A Robust Traffic-Aware City-Scale Multi-Camera Vehicle Tracking Of Vehicles

Duong Nguyen-Ngoc Tran, Long Hoang Pham, Hyung-Joon Jeon,
Huy-Hung Nguyen, Hyung-Min Jeon, Tai Huu-Phuong Tran, Jae Wook Jeon*

Department of Electrical and Computer Engineering,
Sungkyunkwan University, Suwon, South Korea
{duongtran, phlong, joonjeon, huyhung91, hmjeon, taithp, jwjeon}@skku.edu

Abstract

Multi-Target Multi-Camera Tracking (MTMC) has an immense domain of Intelligent Traffic Surveillance System applications. Multifarious tasks manage to apply MTMC trackings, such as crowd analysis and city-scale traffic management. This paper describes our framework using spatial constraints for the Task of the Track 1 multi-camera vehicle tracking in the 2022 AI City Challenge. The framework includes single-camera detection and tracking, vehicle re-identification, and multi-camera track matching. To improve the system's accuracy, we proposed Region-Aware for the precision of vehicle detection and tracking, leading to the effective service of vehicle re-identification models to extract targets and appearance features. We use Crossing-Aware for a tracker to utilize the rich feature to find the tracklets and operate trajectory matching for multi-camera tracklets connection. Finally, the Inter-Camera Matching generated the global IDs for vehicle trajectory. Our method acquired an IDF1 score of 0.8129 on the AI City 2022 Challenge Track 1 public leaderboard.

1. Introduction

Multi-Target Multi-Camera Tracking aims to determine the position of every vehicle at all times from video streams taken in a multi-camera network. The resulting multi-camera trajectories enable applications including visual surveillance, suspicious activity, and anomaly detection. Therefore, MTMC takes a vital role in a traffic surveillance system.

Most MTMCT methods include the following two tasks. The first task is a generation of local tracklets by tracking each detected target within a single camera. The second task is the cross-camera tracklet matching that matches the local tracklet on all cameras to create a complete trajectory for each target in the entire multi-camera network.

In the traffic surveillance system, cameras are often spaced far apart to reduce costs, and their fields of view do not always overlap. The placement of cameras results in a largely unsolved problem: (1) The background clutter and clogged objects cause errors such as incomplete local tracking results under a single camera. (2) The dramatic change in image and surroundings caused by different viewing angles from different cameras makes cross-camera local tracking matching extremely difficult. (3) The number of cameras on which each target appears and the number of targets in the entire multi-camera network is unknown, and thus it becomes even more difficult to deduce the global trajectory of each target.

Better detectors, data association strategies, or single object tracking could solve the issues of background clutter and object occlusions. However, since there are different fields of view from diverse cameras, it is hard to set the hyperparameter of the vehicle detector or tracker to satisfy the outside environment scenario of the camera. We introduce Region-Aware Vehicle Detection (RW) to improve the detector's precision, which produces high accuracy in detection and even tracking. Since the number of people is typically unknown in advance and the amount of data to process is enormous, we proposed Crossing-Aware Single-Camera (CW) Tracking and Inter-Camera (IC) Matching. CW reduces the unimportant MTMC candidate to reduce the tension for associates with all candidate trajectories of all cameras. IC is the Multi-camera Tracklets Matching with utilizing the feature of trajectory and removing the unnecessary trajectories to enhance the quality and quantitive of the system.

In summary, the main contributions of this paper are summarized as follows:

- We introduce Region-Aware Vehicle Detection (RW) to improve precision, leading to improvements in vehicle detection and tracking.
- We present Crossing-Aware Single-Camera Tracking

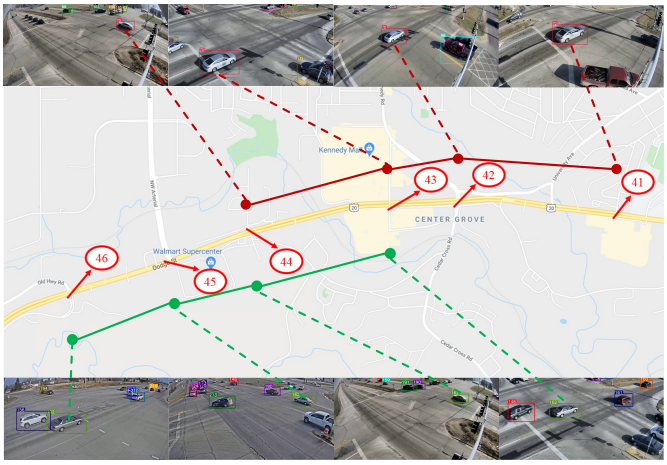


Figure 1. The camera locations, their surrounding roads, and example of result in MTMC framework.

