Generalizable Multi-Camera 3D Pedestrian Detection

João Paulo Lima[2,1], Rafael Roberto[1], Lucas Figueiredo[1], Francisco Simões[2,1], Veronica Teichrieb[1]
[1] Voxar Labs, Centro de Informática, Universidade Federal de Pernambuco
[2] Departamento de Computação, Universidade Federal Rural de Pernambuco
{jpsml, rar3, lsf, fpms, vt}@cin.ufpe.br, {joao.mlima,francisco.simoes}@ufrpe.br

Abstract

We present a multi-camera 3D pedestrian detection method that does not need to train using data from the target scene. We estimate pedestrian location on the ground plane using a novel heuristic based on human body poses and person's bounding boxes from an off-the-shelf monocular detector. We then project these locations onto the world ground plane and fuse them with a new formulation of a clique cover problem. We also propose an optional step for exploiting pedestrian appearance during fusion by using a domain-generalizable person re-identification model. We evaluated the proposed approach on the challenging WILDTRACK dataset. It obtained a MODA of 0.569 and an F-score of 0.78, superior to state-of-the-art generalizable detection techniques.

1. Introduction

Pedestrian detection is a relevant problem in several contexts, such as smart cities, surveillance, monitoring, autonomous driving, and robotics. While several solutions focus only on 2D pedestrian detection [7, 15, 14], estimating the 3D location of pedestrians allows georeferencing them in the environment. This referencing enables location-based services, spatial visualization, and others [6]. Nowadays, it is common that environments have multiple monocular cameras with overlapping fields of view, such as security cameras. Using such a setup makes 3D pedestrian detection easier since it can exploit multi-view constraints and better handle occlusions. Nevertheless, multi-camera 3D pedestrian detection in crowded environments is still a challenging task.

The methods that currently obtain the best results for detecting pedestrians in 3D using multiple cameras need to perform training using data from the target scene [10, 1]. This implies that they need to retrain when the target scene changes, with different multi-camera configurations and environment conditions. The training procedure is usually time-demanding and may require laborious annotation of ground-truth data. Due to this, it is desirable to have a generalizable multi-camera 3D pedestrian detection solution that can be applied out-of-the-box without training with target scene data [16, 23].

In this context, we present a novel method for multi-camera 3D pedestrian detection classified as generalizable. The proposed approach encompasses monocular pedestrian detection, estimation of pedestrian location on the ground plane, fusion of multi-camera pedestrian detections, and person re-identification (re-ID). We summarize our approach in Figure 1. Since it is based on off-the-shelf monocular person detectors and person re-ID models, it does not require training using target scene data.

The contributions of this work are:

1. An approach for estimating pedestrian location on the ground plane from off-the-shelf monocular person detectors that provide both human body poses and bounding boxes (Section 3);

2. A technique for fusing multi-camera pedestrian locations by solving an instance of the clique cover problem from graph theory (Section 4);

3. An alternative to take pedestrian appearance into account in this fusion step based on domain-generalizable person re-ID models (Section 5);

4. Quantitative and qualitative evaluations regarding the proposed method's detection performance with respect to state-of-the-art generalizable multi-camera 3D detection approaches (Section 6).

2. Related Work

Monocular 3D Pedestrian Detection. 3D monocular detectors can retrieve pedestrians' 3D location, requiring a single camera. MonoPair [5] uses trained networks to acquire 3D bounding boxes of detected people (among other targets). It then adds pairwise spatial relationships (based on predicted constraints related to the mid keypoint between targets) to improve the resulting location. MonoLoco [2]

