses of multiple people from a few camera views. The main challenge is to establish consistent correspondences of 2D observations among multiple views, e.g., 2D human-body keypoints in images, which may be noisy and incomplete.

1. Introduction

Recovering 3D human pose and motion from videos has been a long-standing problem in computer vision, which has a variety of applications such as human-computer interaction, video surveillance and sports broadcasting. In particular, this paper focuses on the setting where there are multiple people in a scene, and the observations come from a few calibrated cameras (Figure 1). While remarkable advances have been made in multi-view reconstruction of a human body, there are fewer works that address a more challenging setting where multiple people interact with each other in crowded scenes, in which there are significant occlusions.

Existing methods typically solve this problem in two stages. The first stage detects human-body keypoints or parts in separate 2D views, which are aggregated in the second stage to reconstruct 3D poses. Given the fact that deep-learning based 2D keypoint detection techniques have achieved remarkable performance [8, 30], the remaining challenge is to find cross-view correspondences between detected keypoints as well as which person they belong to. Most previous methods [1, 2, 21, 12] employ a 3D pictorial structure (3DPS) model that implicitly solves the correspondence problem by reasoning about all hypotheses in 3D that are geometrically compatible with 2D detections. However, this 3DPS-based approach is computationally expensive due to the huge state space. In addition, it is not robust particularly when the number of cameras is small, as it only uses multi-view geometry to link the 2D detections

*The authors from Zhejiang University are affiliated with the State Key Lab of CAD&CG and the ZJU-SenseTime Joint Lab of 3D Vision.
†Corresponding author.

across views, or in other words, the appearance cues are ignored.

In this paper, we propose a novel approach for multi-person 3D pose estimation. The proposed approach solves the correspondence problem at the body level by matching detected 2D poses among multiple views, producing clusters of 2D poses where each cluster includes 2D poses of the same person in different views. Then, the 3D pose can be inferred for each person separately from matched 2D poses, which is much faster than joint inference of multiple poses thanks to the reduced state space.

However, matching 2D poses across multiple views is challenging. A typical approach is to use the epipolar constraint to verify if two 2D poses are projections of the same 3D pose for each pair of views [23]. But this approach may fail for the following reasons. First, the detected 2D poses are often inaccurate due to heavy occlusion and truncation, as shown in Figure 2(b), which makes geometric verification difficult. Second, matching each pair of views separately may produce inconsistent correspondences which violate the cycle consistency constraint, that is, two corresponding poses in two views may be matched to different people in another view. Such inconsistency leads to incorrect multi-view reconstructions. Finally, as shown in Figure 2, different sets of people appear in different views and the total number of people is unknown, which brings additional difficulties to the matching problem.

We propose a multi-way matching algorithm to address the aforementioned challenges. Our key ideas are: (i) combing the geometric consistency between 2D poses with the appearance similarity among their associated image patches to reduce matching ambiguities, and (ii) solving the matching problem for all views simultaneously with a cycle-consistency constraint to leverage multi-way information and produce globally consistent correspondences. The matching problem is formulated as a convex optimization problem and an efficient algorithm is developed to solve the induced optimization problem.

In summary, the main contributions of this work are:

- We propose a novel approach for fast and robust multi-person 3D pose estimation. We demonstrate that, instead of jointly inferring multiple 3D poses using a 3DPS model in a huge state space, we can greatly reduce the state space and consequently improve both efficiency and robustness of 3D pose estimation by grouping the detected 2D poses that belong to the same person in all views.

- We propose a multi-way matching algorithm to find the cycle-consistent correspondences of detected 2D poses across multiple views. The proposed matching algorithm is able to prune false detections and deal with partial overlaps between views, without knowing the true number of people in the scene.

- We propose to combine geometric and appearance cues to match the detected 2D poses across views. We show that the appearance information, which is mostly ignored by previous methods, is important to link the 2D detections across views.

- The proposed approach outperforms the state-of-the-art methods by a large margin without using any training data from the evaluated datasets. The code is available at https://zju3dv.github.io/mvpose/.

2. Related work

Multi-view 3D human pose: Markerless motion capture has been investigated in computer vision for a decade. Early works on this problem aim to track the 3D skeleton or geometric model of human body through a multi-view sequence [38, 43, 11]. These tracking-based methods require initialization in the first frame and are prone to local optima and tracking failures. Therefore, more recent works are generally based on a bottom-up scheme where the 3D pose is reconstructed from 2D features detected from images [36, 6, 32]. Recent work [22] shows remarkable results by combing statistical body models with deep learning based 2D detectors.

In this work, we focus on the multi-person 3D pose estimation. Most previous works are based on 3DPS models in which nodes represent 3D locations of body joints and edges encode pairwise relations between them [1, 20, 2, 21, 12]. The state space for each joint is often a 3D grid representing a discretized 3D space. The likelihood of a joint being at some location is given by a joint detector applied to all 2D views and the pairwise potentials between joints are given by skeletal constraints [1, 2] or body parts detected in 2D views [21, 12]. Then, the 3D poses of multiple people are jointly inferred by maximum a posteriori estimation.