(CW), which helps downsize matching space for visual re-identification and tracklets merging.

- We demonstrate the Inter-Camera Matching (IC), which accurately merges trajectory and removes the outlier.
- The comprehensive experiments show the efficiency of the framework.

The rest of this paper is organized as follows. In Section 2, the related works review some method impact on the framework. The detail of the proposed method is presented in a detailed description in Section 3. In Section 4, the experiments show qualitative and benchmark results of the proposed method. Conclusions are mentioned in Section 5.

2. Related Work

A large amount of literature on person Re-ID and MTMC have attracted growing attention in the past few years. In addition, some works tackle vehicle Re-ID due to smart-city-related systems. This section discusses the most relevant research works to the MTMCT tasks in three parts: Multiple Vehicle Tracking, Vehicle Re-identification, and Trajectory Clustering.

2.1. Multiple Vehicle Tracking

Detection model Moving object detection and identification is one of the most fundamental and challenging problems in multiple object tracking [38,39]. The advent of convolution neural networks and deep neural network architectures has made solving the complicated problem in object detection algorithms more convenient and reliable. It avoids manual feature extraction and uses a data-driven approach that automatically allows machines to learn feature expressions. There are two standard classifications of object detection, two-stage detection and one-stage detection. Two-stage frameworks separate the detection process into the region proposal and the classification stage, and the well-known models are Fast R-CNN [7], Faster R-CNN [30], and Mask R-CNN [8]. At the same time, one-stage detectors handle a single feed-forward fully convolutional network that directly provides the bounding boxes and the object classification, and the widely used models are YOLO [29], SSD [19]. Meanwhile, one of the biggest challenges many object detection methods face is the dilemma between speed performance and accuracy. It finds hard to improve both of them simultaneously. At present, there are some high accurate real-time one-stage anchor-based object detectors YOLOv4 [2], EfficientDet [33]. This paper uses YOLOv5 [15], scaled-YOLOv4 [40], and YOLOR [41] for object detection because it gets the highest accuracy in many benchmarks and has fast convergence.

Tracking model Multiple Object Tracking (MOT) plays a vital role in computer vision. video-based systems [25] use them as the core process. Many MOT studies adopt building as the post-process of detection models based on object detection development. The tracking could be run offline in traffic analysis or online, running real-time processing simultaneously with the camera or video input frame. The model can use detection over the entire frame sequence for the offline methods and then global optimizations, including graph-based and hierarchical methods. The standard offline methods have structure as the graph model, which can be enhanced by using minimum cost flow [42], and subgraph decomposition [34]. On the other hand, the online method is the tracking-by-detection paradigm, which uses only current and previous frames to link detection results per frame or track into longer tracks with spatial and temporal consistency. The challenge of the online method is the feature association between tracking objects and detection results. Therefore, to estimate the match between them, the process could use Kalman Filter based [1]. In this paper, we use the SORT [1], DeepSORT [44], JDE [43], and FairMOT [47] which requires no online training, allowing for fast-speed tracking of objects.

2.2. Vehicle Re-identification

Vehicle Re-Identification (Re-ID) is essential in multi-camera traffic flow for intelligent cities systems to retrieve vehicles that emerge in various surveillance. Re-ID features efficiently work on occlusion and viewpoint changes. Therefore, it plays a vital role in tracklet formation and matching in MTMC. There are some approaches to improve the Re-ID model. First, several loss functions, sampling strategies and samples generation methods have been

