Multi-View Multi-Person 3D Pose Estimation with Uncalibrated Camera Networks

Yan Xu
yxu2@andrew.cmu.edu
Kris Kitani
kmkitani@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Existing efforts in multi-view multi-person 3D human pose estimation often rely on 6 DoF camera poses to obtain cross-view body joint matches for solving 3D poses. Some other efforts use networks specifically trained for each dataset to regress 3D human poses. These methods do not generalize well to scenarios in the wild since they require calibrated camera poses or large amounts of training data. We present an approach that requires none of them. Our key insight is to combine (1) the well-developed 2D human detection and description networks that can be pre-trained on open datasets with (2) multi-view geometry and optimization algorithms that generalize to arbitrary settings. Using 2D human appearance embedding as the input, we solve cross-view human matching as an optimization problem with the numbers of cameras and people and the fact that one person cannot be matched to another person in the same view as the constraints. With the cross-view matches, we estimate the camera poses and 3D human poses simultaneously using multi-view geometry and bundle adjustment optimization. On open datasets, our approach reaches smaller pose estimation error than previous works with fewer requirements of camera pose and model training. We also evaluate our approach with three wild datasets with various settings, including indoor and outdoor environments, static and dynamic cameras, etc. It shows excellent generalization ability across different settings. The code is made for public at: https://github.com/yan293/UncalibratedMVMP3DPose.

1 Introduction

Multi-view multi-person (MVMP) 3D human pose estimation is a fundamental problem underneath many tasks, such as Augmented Reality, Virtual Reality, and social activity analysis. Existing efforts generally focus on controlled environments where the 6-DoF camera poses are well calibrated. Despite the progress on numbered public datasets [2, 26], solving the task in a more real-life-like environment without calibrated camera poses has received less attention. This work targets the setting where camera poses are unknown, and dataset-specific model training is not allowed, aiming to attempt toward in-the-wild scenarios.

Under our setting, previous MVMP 3D pose estimation efforts can no longer apply due to their limitations. Previous efforts generally fall into two categories. The first category of methods [48, 55, 63] directly regress 3D human poses from multi-view images/videos

Figure 1: 3D human pose estimation using our approach with two static cameras and one camera mounted on a flying drone. Given multi-view raw images, our approach estimates human pose without requiring camera poses and model training.

using pre-trained neural networks. However, these methods require the same camera setting and scene for training and inference. In other words, they must collect data and train a new network once the camera setting or the scene changes. The second category of methods solve the task in a multi-stage manner [9, 13, 15, 58]. These methods often first obtain cross-view correspondences using appearance feature and geometric constraint [13] and then estimate the 3D human pose by optimizing a 3DPS model [2]. However, both the geometric constraint and the 3DPS model require known camera poses.

This work presents an approach that does not require camera poses for solving cross-view matching. Our insight is to use the number of cameras, the number of people, and the fact that one person cannot be matched to another person in the same view as prior knowledge to constrain the possible solution space of cross-view matching to a small solvable region. Specifically, consider an MVMP system with N cameras and K people. We first detect the people from all images [11] and embed them using a pre-trained person re-identification network [37]. Next, we match these appearance embeddings to K people. The matching must follow three rules: (1) Each person must be observed by at least two cameras to be rid of depth ambiguity, (2) each person must have no more than N matches, and (3) people from the same view must not match. Using these rules, we formulate the cross-view matching problem as constrained optimization and narrow the solution space to a small feasible region from which the correct solution can be easily reached. Sec. 3.1 will expands more details.

Once matched people across views, we can obtain 2D point correspondences by associating their joints and then solve the relative 6-DoF camera poses. However, the solved camera poses can be incorrect due to the low quality of the point correspondences. For example, the key points from some camera views can gather in a small region in the image, leading to local minima of the estimated camera poses. To address this, we introduce a "Camera Pose Self-Validation" process. Specifically, we let camera poses solved from all camera pairs validate each other and use the one that aligns the best with others as the final solution. This way, we can leverage those high-quality correspondences from all camera pairs to help obtain robust camera pose estimates. Sec. 3.2 will detail on this process. With the camera poses, we then solve the 3D human poses, aggregate the solutions from all camera pairs using the fact that the length of a bone keeps fixed in 3D space, and perform a further pose optimization through Bundle Adjustment [54]. Fig. 2 presents an overview of our approach.