As all body joints for all people are considered simultaneously, the entire state space is huge, resulting in heavy computation in inference. Another limitation of this approach is that it only uses multi-view geometry to link 2D evidences, which is sensitive to the setup of cameras. As a result, the performance of this approach degrades significantly when the number of views decreases [21]. Recent work [23] proposes to match 2D poses between views and then reconstructs 3D poses from the 2D poses belonging to the same person. But it only utilizes epipolar geometry to match 2D poses for each pair of views and ignores the cycle consistency constraint among multiple views, which may result in inconsistent correspondences.

Single-view pose estimation: There is a large body of literature on human pose estimation from single images.

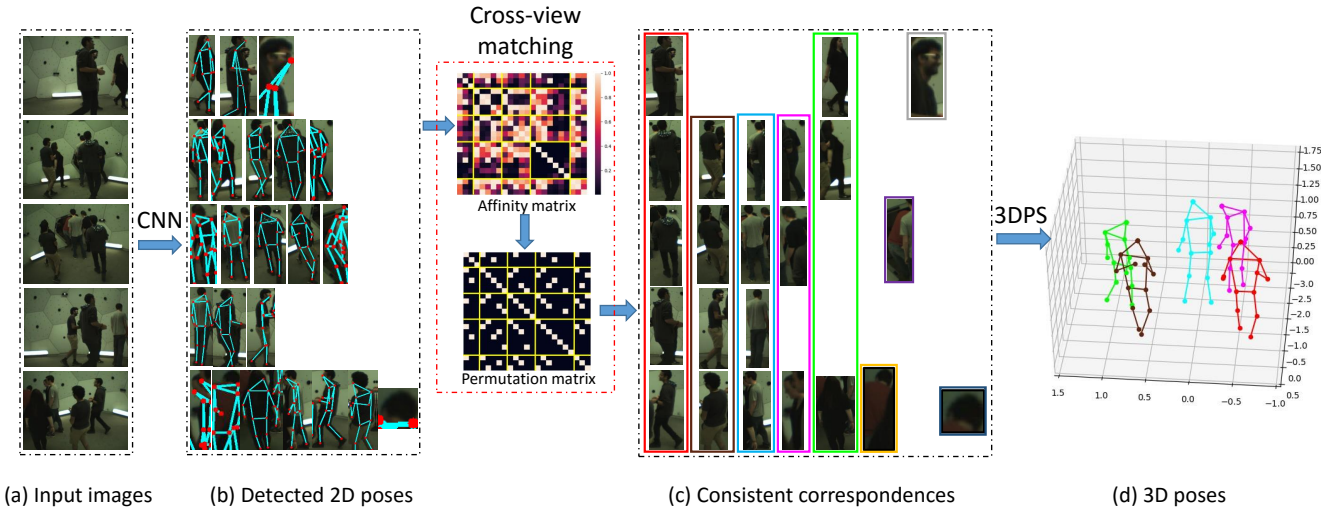

Figure 2: Overview of the proposed approach. Given images from a few calibrated cameras (a), an off-the-shelf human pose detector is used to produce 2D bounding boxes and associated 2D poses in each view, which may be inaccurate and incomplete (b). Then, the detected bounding boxes are clustered by a novel multi-view matching algorithm. Each resulting cluster includes the bounding boxes of the same person in different views (c). The isolated bounding boxes that have no matches in other views are regarded as false detections and discarded. Finally, the 3D pose of each person is reconstructed from the corresponding bounding boxes and associated 2D poses (d).

Single-person pose estimation [41, 34, 42, 30, 17] localizes 2D body keypoints of a person in a cropped image. There are two categories of multi-person pose estimation methods: top-down methods [10, 17, 15, 13] that first detect people in the image and then apply single-person pose estimation to the cropped image of each person, and bottom-up methods [25, 29, 8, 35, 18] that first detect all keypoints and then group them into different people. In general, the top-down methods are more accurate, while the bottom-up methods are relatively faster. In this work, We adopt the Cascaded Pyramid Network [10], a state-of-the-art approach for multi-person pose detection, as an initial step in our pipeline.

The advances in learning-based methods also make it possible to recover 3D human pose from a single RGB image, either lifting the detected 2D poses into 3D [28, 47, 9, 27] or directly regressing 3D poses [40, 37, 39, 45, 31] and even 3D body shapes from RGB [4, 24, 33]. But the reconstruction accuracy of these methods is not comparable with the multi-view results due to the inherit reconstruction ambiguity when only a single view is available.

Person re-ID and multi-image matching: Person re-ID aims to identify the same person in different images [44], which is used as a component in our approach. Multi-image matching is to find feature correspondences among a collection of images [16, 46]. We make use of the recent results on cycle consistency [16] to solve the correspondence problem in multi-view pose estimation.