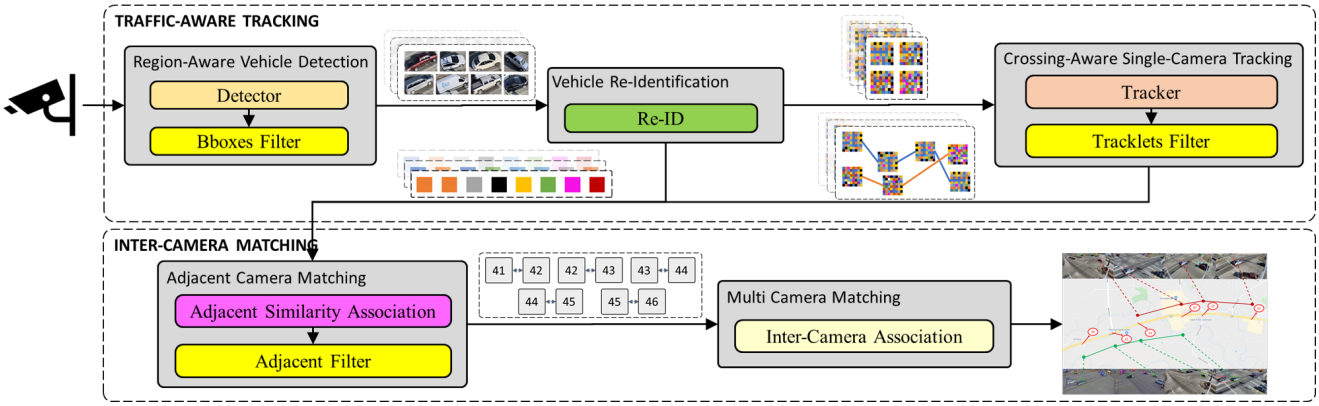
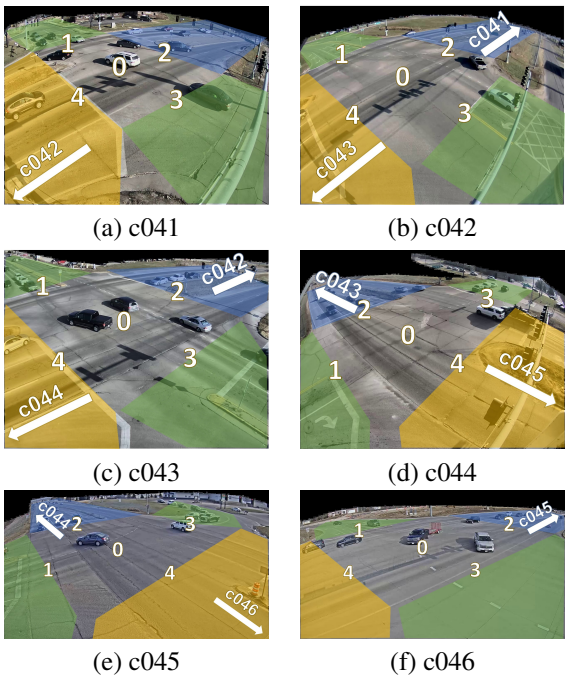


Figure 2. **The pipeline of Multi-Target Multi-Camera**: The MTMC system first runs the detector to acquire the bounding box of the vehicle from each frame of each camera video. After that, we use the Re-ID model to extract the features of the target bounding box, which are fed to the single-camera tracker to induce single-camera tracklets. Finally, we use both the single-camera tracklets and the Re-ID feature for the Inter-camera matching to generate a multi-camera trajectory with IDs.



(a) c041 (b) c042

(c) c043 (d) c044

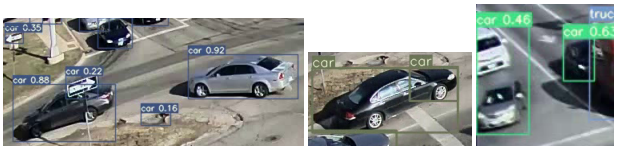
(e) c045 (f) c046

Figure 3. Zone of vehicle multi-camera tracking: the $zone_1$ and $zone_3$ are out of the main road. The $zone_2$ connects to the next camera that has higher ID, and $zone_4$ links to the adjoining camera that has lower ID.

proposed to learn discriminative representations. The well studied person re-identification usually studies loss functions [4], partbased models [32] or unsupervised/semi supervised learning [6]. Second, working with sampling strategies, many useful tricks [17, 24] have been proposed to set strong baselines for the field. In order to learn the robust vehicle representation, many recent works have explored samples generation methods. Last, to enrich the domain of data, the topic has seen multi domain learning [10], largescale datasets [21], synthetic data [48] and so on. With the emergence of transformer-based vision tasks, vehicle re-identification has been greatly improved as in [11]. Due to the promising results of this work [22], we also rely on a global feature learning approach for our vehicle re-ID component.

