Multi-person 3D Pose Estimation in Crowded Scenes Based on Multi-View Geometry

He Chen*[1], Pengfei Guo*[1], Pengfei Li[1], Gim Hee Lee†[2], Gregory Chirikjian†[2,1]

[1] The Johns Hopkins University, USA
[2] National University of Singapore, Singapore
{hchen136, pguo4, pli32, gchirik1}@jhu.edu
{gimhee.lee@comp, mpegre@}.nus.edu.sg

Abstract. Epipolar constraints are at the core of feature matching and depth estimation in current multi-person multi-camera 3D human pose estimation methods. Despite the satisfactory performance of this formulation in sparser crowd scenes, its effectiveness is frequently challenged under denser crowd circumstances mainly due to two sources of ambiguity. The first is the mismatch of human joints resulting from the simple cues provided by the Euclidean distances between joints and epipolar lines. The second is the lack of robustness from the naive formulation of the problem as a least squares minimization. In this paper, we depart from the multi-person 3D pose estimation formulation, and instead reformulate it as crowd pose estimation. Our method consists of two key components: a graph model for fast cross-view matching, and a maximum a posteriori (MAP) estimator for the reconstruction of the 3D human poses. We demonstrate the effectiveness and superiority of our proposed method on four benchmark datasets. Our code is available at: https://github.com/HeCraneChen/3D-Crowd-Pose-Estimation-Based-on-MVG.

Keywords: 3D pose estimation, occlusion, correspondence problem

1 Introduction

Fast 3D human pose estimation for crowded scenes is an important component in many computer vision applications such as autonomous driving, surveillance, and robotics [12,17,26,29,30,31,41,43,47]. However, recovering 3D human pose from crowded real-world setting is a challenging endeavor due to the inherent depth ambiguity caused by 2D to 3D backprojections, self-occlusions, and occlusions by other people in crowded scenes [1,25,38]. A three-step process is commonly used in the multi-person multi-camera 3D pose estimation problem: 1) Detecting human body keypoints or parts in separate 2D views; 2) Matching people across different views; 3) Reconstructing 3D pose by triangulation. Unfortunately, the critical second step of matching people across different views is non-trivial. Well-known matching algorithms such as the Harris corner detector [19] and the Scale

* Equal first author contribution. † Jointly supervised this work.

Invariant Feature Transform (SIFT) [35] give mostly wrong matches even after robust estimation with RANSAC [18]. The problem is further aggravated in the third step when these unreliable matches are used in a vanilla triangulation algorithm to recover the 3D points.

With the rapid development of deep learning, features are extracted more precisely and significant improvements are made for appearance-based feature matching across different viewpoints on the spatial level or different frames on the temporal level [34,40,48]. Despite the improvements, these methods are sub-optimal for the task of people matching across multiple views in crowded scenarios. The reasons are threefold. Firstly, intra-class variation of human body appearance is relatively smaller than objects such as architectural features or graffiti paintings, and thus more outliers can result if the aforementioned methods are deployed directly. Secondly, dense feature matching across whole images is usually computationally inefficient for applications such as autonomous driving, where real-time is one of the primary concerns. Thirdly, appearance-based matching has a lower correctness criterion than people-based matching across multiple views. On the other hand, it is interesting to note that the level of occlusion in the same object can differ drastically among different views. Therefore, it is reasonable to trust the slightly occluded views more than the highly occluded views in the process of triangulation.

In this paper, we propose a 3D crowd human pose estimation method based on multi-view geometry. Specifically, we focus on overcoming the bottlenecks of multi-person 3D pose estimation and pushing it further to dense crowd 3D pose estimation. To this end, we propose the matching of feet across multiple views to improve the accuracy of body joint correspondences. We first modify a 2D pose estimation network, i.e. the joint-candidates single person pose estimation (SPPE) [28] to include additional joints for the feet. Subsequently, we find the best matches of the feet across multiple views, and then extend the correspondences to the other joints using the kinematic chain of the human body. We cast the matching problem as a binary linear program and solve it efficiently with the Jonker-Volgenant algorithm [22]. Finally, we improve the robustness of triangulation by formulating the problem as a maximum a posteriori (MAP) estimation that weighs the likelihood term with the uncertainty of the 2D joint observation and enforces a prior on the average bone lengths of the estimated 3D human poses. We evaluate our proposed method on four challenging benchmark datasets. Experimental results show that our method outperforms all existing algorithms on these datasets.

Our main contributions in this work are summarized as follows:

- Design a simple and efficient people matching mechanism based on feet assignment across different views, which is applicable for dense crowds.
- Propose a more robust triangulation for 3D crowd reconstruction using MAP estimation that accounts for the uncertainty of 2D joint detection and enforces the average 3D bone lengths.

– Define a problem of crowd 3D human pose estimation, and argue its existence as a separate problem from multi-person multi-camera 3D human pose estimation.

2 Related Work

Single-Person Human Pose Estimation. A large amount of literature exists in this field due to the advancement of deep learning. We briefly summarize those for 3D human pose which are more closely related to this work. State-of-the-art methods can be divided into two categories, direct regression methods [10,21,36] and indirect regression methods based on heat maps [20,27,37]. In [37], a coarse-to-fine prediction scheme was developed by analyzing 3D human pose in a volumetric representation. Integral pose [42] unifies the heat map representation and joint regression by replacing the non-differentiable *argmax* with integral operation. Regardless of the good performance, learning 3D pose from a single image is still an ill-posed problem. Instead of finding one exact solution, [27] developed a multimodal mixture density network, so that multiple feasible solutions are found before refining into one solution. The authors of [20] proposed a volumetric aggregation from intermediate 2D backbone feature maps and combines 3D information from multiple 2D views. The aforementioned methods obtained state of the art performance for single person 3D pose estimation, but unfortunately in the multi-person scenario, additional ambiguity makes these methods suboptimal.