We evaluate our approach on three public datasets [2, 26] and three wild datasets. On the public datasets, our approach outperforms SOTAs with fewer requirements of camera poses and data-specific model training. On the wild datasets, our approach shows good generalization ability across various settings, such as indoor and outdoor environments, static and moving cameras, small and large field-of-views, etc. Fig. 1 shows an example under the moving camera setting. Our main contributions are threefold: (1) We present an approach

for MVMP 3D human pose estimation without requiring camera poses and dataset-specific model training; (2) We introduce a constrained optimization formulation for cross-view human matching when epipolar constraints are inapplicable; (3) We proposed a "Camera Pose Self-Validation" process to deal with the low-quality correspondences problem.

2 Related work

Multi-Person 3D Human Pose Estimation We discuss both SVMP (single-view multi-person) and MVMP settings. Existing works in the SVMP setting generally include two categories: Top-down approaches and bottom-up approaches. Top-down approaches [5, 41, 50, 51, 61] detect 2D people, then perform 2D-to-3D lifting [9, 32, 38] or direct regression [33, 44, 45] to obtain 3D poses. Bottom-up approaches [16, 39, 40, 43, 62] first estimate 3D locations for all the joints, then associate the joints to each person using depth information. Since the problem is ill-posed, both approaches usually require large amounts of data to constrain the search domain in a small space. Our approach follows the top-down strategy but targets the multi-view setting and does not require dataset-specific training. Works in the MVMP setting also typically fall into two categories: multi-stage approaches and single-stage approaches. Multi-stage approaches [3, 4, 13, 15] first obtain the 2D poses [8, 12, 34, 52], then match the 2D poses using appearance features [37, 60] and geometry cues [13]. Finally, they solve the 3D pose using multi-view geometry [7, 21] Single-stage approaches [48, 55, 63] solve the problem through end-to-end regression. They usually divide the scene into 3D voxels and localize each person regressed from multi-view data. They then perform a fine-grained regression to obtain joint locations. Some efforts [48, 58] utilize Graph Convolutional Neural Networks [28] and Transformers [56] to improve performance. Multi-stage methods rely on camera poses for cross-view matching. Single-stage approaches need to train a new regressor for each scene. Our approach requires neither.

Pose Estimation in the Wild Because of the difficulty of the task and lack of data, existing efforts mainly focus on the single-person setting. Works commonly use multi-view geometry [10, 14, 25, 29, 49, 57] as supervision for solving the task. They estimate the 3D human pose and relative camera poses simultaneously [10, 14, 49, 57], then back-project the 3D pose to each view and minimize the re-projection errors to update the estimation. They usually limit the feasible camera poses through random sampling [10, 14] or enumeration [49] and restrict the human pose using prior knowledge learned from other datasets [24, 25]. Instead of focusing on the single-person setting, we make an attempt toward the multi-person setting in this work. To clarify, we assume the intrinsic and distortion parameters provided, only the camera poses unknown.

Pose Estimation and Cross-View Matching Using Human Since our method uses human information to estimate camera poses as an intermediate step, we briefly cover related works that use human information for camera pose estimation, which have been well-studied over decades. Some methods [27, 30, 36] use human height distribution learned from large amounts of data as prior knowledge for camera pose estimation. Other methods use head and foot locations [18, 23, 53] or human trajectories [1, 35, 47, 59, 60] as geometric constraints for pose estimation, cross-view matching, and tracking. Similar to these methods, our approach also uses humans in the scene to extract useful geometric information. The difference is that our approach uses 2D human pose detection and associates corresponding human body parts across camera views to obtain 2D-to-2D points matches.

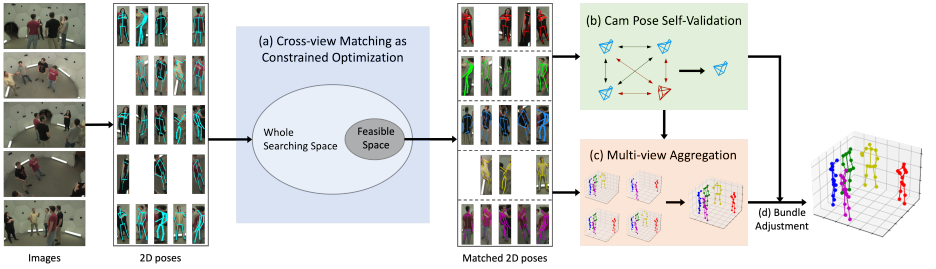


Figure 2: Overview of our approach. Given multi-view images, (a) we first detect 2D poses and solve cross-view human matching as constrained optimization. Then, (b) we estimate camera poses and perform self validation. Finally, (c) we solve 3D human poses, aggregate multi-view information and further optimize through (d) Bundle Adjustment.

