Addressing the Occlusion Problem in Multi-Camera People Tracking with Human Pose Estimation

Jeongho Kim* Wooksu Shin* Hancheol Park* Jongwon Baek
Nota Inc., Republic of Korea
{jeongho.kim,wooksu.shin,hancheol.park,jwbaek}@nota.ai

Abstract

Multi-camera people tracking (MCPT) is a challenging task that is crucial for developing intelligent surveillance applications. In this work, we propose an MCPT system for Challenge Track 1 in the 2023 AI City Challenge. Specifically, we address the issue of occlusion, which causes significant changes in a person's appearance and makes it difficult to estimate their exact location on a global map of a given area. In this paper, we present several solutions that utilize human pose estimation for overcoming this challenge. Our experimental results demonstrate that using human pose estimation significantly improves the performance of our system. Furthermore, we achieved promising results on the official evaluation set, with an IDF1 score of 86.76%. Our code is publicly available at https://github.com/nota-github/AIC2023_Track1_Nota.

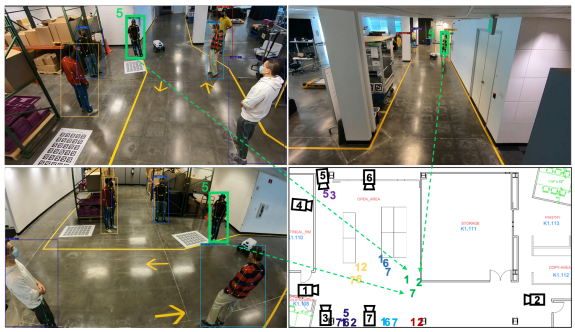

Figure 1. An example of the multi-camera people tracking task. The same person who appears across multiple cameras is represented by numbers of the same color, which are positioned close to each other on the global 2D map. The numbers on the 2D map correspond to the unique identifiers of the cameras. The numbers shown in the other images represent the global ID of the corresponding person.

1. Introduction

Multi-camera people tracking (MCPT) is essential for developing advanced surveillance systems and analyzing human behavior. The aim of an MCPT system is to detect and track people across multiple cameras. As shown in Fig. 1, the system first detects people's positions in each camera using bounding boxes. Then, a single-camera people tracking module produces local tracklets for the identified people within each camera. Finally, the MCPT system matches these tracklets across multiple cameras and assigns global identities to them. Most existing methods typically follow this pipeline [2, 5, 17, 23, 24].

In this paper, we present our MCPT method for Challenge Track 1 of the 2023 AI City Challenge [11]. The goal of this track is to build an MCPT system that works in indoor settings, such as warehouses. The system is evaluated using a combination of real and synthetic datasets. In this track, we also use the conventional pipeline for the MCPT task, but we have focused even more on addressing problems with occluded people to improve the performance of our system.

The occlusion problem poses serious challenges for MCPT systems. Firstly, the similarity score between appearance features from two human objects cannot be accurate for the matching process when a part of the human body is occluded. For instance, comparing appearance features from a full body and only a head would not yield accurate results. In such cases, we must rely more on other information (*e.g.*, such as if two human objects from different cameras are closely located on the 2D map as shown in Fig. 1, they are likely to be the same person). In these situations, appearance information alone may not be sufficient for the matching process.

Secondly, even with positional information on the 2D map, if a part of the body is occluded, the projected coordinates for the person on the 2D map may not be accurate. For example, when projecting the positions of two people captured on CCTV onto a 2D map, if one person's position is projected based on the location of their feet and the other

*These authors contributed equally to this work.

person's position is projected based on the position of their chest due to their lower body being occluded, the positions of the two people on the 2D map will appear farther apart than they actually are (as shown in Fig. 7). This makes it difficult to use location information.

To tackle these challenges, in this work, we propose the use of human pose estimation for MCPT. The output of a pose estimation model enables us to determine which body parts are visible and which are occluded. This can be used to make better decisions about which information to rely on for matching, instead of relying solely on appearance features. Additionally, pose estimation can help estimate the positions of occluded body parts, which improves the accuracy of projection onto a 2D map (see Fig. 7). Our experimental results demonstrate that utilizing human pose estimation significantly improves the performance of our system, with promising results on the official evaluation set achieving an IDF1 score of 86.76%.