2.3. Trajectory Clustering

One of the multiple camera multi-tracking methods is a trajectory clustering problem. Many prior works follow this strategy for MTMC. To build a global graph for multiple cameras, graph-based methods [3] establish connection for multiple tracklets in various cameras and optimize for a MTMC solution. By considering spatial-temporal constraints and traffic rules, [13] implement these conditions into the clustering stage, which results in significantly reducing the searching space. Thus, vehicle re-identification accuracy improves significantly. By learning the transition time distribution for each pair of adjacently connected cameras, The methods [36] run well on the same camera distribution of test data and training data. With a different test set without knowing camera allocation, methods [28] observe some basic rules to constrain the matching field. By using sub-clustering in adjacent cameras, the methods [18] match as many trajectories as possible while still ensuring accuracy. The following Sections show the seriated step-by-step of our framework.



(a) (b) (c)

Figure 4. Examples of wrong bounding box in vehicle detection. (a) The image shows wrong position boxes, which cover traffic signs and obstacles on pavement. (b) The image demonstrates overlap of detection. (c) The image illustrates wrong size of a bounding box.

3. Methodology

3.1. MTMC pipeline

The pipeline of MTMC framework is shown in Figure 2, which we modify from our exist structure [37]. There are five steps in the proposed MTMC: (1) Running object detection, obtaining the bounding box from each frame, and applying Region-Aware (RW) for the bounding box filter. (2) Extracting the appearance feature of each bounding box. (3) Applying the Re-ID feature and bounding box to generate Single-Camera Tracking (SCT) results and utilizing Crossing-Aware (CW) for the tracklet filter. (4) Using the feature of each trajectory to generate the Inter-Camera Matching (IC) results. The detailed process will be described in the Sections below.

3.2. Region-Aware Vehicle Detection

3.2.1 Model Detection

Precise vehicle detection is a provision for following vehicle tracking and matching. We evaluated state-of-the-art detection algorithms, Mask R-CNN [8], YOLOv5 [15], scaled-YOLOv4 [40], and YOLOR [41]. All models are pretrained on the COCO dataset and do not introduce external data to the detection. We fetch the bounding box of the detected object in each video frame and the corresponding confidence using the detection model. Table 4 shows the comparison and impact result of each detector.

For more detail, only 3 of 80 categories in the COCO dataset relate to the vehicle, such as cars, trucks, and buses. To prevent multiple definitions of one object, we perform class-agnostic object detection [14] for all vehicle. We extract a detection bounding box for each camera frame for succeeding vehicle tracking and matching:

$$B_i = \{(b_i, t_i) \,|i \in \nu\} \tag{1}$$

where b_i is the corresponding bounding box information, t_i is the time frame, and ν is the length of video. For more information, After getting the detection results, we have a bounding box $b = (x_c, y_c, w, h, \psi)$. $p_c = (x_c, y_c)$ is the position of center point, (w, h) are the width and height of bounding box, and ψ is the confident score. Finally, we apply the bounding box filter to get a higher precise bounding box, which is illustrated the in following Section.

	c041	c042	c043	c044	c045	c046
$zone_1$	0.15	0.1	0.15	0.15	0.15	0.1
$zone_2$	0.1	0.1	0.15	0.15	0.15	0.15
$zone_3$	0.2	0.2	0.2	0.2	0.2	0.2
$zone_4$	0.3	0.3	0.3	0.3	0.2	0.2

Table 1. Vehicle detection threshold for each zone.

3.2.2 Bounding Boxes Filter

The filter removes the wrong bounding box by using the region threshold, fault size and position (such as static traffic signs, utility holes, and traffic lights), and bounding box overlap. These raw bounding boxes may contain false positives, resulting in the tracking task's lower accuracy. For the region threshold, we determine the zone where the bounding box belongs by considering the position of the center point of the bounding box, $p_c \in zone_i$. Based on Table 1, we revise the threshold of the bounding box to determine if the detection result is kept. Moreover, we remove the wrong size and fault position of bounding box (as shown in Fig. 4).