3. Technical approach

Figure 2 presents an overview of our approach. First, an off-the-shelf 2D human pose detector is adopted to produce bounding boxes and 2D keypoint locations of people in each view (Section 3.1). Given the noisy 2D detections, a multi-way matching algorithm is proposed to establish the correspondences of the detected bounding boxes across views and get rid of the false detections (Section 3.2). Finally, the 3DPS model is used to reconstruct the 3D pose for each person from the corresponding 2D bounding boxes and keypoints (Section 3.3).

3.1. 2D human pose detection

We adopt the recently-proposed Cascaded Pyramid Network [10] trained on the MSCOCO [26] dataset for 2D pose detection in images. The Cascaded Pyramid Network consists of two stages: the GlobalNet estimates human poses roughly whereas the RefineNet gives optimal human poses. Despite its state-of-the-art performance on benchmarks, the detections may be quite noisy as shown in Figure 2(b).

3.2. Multi-view correspondences

Before reconstructing the 3D poses, the detected 2D poses should be matched across views, i.e., we need to find in all views the 2D bounding boxes belonging to the same

person. However, this is a challenging task as we discussed in the introduction.

To solve this problem, we need 1) a proper metric to measure the likelihood that two 2D bounding boxes belong to the same person (a.k.a. affinity), and 2) a matching algorithm to establish the correspondences of bounding boxes across multiple views. In particular, the matching algorithm should not place any assumption about the true number of people in the scene. Moreover, the output of the matching algorithm should be cycle-consistent, i.e. any two corresponding bounding boxes in two images should correspond to the same bounding box in another image.

Problem statement: Before introducing our approach in details, we first briefly describe some notations. Suppose there are V cameras in the scene and p_i detected bounding boxes in view i. For a pair of views (i, j), the affinity scores can be calculated between the two sets of bounding boxes in view i and view j. We use $\boldsymbol{A}_{ij} \in \mathbb{R}^{p_i \times p_j}$ to denote the affinity matrix, whose elements represent the affinity scores. The correspondences to be estimated between the two sets of bounding boxes are represented by a partial permutation matrix $\boldsymbol{P}_{ij} \in \{0, 1\}^{p_i \times p_j}$, which satisfies the doubly stochastic constraints:

$$\mathbf{0} \leq \boldsymbol{P}_{ij}\mathbf{1} \leq \mathbf{1}, \mathbf{0} \leq \boldsymbol{P}_{ij}^T\mathbf{1} \leq \mathbf{1}. \tag{1}$$

The problem is to take $\{\boldsymbol{A}_{ij}|\forall i, j\}$ as input and output the optimal $\{\boldsymbol{P}_{ij}|\forall i, j\}$ that maximizes the corresponding affinities and is also cycle-consistent across multiple views.

Affinity matrix: We propose to combine the appearance similarity and the geometric compatibility to calculate the affinity scores between bounding boxes.

First, we adopt a pre-trained person re-identification (re-ID) network to obtain a descriptor for a bounding box. The re-ID network trained on massive re-ID datasets is expected to be able to extract discriminative appearance features that are relatively invariant to illumination and viewpoint changes. Specifically, we feed the cropped image of each bounding box through the publicly available re-ID model proposed in [44] and extract the feature vector from the "pool5" layer as the descriptor for each bounding box. Then, we compute the Euclidean distance between the descriptors of a bounding box pair and map the distances to values in $(0, 1)$ using the sigmoid function as the appearance affinity score of this bounding box pair.

Besides appearances, another important cue to associate two bounding boxes is that their associated 2D poses should be geometrically consistent. Specifically, the corresponding 2D joint locations should satisfy the epipolar constraint, i.e. a joint in the first view should lie on the epipolar line associated with its correspondence in the second view. Suppose $\boldsymbol{x} \in \mathbb{R}^{N \times 2}$ denotes a 2D pose composed of N joints.

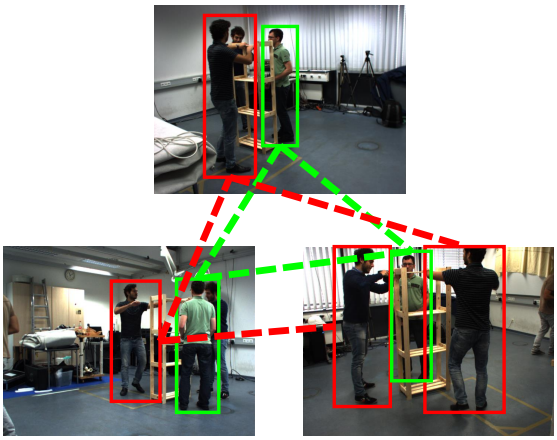

Figure 3: An illustration of cycle consistency. The green lines denote a set of consistent correspondences and the red lines show a set of inconsistent correspondences.

Then, the geometric consistency between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from two views can be measured by the following distance:

$$D_g(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2N}\sum_{n=1}^{N} d_g(\boldsymbol{x}_i^n, \boldsymbol{L}_{ij}(\boldsymbol{x}_j^n)) + d_g(\boldsymbol{x}_j^n, \boldsymbol{L}_{ji}(\boldsymbol{x}_i^n)),$$

where $\boldsymbol{x}_i^n$ denotes the 2D location of the n-th joint of pose i, $\boldsymbol{L}_{ij}(\boldsymbol{x}_j^n)$ the epipolar line associated with $\boldsymbol{x}_j^n$ from the other view, and $d_g(\cdot, \boldsymbol{l})$ the point-to-line distance for l. The distances D_g are also mapped to values in $(0, 1)$ using the sigmoid function as the final geometric affinity scores.