Multi-Person Human Pose Estimation. Several recent works have focused on multi-person scenarios in problem formulation either based on monocular setting [44] or multi-view setting [2,3,4,5,13,24]. Results obtained from the multi-view setting are generally more precise due to the additional information. However, bottlenecks still exist in these multi-view based methods, *i.e.* how to cope with the correspondence problem and how to make the triangulation of depth information sufficiently robust against noise. In [24], epipolar constraints are directly applied for people assignment among different views. This worked perfectly when people in the scene stand far away from each other. However, this constraint is likely to fail when the scenario gets crowded. For instance, if some epipolar line of a particular joint happens to pass through several other people, it is hard to make sure that no other joint is closer to the line than the correct matching joint. The authors of [13] incorporated appearance cues by fusing re-identification with epipolar constraints. However, the two kinds of constraints are still independently considered. The 3D pictorial structure model [2,3] resolves ambiguities of mixed parts, occlusion, and false positives by building multi-view unary potentials, while at the same time integrating prior model by pairwise and ternary potential functions. This motivates our work in using MAP as a formulation to cope with measurement noise in triangulation process.

Previous 'multi-person' methods work on relatively sparse crowds. In [28], crowd pose estimation is firstly defined as a separate research field, but the

problem is defined in 2D. When extending to 3D, more uncertainties are introduced. This encourages us to define the crowd pose estimation problem in 3D and explore a potential solution in this paper.

Feature Matching and Correspondence Problem. Feature correspondence in general raises stricter demand than feature matching due to the fact that both appearance and location need to be taken into consideration. In [6], a globally-optimal inlier set cardinality maximization approach is proposed to jointly estimate optimal camera pose and optimal correspondences. [46] solves the correspondence problem between two images by defining energy function measuring data consistency and spatial regularity. In [14], Point-Line Minimal Problems are thoroughly defined and analyzed. This provides a theoretical guidance to solve the specific problem of point line matching for the people assignment task.

3 Our Method

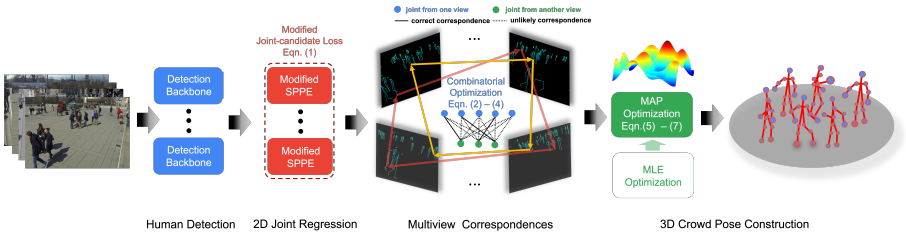

Fig. 1. The pipeline of our proposed approach. See text for more detail.

Figure 1 shows an overview of our approach. Human bounding box proposals are first obtained by an off-the-shelf detection network, and then fed into a modified SPPE network (Sec. 3.1) to estimate the 2D joints. Subsequently, we get the multi-view joint correspondences by solving a combinatorial optimization problem via graph matching (Sec. 3.2). Finally, the 3D crowd poses are reconstructed using a MAP formulation (Sec. 3.3) solved by the trust region method [11].

3.1 2D Pose Estimation

We leverage on the recently proposed CrowdPose network [28] trained on the CrowdPose Dataset [28] for 2D pose detection on the input images. The CrowdPose network follows a top-down framework. It first detects the bounding boxes of individual persons using YOLOv3 [39], and then performs joint-candidate SPPE and a global maximum joints association algorithm to estimate the 2D joints. Similar to other 2D pose estimation methods, the accuracy of the joint

detection drops as it moves farther away from the center of a person (i.e. the 'hip' joint) despite the state-of-the-art performance of [28] on the benchmark datasets. As a result, detection of the 'ankle' joints, which are usually used to represent feet, are especially noisy. To mitigate this problem, we follow [7] in adding 6 additional joints on the feet (3 on each foot) and modify the loss function of the network into the weighted sum of the mean square error (i.e. MSE[.,.]) from the body joints and the feet joints as follows:

$$\mathcal{L} = \frac{1}{I + 6\lambda} \left\{ \sum_{i=1}^{I} \text{MSE}\left[\mathbf{P}_h^i, \mathbf{T}_h^i + \mu \mathbf{C}_h^i\right] + \lambda \sum_{i=I+1}^{I+7} \text{MSE}\left[\mathbf{P}_h^i, \mathbf{T}_h^i + \mu \mathbf{C}_h^i\right] \right\}. \quad (1)$$

I stands for the number of joints of the body part excluding the 6 joints representing the feet (e.g. $I = 17$ for MSCOCO [32]). $\mathbf{P}_h^i$ and $\mathbf{T}_h^i$ represents the output heatmap and the heatmap of the target joints, respectively, for the i^{th} joint of the h^{th} person. $\mathbf{C}_h^i$ represents detections of the same joint type from other persons that might be within the bounding box of the h^{th} person. We include $\mathbf{C}_h^i$ into the loss function to learn a multi-modal heatmap $\mathbf{P}_h^i$. μ is the attention factor in the range of $[0, 1]$ to control the extent of the contribution of $\mathbf{C}_h^i$, which we set to 0.5 in all our implementations. We set $\lambda > 1$, so that the 6 additional joints on the feet receive more attention during training. Our network is trained on the Human Foot Keypoint Dataset [7].

3.2 Multi-view Correspondence with Graph Matching

Previous methods [13,24] apply epipolar constraints to all joints in order to solve the correspondence problem. We argue that this can give a suboptimal solution when the crowd becomes denser. This is because the epipolar line that corresponds to a joint in one view is likely to pass through multiple joints in the other view for a crowded scene. Consequently, this ambiguity renders the Euclidean distance between the epipolar line and joints to be a less ideal metric. We circumvent this challenge by casting the joint correspondence problem into a feet assignment problem. Specifically, we first establish the feet that belong to a same person across the multiple views, and then grow the joint correspondences from the feet using the kinematic chain of the human body.