The remainder of this paper is organized as follows: in Sec. 2, we review exiting work and highlight the differences from these methods. The proposed methods are described in Sec. 3. We demonstrate the effectiveness of our method in Sec. 4 and conclude the paper in Sec. 5.

2. Related Work

Multi-camera multi-target (MCMT) tracking has been extensively studied in various domains, including intelligent traffic systems [8,16,19,25,26] and human surveillance systems [5, 12, 17, 23, 24]. Regardless of the domain, MCMT tracking systems use the conventional pipeline, which includes object detection, single-camera tracking (SCT), and inter-camera association (ICA).

Deep learning-based detectors, such as Faster R-CNN [14] and You Only Look Once (YOLO) [3, 7, 13, 20], are commonly used for object detection, while DeepSORT [22], ByteTrack [28], and BoT-SORT [1] are popularly used for single-camera tracking. These trackers commonly use appearance features from Re-Identification (ReID) models [10, 21, 27] to compare the currently detected objects and previously detected ones. For ICA, appearance features are also considered to match objects across multiple cameras.

In the context of people tracking, various MCMT tracking methods have been proposed [5, 12, 17, 23, 24]. They commonly highlight the occlusion problem as a significant challenge for MCMT tracking systems. Specifically, they address the issue of detection models failing to detect occluded people in frames, resulting in a failure of single-camera tracking. To overcome this issue, they interpolate detected results from pre and post frames of an image where people detection failed [5, 16, 23]. Other methods use information from other cameras to infer undetected people [12, 24].

In this work, we address the occlusion problem from a different perspective, focusing on the issues that arise during the ICA process. Specifically, we address the problems of inaccurate similarity scores between appearance features and the failure of location estimation of each person in each camera on the global 2D map.

3. Proposed Method

In this Section, we describe our proposed methods. As shown in Fig. 2, our MCPT system contains three modules, namely people detection (Sec. 3.1), single-camera people tracking (SCPT) (Sec. 3.2), and inter-camera association (ICA) (Sec. 3.3).

3.1. People Detection

As the first step of our MCPT system, the people detection module is responsible for predicting the locations of individuals in a given image frame. Accurate detections are crucial to avoid errors in subsequent tracking processes. To reduce detection errors, such as failing to detect a large number of people or capturing non-human objects, we considered state-of-the-art YOLO-based detectors ranging from YOLOv5 to v8 [3,7,20], all of which were pre-trained with the COCO dataset. To select the most accurate model for people detection, we evaluated the performance of the largest model from each YOLO detector on the evaluation set of the COCO dataset, with a focus on people detection only. Our evaluation showed that YOLOv8x6 achieved the highest mean Average Precision (mAP) score for people detection, and thus we selected this model for use in our work.

To ensure that this model can also work on synthetic datasets, we fine-tuned the detector using 52,148 images from the training set and 25,338 images from the validation set of the Challenge Track 1 dataset[1]. These images were randomly selected from the original training and validation sets. This fine-tuned model is only used to evaluate our system on samples containing synthetic videos.

We also observed that detection is not effective when people are located in darker areas in the image as described in the first row of Fig. 3. To address the issue of detection failure, we perform gamma correction as a pre-processing step during the inference phase to make the image brighter. With the gamma correction, input pixels are modified according to the following equation:

$$O = (\frac{I}{255})^{\frac{1}{\gamma}} \cdot 255 \tag{1}$$

where I is an input pixel value and O is the output pixel value. More specifically, when the gamma value is greater than 1.0, darker areas become brighter while originally brighter areas become relatively less bright compared to the dark areas. In this work, the gamma γ is set to 2.0.