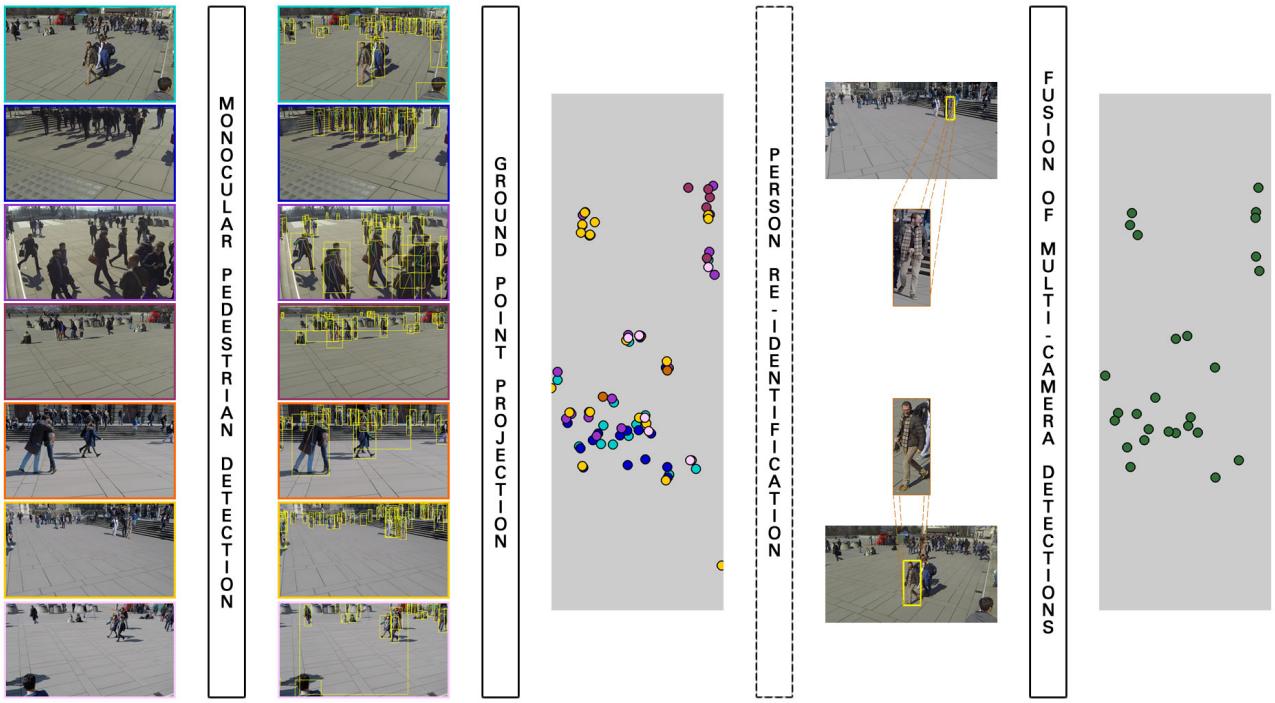

Figure 1. Summary of our generalizable multi-camera 3D pedestrian detection. We detect pedestrians' bounding boxes and body poses in each camera's image (represented by different colors). Then we estimate every detection's ground point on every camera and eliminate those out of the area of interest (grey rectangle). We can compute a descriptor for each pedestrian bounding box aiming at person re-identification. Finally, we solve an instance of the clique cover problem from graph theory to fuse pedestrian detections.

adopts a monocular approach that infers the depth of each person's detected 2D bounding box based on uncertainty estimation. It uses 2D estimated poses to model the ambiguity of the 3D location related to intrinsic characteristics such as the different heights of the tracked population. Hayakawa and Dariush [8] improve MonoLoco's 3D localization approach by introducing an asymmetric loss function. It better handles the pixel's related error of estimated joints from distant pedestrians, achieving improved accuracy in these cases.

Non-Generalizable Multi-View 3D Pedestrian Detection. In contrast with monocular approaches, estimation using multiple cameras has better results in general because it is prone to deal with complex occlusion problems [1]. There are multi-view techniques that consider an additional training (supervised or unsupervised) step with the target scene content to better handle contextual information from the application domain. Given the implicit cost of performing training for each new scene, we can classify them as non-generalizable. In this sense, Baqué et al. [1] present a combination of convolutional neural networks (CNNs) and conditional random fields (CRFs) to handle the pairwise matching ambiguities between the observed pedestrians. Alternatively, MVDet [10] aggregates the multi-view people detection information by applying a feature perspective transform to place all ground heatmaps (and later locations) of pedestrians in the same coordinate space. Similarly, DMCT [21] proposes a perspective-aware network, which produces distorted detection blobs (related to the camera's perspective). This is followed by a fusion procedure for ground-plane occupancy heatmap estimation and the use of a Deep Glimpse Network for person detection.

Generalizable Multi-View 3D Pedestrian Detection. The pipeline to estimate the 3D location of pedestrians in multi-camera scenarios, in a generalizable manner, often employs 2D monocular pedestrian detectors [7, 15, 14] and later fuse their results based on multi-view properties [16, 23, 17]. In this scenario, one way to make 3D pedestrian detection robust to domain shift, therefore generalizable, is to use monocular person detectors that do not need retraining for a specific target domain [19, 9, 12, 7]. Another advantage of monocular detectors re-use is to simplify setup requirements, easing cameras addition/removal/combination [17].

3. Monocular Pedestrian Detection and Ground Point Estimation

Our approach relies on an off-the-shelf monocular pedestrian detector that is not retrained for the target domain. From the monocular detections, we estimate each person's ground point, which is its location on the ground plane. We can then fuse these ground points for estimating the 3D coordinates of pedestrians in the world ground plane.