3 Method

Fig. 2 presents an overview of our approach. It includes multiple stages. Sec. 3.1 introduces the process of solving cross-view human matching as a constrained optimization problem. Sec. 3.2 details camera pose estimation and self-validation. Sec. 3.3 expands on 3D human pose estimation, multi-view information aggregation, and bundle adjustment optimization.

3.1 Cross-View Matching As Constrained Optimization

Given multi-view images, we detect 2D human poses [11] and represent each person with its re-identification (re-ID) feature [37], extracted with a network pre-trained on open datasets [64]. The feature is $L2$-normalized with a dimension of 1×2048.

Consider a MVMP system with N cameras and K people. We refer to a camera by its id i and a person by its id k, where $1 \leq i \leq N$ and $1 \leq k \leq K$. Since one camera may not view all the people, so we use M_i, $1 \leq M_i \leq K$, to represent the number of people observed by camera i. We denote the re-ID feature of the j-th person from camera i as $x_{i,j} \in \mathbb{R}^{2048}$, where $1 \leq i \leq N$, $1 \leq j \leq M_i$. The features of all people from all camera views are, $\mathcal{D} = \{x_{i,j}\}_{i=1,j=1}^{N,M_i}$. Our goal is to match the 2D features of the same person from multiple camera views. The goal can be achieved by finding K cluster centers, $\{C^1, \cdots, C^K\}$, such that the sum of the $L2$ distance between each data point $x_{i,j}$ and its nearest cluster center C^k is minimized. Formally,

$$\min_{C,W} \quad \sum_{i=1}^{N}\sum_{j=1}^{M_i}\sum_{k=1}^{K} W_{i,j}^k \cdot (\frac{1}{2}\|x_{i,j} - C^k\|_2^2) \tag{1}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} W_{i,j}^k = 1, i \in \{1,\cdots,N\}, j \in \{1,\cdots,M_i\}$$

$$W_{i,j}^k \geq 0, i \in \{1,\cdots,N\}, j \in \{1,\cdots,M_i\}, k \in \{1,\cdots,K\}$$

where, $W_{i,j}^k = 1$ if cluster center C^k is closest to data point $x_{i,j}$, otherwise, $W_{i,j}^k = 0$. The solved $C^k \in \mathbb{R}^{2048}$ will be a general representation of person k such that it is the closest cluster center to the 2D features of person k from multiple camera views.

However, the correct solution of Eq. 1 difficult to reach because of the ample solution space. According to Bradley *et al.* [6], when feature dimension $d \geq 10$, local optima are

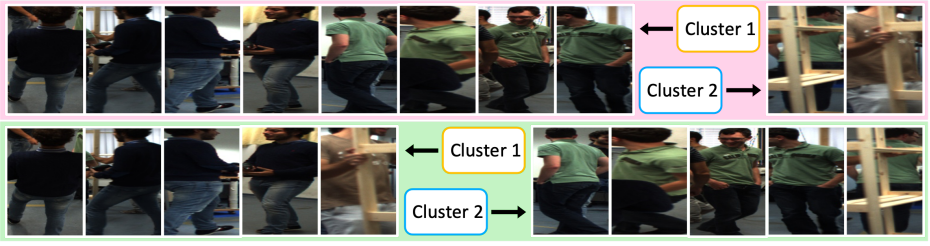


Figure 3: Results of unconstrained matching (up) and our approach (bottom). The unconstrained matching gets stuck at a local optimum. Our approach reaches the correct solution.

likely to happen. Fig. 3 (a) shows an example of local minima. To restrict the solution space to a small feasible region, we found that three implicit rules can be leveraged.