[1]https://www.aicitychallenge.org/2023-data-and-evaluation/

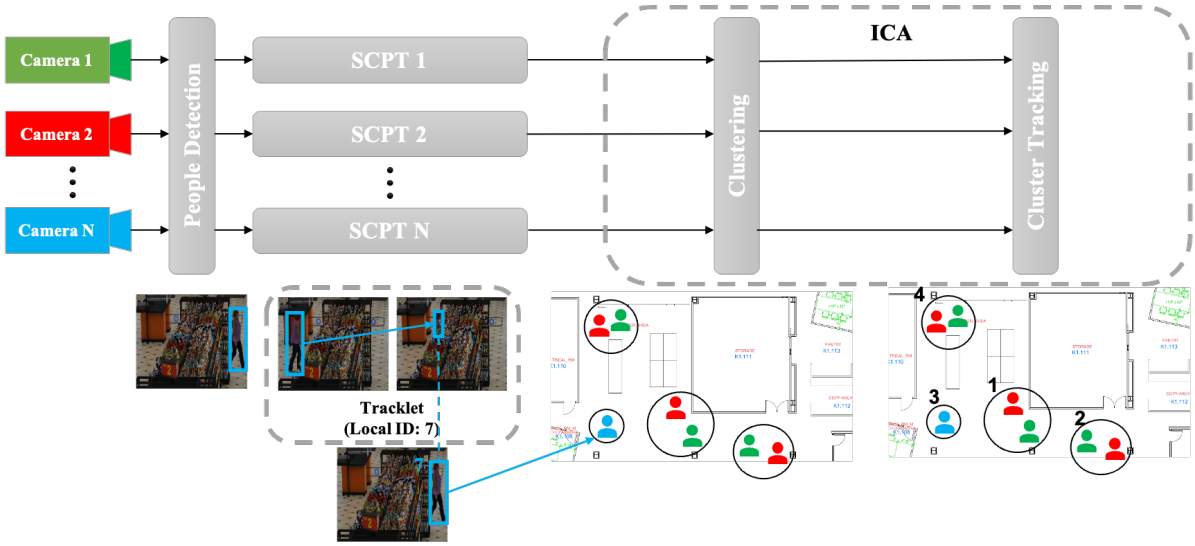

Figure 2. Overview of our proposed method. SCPT and ICA stand for single-camera people tracking and inter-camera association, respectively.

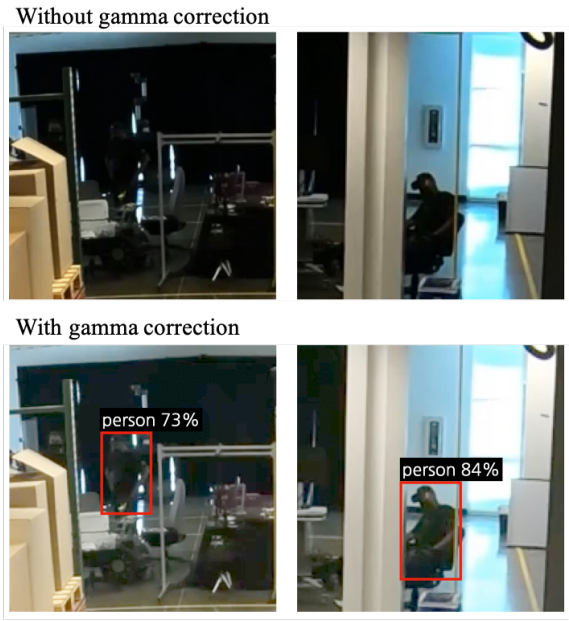

Figure 3. Examples of successfully detecting people with gamma correction

As shown in Fig. 3, in our experiments, we found that this helped our system significantly reduce the number of people that could not be detected.

3.2. Single-Camera People Tracking

The single-camera people tracking module associates people detected at time t with tracklets. In other words, a unique identifier, also known as local tracking ID or local ID, is assigned to each person detected in each camera, as shown in Fig. 2. Any detections that are not associated with existing tracklets are used to initialize new tracklets.

In this work, we use BoT-SORT [1] as a baseline method. BoT-SORT initially associates bounding boxes detected with high confidences (*i.e.*, ≥ 0.6) with tracklets based on the distance scores between them. Two distance metrics are used: Intersection over Union (IoU) ratio between a bounding box that is predicted by the Kalman filter [4] for a tracklet and a detected box at time t, and cosine distance between appearance features of the tracklet and the detected box.