Some monocular person detectors can provide both bounding boxes and human body poses as output [12]. The use of full-body poses allows better handling of occlusions, making it possible to verify which parts of the pedestrian's body are visible. A widely used representation for a body pose is the one employed by the MSCOCO dataset [13], which consists of 17 keypoints that correspond to human body landmarks. We propose a heuristic for estimating ground points from body poses in the MSCOCO format together with person's bounding boxes.

We only use the ankle keypoints, which are the lower ones among all 17 keypoints. We only keep detections that have scores for ankle keypoints both higher than a threshold t_s. However, the ankles are not on the ground plane, so we need to apply an offset δ to the y coordinate of the ankle keypoints to obtain their correspondences on the ground. This offset is given by

$$\delta = bb_{y_{max}} - max(la_y, ra_y), \tag{1}$$

where $bb_{y_{max}}$ is the y coordinate of the bottom edge of the full-body bounding box, and la_y and ra_y are the y coordinates of left and right ankle keypoints, respectively. We estimate the ground point as the midpoint of the line segment whose endpoints are the offsetted ankle keypoints. We illustrate the proposed ground point estimation heuristic in Figure 2.

Figure 2. Midpoint (in yellow) of the two ankles (in blue). We find the estimated ground point (in green) by adding the δ distance to the the midpoint.

4. Fusion of Multi-Camera Detections

Assuming that the cameras are calibrated, the camera frames are undistorted, and that the ground plane corresponds to the $Z = 0$ plane in world coordinates, we compute for each camera a homography $\mathbf{H}$ that maps the image plane onto the world ground plane. Considering a camera with intrinsic parameters matrix $\mathbf{K}$ and extrinsic parameters matrix $[\mathbf{R}|\mathbf{t}]$, the projection of the world ground point $\mathbf{M} = (X, Y, 0)^T$ onto the image ground point $\mathbf{m} = (x, y)^T$ is given by

$$\begin{aligned} s\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} &= \mathbf{K}[\mathbf{R}^1\mathbf{R}^2\mathbf{R}^3\mathbf{t}]\begin{pmatrix} X \\ Y \\ 0 \\ 1 \end{pmatrix} \\ &= \mathbf{K}[\mathbf{R}^1\mathbf{R}^2\mathbf{t}]\begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \\ &= \mathbf{H}^{-1}\begin{pmatrix} X \\ Y \\ 1 \end{pmatrix}, \end{aligned} \tag{2}$$

where $\mathbf{R}^i$ is the i-th column of $\mathbf{R}$. Then we use the computed homographies to project all ground points from all cameras to the world ground plane. If there is a predefined area of interest in the world ground plane, we can use it to discard world ground points outside this area.

We adopt the same two conditions used by López-Cifuentes et al. [16] for fusing world ground points. All world ground points that correspond to the same pedestrian should have the following properties:

1. Come from different cameras since a pedestrian can only appear once per camera frame;
2. Pairwise Euclidean distances lower than a threshold t_g.

As López-Cifuentes et al. [16], we create a graph representing these conditions. The vertices are the world ground

points, and the edges connect vertices associated with world ground points that satisfy both constraints.

We propose to formulate the problem of finding world ground points to be fused as a clique cover problem[1]. A clique is a subset of vertices of a graph such that every two distinct vertices are connected. Vertices of our graph that belong to a clique represent world ground points that can be fused. A clique cover is a partition of the vertices of the graph into cliques. A minimum clique cover is a clique cover that uses as few cliques as possible. The clique cover problem is the problem of finding a minimum clique cover of a graph. A clique cover of a graph G may be seen as a graph coloring of the complement graph of G.

We use the greedy coloring algorithm [11] for coloring the complement of the graph created from the world ground points. We employ the smallest-last vertex ordering strategy with color interchange described by Kosowski and Manuszewski [11]. This prioritizes the fusion of world ground points belonging to pedestrians seen by many cameras at the same time. Vertices assigned with the same color represent world ground points that meet the fusion criteria. We assume that the cameras have overlapping fields of view, so we discard cliques with only one vertex. This helps to decrease the number of false positives since most of the times a pedestrian should appear in more than one camera. The final 3D coordinate of a detected pedestrian is the arithmetic mean of all world ground points represented by a clique found. Figure 3 illustrates the fusion procedure.