Rule 1: Each person is viewed by at least two cameras. *Rule 1* is straightforward since a person needs to be viewed by at least two cameras in order to be reconstructed without depth ambiguities. Using the above definitions, we can formally define *Rule 1* as

$$\sum_{i=1}^{N}\sum_{j=1}^{M_i} W_{i,j}^{k} \geq 2, k \in \{1,\cdots,K\} \tag{2}$$

Rule 2: The number of matches for a person is less than the number of cameras. Since each person can be viewed from at most N cameras, the expected size of any cluster should be smaller than N, as shown in Fig. 4 (b). Formally, we can define *Rule 2* as

$$\sum_{i=1}^{N}\sum_{j=1}^{M_i} W_{i,j}^{k} \leq N, k \in \{1,\cdots,K\} \tag{3}$$

Rule 3: Observations from the same view should not be matched. *Rule 3* means that the members of a cluster should come from different camera views. Violating *Rule 3* means two people in the same image belong to the same cluster. The formulation of *Rule 3* is less straightforward. Fig. 4 (c) and (d) present visual explanations. As the figure shows, for camera i, the sum of $W_{i,j}^{k}$ over people should be no greater than 1 for any C^k. Formally,

$$\sum_{j=1}^{M_i} W_{i,j}^{k} \leq 1, i \in \{1,\cdots,N\}, k \in \{1,\cdots,K\} \tag{4}$$

Fig. 4 presents the graph representations of cross-view matching and the above rules. Combining Eq. 1 2, 3, 4, we have the final constrained optimization formulation for the cross-view matching problem. This constrained optimization problem can be solved following a standard E-M process [42]. Specifically, at step t, with the cluster centers as $C_t^1, C_t^2, \cdots, C_t^K$, assigning each data point equals to a linear optimization problem that can be solved with fast network simplex algorithms [6]. Then, at step $t+1$, C_k^1 can be updated with

$$C_{t+1}^{k} = \begin{cases} \frac{\sum_{i=1}^{N}\sum_{j=1}^{M_i} W_{i,j}^{k,t} \cdot x_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{M_i} W_{i,j}^{k,t}} & \text{if } \sum_{i=1}^{N}\sum_{j=1}^{M_i} W_{i,j}^{k,t} > 0 \\ C_t^k & \text{otherwise} \end{cases} \tag{5}$$

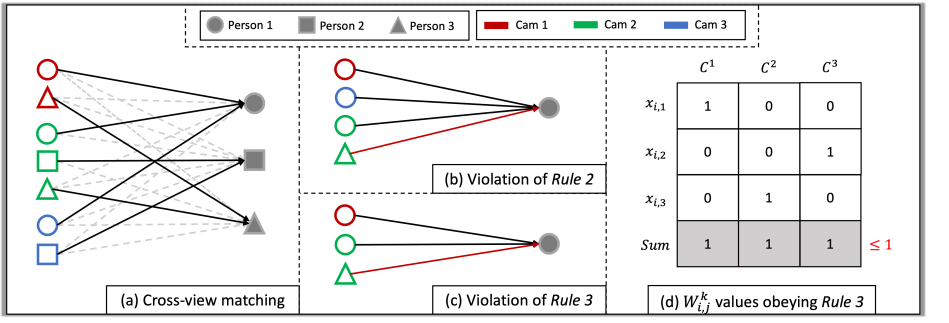


Figure 4: Graph representation of constrained cross-view matching: Solid arrows in (a) are correct matches, dotted arrows are incorrect but possible matches. Red arrows in (b) and (c) are two examples violating *Rule 2* and *Rule 3*. The table in (d) visually explains Eq. 4.

where $W_{i,j}^{k,t}$ is the "assignment" variable at time t. Fig. 3 presents the result of the cross-view 2D matching under occlusion using Eq. 1 and our approach. Using the rules as constraints, our approach successfully get rid of local minima and reaches the correct solution.

3.2 Camera Pose Estimation and Self-Validation

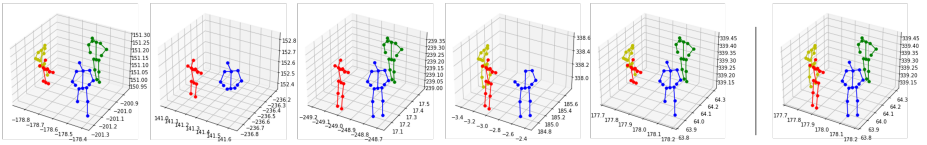
Camera Pose Estimation Once the cross-view human matching is solved, we associate the joints of the matched people from different camera views to get 2D point correspondences. We then can solve the relative pose between camera pairs using these point correspondences. Formally, let the set of point correspondences of camera pair (a,b) be (p_a, p_b). We first solve the essential matrix E_{ab} inside a RANSAC [7] loop. Next, we decompose E_{ab} into a relative rotation matrix R_{ab} and an up-to-scale relative translation t_{ab}. We repeat this process for all camera pairs to get the 3D layout of the camera network.