The appearance feature for each bounding box detected by our object detector is extracted using ReID models. In this study, we use three ReID models to extract three appearance feature vectors for each bounding box, and then average them as shown in Fig. 4. This averaging of features was found to be empirically effective in our experiments. For our evaluations on real videos, we used ResNet50-IBN [21], ResNet101-IBN [10], and ResNeSt-50 [27], which were pre-trained with Market1501, DukeMTMC, and MSMT17 datasets, respectively, as ReID models. For the synthetic videos, we fine-tuned the ResNet50-IBN [21] using the training and validation datasets of Challenge Track 1 dataset and used only this model.

The appearance feature of a tracklet is updated in an exponential moving average (EMA) fashion as follows:

$$e_i^t = \alpha e_i^{t-1} + (1 - \alpha) f_i^t \tag{2}$$

where e_i^t is the appearance state of the i-th tracklet at time t and f_i^t is the ReID feature of the matched detection at time t. The α is a momentum term and is set to 0.9. BoT-SORT uses the Hungarian algorithm [6] to associate de-

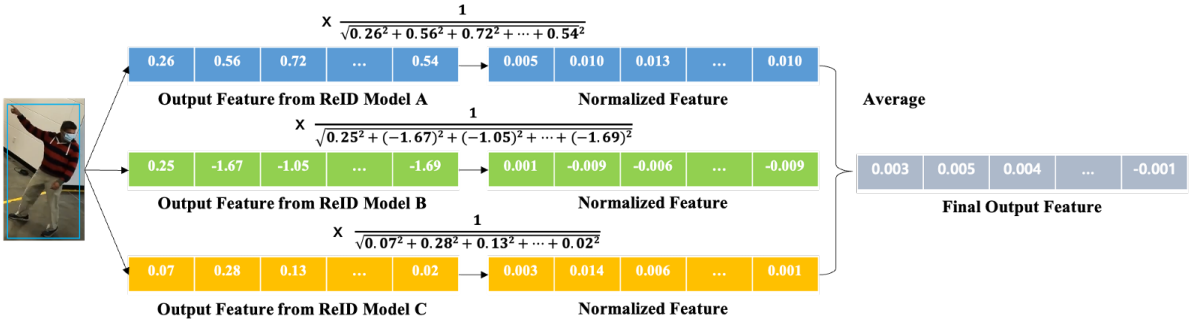

Figure 4. The method of computing an appearance feature vector using ReID models in our work

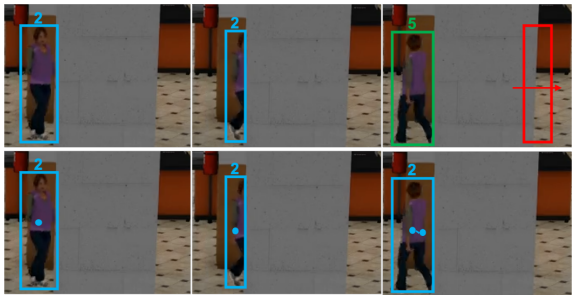

Figure 5. In the first row, an incorrect local ID was assigned as a result of calculating the IoU ratio with the bounding box predicted by the Kalman filter. The red box indicates the bounding box that is predicted by Kalman filter. In the second row, the correct ID was assigned by using the Euclidean distance between the bounding boxes detected at the previous time step and the current time step.

tected bounding boxes at time t with tracklets based on their distance scores.

Afterwards, the BoT-SORT tracker associates bounding boxes with low confidences (*i.e.*, $0.1 \leq$ and ≤ 0.6) with the unmatched tracklets. In this step, BoT-SORT only uses the IoU ratio because low-confidence detections often contain occlusions, which make appearance features inaccurate [28]. After this step, detections with high-confidence (*i.e.*, ≥ 0.7) that still do not match are initialized as new tracklets.

We observed that unlike objects like vehicles, there are many situations where it is difficult to rely on the prediction results of the Kalman filter for humans because their movements are not linear as shown in Fig. 5. Therefore, we use Euclidian distance between the detected bounding box at time t and the last bounding box that is detected before the time t in a tracklet, instead of using IoU ratio.