Based on the fact that a pair of correctly detected and matched 2D poses must satisfy the geometric constraint (D_g is small), we combine the two affinity matrices as follows:

$$\boldsymbol{A}_{ij}(\cdot) = \begin{cases} \sqrt{\boldsymbol{A}_{ij}^a(\cdot) \times \boldsymbol{A}_{ij}^g(\cdot)}, & \text{if } D_g \leq th, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\boldsymbol{A}_{ij}(\cdot)$, $\boldsymbol{A}_{ij}^a(\cdot)$, and $\boldsymbol{A}_{ij}^g(\cdot) \in [0, 1]$ denote values of the fused affinity matrix, appearance affinity matrix, and geometry affinity matrix of view pair (i, j), respectively. th denotes a threshold. Experimental results demonstrate that this simple combination of appearance and geometry is superior to merely using one of them.

Multi-way matching with cycle consistency: If there are only two views to match, one can simply maximize $\langle \boldsymbol{P}_{ij}, \boldsymbol{A}_{ij} \rangle$ and find the optimal matching by the Hungarian algorithm. But when there are multiple views, solving the matching problem separately for each pair of views ignores the cycle-consistency constraint and may lead to inconsistent results. Figure 3 shows an example, where the correspondences in red are inconsistent and the ones in green are cycle-consistent as they form a closed cycle.

We make use of the results in [16] to solve this problem. Suppose the correspondences among all $m = \sum_{i=1}^{V} p_i$ detected bounding boxes in all views are denoted by $\boldsymbol{P} \in \{0,1\}^{m \times m}$:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} & \cdots & \boldsymbol{P}_{1n} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} & \cdots & \boldsymbol{P}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{P}_{n1} & \cdots & \cdots & \boldsymbol{P}_{nn} \end{pmatrix}, \tag{3}$$

where $\boldsymbol{P}_{ii}$ should be identity. Then, it can be shown that the cycle consistency constraint is satisfied if and only if

$$\text{rank}(\boldsymbol{P}) \leq s, \ \boldsymbol{P} \succeq 0, \tag{4}$$

where s is the underlying number of people in the scene. The intuition is that, if the correspondences are cycle-consistent, $\boldsymbol{P}$ can be factorized as $\boldsymbol{Y}\boldsymbol{Y}^T$ where $\boldsymbol{Y} \in \mathbb{R}^{m \times s}$ denotes the correspondences between all 2D bounding boxes and 3D people.

As s is unknown in advance, we propose to minimize the following objective function to estimate the low-rank and positive semidefinite matrix $\boldsymbol{P}$:

$$\begin{aligned} f(\boldsymbol{P}) &= -\sum_{i=1}^{n}\sum_{j=1}^{n} \langle \boldsymbol{A}_{ij}, \boldsymbol{P}_{ij} \rangle + \lambda \cdot \text{rank}(\boldsymbol{P}), \\ &= -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \cdot \text{rank}(\boldsymbol{P}), \end{aligned} \tag{5}$$

where $\boldsymbol{A}$ is concatenation of all $\boldsymbol{A}_{ij}$ similar to the form in (3), λ denotes the weight of low-rank constraint.

The benefits of formulating the problem in this way are two-fold. First, the cycle consistency constraint aggregates the multi-way information to improve the matching and prune the false detections, which can hardly be realized if only two views are considered. Second, the rank minimization will automatically recover a rank (the number of people in the scene) that can best explain the observations.

Optimization: To make the optimization tractable, we have to make appropriate relaxations. Instead of minimizing the rank, which is a discrete operator, we minimize the nuclear norm $\|\boldsymbol{P}\|_*$, which is the tightest convex surrogate of rank [14]. We replace the integer constraint on $\boldsymbol{P}$ by saying that P is a real matrix with values in $[0, 1]$:

$$0 \leq \boldsymbol{P} \leq 1, \tag{6}$$

which is a common practice in matching algorithms. We remove the semidefinite constraint and only require $\boldsymbol{P}$ to be symmetric:

$$\boldsymbol{P}_{ij} = \boldsymbol{P}_{ji}^T, \ \ 1 \leq i, j \leq n, i \neq j, \tag{7}$$

$$\boldsymbol{P}_{ii} = \boldsymbol{I}_{p_i}, \quad 1 \leq i \leq n. \tag{8}$$

Finally, we solve the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{P}} \ & -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{P}\|_*, \\ \text{s.t.} \ \ & \boldsymbol{P} \in \mathcal{C}, \end{aligned} \tag{9}$$

where $\mathcal{C}$ denotes the set of matrices satisfying the constraints (1), (6), (7), and (8).