Camera Pose Self-Validation Simply solving the relative poses between all camera pairs is not enough. The point correspondences between some camera pairs can be "low quality". For example, they can be less in number, or the points gather in a small region in the image. In this case, the solved camera pose easily gets stuck to non-optimal local minima. We use a "Camera Pose Self-Validation" process to address this.

Define the relative pose between camera pair (a,b) as M_{ab}, which could be a local optimum. To solve this, we first introduce another camera c and solve the relative pose M_{ac} and M_{cb}. Next, we indirectly obtain the relative pose of (a,b) using M_{ac} and M_{cb} and denote it as M_{acb}. M_{acb} can be a better estimation than M_{ab} since (a,c) and (c,b) have higher-quality point correspondences than (a,b). In this way, we can obtain a set of direct and indirect pose estimations of (a,b): $\mathcal{M} = \{M_{ab}, M_{acb}, \cdots\}$. The idea is to let these estimations validate each other. Specifically, we use $\hat{M} \in \mathcal{M}$, that has the smallest average alignment angle with other elements in $\mathcal{M}$, as the final estimation of the relative pose between (a,b). We observe that this self-validation process greatly improves the robustness of camera pose estimation.

3.3 Multi-View Aggregation and Bundle Adjustment

Multi-View 3D Pose Aggregation After solving relative camera poses, we can triangulate the 3D human poses from any camera pair. However, not all people are visible to all cameras



(a) 3D poses estimation from different camera pairs (b) Aggregation result

Figure 5: Illustration of multi-view aggregation. (a) shows 3D human poses estimated from different camera pairs, (b) shows the pose estimation result after multi-view aggregation.

due to occlusion. We need to aggregate information from all camera views to estimate the full-body pose of every person in the scene. We follow a two-step multi-view information aggregation process to do that. First, we convert the 3D poses estimated from all camera pairs to one common camera coordinate system using the relative camera poses. Second, we use the fact that the length of a bone is fixed in the 3D space to solve the scale ambiguities and merge the human poses from different camera pairs. Fig. 5 presents the result of multi-view aggregation. The 3D human poses of all people are correctly estimated under occlusion.

Bundle Adjustment The last step of our approach is Bundle Adjustment [54]. The multi-view aggregation step simply puts 3D poses solved from different camera pairs together. However, this may cause some poses to look unnatural or the relative positions between different people to be incorrect. We thus use Bundle Adjustment to further fine-tune the 3D human poses. Let the intrinsic, extrinsic, and distortion parameters of camera i be K_i, M_i, D_i, the 2D pose of person k in camera i be $p_{i,k}$, and the 3D pose of person k be P_k. Assume that the intrinsic and distortion parameters known for all cameras, we aim to minimize the L-2 distance between the 2D poses and 3D re-projection for all camera views:

$$\min_{M,P} \quad \sum_{i=1}^{N}\sum_{k=1}^{K} A_{i,k} \cdot \frac{1}{2}\|p_{i,j} - \pi(M_i, P_k; K_i, D_i)\|_2^2 \tag{6}$$

where, $\pi(\cdot)$ is the perspective projection function, $A_{i,k} = 1$ if person k is visible from camera i, otherwise, $A_{i,k} = 0$. Note that a camera may only view some joints of a person, so the visibility of each joint needs to be considered independently. We use Eq. 6 for the ease of understanding. We solve Eq. 6 using Levenberg–Marquardt algorithm [31], which gives globally optimized 3D human pose and 6 DoF relative camera pose.

4 Experiment

We evaluate our approach on three open datasets and compare it with previous works. Beyond this, we also evaluate our approach on three self-collected "wild" datasets.

4.1 Open Datasets

Campus dataset [2] captures three people walking and interacting with each other in an outdoor environment from three calibrated cameras. We follow the same evaluation protocol as previous works [2, 13, 15, 46] and use the Percentage of Correct Parts (PCP) as the metric. *Note that* the estimated 3D human poses using our approach are with respect to one of the

cameras. To compare with other works, we use the ground truth camera pose provided by the dataset to convert our results to the same world coordinate system defined by the dataset.

Shelf Dataset [2] captures four people disassembling a shelf from five calibrated cameras. We also use PCP as the evaluation metric for the Shelf dataset.

CMU Panoptic Dataset [26] captures people doing various activities in an indoor studio. Following previous works [13, 46], we qualitatively evaluate our approach on this dataset.

Table 1: Comparison with other methods on the Campus and Shelf datasets. The reported numbers are PCP values. Results of other methods are taken from according papers.