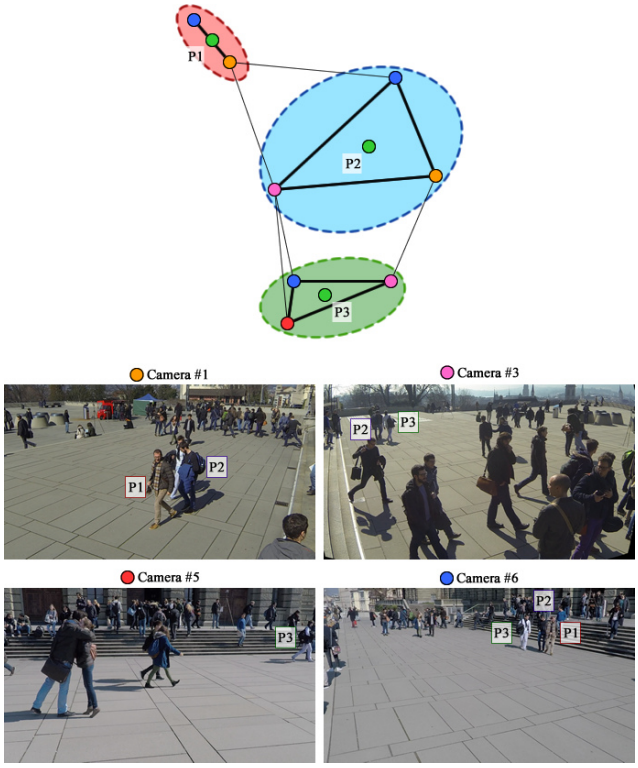

Figure 3. Graph illustrating the detection of three persons. Each node represents the detection by one camera, which can be identified by its color. The strong edges in the graph denote different cliques, which we use to compute the person position (green circles over the graph).

5. Person Re-Identification

So far, we only used the proximity of world ground points for fusing them. As an optional step, we can also exploit appearance cues for guiding the fusion process.

One way to represent pedestrian appearance is by computing discriminative descriptors invariant to viewing direction and background conditions. The person re-ID task tackles this problem. However, person re-ID models often suffer from the domain shift problem, which means that a person re-ID model trained on one source dataset usually presents a degraded performance on an unseen target dataset. In order to keep multi-camera 3D pedestrian detection generalizable, we need to use a domain-generalizable person re-ID model such as OSNet-IAP [20] and OSNet-AIN [22]. These models better handle the domain shift issue so that, once trained, they can be deployed without any retraining.

We propose to also use descriptors provided by a domain-generalizable person re-ID model to help the fusion of world ground points. The person re-ID model takes as input the pedestrian bounding box cropped out of the respective camera frame and outputs a high-dimensional vector as a descriptor. We can then have an additional condition that all world ground points belonging to the same pedestrian should satisfy:

3. Pairwise descriptor distances lower than a threshold t_d.

When using person re-ID, we should take into account all the three criteria presented for creating the fusion graph (following the approach explained in Section 4).

6. Experiments

We evaluated the proposed method under a multi-camera 3D pedestrian detection scenario with a crowded setup. We present in the following subsections the details of the experiments carried out and the results obtained.

6.1. Dataset and Metrics

We used the publicly available and challenging WILDTRACK dataset[2] [3], which was acquired using seven static cameras with overlapping fields of view in a crowded public open area. It provides both intrinsic and extrinsic calibration for each camera and synchronized frames with a

[1] https://en.wikipedia.org/wiki/Clique_cover

[2] https://www.epfl.ch/labs/cvlab/data/data-wildtrack/

1920 × 1080 resolution. Ground-truth 3D locations of pedestrians are available for 400 frames at 2 fps, covering an area of interest of 12 × 36 m, with an average of 23.8 people per frame and a total of 9518 annotations.

We followed the evaluation protocol proposed by Chavdarova et al. [3], which uses the subsequent metrics: Multiple Object Detection Accuracy (MODA), Multiple Object Detection Precision (MODP), precision (Prcn), and recall (Rcll). The 3D detections are assigned to ground truth using Hungarian matching and only if they are closer than $0.5m$. We consider MODA as the primary performance indicator since it takes into account both false negatives and false positives.

6.2. Environment Setup

The hardware used in the evaluations was a laptop with an Intel Core i7-7700HQ @2.80 GHz processor, 32 GB RAM, and an NVIDIA GeForce GTX 1060 graphics card.

We used AlphaPose[3] [12] for human body pose estimation, which employs YOLOv3[4] [18] for person bounding box detection. We performed greedy graph coloring with the algorithm implementation available in the NetworkX[5] package for Python. For person re-ID, we used an OpenVINO[6] model based on the omni-scale network (OSNet) backbone with Linear Context Transform (LCT) blocks [20]. It outputs 256-dimensional descriptors that we compare using the cosine distance.

In all experiments we used empirically obtained threshold values for the proposed method as following: keypoint score threshold $t_s = 0.4$, ground point distance threshold $t_g = 0.7m$ and descriptor distance threshold $t_d = 1.0$. We report the results of competing approaches using the optimal parameter values found.