Note that the problem in (9) is convex and we use the alternating direction method of multipliers (ADMM) [5] to solve it. The problem is first rewritten as follows by introducing an auxiliary variable $\boldsymbol{Q}$:

$$\begin{aligned} \min_{\boldsymbol{P},\boldsymbol{Q}} \ & -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{Q}\|_*, \\ \text{s.t.} \ \ & \boldsymbol{P} = \boldsymbol{Q}, \ \boldsymbol{P} \in \mathcal{C}. \end{aligned} \tag{10}$$

Then, the augmented Lagrangian of (10) is:

$$\begin{aligned} \mathcal{L}_\rho(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{Y}) = & -\langle \boldsymbol{A}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{Q}\|_* + \langle \boldsymbol{Y}, \boldsymbol{P} - \boldsymbol{Q} \rangle \\ & + \frac{\rho}{2} \|\boldsymbol{P} - \boldsymbol{Q}\|_F^2, \end{aligned} \tag{11}$$

where $\boldsymbol{Y}$ denotes the dual variable and ρ denotes a penalty parameter. Each primal variable and the dual variable are alternately updated until convergence. The overall algorithm is shown in Algorithm 1, where $\mathcal{D}$ denotes the operator for singular value thresholding [7] and $\mathcal{P}_{\mathcal{C}}(\cdot)$ denotes the orthogonal projection to $\mathcal{C}$.

Algorithm 1: Consistent Multi-Way Matching

Input: Affinity matrix $\boldsymbol{A}$
Output: Consistent correspondences $\boldsymbol{P}$

```
randomly initialize P and Y = 0 ;
while not converged do
    Q ← D_{1/ρ}(1/ρ Y + P) ;
    P ← P_C(Q − 1/ρ (Y − A)) ;
    Y ← Y^k + ρ(P − Q) ;
end
quantize P with a threshold equal to 0.5.
```

The output $\boldsymbol{P}$ gives us the cycle-consistent correspondences of bounding boxes across all views. Figure 2 shows an example. The bounding boxes with no matches in other views are regarded as false detections and discarded.

3.3. 3D pose reconstruction

Given the estimated 2D poses of the same person in different views, we reconstruct the 3D pose. This can be simply done by triangulation, but the gross errors in 2D pose estimation may largely degrade the reconstruction. In order to fully integrate uncertainties in 2D pose estimation and incorporate the structural prior on human skeletons, we make

use of the 3DPS model and propose an approximate algorithm for efficient inference.

3D pictorial structure: We use a joint-based representation of 3D poses, i.e., $\boldsymbol{T} = \{\boldsymbol{t}_i | i = 1, ..., N\}$, where $\boldsymbol{t}_i \in \mathbb{R}^3$ denotes the location of joint i. Given 2D images from multiple views $I = \{I_v | v = 1, ..., V\}$, the posterior distribution of 3D poses can be written as:

$$p(\boldsymbol{T}|I) \propto \prod_{v=1}^{V} \prod_{i=1}^{N} p(I_v|\pi_v(\boldsymbol{t}_i)) \prod_{(i,j)\in\varepsilon} p(\boldsymbol{t}_i, \boldsymbol{t}_j), \quad (12)$$

where $\pi_v(\boldsymbol{t}_i)$ denotes the 2D projection of $\boldsymbol{t}_i$ in the v-th view and the likelihood $p(I_v|\pi_v(\boldsymbol{t}_i))$ is given by the 2D heat map output by the CNN-based 2D pose detector [10], which characterizes the 2D spatial distribution of each joint.

The prior term $p(\boldsymbol{t}_i, \boldsymbol{t}_j)$ denotes the structural dependency between joint $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$, which implicitly constrains the bone length between them. Here, we use a Gaussian distribution to model the prior on bone length:

$$p(\boldsymbol{t}_i, \boldsymbol{t}_j) \propto N(\|\boldsymbol{t}_i - \boldsymbol{t}_j\| | L_{ij}, \sigma_{ij}), \quad (13)$$

where $\|\boldsymbol{t}_i - \boldsymbol{t}_j\|$ denotes the Euclidean distance between joint $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$, L_{ij} and σ_{ij} denote the mean and standard deviation respectively, learned from the Human3.6M dataset [19].

Inference: The typical strategy to maximize $p(\boldsymbol{T}|I)$ is first discretizing the state space as a uniform 3D grid, and applying the max-product algorithm [6, 32]. However, the complexity of the max-product algorithm grows fast with the dimension of the state space.

Instead of using grid sampling, we set the state space for each 3D joint to be the 3D proposals triangulated from all pairs of corresponding 2D joints. As long as a joint is correctly detected in two views, its true 3D location is included in the proposals. In this way, the state space is largely reduced, resulting in much faster inference without sacrificing the accuracy.

4. Empirical evaluation

We evaluate the proposed approach on three public datasets including both indoor and outdoor scenes and compare it with previous works as well as several variants of the proposed approach.