Campus	CamPose	Training	Actor 1	Actor 2	Actor 3	Average
Huang *et al.* [22]	✓	✓	98.0	94.8	97.4	96.7
Tu *et al.* [55]	✓	✓	97.6	93.8	98.8	96.7
Zhang *et al.* [63]	✓	✓	98.2	94.1	97.4	96.6
Reddy *et al.* [48]	✓	✓	97.9	**95.2**	99.1	97.4
Belagiannis *et al.* [4]	✓	-	93.5	75.7	84.4	84.5
Ershadi *et al.* [15]	✓	-	94.2	92.9	84.6	90.6
Dong *et al.* [13]	✓	-	97.6	93.3	98.0	96.3
Perez-Yus *et al.* [46]	✓	-	98.4	93.4	98.3	96.7
Ours	-	-	**99.0**	94.7	**99.6**	**97.8**
Shelf	**CamPose**	**Training**	**Actor 1**	**Actor 2**	**Actor 3**	**Average**
Huang *et al.* [22]	✓	✓	98.8	96.2	97.2	97.4
Tu *et al.* [55]	✓	✓	99.3	94.1	97.6	97.0
Zhang *et al.* [63]	✓	✓	99.3	95.1	97.8	97.4
Reddy *et al.* [48]	✓	✓	99.1	96.3	98.3	98.2
Wu *et al.* [58]	✓	✓	99.3	96.5	97.3	97.7
Belagiannis *et al.* [4]	✓	-	75.3	69.7	87.6	77.5
Ershadi *et al.* [15]	✓	-	93.3	75.9	94.8	88.0
Dong *et al.* [13]	✓	-	98.8	94.1	97.8	96.9
Perez-Yus *et al.* [46]	✓	-	98.9	92.3	97.8	96.5
Ours	-	-	**99.6**	**95.2**	**98.5**	**97.8**

Table 2: Variants of our approach on the Campus and Shelf datasets. *Oracle* knows the GT camera poses and evaluates cross-view matching. *One-Step* simultaneously estimates camera and human poses using single-frame images. *Two-Step* estimates camera poses first.

Variants	Campus				Shelf			
	Actor 1	Actor 2	Actor 3	Average	Actor 1	Actor 2	Actor 3	Average
Oracle	99.0	96.7	99.6	98.4	100.0	100.0	99.6	99.9
One-Step	98.8	64.1	79.5	80.8	**99.6**	**95.2**	**98.5**	**97.8**
Two-Step	**99.0**	**94.7**	**99.6**	**97.8**	99.6	95.2	98.5	97.8

4.2 Comparison with state-of-the-art

Following previous multi-stage works [4, 13, 15, 46], we quantitatively evaluate our approach on the Shelf and Campus datasets. Existing multi-stage works generally use the 3DPS model [2] for 3D human pose estimation, which implicitly requires camera poses for

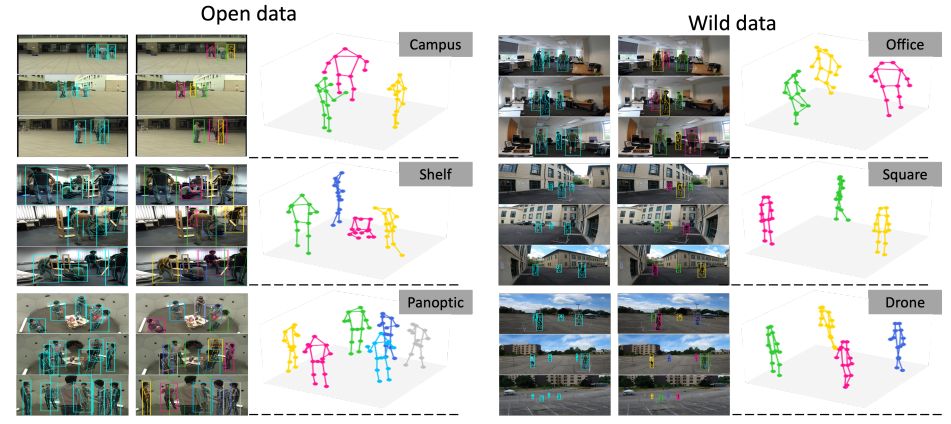


Figure 6: Qualitative evaluation of our approach on open and wild datasets. For each dataset, the left and middle columns are the 2D poses before and after cross-view matching, the right column shows the estimated 3D human poses.

geometric constraints. Compared with them, our approach does not assume known camera poses. Tab.1 presents the evaluation results of our approach and other method. We also include four deep learning-based methods [22, 48, 55, 63] in Tab.1 for a more comprehensive comparison. The result shows that our approach reaches state-of-the-art performance. We did not use camera poses or model training during our experiments. The result also reflects the effectiveness of our cross-view human matching method.