3.3. Inter-Camera Association

The ultimate goal of our task is to assign a global ID, which corresponds to the local ID assigned to each person detected in a single camera at time t, allowing for tracking of the same object with the same ID across different cameras. This tracking process is performed through two main steps, namely clustering and cluster tracking, as shown in Fig. 6.

In the first step (*i.e.*, clustering), we group together the bounding boxes detected in different cameras at time t for the same individuals. To identify the same person that appears across multiple cameras, we use two pieces of information: location and appearance. Firstly, we use the precalculated homography matrix to project the locations of people captured by each camera onto a 2-dimensional map, which we will refer to as the "global map" from now on (see Fig. 1). This map represents the entire space where the cameras are installed, and the process of projecting locations onto it is known as a projective transformation. If the coordinates of a keypoint in the image frame captured by the camera are (x, y), and the corresponding coordinates in the global map are (x', y'), then the homography matrix is a matrix that transforms (x, y) into (x', y'). We computed the homography matrix for each camera by specifying 4 to 8 key points, where the key points are the bottom center points of specific objects. With the location information on the global 2D map, we can identify the same person across multiple cameras. However, if some parts of a person's body are occluded, their position on the 2D map can be inaccurate, which can cause errors in the clustering step.

As shown in the top image of Fig. 7, if one person's position is projected based on the location of their feet and the other person's position is projected based on the position of their chest due to their lower body being occluded, the positions of the two people on the 2D map will appear farther apart than they actually are. To address this problem, we introduce human pose estimation in this step to infer the correct key points for detected people. To make these estimations, we applied a human pose estimation model to our training and validation sets and discovered the following rules: if only the area from the head to the hip is visible, we found that the location of the accurate key point is approximately two times the height of the observed area; if