Feet Assignment. We propose to use feet assignment to realize people matching as shown in Figure 2(a). The core intuition is that prior information, appearance constraints, location constraints are naturally fused in such setting. We use the fact that at least one foot is on the ground when a person is walking as the prior information. The detected joints of the feet as described in Sec. 3.1 are used as the appearance information. To incorporate location constraints, we use the homographies between all view pairs to rectify the ground planes among different views into a common reference. We denote the homography between the ground planes of view j and k as $H_{j,k}$. Consequently, we can directly compare the joints of the feet across different views. We get ≥ 4 point correspondences between the ground plane of each pair of view j and k to compute $H_{j,k}$. It is

interesting to note that applying the homography to all pixels in the image, we might get a twisted image which appear to be strange at first glance. This is based on the prior that this 'the world is 3D'. However, if we change the prior into 'the world is 2D', and treat everything as chalk art drawn on the ground, then everything in the rectified image starts to look reasonable. In this light, the problem of joint matching boils down to feet assignment.

Graph Building. A naive search for the optimal feet assignment is intractable due to the large combinatorial search space. To improve the efficiency of the search, we build a complete bipartite graph from the feet across two views and solve it as a linear assignment problem. Let $\mathcal{V}_j = \{v_{ij} : \forall i \in \{1, \dots, a_j\}\}$ denote the set of pair-of-feet in view j. v_{ij} is the detected pair-of-feet with index i in view j, and a_j is the total number of detected pair-of-feet in view j. We further denote the set of edges in the complete bipartite graph for the pair of views j and k as $\mathcal{E}_{j,k} = \{e_{l,m} : \forall l \in \{1, \dots, a_j\}, m \in \{1, \dots, a_k\}\}$. The complete bipartite graph for each pair of views can then be formally written as:

$$\mathcal{K}_{a_j,a_k} = ((\mathcal{V}_j, \mathcal{V}_k), \mathcal{E}_{j,k}), \tag{2}$$

Optimal Cross-view Matching. Based on this construction, our goal becomes finding a subgraph $\mathcal{G} \subset \mathcal{K}_{a_j,a_k}$ by eliminating edges in the graph that represent the unlikely correspondences. We solve this edge elimination problem as a binary linear program that minimizes the total edge costs subjected to a set of linear constraints, i.e.

$$\begin{aligned} \min_{\mathbf{d}} \quad & \sum_{l=1}^{a_j} \sum_{m=1}^{a_k} c_{l,m} \cdot d_{l,m} \\ \text{s.t.} \quad & \sum_{l=1}^{a_j} c_{l,m} \leq 1, \quad \sum_{m=1}^{a_k} c_{l,m} \leq 1, \\ & \sum_{l=1}^{a_j} d_{l,m} = 1, \quad \sum_{m=1}^{a_k} d_{l,m} = 1, \quad \mathbf{d} \in \{0,1\}^{a_j \times a_k}. \end{aligned} \tag{3}$$

$d_{l,m} \in \mathbf{d}$ is a binary variable that represents the selection of the edge $e_{l,m}$ when it is equals to 1. $c_{l,m}$ is the cost of selecting the edge $e_{l,m}$, which we define as:

$$c_{l,m} = k_1 \cdot \left| p_l - H_{j,k} \cdot p_m \right| + k_2 \cdot \left| |\mathbf{v}_l| - |\mathbf{v}_m| \right| + k_3 \cdot \left(\frac{\mathbf{v}_l \times \mathbf{v}_m}{|\mathbf{v}_l| \cdot |\mathbf{v}_m|} \right), \tag{4}$$

where p_l and p_m respectively represents the location of two pairs of feet, $H_{j,k}$ represents the homography matrix between the two views j and k, $\mathbf{v}_l$ and $\mathbf{v}_m$ represent vectors of strides. k_1, k_2, k_3 are hyper parameters to adjust the importance between the foot location, stride size, and stride direction. The metric is visualized in Figure 2(b).

Solver. We use the Jonker-Volgenant algorithm [22] as the solver to find the solution to the two-view feet assignment problem formulated in Eq. 3. We ensure consistency of the assignment across multiple views by resolving the conflict in the correspondences with priority given to edges with lower edge cost as defined in Eq. 4. A directed graph where the skeleton is a spanning union of disjoint cycles is obtained when the matching across n views is successful. Our matching algorithm has a time complexity of $\mathcal{O}((2N)^3) = \mathcal{O}(8N^3)$, where N is the number of persons per image. In contrast, the $O(n^4)$ implementation of Hungarian algorithm has a total time complexity of $\mathcal{O}((17N)^4)$ on 17 joints. Although the constant term is usually considered unimportant for time complexity analysis, it cannot be neglected in this study since $N < 30$ usually holds. Thus, our method is significantly faster.

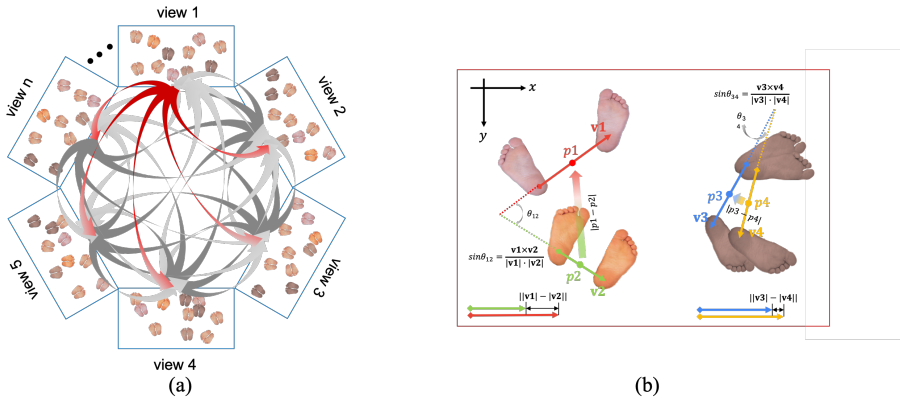

Fig. 2. People matching using feet assignment. (a) The matching process across n views, and (b) visualization of edge cost defined in Eq. 4.