6.3. Detection Performance Evaluation

First, we evaluated different approaches for monocular pedestrian detection and ground point estimation:

- **BB only**, which uses only the bounding boxes provided by the YOLOv3 detector and estimates ground points as Zhu [23] and López-Cifuentes et al. [16] by taking the center of the bottom edge of the bounding boxes;
- **BP & BB**, which is the proposed approach described in Section 3 based on both body poses and bounding boxes.

We tested both strategies together with the proposed approach detailed in Section 4 for the fusion of multi-camera detections based on the clique cover problem (CC). As can be seen in Table 1, the proposed BP & BB method presented significantly better results than the BB only approach. Figure 4 shows examples of results obtained using BB only + CC and BP & BB + CC. The proposed BP & BB approach presented far fewer false positives than the BB only method for this given frame.

Method	MODA	MODP	Prcn	Rcll
BB only + CC	0.147	0.587	0.560	0.689
BP & BB + CC	**0.569**	**0.673**	**0.808**	**0.746**

Table 1. Performance evaluation of different strategies for monocular pedestrian detection and ground point estimation: using only bounding boxes (BB only) and the proposed approach using both body poses and bounding boxes (BP & BB). We employ the proposed clique cover method (CC) for the fusion of multi-camera detections in both strategies.

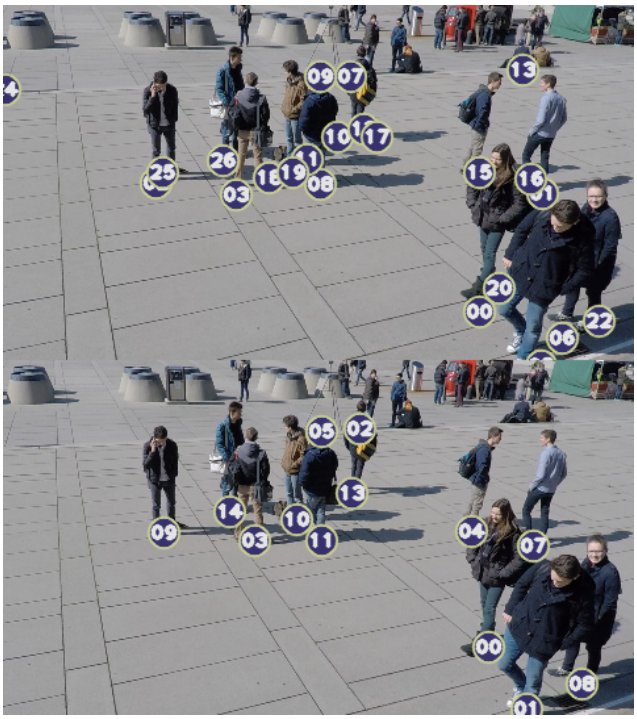

Figure 4. 3D detections projected onto frame #1650 from camera 7 of the WILDTRACK dataset. Top: BB only + CC. Bottom: BP & BB + CC.

In the next experiment, we compared using the BP & BB method together with different strategies for the fusion of multi-camera detections:

- **AH**, which computes an average heatmap as described by You and Jiang [21]. It obtains each camera's heatmaps by considering a non-normalized Gaussian kernel in world ground plane coordinates centered at each ground point with a radius of $0.8m$ and $\sigma = 10.1$.

[3] https://github.com/MVIG-SJTU/AlphaPose
[4] https://pjreddie.com/darknet/yolo/
[5] https://networkx.org/
[6] https://docs.openvinotoolkit.org/

Then it retrieves the detections as local maxima on the average heatmap with a minimum allowed distance of $0.5m$ and minimum value of 0.3;

- **GH**, which is the greedy heuristic presented by Zhu [23] using a distance of $0.8m$ for combining detections;
- **CC**, which is the proposed approach detailed in Section 4 based on the clique cover problem.

Table 2 shows that the proposed CC method obtained the best results with respect to MODA, MODP, and precision. The recall obtained by CC was higher than by AH and slightly lower than by GH. Figure 5 depicts examples of results obtained using BP & BB + AH, BP & BB + GH and BP & BB + CC. While one false negative appears in the presented view for both AH and GH, the proposed CC approach correctly detects all persons visible in this view that are inside the area of interest.

Method	MODA	MODP	Prcn	Rcll
BP & BB + AH	0.262	0.670	0.608	0.736
BP & BB + GH	0.564	0.670	0.802	**0.749**
BP & BB + CC	**0.569**	**0.673**	**0.808**	0.746

Table 2. Performance evaluation of different strategies for fusion of multi-camera detections: using an average heatmap (AH), a greedy heuristic (GH) and the proposed approach based on the clique cover problem (CC). In all strategies, we employ the proposed method for monocular pedestrian detection and ground point estimation using both body poses and bounding boxes (BP & BB).

