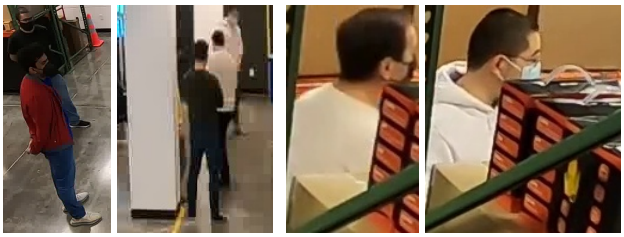
ReidTrack: Reid-only Multi-target Multi-camera Tracking

Andreas Specker[1,2,3] Jürgen Beyerer[2,1,3]
[1]Karlsruhe Institute of Technology
[2]Fraunhofer IOSB
[3]Fraunhofer Center for Machine Learning
{andreas.specker,juergen.beyerer}@iosb.fraunhofer.de

Abstract

Multi-target multi-camera tracking of persons in indoor scenarios such as retail stores or warehouses enables efficient placement of products and improvement of working processes. In this work, we propose the ReidTrack framework, which performs the task solely based on peoples' visual appearances. In theory, accurate person re-identification is able to solve the whole task without the need for additional and complex scene models or post-processing steps. ReidTrack is based on clustering appearance embeddings with a mechanism to avoid identity switches caused by detection bounding boxes showing the body parts of multiple individuals. With only a robust person re-identification model and the real-time detector YOLOv8 and without any auxiliary information, such as complex scene models, our approach ranks fourth concerning Track 1 of the 2023 AI City Challenge.



(a) Overlapping persons (b) Occlusions

(c) Partly visible persons

Figure 1. **Person Re-identification Challenges in Indoor Scenes** – Person-to-person occlusions, heavy occlusions by object, and only partly visible people pose severe challenges for robust person re-id-based tracking in indoor environments.

1. Introduction

The aim of Multi-target Multi-camera Tracking (MTMCT) is to track the routes of multiple targets, *e.g.*, persons, vehicles, or other objects, within a scene surveyed by multiple cameras. The cameras might have overlapping as well as non-overlapping fields of view. MTMCT is a crucial task in many applications. For instance, tracking vehicles' routes in cities enables automatic traffic monitoring and improved signal time planning. Furthermore, indoor cross-camera person tracking systems show huge potential in retail and manufacturing. Retail stores equipped with such systems might gain important insights about their customers, which may help to optimize the locations of products and to improve the routes inside the shops. Concerning the application in manufacturing and warehouses, MTMCT is helpful to enhance efficiency, refine workflows, and reduce the distance to be covered by workers. As a solution for the AI City Challenge 2023, this work deals with person tracking in various indoor scenarios ranging from supermarkets to warehouses.

In general, the task of MTMCT can be divided into two main parts: Multi-target Single-camera Tracking (MTSCT), also known as Multi-object Tracking (MOT), and cross-camera association. MTSCT detects persons in the camera frames and connects the detections in subsequent frames to create single-camera tracklets, *i.e.*, the trajectories of persons within a single camera view. Afterward, cross-camera association links single-camera tracklets seen in different cameras to form multi-camera tracks that represent an individual's movements through the entire camera network. Most MOT approaches in MTMCT systems leverage the *tracking-by-detection* paradigm. Following this procedure, a person detector is applied first to locate the relevant objects in each video frame, and subsequently, these detections are associated if they depict the same person. The association step typically relies on bounding box positions, motion information, and the overall visual appearance. Whereas many works exist that focus on trackers

without the use of visual features, tracking solely based on information about the looks of people is rarely studied, although being the only reliable type of information to re-identify a person after occlusion or to connect the tracklets across cameras with non-overlapping fields of view. To close this research gap and compare the accuracy to other state-of-the-art methods, we follow a person Person re-identification (re-id)-only approach, named ReidTrack, in this work. Person re-id aims at finding further occurrences of a person shown in an image by generating so-called image embeddings encoding the visual appearance of human beings. The similarity of person images can be assessed by computing a distance measure in the learned feature space. Our core idea is to cluster embeddings within one camera view to create single-camera tracklets and cluster them across cameras to obtain multi-camera tracks. Since the goal is to evaluate a re-id-only system, we waive scene models or complex post-processing steps. For instance, we do not make use of any computation-heavy explicit handling of overlapping camera views via, *e.g.*, homographies, or other information about the camera network's topology. Even if, in theory, perfect re-id leads to precise MTMCT, severe challenges must be considered. Examples are shown in Fig. 1. A particular challenge are persons occluding each other, as shown in Fig. 1a. As a result, the body parts of multiple individuals are present in detection bounding boxes. Therefore, extracted visual information might include misleading cues. Moreover, occlusions caused by several obstacles might interrupt single-camera tracklets of persons. Especially in supermarkets, shelves often obscure the view of the cameras. However, in contrast to, *e.g.*, inaccurate motion prediction, re-id can efficiently bridge the occurrence of individuals before and after the occlusion. Nevertheless, partly occluded people are challenging since the re-id feature should not encode the visual appearance of the obstacle but only focus on the person to avoid the later matching of detections based on irrelevant objects. Lastly, associating partly visible persons, *e.g.*, the head, with full-body detections is challenging. Such problems are triggered by the abovementioned challenges or at the camera frame edges. Person re-id approaches must be robust against this and achieve decent accuracy in such cases. We address these challenges by excluding overlapping bounding boxes with misleading information during the first association of detections and by applying data augmentation techniques to obtain robust re-id models.