3.3 3D Crowd Pose Reconstruction

Under the assumption that the camera parameters are known, we can reconstruct the 3D human poses by triangulation of the joint correspondences across the multiple views obtained from the previous section. One naive method of triangulation is to directly minimize the squared sum of perpendicular distances between the epipolar line and the detected joint. We refer to this naive method as the vanilla triangulation method. This is a classical method that works well in single person scenarios. However, in occluded scenes, the 2D joints are noisy and might have shifts of a few pixels. Consequently, this breaks the correspondence across multiple views and causes the 3D reconstructed points to be unreliable. We formulate a MAP optimization to mitigate the problem from the unreliable correspondences, where we model the likelihood with the 2D measurement uncertainty and use the prior term to constrain the bone lengths of the estimated body poses.

MAP Optimization. The ultimate goal of the proposed method is to estimate 3D coordinates of human joints. We formulate this as a MAP over the latent 3D poses $\mathbf{Q}$, i.e.

$$\mathbf{Q}_{\text{MAP}} = \underset{\mathbf{Q}}{\text{argmax}} \prod_{i=1}^{N} P(Q_i) \prod_{j=1}^{M} \prod_{k=1}^{O} P(q_{ijk} \mid \mathcal{P}_k, Q_{ij}), \tag{5}$$

where N is the total number of persons in the scene, M is the number of joints per person, and O is the total number of camera views. q_{ijk} is the j^{th} 2D joint of the i^{th} person in the k^{th} camera view. $Q_{ij} \in Q_i$ is the j^{th} 3D joint from the 3D pose $Q_i \in \mathbf{Q}$ of the i^{th} person in the scene. $\mathcal{P}_k$ is the projection matrix of the k^{th} camera. The likelihood term is given by the following Gaussian distribution:

$$P(q_{ijk} \mid \mathcal{P}_k, Q_{ij}) = \frac{1}{2\pi\sigma_{ijk}} \exp\left\{ -\frac{\|q_{ijk} - \alpha(\mathcal{P}_k, Q_{ij})\|^2}{2{\sigma_{ijk}}^2} \right\}, \tag{6}$$

where $\sigma_{ijk} = f(s^i_{bbox}, s^k_{heatmap}, q_{ijk})$ is the uncertainty of the j^{th} 2D joint q_{ijk} computed from the bounding box s^i_{bbox} of the i^{th} person and the output heatmap of the image from the k^{th} view. $\|q_{ijk} - \alpha(\mathcal{P}_k, Q_{ij})\|$ is the reprojection error computed from the 2D joint q_{ijk} and the normalized coordinates of the 3D joint Q_{ij} projected into the image of the k^{th} view given by $\alpha(.,.)$. The prior term is defined as:

$$P(Q_i) = \prod_{l=1}^{L} \frac{1}{2\pi\sigma_l} \exp\left\{ -\frac{\left\|b^l_{\text{ref}} - b^l_i\right\|^2}{2\sigma_l^2} \right\}, \tag{7}$$

where b^l_i represents the l^{th} bone length between two 3D joints in the i^{th} person, and b^l_{ref} represents the average length of the l^{th} bone. L is the total number of bones in the human body representation. σ_l is the standard deviation in the length of the l^{th} bone. Intuitively, the prior term enforces the bone lengths of the estimated 3D human pose to be close to the average lengths.

Initialization and solver. We initialize the iterative MAP optimization with the vanilla triangulation. Subsequently, we use the trust region method [11] as a solver for the MAP optimization. In addition, we empirically observe that performing the maximum likelihood estimation (MLE) with the initialized values as an intermediate step before MAP improves the final estimation of the 3d human poses.

4 Experiments

We evaluate our proposed method on four public datasets. These datasets consist of scenarios that include autonomous driving and surveillance with challenging situations such as moving camera and heavy occlusions.

4.1 Datasets

LOEWENPLATZ [15]. This is a dataset of driving recorder scenario captured in Zurich with two calibrated cameras. The dataset represents common scenarios that autonomous driving cars are likely to experience everyday.

Chariot Mk I [16]. This is a dataset captured by hand-held cameras. The cameras are moving and shaking, which resemble real-life scenarios from the perspective of the pedestrians.

Wildtrack [9]: This dataset emulates surveillance scenarios with the set-up of 7 fixed cameras. All cameras are fully calibrated, i.e. known intrinsics and extrinsics camera parameters. Occlusion is severe in each view of this dataset.

CMU Panoptic Dataset [23]: This dataset is captured in a studio and provides precise 3D ground truth in MSCOCO [32] format. In this paper, we evaluate the performance of our method quantitatively on the 'Ultimatum' sequences with complete 3D human pose annotations. This sequence consists of relatively more active and complicated social scenarios for human pose estimation than other sequences.

Table 1. Quantitative results for the Chariot Mk I, LOEWENPLATZ, and Wildtrack datasets using the evaluation metrics from MSCOCO [32].

Chariot Mk I	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}	**AR**	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Belagiannis *et al.* [3]	48.1	64.8	59.3	63.7	64.6	58.1	62.7	55.9	54.4	61.9
Dong *et al.* [13]	69.3	87.4	73.6	77.5	75.4	71.9	87.5	81.7	78.1	80.0
Ours w/ Vanilla Trigulation	60.0	90.8	72.2	65.4	77.6	72.3	95.3	83.0	76.6	81.8
Ours w/ Proposed MAP	**89.8**	**98.9**	**92.7**	**91.7**	**99.5**	**93.9**	**99.8**	**96.0**	**95.4**	**99.6**
LOEWENPLATZ	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}	**AR**	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Belagiannis *et al.* [3]	49.3	63.7	58.2	63.2	56.9	61.9	84.3	64.3	73.7	55.3
Dong *et al.* [13]	62.1	88.3	63.5	61.3	72.5	80.3	87.2	77.9	81.7	84.6
Ours w/ Vanilla Trigulation	66.7	93.8	73.1	71.6	84.4	78.2	96.7	84.5	80.1	88.9
Ours w/ Proposed Optimization	**81.8**	**97.1**	**88.7**	**83.3**	**90.8**	**88.9**	**98.5**	**93.5**	**90.0**	**94.4**
Wildtrack	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}	**AR**	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Belagiannis *et al.* [3]	44.1	53.4	46.0	19.4	47.8	64.1	79.1	61.4	20.9	55.4
Dong *et al.* [13]	55.6	78.4	53.1	34.9	60.0	73.4	87.8	68.1	38.1	77.6
Ours w/ Vanilla Trigulation	55.3	79.6	50.6	33.2	60.1	77.3	88.7	72.9	38.6	78.4
Ours w/ Proposed MAP	**70.0**	**90.2**	**73.6**	**44.7**	**76.4**	**78.3**	**93.6**	**82.4**	**55.5**	**83.7**