4.3 Evaluation on wild data

We follow previous works [4, 13, 15, 46] and qualitatively evaluate our approach on the Panoptic dataset. In addition, we also collected three wild datasets to evaluate the generalization ability of our approach. The three datasets include one indoor dataset with three static GoPro HERO8 cameras, one outdoor dataset with four static GoPro HERO8 cameras, and another outdoor dataset with two static GoPro HERO8 cameras and one dynamic drone camera. We name the datasets **Office**, **Square**, and **Drone**.

Fig.6 presents results of our approach on the three open datasets and the three wild datasets. We assume unknown camera poses and use the same hyper-parameters for all the datasets. The six datasets include a variety of settings, such as indoor and outdoor environments, small and large field-of-views, strong and weak lighting conditions, high- and low-resolution images, static and moving cameras, etc. Our approach has shown good generalization ability across these settings.

4.4 Ablation

Usage of Method Tab. 2 presents ablation study results as guidance for using our approach. Tab. 2 presents results of three variants of our approach. “Oracle" assumes the camera poses are known and fixed for all video frames and only optimizes the human poses. Since the GT camera poses are optimized together with human poses for each frame. The GT and “Oracle" camera and human poses will differ slightly for each frame. The “One-Step" variant

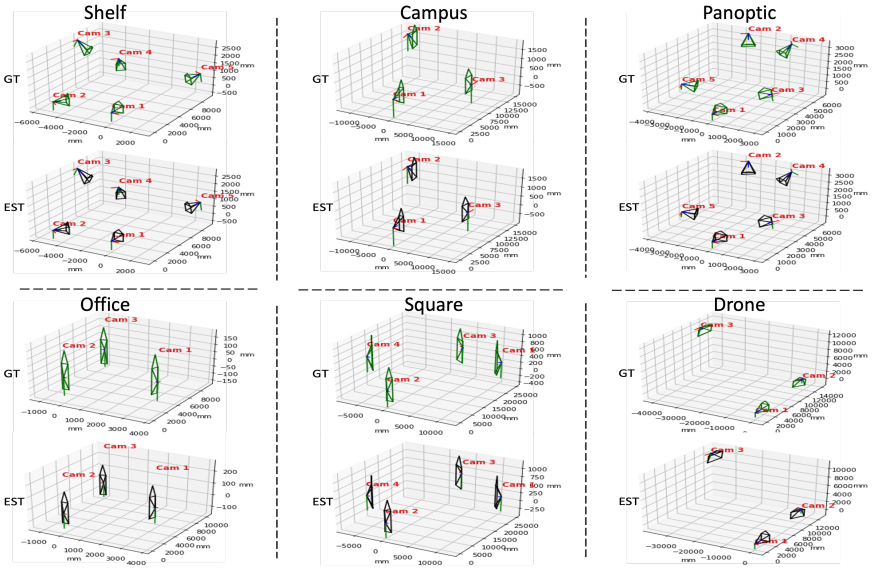


Figure 7: Qualitative comparison between the ground truth camera poses (top row) and the estimated camera poses using our approach (bottom row).

simultaneously estimates 6 DoF camera poses and 3D human poses using single-frame multi-view images. This variant suits scenarios when the number of people is large and people are close to the cameras. We use this variant for the Shelf dataset. The "Two-Step" variant first estimates the camera poses using multi-frame images, then solves the human poses. We use this variant for the Campus dataset since the people are away from the cameras, making the point correspondences low quality.

Camera Pose Estimation Camera pose estimation is an intermediate step of our approach. The result of camera pose estimation will directly impact 3D human pose estimation. For a more comprehensive understanding of the camera pose estimation, we present quantitative comparison between GT and estimated camera poses in Fig. 7. Fig.7 compares the ground truth and estimated camera poses. Since we assume the camera poses (w.r.t a pre-defined world origin) unknown, in each dataset, we use one camera as the world origin: Shelf (camera 1), Campus (camera 1), Panoptic (camera 1), Office (camera 2), Square (camera 2), Drone (camera 1). The estimated camera poses are close to the GTs across six scenes.