Our main contributions can be summarized as follows:

- We develop the first re-id-only MTMCT system and prove its efficiency.
- We propose a handling mechanism for person-to-person overlaps since these are a primary error source in re-id-based tracking approaches.
- Our ReidTrack framework achieves competitive results in Track 1 of the 2023 AI City Challenge [36] without using any information about the scene or manually defined scene models.

2. Related Work

2.1. Person Detection

Person detection is a sub-category of object detection because it focuses on a single class. In general, one-stage and two-stage approaches are distinguished. The benefit of one-stage approaches, *e.g.*, SSD [28], RetinaNet [24], or YOLO [40] and its variants [6, 16, 21, 41, 57] is their efficient computation since localization and classification of objects is done at once. In contrast, two-stage methods such as Faster R-CNN [43], Feature Pyramid Networks [23], and Cascade R-CNN [7] generate region candidates first and subsequently classify them in a separate stage. As a result, these methods are usually more accurate, but at the cost of computation time. In recent years, several challenges [34, 35, 37, 68] have been conducted, for which high-ranking participants often employed YOLO-related approaches [17, 27, 29, 48, 55, 56] due to the excellent trade-off between speed and accuracy. We also follow this approach since MTMCT is a complex and computation-heavy task that nevertheless needs many computing resources.

2.2. Person Re-identification

In general, person re-id approaches either learn a single global feature [18, 31, 63], apply implicit attention mechanisms [9, 66], learn multiple features for particular body parts [46], or leverage auxiliary information [26, 49] such as semantic attributes. However, literature, especially concerning real-world application of models, indicates that learning a global feature without any further knowledge or cues achieves state-of-the-art results at a favorable trade-off between accuracy and inference time [30, 32]. As a result, we investigate global feature learning approaches to determine a suitable re-id component for the ReidTrack framework in this work.

2.3. Single-camera Tracking

Most Single-camera Tracking (SCT) methods rely on the tracking-by-detection paradigm, which considers the task two sub-problems [1, 3–5, 12, 42, 45, 51, 53, 58–62, 65]. First, objects are detected and subsequently associated using a similarity measure. Typically, multiple information about the objects is combined, such as the location of objects [4], their motion [3, 65], visual appearance features [1, 5, 12, 51, 53, 58], or body poses [53, 60]. Nonetheless, recent works fuse detection and tracking to increase the focus on the temporal context in videos. Object de-

tectors are extended to full trackers [2, 67], existing tracks are used as prior knowledge for the detection in subsequent frames [14, 67], or 3D CNNs are leveraged to detect entire tracklets directly [38]. Furthermore, the use of transformer-based architectures gains increasingly importance [8,33,52]. However, re-id-only tracking as introduced in this work is rarely studied in the literature so far.

2.4. Multi-camera Tracking

MTMCT system usually consist of a detection, Multi-target Single-camera (MTSC), and cross-camera association stages [20, 22, 44, 54]. In general, methods are distinguished into online and offline approaches. Online approaches [15, 47, 64] solve the task frame-by-frame, whereas offline trackers perform the task as a post-processing step using the output of the single-camera tracking [19, 22, 29, 48, 50, 62]. Since offline trackers achieve higher tracking accuracy, these approaches are typically applied in challenges [34, 35, 37]. Recent advances indicate that prior information about the scene's setup, *e.g.*, the positions and orientations of cameras, is a crucial clue for MTMCT [19, 20, 22, 39, 47, 48, 50]. It enables avoiding infeasible inter-camera associations, dealing with overlapping camera fields of view, or using transition times as complementary information to compute the similarity metric. However, such scene models need much manual supervision and might not be available in real-world situations if, for instance, the cameras are moving. Due to this reason, we solely rely on visual appearance information in this work.

3. Methodology

Given a multi-camera network consisting of C cameras, each video V_c included in the set of video streams $\mathcal{V} = \{V_1, \dots, V_C\}$ is processed separately by the single-camera tracking pipeline shown in Fig. 2. The resulting set of single-camera tracklets $\mathcal{T}^{SC} = \{\mathcal{T}_1^{SC}, \dots, \mathcal{T}_C^{SC}\}$ then serves as input to the multi-camera association stage, which is depicted in Fig. 3. Both core components and their submodules are thoroughly described in the following sections.

3.1. Single-camera Tracking

Fig. 2 introduces the operating principle of the single-camera tracking. As a first step, a detector is applied to locate persons within the video frames. Received detections $\mathcal{D}_c$ are then split into overlapping and non-overlapping detections, *i.e.*, $\mathcal{D}_c^{OV}$ and $\mathcal{D}_c^{NOV}$, respectively. In this case, overlapping detections refer to multiple individuals visible in bounding boxes. The presence of several persons within a bounding box may mislead the person embedding and thus deteriorate performance [48, 50]. Due to this reason, we consider only bounding boxes showing single persons in the first step to avoid association errors such as identity switches. Splitting is done based on the Intersection over Union (IoU) between the detections in a frame. If the IoU exceeds a threshold τ_{OV}, related detections are considered overlapping detections. Whereas re-id features are extracted for all detections, only non-overlapping detections are used for the first association round. To reduce the memory footprint of clustering and to speed up the single-camera tracking process, we include an IoU tracking stage before re-id-based association. Directly clustering the embeddings of all detections might lead to colossal distance matrices of size $|\mathcal{D}_c^{NOV}| \times |\mathcal{D}_c^{NOV}|$, with $|\cdot|$ being the set cardinality. In particular, this would result in infeasible hardware requirements for long videos in which many people occur. The experiments will explore the effects of IoU tracking in more detail. The IoU tracker [4] connects detections between subsequent frames if the overlap is larger than a specified threshold τ_{IoU}. Resulting tracks $\mathcal{T}_c^{IoU}$ are then passed to the re-id-based single-camera clustering. Finally, the previously excluded overlapping detections $\mathcal{D}_c^{OV}$ are assigned to the $\mathcal{T}_c^{C}$ obtained from the clustering step. The detection, feature extraction, association, and overlapping detection assignment components are thoroughly introduced in the following.