We also evaluated the effect of using person re-ID as proposed in Section 5. The addition of person re-ID brought almost no changes to MODA, MODP, precision, and recall. Due to this, we show in Table 3 the number of true positives, false positives, and false negatives. It is worth noting that using person re-ID caused a slight increase of false positives. Figure 6 depicts examples of results obtained using BP & BB + CC and BP & BB + CC + Re-ID. The approach that employs re-ID could not correctly connect all graph nodes belonging to the same person. This caused a noticeable shift in the location of detection #14, resulting in one false negative and one false positive.

Table 4 compares the results obtained with the best configuration of the proposed multi-camera 3D pedestrian detection method to state-of-the-art approaches that can be classified as generalizable. The results of RCNN-projected, POM-CNN and Pre-DeepMCD are the ones reported by Chavdarova et al. [3], and the results of López-Cifuentes et al. 2018 [16] and Zhu 2019 [23] are the ones reported in their respective works. Since some methods only reported F-score instead of precision and recall, we also added this metric to the evaluation. Our technique outperformed all other approaches regarding MODA and F-score.

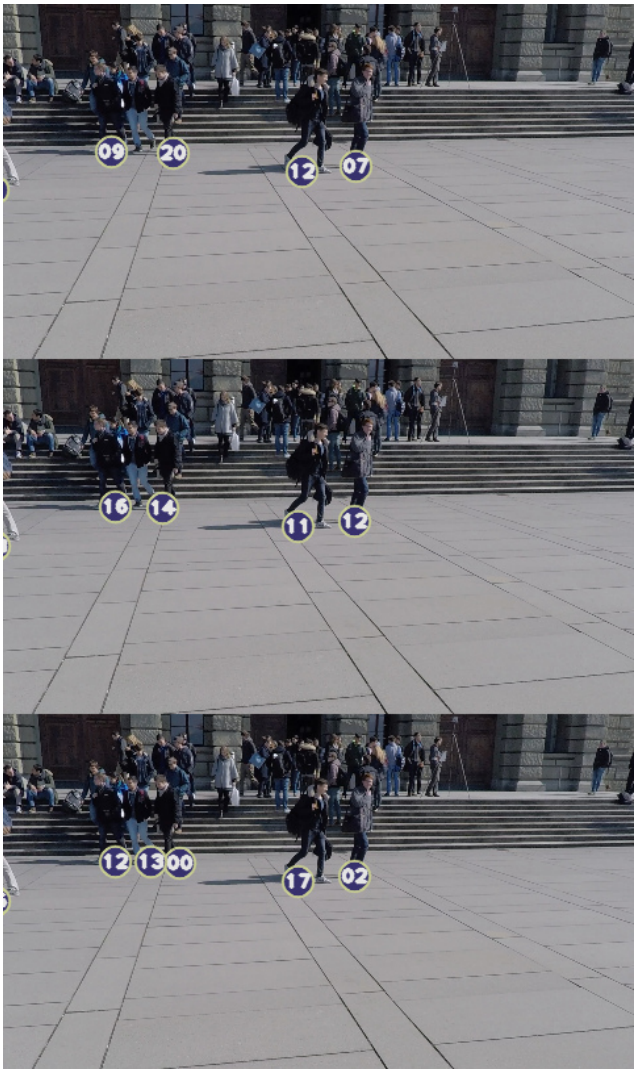

Figure 5. 3D detections projected onto frame #340 from camera 5 of the WILDTRACK dataset. Top: BP & BB + AH. Middle: BP & BB + GH. Bottom: BP & BB + CC.

Method	TP	FP	FN
BP & BB + CC	7096	**1683**	2422
BP & BB + CC + Re-ID	7096	1685	2422

Table 3. Performance evaluation of the proposed multi-camera 3D pedestrian detection method without (BP & BB + CC) and with (BP & BB + CC + Re-ID) the proposed person re-ID approach with respect to number of true positives (TP), false positives (FP) and false negatives (FN).

Figure 8 and supplementary material depict a visualization from all views and from the world ground plane of a result obtained using BP & BB + CC. We can note that the proposed method could correctly detect the 3D locations of

Figure 6. 3D detections projected onto frame #800 from camera 2 of the WILDTRACK dataset. Top: BP & BB + CC. Bottom: BP & BB + CC + Re-ID.

all pedestrians inside the area of interest.

6.4. Execution Time Analysis

Table 5 presents a time performance analysis of a non-optimized version of the proposed approach. Fusion of multi-camera detections is executed only once per frame, while the other procedures are executed once per camera for each frame. The bottleneck is the monocular pedestrian detection and ground point estimation step.

6.5. Limitations

Similar to many other existing multi-camera 3D pedestrian detection methods in the literature [10, 21, 23, 16, 1], our proposed approach is restricted to the ground plane. Therefore, it cannot correctly estimate the 3D location of people who are not standing on the ground (e.g., jumping). This may limit its application to domains such as sports analytics.