4.1. Datasets

The following three datasets are used for evaluation:

Campus [1]: It is a dataset consisting of three people interacting with each other in an outdoor environment, captured with three calibrated cameras. We follow the same evaluation protocol as in previous works [1, 3, 2, 12] and use the percentage of correctly estimated parts (PCP) to measure the accuracy of 3D location of the body parts.

Shelf [1]: Compared with Campus, this dataset is more complex, which consists of four people disassembling a shelf at a close range. There are five calibrated cameras around them, but each view suffers from heavy occlusion. The evaluation protocol follows the prior work and the evaluation metric is also 3D PCP.

CMU Panoptic [20]: This dataset is captured in a studio with hundreds of cameras, which contains multiple people engaging in social activities. For the lack of ground truth, we qualitatively evaluate our approach on the CMU Panoptic dataset.

4.2. Ablation analysis

We first give an ablation analysis to justify the algorithm design in the proposed approach. The Campus and Shelf datasets are used for evaluation.

Appearance or geometry? As described in section 3.2, our approach combines appearance and geometry information to construct the affinity matrix. Here, we compare it with the alternatives using appearance or geometry alone. The detailed results are presented in Table 1.

On the Campus, using appearance only achieves competitive results, since the appearance difference between actors is large. The result of using geometry only is worse because the cameras are far from the people, which degrades the discrimination ability of the epipolar constraint. On the Shelf, the performance of using appearance alone drops a lot. Especially, the result of actor 2 is erroneous, since his appearance is similar to another person. In this case, the combination of appearance and geometry greatly improve the performance.

Direct triangulation or 3DPS? Given the matched 2D poses in all views, we use a 3DPS model to infer the final 3D poses, which is able to integrate the structural prior on human skeletons. A simple alternative is to reconstruct 3D pose by triangulation, i.e., finding the 3D pose that has the minimum reprojection errors in all views. The result of this baseline method ('NO 3DPS') is presented in Table 1.

The result shows that when the number of cameras in the scene is relatively small, for example, in the Campus dataset (three cameras), using 3DPS can greatly improve the performance. When a person is often occluded in many views, for example, actor 2 in the Shelf dataset, the 3DPS model can also be helpful.

Matching or no matching? Our approach first matches 2D poses across views and then applies the 3DPS model to each cluster of matched 2D poses. An alternative approach

Campus	Actor 1	Actor 2	Actor 3	Average
Ours	**97.6**	**93.3**	**98.0**	**96.3**
Appearance	**97.6**	**93.3**	96.5	95.8
Geometry	97.4	90.1	89.4	92.3
No 3DPS	90.6	89.2	97.7	92.5
No matching	84.8	89.0	71.5	81.8
Shelf	Actor 1	Actor 2	Actor 3	Average
Ours	**98.8**	**94.1**	**97.8**	**96.9**
Appearance	98.6	60.5	94.3	84.5
Geometry	97.2	79.5	96.5	91.1
No 3DPS	97.9	89.5	**97.8**	95.1
No matching	98.1	91.1	92.8	94.0

Table 1: Ablative study on the Campus and Shelf datasets. Appearance and geometry denote the different types of affinity matrices, i.e., using appearance only and using geometry only. 'No 3DPS' uses triangulation instead of the 3DPS model to reconstruct 3D poses. 'No matching' represents the 3DPS model without bounding box matching, an approach typically used in previous methods [2, 21]. We re-implement this approach with the state-of-the-art 2D pose detector. The numbers are the percentage of correctly estimated parts (PCP).

Campus	Actor 1	Actor 2	Actor 3	Average
Belagiannis *et al*. [1]	82.0	72.4	73.7	75.8
Belagiannis *et al*. [3]	83.0	73.0	78.0	78.0
Belagiannis *et al*. [2]	93.5	75.7	84.4	84.5
Ershadi-Nasab *et al*. [12]	94.2	92.9	84.6	90.6
Ours w/o 3DPS	90.6	89.2	97.7	92.5
Ours	**97.6**	**93.3**	**98.0**	**96.3**
Shelf	Actor 1	Actor 2	Actor 3	Average
Belagiannis *et al*. [1]	66.1	65.0	83.2	71.4
Belagiannis *et al*. [3]	75.0	67.0	86.0	76.0
Belagiannis *et al*. [2]	75.3	69.7	87.6	77.5
Ershadi-Nasab *et al*. [12]	93.3	75.9	94.8	88.0
Ours w/o 3DPS	97.9	89.5	**97.8**	95.1
Ours	**98.8**	**94.1**	**97.8**	**96.9**

Table 2: Quantitative comparison on the Campus and Shelf datasets. The numbers are percentage of correctly estimated parts (PCP). The results of other methods are taken from respective papers. 'Ours w/o 3DPS' means using triangulation instead of the 3DPS model to reconstruct 3D poses from matched 2D poses.

in most previous works [2, 21] is to directly apply the 3DPS model to infer multiple 3D poses from all detected 2D poses without matching. Here, we give a comparison between them. As Belagiannis *et al*. [2] did not use the most recent CNN-based keypoint detectors and Joo *et al*. [21] did not report results on public benchmarks, we re-implement their approach with the state-of-the-art 2D pose detector [8] for a fair comparison. The implementation details are given in the supplementary materials. Table 1 shows that the 3DPS without matching obtained decent results on the Self dataset but performed much worse on the Campus dataset, where there are only three cameras. The main reason is that the 3DPS model implicitly uses multi-view geometry to link the 2D detections across views but ignores the appearance cues. When using a sparse set of camera views, the multi-view geometric consistency alone is sometimes insufficient to differentiate the correct and false correspondences, which leads to false 3D pose estimation. This observation coincides with the other results in Table 1 as well as the observation in [21]. The proposed approach explicitly leverage the appearance cues to find cross-view correspondences, leading to more robust results. Moreover, the matching step significantly reduces the size of state space and makes the 3DPS model inference much faster.