3.1.1 Detection

Due to the great complexity of MTMCT approaches and thus long computation times, we have decided to favor a fast single-stage detector over slower but more accurate two-stage models. Specifically, YOLOv8 (YOLOv8)x [21] is utilized as a detector since it offers an appropriate tradeoff between speed and accuracy. Since only training data for the synthetic domain is available (see Sec. 4), we employ two different variants. A COCO [25] pre-trained model is fine-tuned on the challenge's training data for the synthetic scenarios. In contrast, the COCO pre-trained model is applied directly to the test scenario consisting of real videos. Regarding image resolution, the experimental results do not indicate a substantial benefit from using the original Full HD resolution of the videos. So, to further speed up the computation of the ReidTrack framework, video frames are downscaled to 640×480 for detection. Besides, 0.25 is chosen as the confidence threshold, and the non-maximum suppression IoU threshold is set to 0.7 to avoid false positive detections.

3.1.2 Feature extraction

The feature extraction stage generates appearance embeddings for cropped bounding boxes. Embeddings are necessary for single-camera clustering as well as for matching overlapping detections to tracks. So, extraction is performed for non-overlapping $\mathcal{D}_c^{NOV}$ and overlapping $\mathcal{D}_c^{OV}$ detections. The resulting sets of features are referred to as $\mathcal{F}_c^{NOV}$ and $\mathcal{F}_c^{OV}$, respectively. $\mathcal{F}_c^{NOV}$ serve as input to

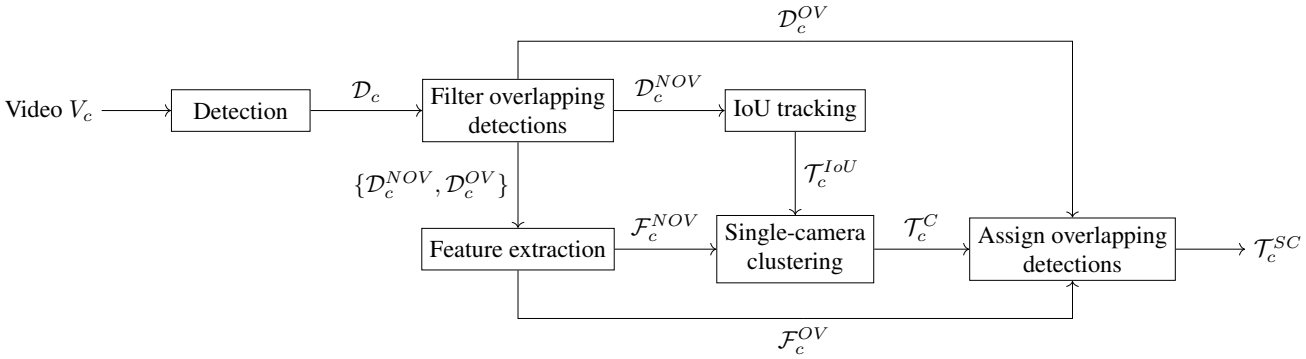


Figure 2. **Single-camera Tracking** – First, a detector localizes the persons in the frames. To avoid identity switches caused by person-to-person occlusions, overlapping detections are filtered, and only non-overlapping detections are used for association using an IoU tracker and clustering. The IoU tracking is not necessary but reduces the hardware requirements. After clustering the visual embeddings to tracklets, the previously discarded overlapping detections are associated.

the clustering-based association and $\mathcal{F}_c^{OV}$ are leveraged for the assignment of overlapping detections. The person re-id model AGW [63] is applied as feature extractor. Similar to the detection stage, we use two variants for real and synthetic data. However, both models are solely trained using synthetic images. The model for the real test scenario is trained with the MOTSynth [13] dataset. More details about the dataset and its pre-processing are given in Sec. 4. Due to the vast amount of available training data, ResNet-101 is chosen as the backbone. Furthermore, AutoAugment [10] is applied as additional data augmentation technique to avoid overfitting and to ensure strong generalization performance. Analogous to detection, the model used for the synthetic test scenarios is fine-tuned on the synthetic AI City training data. Therefore, we construct a training dataset whose composition is described in Sec. 4. Since less diverse imagery is available, ResNet-50 is suitable as the backbone. Larger models are not beneficial and may overfit the training data. Finally, horizontal image flipping is applied as test time augmentation to assure robustness against the walking direction.

3.1.3 Single-camera Clustering

The single-camera clustering stage is the core component. It takes the tracklets obtained by the IoU tracking $\mathcal{T}_c^{IoU}$ and according re-id features $\mathcal{F}_c^{NOV}$ as input and clusters tracklets using the embedding vectors. For this, first mean embeddings are computed for each tracklet in $\mathcal{T}_c^{IoU}$, and then the Cosine distances between all pairs of tracklets are calculated. Based on this distance matrix, tracklets are clustered hierarchically, *i.e.*, the two nearest clusters are merged iteratively until specific conditions are met. In detail, three rounds of clustering are applied that differ concerning the clustering conditions:

1. In the first clustering round, the primary restriction, in addition to the distance threshold τ_{C1}, is that tracklets are not allowed to overlap in time. A person cannot appear in several different places at the same time.