4.2 Results

Quantitative Results. We adopt the key point evaluation metrics of MSCOCO [32], i.e. the average precision (AP), average recall (AR) and their variants. Specifically, the variants of AP and AR are specified by the Object Keypoint Similarity (OKS) that plays the same role as the Intersection over Union (IoU) in object detection. It measures the scale of the object, and the distance between predicted joints and ground truth points. The AP at OKS=.50:.05:.95 (primary challenge metric in MSCOCO [32] competitions) is used to measure the reprojection errors. Table 1 shows that our method outperforms the state-of-the-art

algorithms on the Chariot Mk I, LOEWENPLATZ, and Wildtrack datasets using the evalution metrics from MSCOCO [32]. Table 2 shows the comparative performance for 2D key point detection of our modified body+foot candidate-joint SPPE network on the MSCOCO dataset [32]. Our method achieves a comparable performance with the best performing [8]. Furthermore, our method outperforms on AP@0.5:0.95 for medium objects, which is more valuable for our framework with the feet detection, matching and optimization stages. CMU Panoptic Dataset provides the 3D ground truth. Therefore, we use two metrics, i.e. mean per joint position error (MPJPE) and percentage of correct parts (PCP) instead of the reprojection error for direct evaluation. The results are shown in Table 3 and Table 4.

Table 2. Quantitative comparison of key point detection experiments on COCO body+foot validation set [7].

Method	**AP**	AP^{50}	AP^{75}	AP^{M}	AP^{L}
GT Bbox + CPM [45]	62.7	86.0	69.3	58.5	70.6
SSD [33] + CPM [45]	52.7	71.1	57.2	47.0	64.2
Cao *et al.* [8]	65.3	**85.2**	71.3	62.2	**70.7**
Ours	**65.3**	80.1	**72.2**	**74.1**	68.3

Table 3. Quantitative results for the proposed method on different joints of human body in CMU Panoptic Dataset (Ultimatum sequences, four cameras) using MPJPE (mm).

Metric	**Average**	Head	Shoulder	Elbow	Wrist	Hip	Knee	Foot
MPJPE	50.0	45.1	43.6	55.6	60.7	25.3	53.2	66.0

Table 4. Quantitative results for the proposed method on different body parts in CMU Panoptic Dataset (Ultimatum sequences, four cameras) using the PCP metric.

Metric	**PCP**	Head	Torso	Upper arms	Lower arms	Upper legs	Lower legs
percentage	91.3	74.5	100.0	93.8	80.0	100.0	99.3

Table 5. Ablation study of MLE as an intermediate step on WildTrack dataset.

Method	ave	min	max	var
Ours w/o MLE	64.75	18.06	316.7	50.69
Ours w/ MLE	**38.55**	**2.18**	**219.29**	**27.52**

Qualitative Results. Figure 3, 4, and 5 show the qualitative results on the Wildtrack [9], CMU Panoptic [23], and LOEWENPLATZ [15] datasets, respectively. In Figure 3, our approach gives good quality 3D reconstructions of the

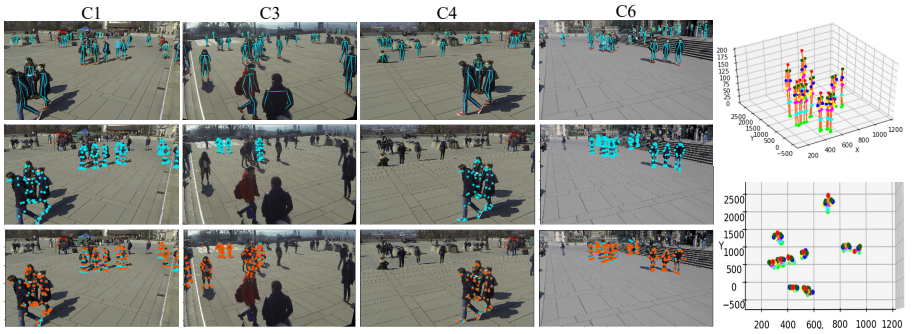

Fig. 3. Qualitative results on Wildtrack dataset. (First four columns) First row shows results of our modified candidate joint SPPE with attention on the feet; Second row shows the ground truth 2D joints (blue dots); Third row shows the reprojection of our estimated 3D joints (orange dots) overlaid on the ground truths (blue dots). The last column shows the (top) estimated 3D crowd human poses and its (bottom) top view.

Table 6. Foot keypoint analysis on COCO foot validation set.

Method	**AP**	**AR**	AP^{75}	AR^{75}
Cao et al. [8]	77.9	82.5	82.1	85.6
Our	80.1	82.0	85.5	87.4

Table 7. Evaluation of correspondence process on CMU Panoptic Dataset.