4.3. Comparison with state-of-the-art

We compare with the following baseline methods. Belagiannis *et al*. [1, 3] were among the first to introduce 3DPS model-based multi-person pose estimation and their method was extended to the video case to leverage temporal consistency [2]. Ershadi-Nasab *et al*. [12] is a very recent method that proposed to cluster the 3D candidate joints to reduce the state space.

The results on the Campus and Shelf datasets are presented in Table 2. Note that the 2D pose detector [10] and the reID network [44] used in our approach are the released pre-trianed models without any fine-tuning on the evaluated datasets. Even with the generic models, our approach outperforms the state-of-the-art methods by a large margin. In particular, our approach significantly improves the performance on Actor 3 in the Campus dataset and Actor 2 in the Shelf dataset, which suffer from severe occlusion. We also include our results without the 3DPS model but using triangulation to reconstruct 3D poses from matched 2D poses. Due to the robust and consistent matching, direct triangulation also obtains better performance than previous methods.

4.4. Qualitative evaluation

Figure 4 shows some representative results of the proposed approach on the Shelf and CMU Panoptic dataset. Taking inaccurate 2D detections as input, our approach is able to establish their correspondences across views, identify the number of people in the scene automatically, and finally reconstruct their 3D poses. The final 2D pose estimates obtained by projecting the 3D poses back to 2D views are also much more accurate than the original detections.