2. The second round omits the time condition since, due to the low detection size, multi-detections of pedestrians occur. Especially in situations where, *e.g.*, the head and the feet are visible due to occlusions caused by shelves, multiple bounding boxes might be detected: one for the head and one for the entire body. As a result, multiple tracklets might overlap in time, although belonging to the same person. Therefore, only a strict threshold τ_{C2} for the visual similarity is applied.

3. Last, short tracklets are associated with long tracklets. Analogous to round 2, the clustering is stopped solely based on an appearance distance threshold τ_{C3}. The threshold is slightly higher than the one in round 2 because the idea is to connect short tracklets to longer ones. For instance, if a person walks through the camera view at the frame borders and is only partly visible for a short period, and thus the visual similarity is different from the longer tracklet of the same individual depicting the entire body.

The resulting merged tracklets $\mathcal{T}_c^C$ are forwarded to the last processing step, assigning the previously discarded overlapping detections to those tracklets.

3.1.4 Assignment of Overlapping Detections

As a last step, overlapping detections $\mathcal{D}_c^{OV}$ need to be assigned to single-camera tracklets $\mathcal{T}_c^C$. Overlapping detections were ignored during clustering to avoid identity switches caused by bounding boxes with body parts of multiple identities. Two mechanisms are applied to assign the detection to tracklets. First, gaps in tracklets are closed

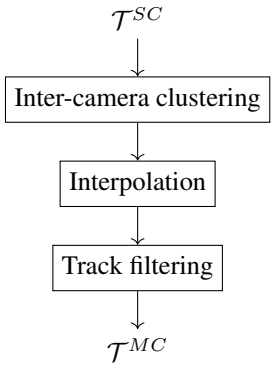


Figure 3. **Cross-camera Association** – The cross-camera association stage of ReidTrack consists of an inter-camera clustering and two post-processing steps.

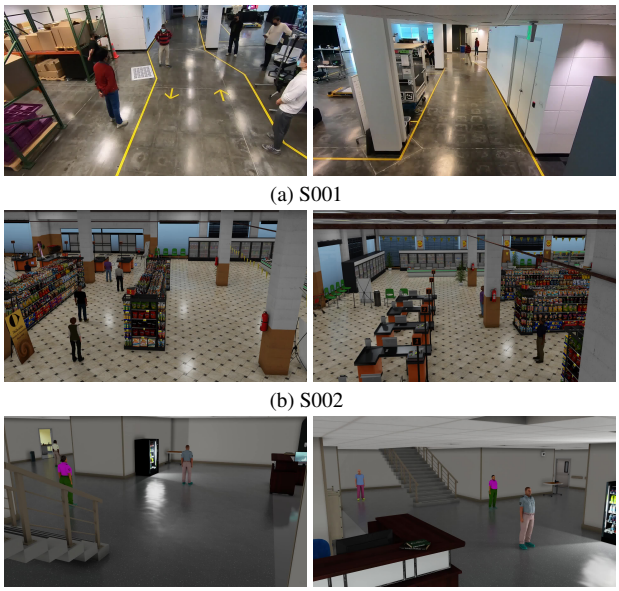
by interpolating the bounding boxes and determining the best matching overlapping detection according to the IoU in each frame. In cases there is no detection with any overlap, no assignment is made. Second, associating overlapping detections with tracklets is performed based on the visual appearance. Overlapping detections are assigned to the tracklet with the closest visual match if a threshold τ_{re-id} is not exceeded, and there is no already existing detection for the same timestamp.

3.2. Cross-camera Association

After tracking persons within the single-camera views, tracklets are associated across cameras to generate multi-camera tracks. The inter-camera association procedure of ReidTrack is visualized in Fig. 3. Analogous to single-camera tracking, hierarchical clustering of the tracklets' mean appearance features is the core component of the association. In the cross-camera case, two rounds of clustering are performed. As a general restriction, each multi-camera track may contain only one tracklet per camera. Furthermore, both rounds of clustering are conducted using the same re-id threshold τ_{MC}. The difference between the clustering rounds is that only tracklets with a minimum length of 10,000 frames are merged during the first round. The idea is to ensure that the longest tracklets are associated if multiple track fragments belong to the same person in a camera. Subsequently, small gaps with a maximum size of 10 frames are bridged by interpolation. Finally, short tracks and tracks that only appear in a single camera are filtered since such tracks most likely are false positives.

4. Experiments

This section introduces the experimental setup and the experimental results. First, the datasets used are presented, followed by the hyperparameters and the discussion of the obtained results.



(a) S001

(b) S002

(c) S013

Figure 4. **AI City Challenge 2023 Dataset** – Selected camera views of three multi-camera scenes of the challenge dataset are shown. The dataset comprises real-world as well as synthetic video data.

4.1. Datasets

Two datasets are used in this work. Besides the AI City Challenge 2023 dataset, the fully-synthetic MOTSynth dataset was leveraged to train the person re-id model for generalization on real data.