Dataset	RANSAC	EC	Ours
Precision	46.0	86.5	93.7
Time Complexity	NA	$O((17N)^4)$	$O((2N)^3)$

human poses even when heavy occlusion happens in the crowded scene. To validate effectiveness of the proposed method, we choose crowded scenes with at least 5 people appearing in each frame as shown Figure 4. We further show the qualitative visualizations of the estimated 3D human pose of several single persons from our method with the ground truth. Location information is used to match estimated pose with ground truth of each individual person. Orange represents estimated skeleton and blue represents ground truth. We zoom in each skeleton to clearly show details. As can be observed, the blue skeleton and orange skeleton has a slight offset. Nonetheless, this offset is in a tolerable range. In Figure 5, we evaluate our method under the setting of autonomous driving. The car went straight, turned left, and stopped at a crosswalk. We can see that our proposed method gives good 3D human pose estimations in different road scenes from a moving camera.

Ablation Study. We perform ablation studies to show the effectiveness of our proposed loss function Eq. 1 for 2D pose estimation, and the MLE as an intermediate step. We define an error distance between the reprojection of a 3D point and its corresponding 2D ground truth for quantitative evaluation. Comparison is carried out between the results from MAP with and without MLE as an intermediate step on the WildTrack dataset. In Table 5, we show the average, minimum, maximum, and variance of the reprojection error distance. Figure 7 shows the histogram of error distribution in pixel unit. We can see that

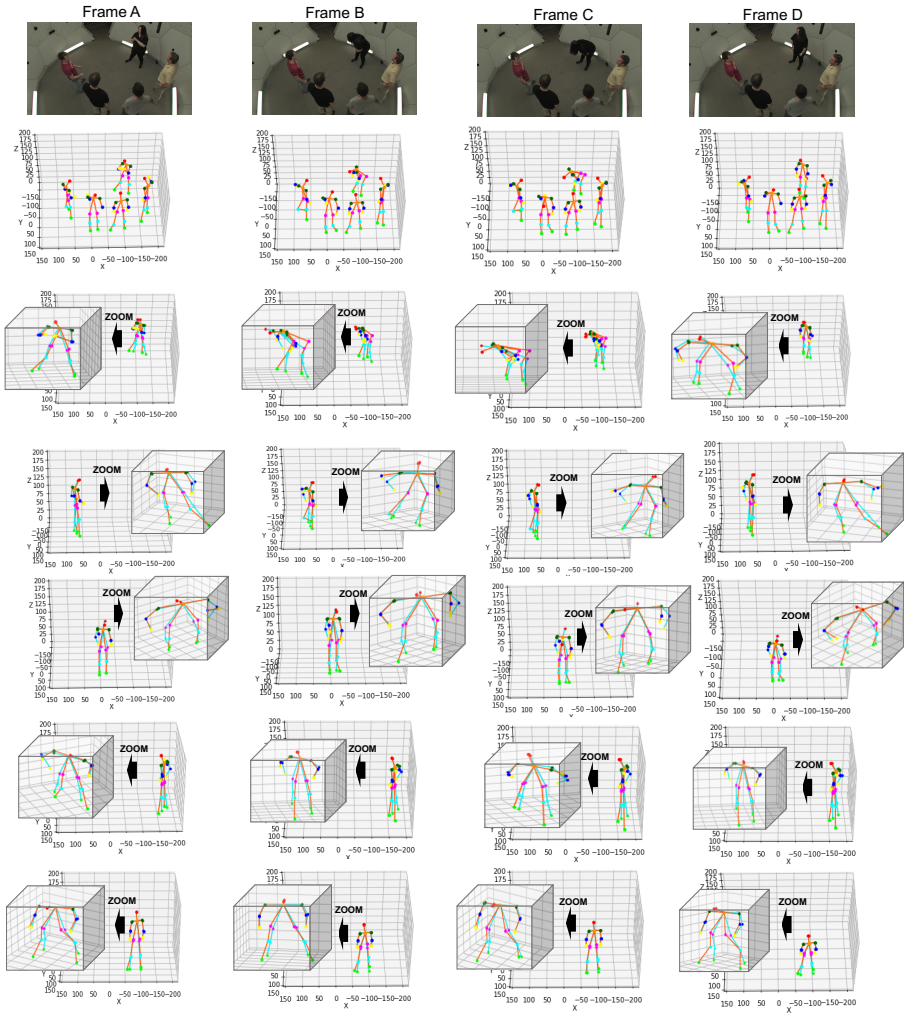

Fig. 4. Qualitative results on CMU Panoptic dataset. The first row shows images from the 4 cameras in the setup. The second row shows 3D crowd pose. The third to seventh row visualize the estimated 3D pose of each person (orange skeleton) and its corresponding ground truth (blue skeleton).

the smaller errors of the estimated 3D poses are obtained with the MLE as an intermediate step. Figure 6 demonstrates the effectiveness of our proposed loss function Eq. 1 for 2D pose estimation. As can be seen in the figure, our network detects the 'big toe', 'small toe' and 'heel' instead of the usual 'ankle' for the representation of a foot. The increased attention of the feet joints improves the estimation of the feet in highly occluded scene, and consequently facilities our matching algorithm. Comparison of the foot keypoints on the COCO foot