5 Summary

In this work, we have presented an approach for multi-view multi-person 3D human pose estimation for camera networks of which the 6 DoF camera poses are uncalibrated. We have introduced a constrained optimization formulation for solving the cross-view 2D matching problem when geometric constraints are unavailable. We have introduced a camera pose self-validation process to deal with the low-quality correspondences. We evaluated our approach on three public and three wild datasets and provided guidance on the usage of our approach.

This work was funded in part by NSF NRI (202417) and Department of Homeland Security (2017-DN-077-ER0001).

References

[1] Nadeem Anjum. Camera localization in distributed networks using trajectory estimation. *Journal of Electrical and Computer Engineering*, 2011:13, 2011.

[2] 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.

[3] Vasileios Belagiannis, Xinchao Wang, Bernt Schiele, Pascal Fua, Slobodan Ilic, and Nassir Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. In *European Conference on Computer Vision*, pages 742–754. Springer, 2014.

[4] 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.

[5] 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.

[6] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000.

[7] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3618–3625, 2013.

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[9] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017.

[10] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019.

[11] 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 *CVPR*, 2020.

[12] 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.

[13] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019.

[14] Dylan Drover, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[15] 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.

[16] 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.

[17] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[18] Junzhi Guan, Francis Deboeverie, Maarten Slembrouck, Dirk Van Haerenborgh, Dimitri Van Cauwelaert, Peter Veelaert, and Wilfried Philips. Extrinsic calibration of camera networks based on pedestrians. *Sensors*, 16(5):654, 2016.

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

[20] Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997.

[21] Richard I Hartley and Peter Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997.

[22] 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. In *European Conference on Computer Vision*, pages 477–493. Springer, 2020.

[23] Shiyao Huang, Xianghua Ying, Jiangpeng Rong, Zeyu Shang, and Hongbin Zha. Camera calibration from periodic motion of a pedestrian. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3025–3033, 2016.

[24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.

[25] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020.

[26] 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.

[27] Imran Junejo and Hassan Foroosh. Robust auto-calibration from pedestrians. In *null*, page 92. IEEE, 2006.

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[29] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019.

[30] Nils Krahnstoever and Paulo RS Mendonça. Autocalibration from tracks of walking people. In *in Proc. British Machine Vision Conference (BMVC*. Citeseer, 2006.

[31] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

[32] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9887–9895, 2019.

[33] Sijin Li, Weichen Zhang, and Antoni B Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2848–2856, 2015.

[34] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[35] Yu-Jhe Li, Xinshuo Weng, Yan Xu, and Kris M Kitani. Visio-temporal attention for multi-camera multi-target association. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2021.

[36] Jingchen Liu, Robert T Collins, and Yanxi Liu. Surveillance camera autocalibration based on pedestrian height distributions. In *British Machine Vision Conference (BMVC)*, volume 2, 2011.

[37] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[38] 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.

[39] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018.

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

[41] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 10133–10142, 2019.

[42] Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

[43] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960, 2019.

[44] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017.

[45] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018.

[46] Alejandro Perez-Yus and Antonio Agudo. Matching and recovering 3d people from multiple views. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3622–3631, 2022.

[47] Ali Rahimi, Brian Dunagan, and Trevor Darrell. Simultaneous calibration and tracking with a network of non-overlapping sensors. In *null*, pages 187–194. IEEE, 2004.

[48] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15190–15200, 2021.

[49] Guillaume Rochette, Chris Russell, and Richard Bowden. Weakly-supervised 3d pose estimation from a single image using multi-view consistency. *arXiv preprint arXiv:1909.06119*, 2019.

[50] 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.

[51] 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.

[52] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[53] Zheng Tang, Yen-Shuo Lin, Kuan-Hui Lee, Jenq-Neng Hwang, Jen-Hui Chuang, and Zhijun Fang. Camera self-calibration from tracking of moving persons. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 265–270. IEEE, 2016.

[54] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[55] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision*, pages 197–212. Springer, 2020.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[57] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2021.

[58] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3d multi-person pose estimation using multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11148–11157, 2021.

[59] Yan Xu, Vivek Roy, and Kris Kitani. Estimating 3d camera pose from 2d pedestrian trajectories. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2568–2577. IEEE, 2020.

[60] Yan Xu, Yu-Jhe Li, Xinshuo Weng, and Kris Kitani. Wide-baseline multi-camera calibration using person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13134–13143, 2021.

[61] 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.

[62] 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, 2018.

[63] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[64] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.