4.1.1 AI City Challenge 2023

The challenge dataset for Track 1 of the 2023 AI City Challenge [36] is a large-scale MTMCT that consists of both synthetically generated and real imagery. The synthetic data was created using the NVIDIA Omniverse Platform. In total, 22 different indoor settings are captured with multiple cameras each, from which one test scenario shows real-world data from a warehouse. Cameras are positioned to include overlapping as well as non-overlapping fields of view. Video resolution is Full HD recorded with 30 frames per second. Sample frames from different scenarios are visualized in Fig. 4. The data is split into ten training, five validation, and seven test settings. However, to increase the amount of training data, we only use three scenarios for validation (S013, S017, S020) and extend the training split by the other two validation scenarios.

For training the detection model, every 30th frame is utilized. Concerning person re-id, only persons in every 128th frame are extracted since the diversity in the data is limited by the comparatively low number of different identi-

ties. One bounding box per person and video is randomly selected as a query for validation.

4.1.2 MOTSynth

As mentioned above, the number of identities included in the AI City Challenge is limited, so we argue that the MOTSynth [13] is better suited to generalize to the real-world warehouse setting of the AI City Challenge dataset. Furthermore, the visual appearance of people is more realistic. MOTSynth is a large-scale synthetic dataset created with the engine of the video game GTA V. It provides annotations for several tasks, such as pedestrian detection, re-identification, segmentation, and tracking. More than 1.3M Full HD video frames recorded at 25 frames per second with about 40M annotated bounding boxes are available. For training and evaluating our re-id model, we employ the official MOTSynth training split and the MOT17 [11] training data for evaluation of the generalization ability. Every 60th frame is sampled to create the re-id dataset. Since the MOT17 dataset comprises multiple videos, we report the mean values and standard deviation of evaluation metrics across the videos.

4.2. Training Parameters & Hyperparameters

For fine-tuning our detectors, we apply the default parameters proposed by YOLOv8. We only reduce the learning rate to 1e-3. Concerning person re-id, the parameters from the AGW paper [63] are applied unless otherwise stated.

The hyperparameters of ReidTrack are chosen as follows. The values for the real scene of the AI City test set are given in brackets if they differ. Detections are considered overlapping if the IoU exceeds $\tau_{OV} = 0.6$. As threshold τ_{IoU} for the IoU tracker, the same value is chosen. Since the re-id model was fine-tuned on the AI City dataset, embeddings distances are lower for synthetic scenarios and thus thresholds differ for the second round. In detail, clustering thresholds are set to $\tau_{C1} = 0.9$, $\tau_{C2} = 0.05$ (0.3), and $\tau_{C3} = 0.35$, respectively. The assignment of overlapping detections is performed using 0.07 (0.6) as threshold τ_{re-id}. Finally, inter-camera association parameters are chosen as $\tau_{MC} = 0.05$ (0.5) and $\tau_{LEN} = 10{,}000$.

4.3. Results & Discussion

Detection Detection results are given in Tab. 1. One can observe that higher image resolution is not beneficial on the AI City Challenge dataset. This finding applies to the models with and without fine-tuning (FT) on the challenge data. In general, using COCO-pre-trained models only leads to decent accuracy. However, significant improvements can be achieved by fine-tuning, especially for the YOLOv8x with 640 pixels (px) resolution. Thus, we apply this fine-tuned model as the detector for synthetic test scenarios. The same architecture is used for the real-world warehouse scenario but without fine-tuning.

Model	Size	FT	Precision	Recall	mAP50	mAP50-95
YOLOv8m	640px		95.0	93.9	95.8	74.2
YOLOv8l	640px		95.5	94.2	96.0	74.4
YOLOv8x	640px		95.4	94.1	96.1	74.4
YOLOv8x	1280px		95.1	93.1	95.4	74.3
YOLOv8l	640px	✓	97.9	96.1	98.3	87.9
YOLOv8x	640px	✓	**98.8**	**98.0**	**99.1**	**94.7**
YOLOv8x	1280px	✓	98.1	97.2	99.0	92.5

Table 1. **Detection Results AI City Challenge –** Comparison of YOLOv8 models of different size, image resolution, and with and without fine-tuning on the AI City Challenge data.

Approach	Backbone	R-1	mAP
SBS [18]	ResNet-50	92.8±5.5	85.6±11.8
BOT [32]	ResNet-50	93.4±5.5	86.0±8.2
AGW [63]	ResNet-50	94.7±5.3	85.8±9.5
BOT [32]	ResNet-101	95.0±3.7	86.6±8.9
AGW [63]	ResNet-101	96.1±3.5	86.6±9.3
AGW [63]+AA [10]	ResNet-101	**97.5±4.1**	**88.4±9.9**

Table 2. **Person Re-identification Results MOTSynth/MOT17 –** Person re-id results for training on MOTSynth and evaluating on the training data of MOT17.

Approach	Backbone	R-1	mAP
AGW [63]	ResNet-50	92.3	88.3
AGW [63]	ResNet-101	91.3	86.4
AGW [63] + BS32	ResNet-50	**94.2**	**89.3**

Table 3. **Person Re-identification Results AI City Challenge –** Person re-id results achieved on the validation split of the AI City Challenge dataset.

Person re-id Tab. 2 presents the results for training the re-id models on the synthetic MOTSynth and examining the generalization performance on the MOT17 dataset, which contains real data. The comparison of the three popular baseline models SBS [18], BOT [32], and AGW [63] indicates that the use of the AGW approach is beneficial, in particular with respect to the Rank-1 (R-1) accuracy. This measure is the most important for the MTMCT task since the clustering fuses the most similar tracklets according to the embedding distance, which correlates with the R-1 accuracy. Exchanging the ResNet-50 backbone with the bigger ResNet-101 improves the results. Further experiments found that using AutoAugment (AA) during training further enhances the performance. Finally, an R-1 accuracy of 97.5% and a Mean Average Precision (mAP) score of

Approach	IDF1	IDP	IDR
Ours	**96.0±3.0**	94.7±4.0	**97.5±2.9**
w/o detector fine-tuning	92.2±5.6	91.8±8.2	92.9±3.9
w/o round 2&3 clustering	94.7±4.8	93.3± 5.6	96.1±4.5
w/o overlapping detections assignment	94.6±2.6	**95.1±3.8**	94.2±2.4

Table 4. **Single-camera Tracking Results –** Ablation study for the components of the single-camera tracking of ReidTrack.