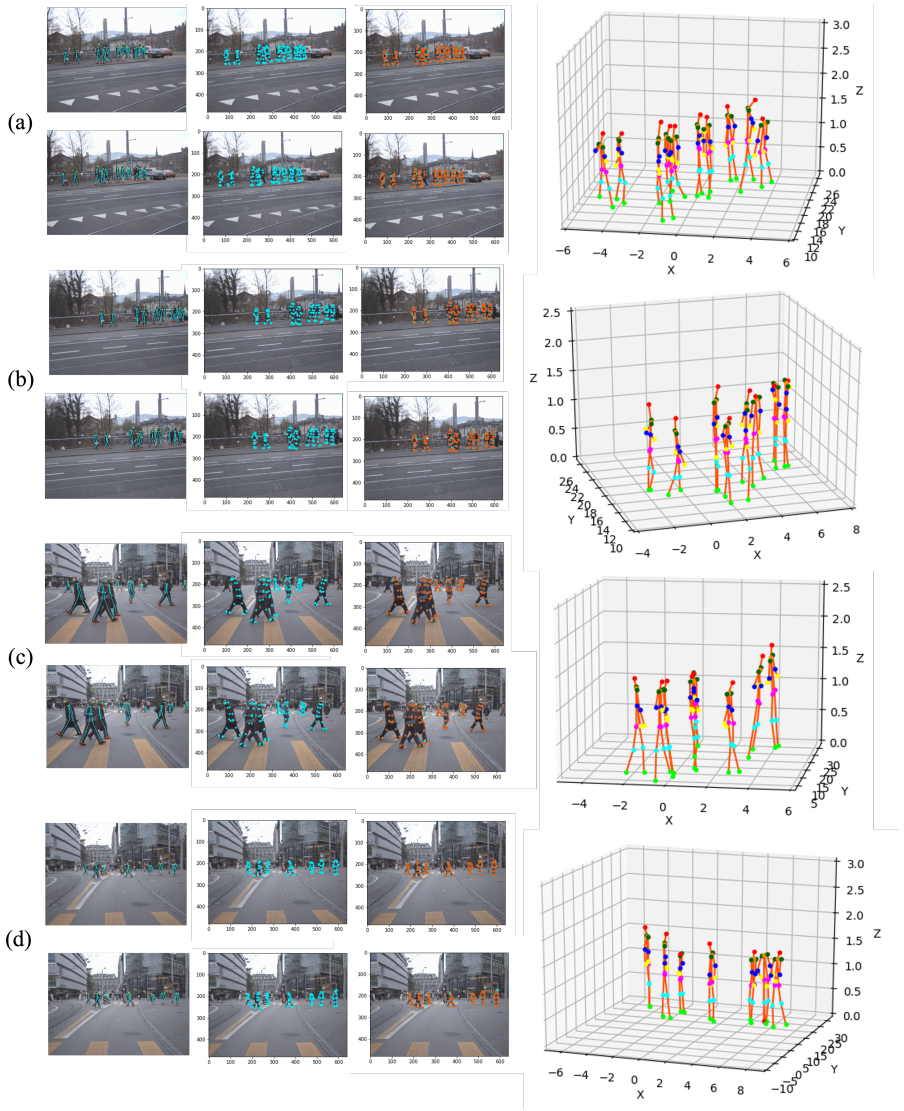

Fig. 5. Qualitative results on the LOEWENPLATZ dataset. The right most column shows the estimated 3D poses of scene (a)-(d). The first column shows the 2D skeletons detected by our modified SPPE network, the second column shows the ground truths of the 2D joints (blue dots), and the third column shows the reprojection of our estimated 3D joints (orange dots) and overlaid on the ground truths (blue dots).

validation set is shown in Table 6. To ablate the correspondence procedure, we conduct evaluations of correspondence process on the CMU Panoptic dataset in Table 7, where EC denotes Epipolar Constraint.

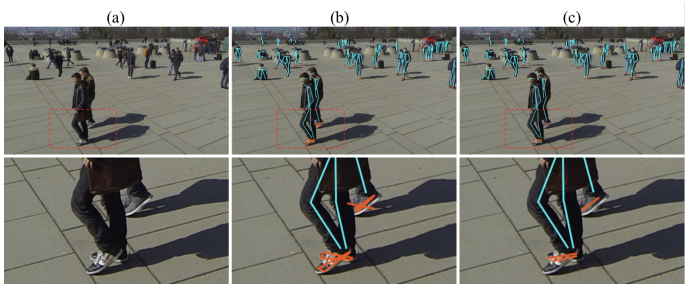

Fig. 6. Qualitative demonstration of our proposed loss function in Eq. 1. The figure shows the (a) original image, and the pose estimation result (b) **with** and (c) **without** the loss term on the feet joints in Eq. 1. The second row shows the corresponding zoomed-in images.

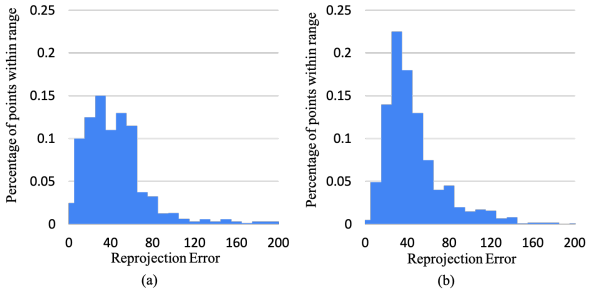

Fig. 7. Error distributions:(a) without and (b) with MLE as an intermediate step.

5 Conclusions

In this work, we propose a simple and effective approach for multi-person 3D pose estimation applicable to dense crowds. Matching of feet across multiple views improves the accuracy of body joint correspondences. A graph model is used for fast cross-view matching based on accurate estimation of foot joints. We cast the bipartite matching problem as a binary linear program and solve it efficiently with the Jonker-Volgenant algorithm. The robustness of triangulation is improved by using a MAP estimation that weighs the likelihood term with the uncertainty of the 2D joint observation and enforces a prior on the average bone lengths of the estimated 3D human poses. Experimental results show that our method outperforms all existing algorithms on four public datasets.

Acknowledgements. The authors would like to thank Yawei Li and Weixiao Liu for useful discussion. This work is supported in parts by the Office of Naval Research Award N00014-17-1-2142 and the Singapore MOE Tier 1 grant R-252-000-A65-114.