Our method also fails when the pedestrian suffers severe occlusions that prevent ankle keypoints from being reliably detected. Figure 7 illustrates such a situation, where AlphaPose was not able to estimate the skeleton of a pedestrian in one of the views and missed the ankle joints in the other two views where he appears. Since we have no ground point for this pedestrian, we end up with a false negative.

Figure 7. Failure case of the proposed method due to severe occlusions. Top row: patches from views of WILDTRACK frame #1990 where non-detected pedestrian #1120 appears, with red circles representing his projected ground truth location. Bottom row: body pose estimation results for each respective image.

7. Conclusion

We presented a new approach for multi-camera 3D pedestrian detection that is generalizable, not requiring scene-dependent training. The proposed method for ground point estimation based on human body poses and bounding boxes proved superior to the commonly used midpoint of the bounding box base. The novel technique for the fusion of ground points as a clique cover problem obtained better results than existing techniques in the literature. The suggested use of a domain-generalizable person re-ID model for giving additional cues to ground point fusion did not

Method	MODA	MODP	Precision	Recall	F-Score
RCNN-projected	0.113	0.184	0.680	0.430	0.53
POM-CNN	0.232	0.305	0.750	0.550	0.63
Pre-DeepMCD	0.334	0.528	0.930	0.360	0.52
López-Cifuentes et al. 2018	0.390	0.550	-	-	0.69
Zhu 2019	0.540	**0.820**	-	-	0.77
BP & BB + CC (ours)	**0.569**	0.673	0.808	0.746	**0.78**

Table 4. Performance comparison of the proposed multi-camera 3D pedestrian detection approach (BP & BB + CC) with state-of-the-art methods not trained on the target dataset.