Approach	IDF1	IDP	IDR
Ours	97.7±1.6	96.2±3.3	**99.2±0.4**
w/o IoU tracking	**97.7±1.1**	**97.1±2.6**	98.3±0.6

Table 5. **Impact of IoU Tracking –** Single-camera tracking results for the validation scenario S013 of the AI City Challenge dataset with and without IoU tracking. The influence of IoU tracking concerning accuracy is negligible. However, it significantly speeds up the single-camera tracking stage and reduces the hardware requirements of ReidTrack.

88.4% is achieved. Since this is the generalization performance on unseen real data, the model should also perform well on the real scenario of the AI City dataset. We have conducted similar studies on the synthetic scenarios on the AI City dataset. Results for training and evaluation on the AI City data can be found in Tab. 3. Since less training data is available, especially concerning the diversity of peoples' appearances, using the larger ResNet-101 backbone is not advantageous. It leads to worse results according to both metrics. Similar to the previous experiments, there were no significant improvements by adjusting the training schedule, loss parameters, or the addition of further data augmentation methods. The only parameter adaption that led to an apparent increase was reducing the training batch size (BS) from 64 to 32. R-1 could be enhanced by about 2 percentage points and mAP by 1 percentage point, respectively.

Single-camera Tracking An ablation study concerning the components of the single-camera tracking approach is provided in Tab. 4. Without fine-tuning the detector using AI City data, the IDF1 is almost 4 percentage points worse. This finding highlights the importance of an appropriate detector for accurate tracking. The impact of leaving out the last two rounds of clustering or dropping the assignment of overlapping detections on the IDF1 is similar. However, while both IDP and IDR decrease by 1 percentage point with only one clustering step, ignoring overlapping detections deteriorates the recall by more than 4 percentage points. It increases the precision at the same time. This finding indicates that false positive overlapping detections are also merged into the tracks. A possible explanation is that persons are detected even if less than 60% of the body or a person's head is visible, which is the annotation specification of the AI City dataset. Last, Tab. 5 shows the difference between ReidTrack with and without the IoU tracking stage. The scores given were obtained using only the validation scenario S013. The results indicate no impact of IoU tracking concerning the IDF1 apart from a slight increase in standard deviation. While ReidTrack with IoU tracking is advantageous regarding the IDR, skipping this step leads to more precise tracklets. This finding proves the choice of applying IoU tracking to speed up computation and decrease the memory footprint of ReidTrack.

Approach	IDF1	IDP	IDR
Ours	**97.0±1.0**	96.0±0.6	**98.0±1.8**
w/o interpolation	**97.0±1.0**	**96.2±0.5**	97.9±1.8
w/o track filtering	97.0±1.1	95.9±0.6	**98.0±1.8**

Table 6. **Multi-camera Tracking Results –** Ablation study for the components of the multi-camera tracking of ReidTrack.

Module	Computation Time		
	Per frame/person	Per video	S013
Detection	19.2ms	5.8min	28.8min
Feature extraction	12.9ms	1.3min	6.5min
Filter overlapping	–	1.5s	7.5s
IoU tracking	–	0.8s	4.0s
Single-camera clustering	–	1.6s	8.0s
Assign overlapping	–	0.5s	2.5s
Inter-camera clustering	–	–	11.1ms
Interpolation	–	–	6.3s
Filtering	–	–	0.3s
ReidTrack (sequential)	–	–	∼36min
ReidTrack (parallel)	–	–	∼7.5min
Tracking latency	–	–	∼11.0s

Table 7. **Computation Times –** Computation times of the ReidTrack components for scenario S013 of the AI City Challenge dataset. Each of the five videos is 10min long, *i.e.*, in total, 50min of video data is processed. In total, the ReidTrack framework can track persons in the entire scenario in only 7.5min and thus is real-time capable. The latency of multi-camera tracking results behind the incoming video streams amounts to about 11s.

Multi-camera Tracking The results in Tab. 6 indicate only minor effects on the evaluation metrics by the interpolation and the track filtering module. However, we argue that, especially in the real-world test setting, these modules might be advantageous since the domain gap between synthetic and real video data leads to more detection errors, *i.e.*, false positives and false negatives. Interpolation bridges short gaps of missing detections, and track filtering is able to remove false positive tracks of, *e.g.*, background objects.