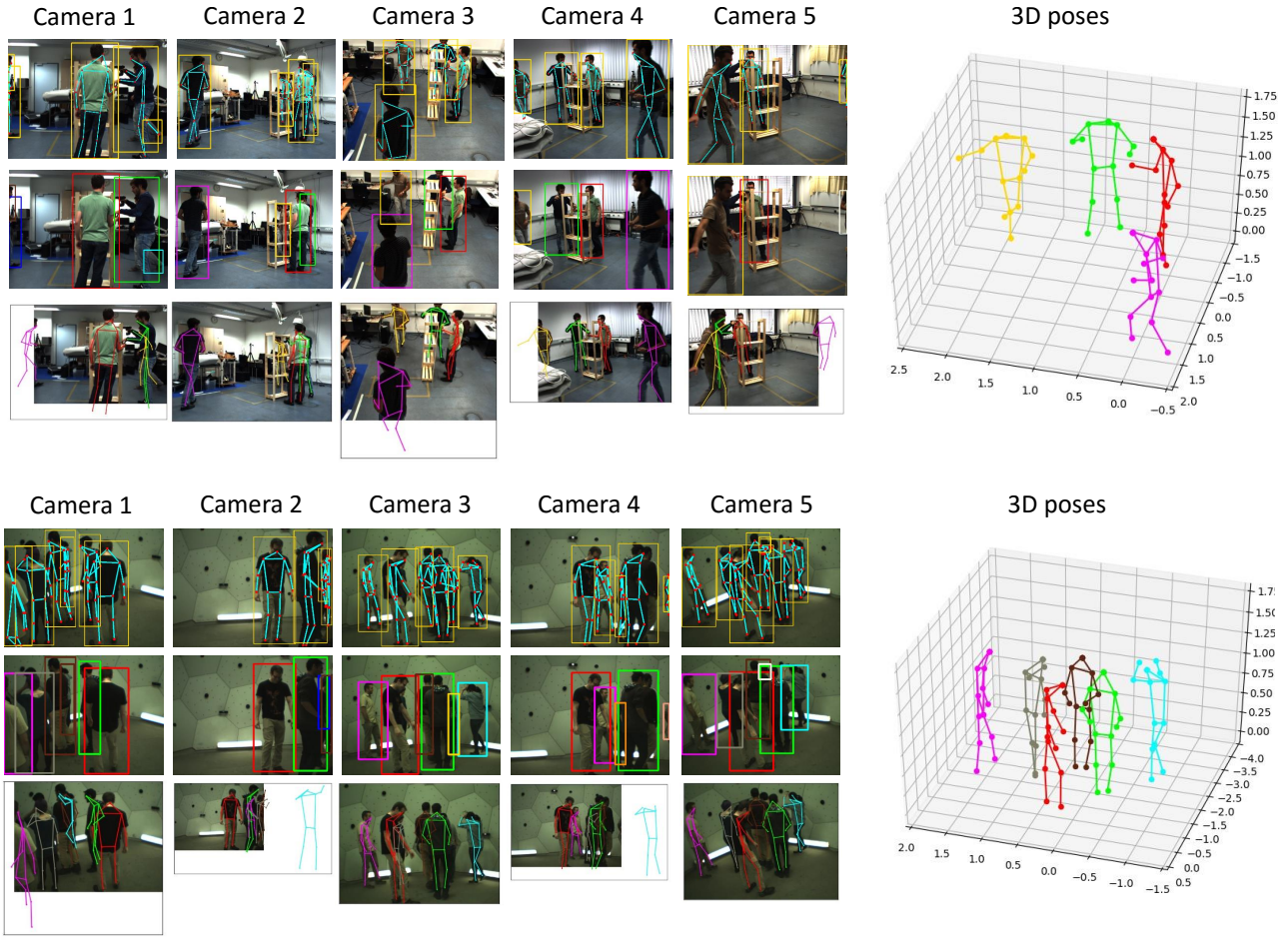

Figure 4: Qualitative results on the Shelf (top) and CMU panoptic (bottom) datasets. The first row shows the 2D bounding boxes and pose detections. The second row shows the result of our matching algorithm where the colors indicate the correspondences of bounding boxes across views. The third row shows the 2D projections of the estimated 3D poses.

4.5. Running time

We report running time of our algorithm on the sequences with four people and five views in the Shelf dataset, tested on a desktop with an Intel i7 3.60 GHz CPU and a GeForce 1080Ti GPU. Our unoptimized implementation on average takes 25 ms for running reID and constructing affinity matrices, 20 ms for the multi-way matching algorithm, and 60 ms for 3D pose inference. Moreover, the results in Table 2 show that our approach without the 3DPS model also obtains very competitive performance, which is able to achieve real-time performance at $>$ 20fps.

5. Summary

In this paper, we propose a novel approach to multi-view 3D pose estimation that can fastly and robustly recover 3D poses of a crowd of people with a few cameras. Compared with the previous 3DPS based methods, our key idea is to use a multi-way matching algorithm to cluster the detected 2D poses to reduce the state space of the 3DPS model and thus improves both efficiency and robustness. We also demonstrate that the 3D poses can be reliably reconstructed from clustered 2D poses by triangulation even without using the 3DPS model. This shows the effectiveness of the proposed multi-way matching algorithm, which leverages the combination of geometric and appearance cues as well as the cycle-consistency constraint for matching 2D poses across multiple views.

Acknowledgements: The authors from Zhejiang University would like to acknowledge support from NSFC (No.61806176), Fundamental Research Funds for the Central Universities and ZJU-SenseTime Joint Lab of 3D Vision. Qixing Huang would like to acknowledge support from NSF DMS-1700234, NSF CIP-1729486, NSF IIS-1618648, a gift from Snap Research and a GPU donation from Nvidia Inc.

References

[1] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 1, 2, 6, 7

[2] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *T-PAMI*, 38(10):1929–1942, 2016. 1, 2, 6, 7

[3] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, and N. Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. In *ECCV workshop*, 2014. 6, 7

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 3

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 5

[6] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 2, 6

[7] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. 5

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 3, 7

[9] C.-H. Chen and D. Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017. 3

[10] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. *CVPR*, 2018. 3, 6, 7

[11] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015. 2

[12] S. Ershadi-Nasab, E. Noury, S. Kasaei, and E. Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018. 1, 2, 6, 7

[13] H. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[14] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002. 5

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *T-PAMI*, 2018. 3

[16] Q.-X. Huang and L. Guibas. Consistent shape maps via semidefinite programming. In *Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing*, pages 177–186. Eurographics Association, 2013. 3, 5

[17] S. Huang, M. Gong, and D. Tao. A coarse-fine network for keypoint localization. In *ICCV*, 2017. 3

[18] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: pages, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 3

[19] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *T-PAMI*, 36(7):1325–1339, 2014. 6

[20] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2, 6

[21] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *T-PAMI*, 2017. 1, 2, 7

[22] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[23] A. Kadkhodamohammadi and N. Padoy. A generalizable approach for multi-view 3d human pose regression. *CoRR*, abs/1804.10462, 2018. 2

[24] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[25] M. Kocabas, S. Karagoz, and E. Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018. 3

[26] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[27] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 3

[28] F. Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 3

[29] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017. 3

[30] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 3

[31] G. Pavlakos, X. Zhou, and K. Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, 2018. 3

[32] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 2017. 2, 6

[33] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018. 3

[34] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 3

[35] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 3

[36] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15–48, 2012. 2

[37] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *ICCV*, 2017. 3

[38] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010. 2

[39] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, 2017. 3

[40] D. Tome, C. Russell, and L. Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR*, 2017. 3

[41] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 3

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 3

[43] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *NIPS*, 2011. 2

[44] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 3, 4, 7

[45] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 3

[46] X. Zhou, M. Zhu, and K. Daniilidis. Multi-image matching via fast alternating minimization. In *ICCV*, 2015. 3

[47] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016. 3