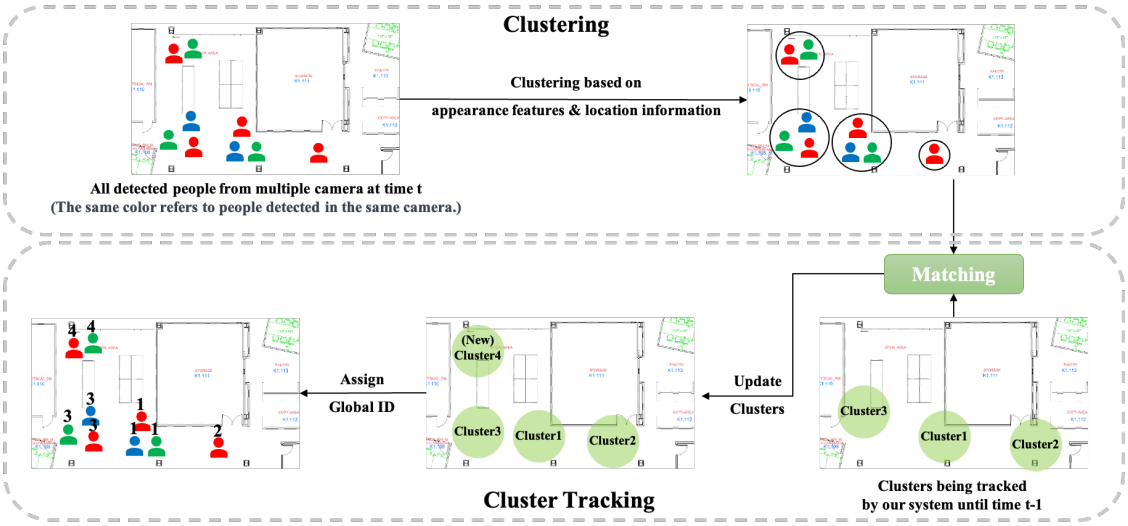

Figure 6. Overview of inter-camera association

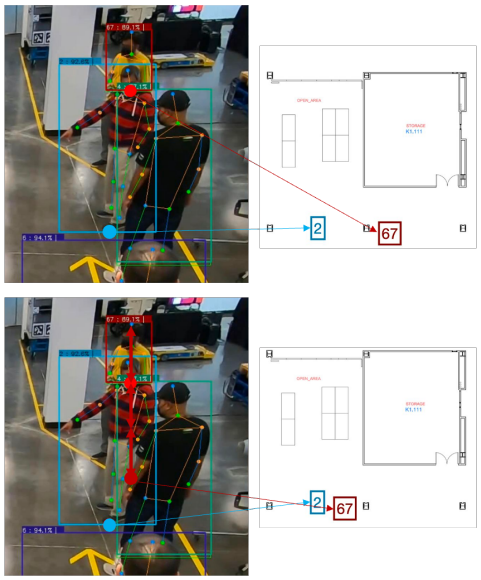

Figure 7. The above image shows a situation where an inaccurate position was estimated on a 2D global map due to the use of incorrect key points caused by occlusion. The below image shows a situation where more accurate position estimation was achieved on the 2D global map by using accurate key points through human pose estimation.

only the area from the head to the elbow is visible, we estimate the correct location to be three times the observed height; if the area up to the shoulder is visible, five times the observed height; and if the area up to the knee is visible, 1.6 times the observed height. In this work, we used HRNet [18] as a human pose estimation model, which was pre-trained with the CrowdPose dataset [9].

Once projective transformation is completed, appearance features of people captured across multiple cameras are extracted using ReID models, as described in the previous Section. To match people detected from two different cameras, we use both location information from the global map and appearance features. We compute a similarity score between person p_i captured in camera i and person p_j captured in camera j as an element of a matrix that is used in the Hungarian algorithm. The similarity score is calculated using the following equation:

$$sim(p_i, p_j) = \lambda cos(p_i, p_j) + (1 - \lambda) dist(p_i, p_j) \quad (3)$$

where $cos(p_i, p_j)$ is the cosine similarity between appearance features of p_i and p_j, and $dist(p_i, p_j)$ is the Euclidean distance between p_i and p_j on the global map. The value of λ is set to (0.5 $\times$ proportional to the number of observed key points compared to all key points used in human pose estimation). In other words, if some parts of the body are occluded, we assign greater importance to the location information rather than appearance similarity.

In the subsequent stage of our ICA module (*i.e.*, cluster tracking), tracklets of clusters are matched to the clusters formed at time t. This means that global IDs defined up to time $t-1$ are assigned to the newly generated clusters. To achieve this, we store N bounding boxes in each tracklet, which are observed at the start of tracking. For real videos, we set N to 50, while for synthetic videos, we set it to 25. In order to match the tracklets and the clusters, appearance features are utilized. We also use human pose estimation to filter out detections whose body parts are occluded from tracklets, as occlusion can result in inaccurate similarity scores between appearance features. The similarity between a cluster and a cluster trackets is calcu-

	# Places	# Cameras	Frames/video
Train	10	58	18,010
Val	5	28	18,010
Test	7	43	51,769∼58,650 (real) 18,010 (synthetic)

Table 1. The statistics of the datasets for Challenge Track 1

lated using the average appearance similarity between all n bounding boxes of the individuals in the cluster and all m bounding boxes in the cluster tracklet. Similarly, matching between clusters and cluster tracklets is executed via the Hungarian algorithm. New global IDs are also assigned to the unmatched clusters.

4. Experiments

4.1. Datasets and Evaluation Metrics

In our experiments, we use the dataset for Challenge Track 1. The training and validation sets consist solely of synthetic videos, whereas the evaluation set contains both synthetic and real videos. The dataset statistics are described in Tab. 1.

The primary metric used in this study is IDF1 [15]. IDF1 measures the ratio of correctly identified detections over the average number of ground-truth and computed detections. As complementary metrics, we also use IDP and IDR, which are reported specifically for our ablation study to demonstrate the effectiveness of using human pose estimation.

4.2. Experimental Results

The effectiveness of using human pose estimation. As described in Tab. 2, incorporating human pose estimation improves the performance of our MCPT system. In particular, using pose estimation for clustering significantly improves the performance by discouraging the system from using inaccurate similarity scores between appearance features.

Comparison with other teams: We submitted the results of our proposed system to the Challenge Track 1 of the 2023 AI City Challenge for official evaluation. Our system achieved an IDF1 score of 86.76% and ranked 10th out of 27 participating teams, as shown in Tab. 3.