Computation Time Another critical aspect of tracking systems is the computation time. Due to the significant complexity of several modules, multi-camera tracking approaches are often slow and unable to process multiple

Rank	Team ID	IDF1	Rank	Team ID	IDF1
1	6	95.36	11	163	88.70
2	9	94.17	12	208	88.05
3	41	93.31	13	38	86.76
4	**Ours**	**92.84**	14	85	84.71
5	10	92.33	15	141	83.43
6	113	92.07	16	19	83.26
7	133	91.09	17	72	75.68
8	34	91.04	18	30	74.47
9	82	89.81	19	47	74.47
10	151	89.68	20	48	74.17

Table 8. **Challenge results –** Challenge results on the official test set.

video streams in real time. The computation times of our ReidTrack framework for processing the scene S013 are analyzed in Tab. 7. Each of the five videos lasts for 10min at 30 FPS. The test system includes Intel Xeon 4210R as the CPU, 256GB RAM, and Nvidia RTX 3090 graphics cards. The detector and the re-id feature extractor consume most of the time. However, the processing is still faster than the duration of the videos, which proves the real-time capability. In comparison, the other components only have minor shares. When investigating the whole ReidTrack system's total runtime, measurements for sequential processing show a processing time of 36min for the entire scene consisting of 50min of video data. If single-camera tracking is carried out in parallel, *i.e.*, all cameras are processed simultaneously, only 7.5min are required to obtain the multi-camera tracking results on the test server. Finally, we examine the latency of tracking since the single-camera and multi-camera clustering stages of the proposed approach are offline methods. If the camera streams are processed in parallel in an online manner, computation results are available after approximately 11s in the evaluated scenario. In other words, multi-camera tracking results for the last 10min are available with an 11s delay. In conclusion, these findings highlight the remarkable efficiency of the proposed method.

AI City Challenge Results The public leaderboard results of Track 1 of the 2023 AI City Challenge is given in Tab. 8. Our re-id-based framework named ReidTrack ranks fourth, albeit omitting external knowledge such as a scene model and computation-heavy but more accurate two-stage detectors.

Qualitative Results Finally, two qualitative examples in Fig. 5 prove the effectiveness of excluding overlapping bounding boxes first and associating them after single-camera clustering. Furthermore, it can be observed that the re-id model can correctly associate bounding boxes if only parts of an individual are visible. The examples show selected bounding boxes of two single-camera tracks. In the first example, bounding boxes are correctly assigned even if only a tiny part of the person is visible due to occlusion by another human. Similarly, the second image from the left in the second example is assigned to the proper track. However, only the head and arm can be seen.

(a)

(b)

Figure 5. **Qualitative Results** – The qualitative examples prove the proposed approach's effectiveness in dealing with persons overlapping each other and the matching of partly visible persons. Each example shows selected bounding boxes for a track in camera 2 of the real-world test scenario. The left images show a clear view of the target person, while the others depict challenging bounding boxes, which are nevertheless correctly associated.

5. Conclusion

In this work, we have proposed the ReidTrack framework that solves the MTMCT task solely using person re-id. The core components are single- and multi-camera clustering of extracted re-id embeddings. Furthermore, a simple procedure is proposed to avoid identity switches in single-camera tracklets caused by multiple persons within the same bounding box. Such bounding boxes are ignored first during clustering and associated with tracklets in the final step of the single-camera tracking pipeline. We provide extensive ablation studies of the presented approach and discuss the impact of single components. In summary, ReidTrack achieves an IDF1 score of 92.84%, which equals the fourth position in Track 1 of the AI City Challenge 2023.

References

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

[2] P. Bergmann, T. Meinhardt, and L. Leal-Taixé. Tracking without bells and whistles. In *Int. Conf. Comput. Vis.*, pages

941–951, 2019. 3

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process.*, pages 3464–3468, 2016. 2

[4] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *International Workshop on Traffic and Street Surveillance for Safety and Security at IEEE AVSS 2017*, Lecce, Italy, Aug. 2017. 2, 3

[5] E. Bochinski, T. Senst, and T. Sikora. Extending iou based multi-object tracking by visual information. In *IEEE Int. Conf. Adv. Video Sign. Surv.*, 2018. 2

[6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 2

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[9] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang. Abd-net: Attentive but diverse person re-identification. In *Int. Conf. Comput. Vis.*, pages 8351–8361, 2019. 2

[10] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 4, 6

[11] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129(4):845–881, 2021. 6

[12] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 2

[13] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Aljoša Ošep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *International Conference on Computer Vision (ICCV)*, 2021. 4, 6

[14] C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. In *Int. Conf. Comput. Vis.*, pages 3057–3065, 2017. 3

[15] Bipin Gaikwad and Abhijit Karmakar. Smart surveillance system for real-time multi-person multi-camera tracking at the edge. *Journal of Real-Time Image Processing*, 02 2021. 3

[16] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[17] Synh Viet-Uyen Ha, Nhat Minh Chung, Tien-Cuong Nguyen, and Hung Ngoc Phan. Tiny-pirate: A tiny model with parallelized intelligence for real-time analysis as a traffic counter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128, 2021. 2

[18] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 2, 6

[19] Y. He, J. Han, W. Yu, X. Hong, X. Wei, and Y. Gong. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 576–577, 2020. 3

[20] H.-M. Hsu, T.-W. Huang, G. Wang, J. Cai, Z. Lei, and J.-N. Hwang. Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 416–424, 2019. 3

[21] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 1 2023. 2, 3

[22] P. Köhl, A. Specker, A. Schumann, and J. Beyerer. The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1042–1043, 2020. 3

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[26] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95:151–161, 2019. 2

[27] Chong Liu, Yuqi Zhang, Hao Luo, Jiasheng Tang, Weihua Chen, Xianzhe Xu, Fan Wang, Hao Li, and Yi-Dong Shen. City-scale multi-camera vehicle tracking guided by crossroad zones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4137, 2021. 2

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, et al. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[29] Jincheng Lu, Meng Xia, Xu Gao, Xipeng Yang, Tianran Tao, Hao Meng, Wei Zhang, Xiao Tan, Yifeng Shi, Guanbin Li, et al. Robust and online vehicle counting at crowded intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4002–4008, 2021. 2, 3