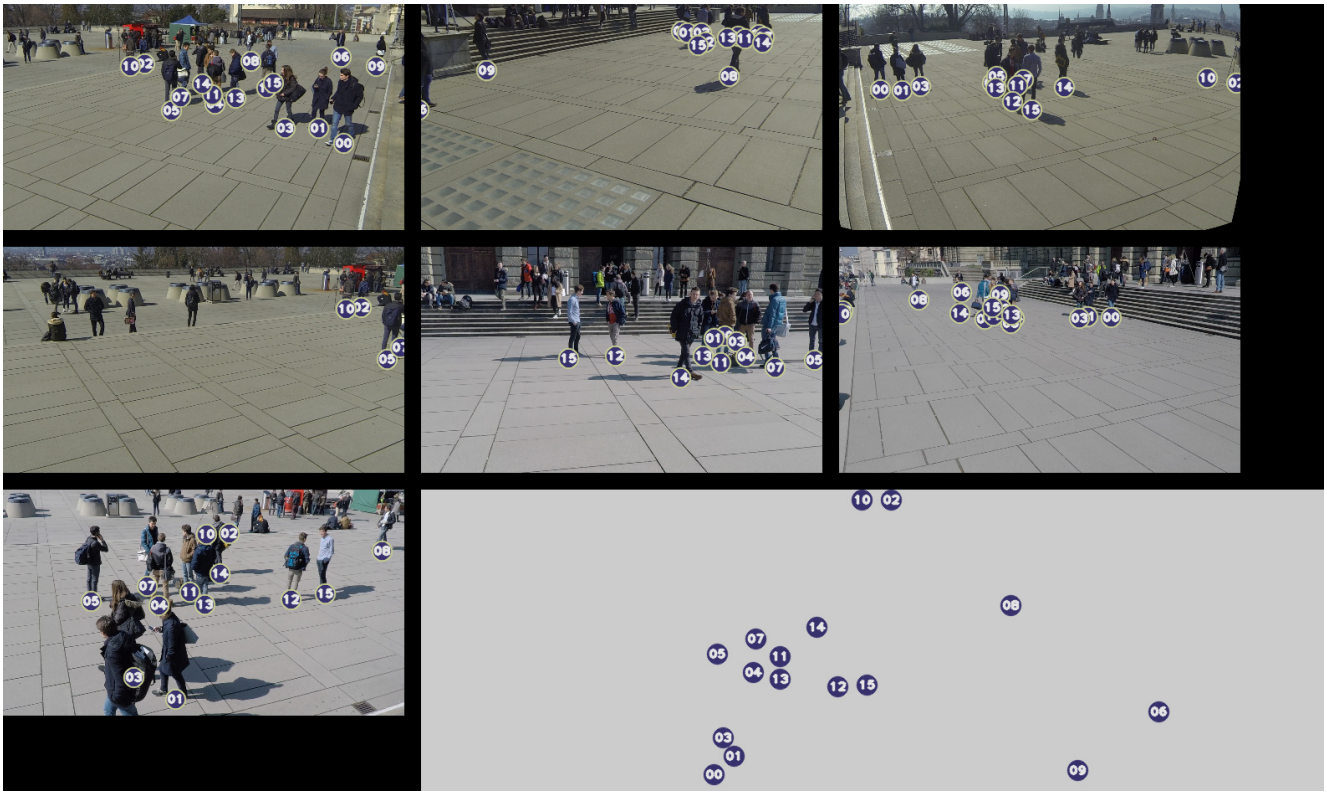

Figure 8. 3D pedestrian detection results obtained with the proposed BP & BB + CC method for frame #1685 of the WILDTRACK dataset. Blue circles represent detected pedestrians. At the bottom right, we show their locations on the world ground plane. From left to right, top to bottom, we see the frames from cameras 1 to 7 with the projection of all 3D detections. Detections with equal number labels refer to the same pedestrian.

Procedure	Time (ms)
Monocular pedestrian detection and ground point estimation (per camera)	829.1 ± 158.3
Ground point projection and area of interest filtering (per camera)	0.1 ± 0.3
Person re-ID descriptors computation (per camera)	186.9 ± 67.6
Fusion of multi-camera detections (per frame)	224.0 ± 71.3

Table 5. Mean and standard deviation of time spent by each procedure of the proposed method for each frame.

bring any improvements. Our approach outperformed state-of-the-art generalizable detection methods.

As future work, we intend to evaluate other heuristics for person re-ID purposes. We also plan to investigate the use of a multi-person 3D pose estimation method [4] as an auxiliary for multi-camera 3D pedestrian detection. We believe this may help to cope with limitations such as restriction to the ground plane and sensitivity to severe occlusions.

Acknowledgements

The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (process 425401/2018-9) for partially funding this research.

References

[1] P. Baqué, F. Fleuret, and P. Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 271–279, 2017. 1, 2, 7

[2] L. Bertoni, S. Kreiss, and A. Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6860–6870, 2019. 1

[3] T. Chavdarova, P. Baqué, S. Bouquet, A. Maksai, C. Jose, T. Bagautdinov, L. Lettry, P. Fua, L. Van Gool, and F. Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018. 4, 5, 6

[4] He Chen, Pengfei Guo, Pengfei Li, Gim Hee Lee, and Gregory Chirikjian. Multi-person 3d pose estimation in crowded

scenes based on multi-view geometry. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 541–557, Cham, 2020. Springer International Publishing. 8

[5] Y. Chen, L. Tai, K. Sun, and M. Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12090–12099, 2020. 1

[6] Andreas Hackeloeer, Klaas Klasing, Jukka M. Krisp, and Liqiu Meng. Georeferencing: a review of methods and applications. *Annals of GIS*, 20(1):61–69, 2014. 1

[7] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. *arXiv preprint arXiv:2003.08799*, 2020. 1, 2

[8] J. Hayakawa and B. Dariush. Recognition and 3d localization of pedestrian actions from monocular video. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7, 2020. 2

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[10] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 1–18, Cham, 2020. Springer International Publishing. 1, 2, 7

[11] Adrian Kosowski and Krzysztof Manuszewski. Classical coloring of graphs. *Contemporary Mathematics*, 352:1–20, 2004. 4

[12] J. Li, C. Wang, H. Zhu, Y. Mao, H. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10855–10864, 2019. 2, 3, 5

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 3

[14] W. Liu, S. Liao, and W. Hu. Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding. *IEEE Transactions on Image Processing*, 29:1413–1425, 2020. 1, 2

[15] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5182–5191, 2019. 1, 2

[16] Alejandro López-Cifuentes, Marcos Escudero-Viñolo, Jesús Bescós, and Pablo Carballeira. Semantic driven multi-camera pedestrian detection. *arXiv preprint arXiv:1812.10779*, 2018. 1, 2, 3, 5, 6, 7

[17] J. Ong, B. T. Vo, B. N. Vo, D. Y. Kim, and S. Nordholm. A bayesian filter for multi-view 3d multi-object tracking with occlusion handling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. 2

[20] Vladislav Sovrasov and Dmitry Sidnev. Building computationally efficient and well-generalizing person re-identification models with metric learning. *arXiv preprint arXiv:2003.07618*, 2020. 4, 5

[21] Quanzeng You and Hao Jiang. Real-time 3d deep multi-camera tracking. *arXiv preprint arXiv:2003.11753*, 2020. 2, 5, 7

[22] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification. *arXiv preprint arXiv:1910.06827*, 2019. 4

[23] Chuting Zhu. Multi-camera people detection and tracking, 2019. 1, 2, 5, 6, 7