5. Conclusion

In this paper, we have presented our multi-camera people tracking system and addressed the issue of occlusion by utilizing human pose estimation. The experimental results indicate that incorporating human pose estimation significantly improves the inter-camera association module. Additionally, our method produces promising results on the official evaluation set. However, introducing a new deep learning model may affect the real-time performance of the system. To tackle this issue in future work, we plan to employ deep compression techniques to reduce the latency caused by heavy deep learning models.

Method	IDF1	IDP	IDR
Baseline	91.89	91.58	92.20
+ For PT	92.27	91.97	92.57
+ For C	95.42	95.04	92.80
+ For CT	95.48	95.10	92.86

Table 2. The results of ablation study on using a human pose estimation. PT, C, and CT stand for projective transformation, clustering, and cluter tracking, respectively.

Rank	Team ID	IDF1
1	6	95.36
2	9	94.17
3	41	93.31
...	...	...
10	**38 (Ours)**	**86.76**
11	47	74.47
12	24	71.22
...	...	...
24	161	13.95
25	172	10.37
26	57	8.69

Table 3. Public leaderboard for the Challenge Track 1

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 2, 3

[2] Rabah Iguernaissi, Djamal Merad, Kheireddine Aziz, and Pierre Drap. People tracking in multi-camera systems: a review. *Multimedia Tools and Applications*, 78:10773–10793, 2019. 1

[3] Glenn Jocher. Yolov8, 2023. https://github.com/ultralytics/ultralytics. 2

[4] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. 3

[5] Philipp Köhl, Andreas Specker, Arne Schumann, and Jürgen Beyerer. The mta dataset for multi target multi camera pedestrian tracking by weighted distance aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 1, 2

[6] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 3

[7] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 2

[8] Fei Li, Zhen Wang, Ding Nie, Shiyi Zhang, Xingqun Jiang, Xingxing Zhao, and Peng Hu. Multi-camera vehicle tracking system for ai city challenge 2022. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3265–3273, 2022. 2

[9] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[10] 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 Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 2, 3

[11] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Sanjita Prajapati, Alice Li, Shangru Li, Krishna Kunadharaju, Shenxin Jiang, and Rama Chellappa. The 7th AI City Challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 1

[12] Duy M. H. Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, and Paul Swoboda. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8866–8875, 2022. 2

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 2

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2

[15] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, pages 17–35, 2016. 6

[16] Andreas Specker, Lucas Florin, Mickael Cormier, and Jurgen Beyerer. Improving multi-target multi-camera tracking by track refinement and completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3199–3209, 2022. 2

[17] Daniel Stadler and Jurgen Beyerer. Modelling ambiguous assignments for multi-person tracking in crowds. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 133–142, 2022. 1, 2

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[19] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Hyung-Joon Jeon, Huy-Hung Nguyen, Hyung-Min Jeon, Tai Huu-Phuong Tran, and Jae Wook Jeon. A robust traffic-aware city-scale multi-camera vehicle tracking of vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3150–3159, 2022. 2

[20] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 2

[21] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia (MM)*, 2018. 2, 3

[22] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, 2017. 2

[23] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4256–4265, 2016. 1, 2

[24] Fan Yang, Shigeyuki Odashima, Sosuke Yamao, Hiroaki Fujimoto, ShoichiMasui, and Shan Jiang. A unified multi-view multi-person tracking framework. *arXiv preprint arXiv:2302.03820*, 2023. 1, 2

[25] Xipeng Yang, Jin Ye, Jincheng Lu, Chenting Gong, Minyue Jiang, Xiangru Lin, Wei Zhang, Xiao Tan, Yingying Li, Xiaoqing Ye, and Errui Ding. Box-grained reranking matching for multi-camera multi-target tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3096–3106, 2022. 2

[26] Hui Yao, Zhizhao Duan, Zhen Xie, Jingbo Chen, Xi Wu, Duo Xu, and Yutao Gao. City-scale multi-camera vehicle tracking based on space-time-appearance features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3310–3318, 2022. 2

[27] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2872–2893, 2022. 2, 3

[28] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 4