[30] Hao Luo, Weihua Chen, Xu Xianzhe, Gu Jianyang, Yuqi Zhang, Chong Liu, Jiang Qiyi, Shuting He, Fan Wang, and

Hao Li. An empirical study of vehicle re-identification on the ai city challenge. In *Proc. CVPR Workshops*, 2021. 2

[31] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[32] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 6

[33] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 3

[34] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Yue Yao, Liang Zheng, Pranamesh Chakraborty, Christian E. Lopez, Anuj Sharma, Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. The 5th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021. 2, 3

[35] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M.-C. Chang, X. Yang, L. Zheng, A. Sharma, R. Chellappa, and P. Chakraborty. The 4th ai city challenge. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 626–627, 2020. 2, 3

[36] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2023. 2, 5

[37] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Qi Feng, Vitaly Ablavsky, Stan Sclaroff, Pranamesh Chakraborty, Alice Li, Shangru Li, and Rama Chellappa. The 6th ai city challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3347–3356, June 2022. 2, 3

[38] B. Pang, Y. Li, Y. Zhang, M. Li, and C. Lu. Tubetk: Adopting tubes to track multi-object in a one-step training model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6308–6318, 2020. 3

[39] Y. Qian, L. Yu, W. Liu, and A. G. Hauptmann. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 588–589, 2020. 3

[40] 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*, pages 779–788, 2016. 2

[41] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[42] Pengfei Ren, Kang Lu, Yu Yang, Yun Yang, Guangze Sun, Wei Wang, Gang Wang, Junliang Cao, Zhifeng Zhao, and Wei Liu. Multi-camera vehicle tracking system based on spatial-temporal filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4213–4219, 2021. 2

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian 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, 2016. 2

[44] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6036–6046, 2018. 3

[45] Kyujin Shim, Sungjoon Yoon, Kangwook Ko, and Changick Kim. Multi-target multi-camera vehicle tracking for city-scale traffic management. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4193–4200, 2021. 2

[46] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body Part-Based Representation Learning for Occluded Person Re-Identification. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV23)*, nov 2023. 2

[47] Andreas Specker and Jürgen Beyerer. Toward accurate online multi-target multi-camera tracking in real-time. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 533–537, 2022. 3

[48] Andreas Specker, Lucas Florin, Mickael Cormier, and Jürgen Beyerer. Improving multi-target multi-camera tracking by track refinement and completion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3198–3208. IEEE, 6/19/2022 - 6/20/2022. 2, 3

[49] A. Specker, A. Schumann, and J. Beyerer. A multitask model for person re-identification and attribute recognition using semantic regions. In *Art. Intell. and Mach. Learn. in Def. Appl.*, 2020. 2

[50] Andreas Specker, Daniel Stadler, Lucas Florin, and Jurgen Beyerer. An occlusion-aware multi-target multi-camera tracking system. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 4173–4182, 2021. 3

[51] D. Stadler, L. W. Sommer, and J. Beyerer. Pas tracker: Position-, appearance- and size-aware multi-object tracking in drone videos. In *Eur. Conf. Comput. Vis. Worksh.*, pages 604–620, 2020. 2

[52] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 3

[53] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3701–3710, 2017. 2

[54] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras

using constrained dominant sets. *arXiv:1706.06196*, 2017. 3

[55] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Huy-Hung Nguyen, Tai Huu-Phuong Tran, Hyung-Joon Jeon, and Jae Wook Jeon. A region-and-trajectory movement matching for multiple turn-counts at road intersection on edge device. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4087–4094, 2021. 2

[56] Vu-Hoang Tran, Le-Hoai-Hieu Dang, Chinh-Nghiep Nguyen, Ngoc-Hoang-Lam Le, Khanh-Phong Bui, Lam-Truong Dam, Quang-Thang Le, and Dinh-Hiep Huynh. Real-time and robust system for counting movement-specific vehicle at crowded intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4228–4235, 2021. 2

[57] 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

[58] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE Int. Conf. Image Process.*, pages 3645–3649, 2017. 2

[59] Minghu Wu, Yeqiang Qian, Chunxiang Wang, and Ming Yang. A multi-camera vehicle tracking system based on city-scale vehicle re-id and spatial-temporal information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4077–4086, 2021. 2

[60] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *Eur. Conf. Comput. Vis.*, pages 472–487, 2018. 2

[61] Kai-Siang Yang, Yu-Kai Chen, Tsai-Shien Chen, Chih-Ting Liu, and Shao-Yi Chien. Tracklet-refined multi-camera tracking based on balanced cross-domain re-identification for vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3983–3992, 2021. 2

[62] Jin Ye, Xipeng Yang, Shuai Kang, Yue He, Weiming Zhang, Leping Huang, Minyue Jiang, Wei Zhang, Yifeng Shi, Meng Xia, et al. A robust mtmc tracking system for ai-city challenge 2021. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4044–4053, 2021. 2, 3

[63] 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*, 2021. 2, 4, 6

[64] Xindi Zhang and Ebroul Izquierdo. Real-time multi-target multi-camera tracking with spatial-temporal information. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. 3

[65] 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. 2022. 2

[66] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen. Relation-aware global attention for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3186–3195, 2020. 2

[67] X. Zhou, V. Koltun, and P. Krähenbühl. Tracking objects as points. In *Eur. Conf. Comput. Vis.*, pages 474–490, 2020. 3

[68] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2