References

1. Baqué, P., Fleuret, F., Fua, P.: Deep occlusion reasoning for multi-camera multi-target detection. In: Proc. ICCV. pp. 271–279 (2017)
2. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3D pictorial structures for multiple human pose estimation. In: Proc. CVPR. pp. 1669–1676 (2014)
3. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3D pictorial structures revisited: Multiple human pose estimation. IEEE Trans. PAMI **38**(10), 1929–1942 (2015)
4. Belagiannis, V., Wang, X., Schiele, B., Fua, P., Ilic, S., Navab, N.: Multiple human pose estimation with temporally consistent 3D pictorial structures. In: Proc. ECCV. pp. 742–754. Springer (2014)
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3D human pose and shape from a single image. In: Proc. ECCV. pp. 561–578. Springer (2016)
6. Campbell, D., Petersson, L., Kneip, L., Li, H.: Globally-optimal inlier set maximisation for simultaneous camera pose and feature correspondence. In: Proc. ICCV. pp. 1–10 (2017)
7. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
8. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proc. CVPR. pp. 7291–7299 (2017)
9. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proc. CVPR. pp. 5030–5039 (2018)
10. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation+ matching. In: Proc. CVPR. pp. 7035–7043 (2017)
11. Conn, A.R., Gould, N.I., Toint, P.L.: Trust region methods, vol. 1. Siam (2000)
12. Dinesh Reddy, N., Vo, M., Narasimhan, S.G.: Carfusion: Combining point tracking and part detection for dynamic 3D reconstruction of vehicles. In: Proc. CVPR. pp. 1906–1915 (2018)
13. Dong, J., Jiang, W., Huang, Q., Bao, H., Zhou, X.: Fast and robust multi-person 3D pose estimation from multiple views. In: Proc. CVPR. pp. 7792–7801 (2019)
14. Duff, T., Kohn, K., Leykin, A., Pajdla, T.: Plmp-point-line minimal problems in complete multi-view visibility. In: Proc. ICCV. pp. 1675–1684 (2019)
15. Ess, A., Leibe, B., Schindler, K., Gool, L.V.: Robust multiperson tracking from a mobile platform. In: IEEE Trans. PAMI. pp. 1831–1846 (2009)
16. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: Proc. CVPR. pp. 1–8. IEEE (2008)
17. Fernando, T., Denman, S., Sridharan, S., Fookes, C.: Tracking by prediction: A deep generative model for mutli-person localisation and tracking. In: Proc. WACV. pp. 1122–1132. IEEE (2018)
18. Fischler, M.A., Bolles, R.C.: 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)
19. Harris, C.G., Stephens, M., et al.: A combined corner and edge detector. In: Alvey vision conference. vol. 15, pp. 10–5244. Citeseer (1988)

20. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proc. ICCV. pp. 7718–7727 (2019)
21. Jahangiri, E., Yuille, A.L.: Generating multiple diverse hypotheses for human 3D pose consistent with 2d joint detections. In: Proc. ICCVW. pp. 805–814 (2017)
22. Jonker, R., Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing **38**(4), 325–340 (1987)
23. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proc. ICCV (2015)
24. Kadkhodamohammadi, A., Padoy, N.: A generalizable approach for multi-view 3D human pose regression. arXiv preprint arXiv:1804.10462 (2018)
25. Korman, S., Milam, M., Soatto, S.: Oatm: Occlusion aware template matching by consensus set maximization. In: Proc. CVPR. pp. 2675–2683 (2018)
26. Kubo, H., Jayasuriya, S., Iwaguchi, T., Funatomi, T., Mukaigawa, Y., Narasimhan, S.G.: Programmable non-epipolar indirect light transport: Capture and analysis. IEEE Trans. VCG (2019)
27. Li, C., Lee, G.H.: Generating multiple hypotheses for 3D human pose estimation with mixture density network. In: Proc. CVPR. pp. 9887–9895 (2019)
28. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proc. CVPR. pp. 10863–10872 (2019)
29. Li, Y., Agustsson, E., Gu, S., Timofte, R., Van Gool, L.: CARN: Convolutional anchored regression network for fast and accurate single image super-resolution. In: Proc. ECCV. pp. 0–0 (2018)
30. Li, Y., Gu, S., Mayer, C., Van Gool, L., Timofte, R.: Group sparsity: The hinge between filter pruning and decomposition for network compression. In: Proc. CVPR (2020)
31. Li, Y., Tsiminaki, V., Timofte, R., Pollefeys, M., Van Gool, L.: 3D appearance super-resolution with deep learning. In: Proc. CVPR. pp. 9671–9680 (2019)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proc. ECCV. pp. 740–755. Springer (2014)
33. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Proc. ECCV. pp. 21–37. Springer (2016)
34. Liu, X., Zheng, Y., Killeen, B., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: Extremely dense point correspondences using a learned feature descriptor. In: Proc. CVPR (2020)
35. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proc. ICCV. vol. 2, pp. 1150–1157 (1999)
36. Mahendran, S., Ali, H., Vidal, R.: 3D pose regression using convolutional neural networks. In: Proc. ICCVW. pp. 2174–2182 (2017)
37. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proc. CVPR (2017)
38. Reddy, N.D., Vo, M., Narasimhan, S.G.: Occlusion-Net: 2D/3D occluded keypoint localization using graph networks. In: Proc. CVPR. pp. 7326–7335 (2019)
39. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
40. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proc. ICCV. pp. 1002–1012 (2019)

41. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: Proc. ICCV. pp. 951–958. IEEE (2011)
42. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proc. ECCV. pp. 529–545 (2018)
43. Vo, M., Yumer, E., Sunkavalli, K., Hadap, S., Sheikh, Y., Narasimhan, S.G.: Self-supervised multi-view person association and its applications. IEEE Trans. PAMI (2020)
44. Wang, C., Wang, Y., Lin, Z., Yuille, A.L.: Robust 3D human pose estimation from single images or video sequences. IEEE Trans. PAMI **41**(5), 1227–1241 (2018)
45. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proc. CVPR. pp. 4724–4732 (2016)
46. Windheuser, T., Cremers, D.: A convex solution to spatially-regularized correspondence problems. In: Proc. ECCV. pp. 853–868. Springer (2016)
47. Xin, S., Nousias, S., Kutulakos, K.N., Sankaranarayanan, A.C., Narasimhan, S.G., Gkioulekas, I.: A theory of fermat paths for non-line-of-sight shape reconstruction. In: Proc. CVPR. pp. 6800–6809 (2019)
48. Yew, Z.J., Lee, G.H.: 3DFeat-Net: Weakly supervised local 3D features for point cloud registration. In: Proc. ECCV. pp. 630–646. Springer (2018)