3.3. Vehicle Re-identification

To match the vehicle, we need to compare the feature of each bounding box in a single camera or trajectory in multi-cameras for right matching. We use strong baseline [23, 24] and good re-identification method [22] to train the model and run the extraction feature from the bounding box we cropped. We train models on both the CityFlow dataset [35] and Synthetic Vehicle X dataset [46]. We train the re-id models with Cross-Entropy loss and Triplet loss. The Cross-Entropy loss is formulated as follow:

$$L_{ce} = -\frac{1}{N}\sum_{i=1}^{N} y_i log\left(\hat{y}_i\right)) \tag{2}$$

where y is the actual label, and $\hat{y}_i$ is the classifier's output. The loss is the negative of the first before multiplying by the logarithm of the second. Also, N is the number of examples. The triplet loss can be formulated as:

$$L_{tri} = \sum_{i=1}^{N}\left[\|f_i^a - f_i^p\|_2^2 - \|f_i^a - f_i^n\|_2^2 + \alpha\right]_+ \tag{3}$$

f denotes the learned embedding function applied to all data points, α is the margin of triplet loss, and $[l]_+$ is the same as

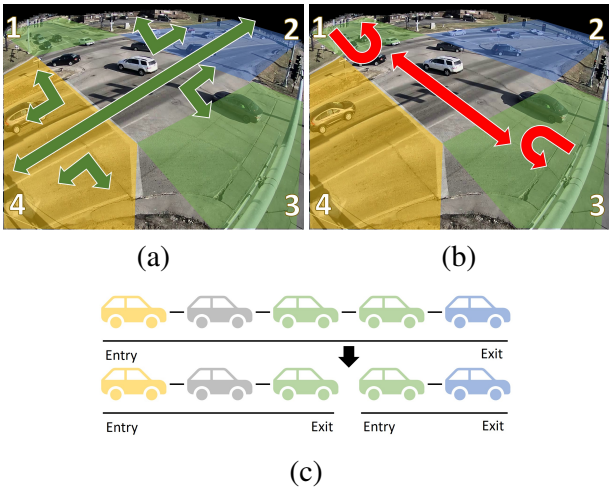


Figure 5. Visualization of the determination of the direction of vehicles. (a) The image shows the correct direction of the trajectory with the green two side arrow. (b) The image demonstrates the incorrect trajectory direction with a red arrow. (c) The example of break motion tracking trajectory into two correct tracklets.

$max(0, l)$. Moreover, f^a, f^p, f^n are the anchor, positive, and negative samples in the feature space, respectively. Finally, we extract a detection bounding box for each camera frame for succeeding vehicle tracking and matching:

$$F_i = \{(b_i, f_i, t_i) \,|i \in \nu\} \tag{4}$$

where f_i is the corresponding Re-ID feature. Table 6 shows the comparison and impact result of each re-identification model.

3.4. Crossing-Aware Single-Camera Tracking

3.4.1 Vehicle Tracking

We use a tracking-by-detection scheme in the video frame to link all detected vehicles to several trajectory candidates for single-camera tracking of multi-vehicle targets in single-camera tracking. Beside using FairMOT [47] as the baseline tracker, we also test several of the single-camera tracking model (SORT [1], DeepSORT [44], JDE [43]), which is a integrated SCT model for detection and tracking paradigm. We modify all the tracker builder and track management parts, namely Kalman Filter [16], into the vehicle tracking version. The modified models use the Re-id feature to construct the trajectory for each vehicle. For further details, as shown in Figure 2, we trim the related target image from the detection results and use the Re-ID model to extract the corresponding vehicle Re-ID features. After that, the trackers use information matrices of the bounding box and vehicle Re-ID features to assign corresponding tracklet IDs with vehicle detection. Finally, the tracker generates a set

	c041 - c042	c042 - c043	c043 - c044	c044 - c045	c045 - c046
$2 \rightarrow 4$	0.7	0.5	0.6	0.5	0.6
$4 \rightarrow 2$	0.5	0.6	0.5	0.5	0.5

Table 2. Clustering threshold for MTMC.

tracklets:

$$T_{id} = \{(b_{id,i}, f_{id,i}, t_{id,i}) \,|i \in \nu\} \tag{5}$$

where, T_{id} is the tracklet corresponding to id. Table 5 shows the comparison and impact result of each tracker.

3.4.2 Tracklets Filter

After getting the tracklets from a tracker, we filter the inconsiderable tracklets from the raw tracklets. The unimportant trajectory belongs to a vehicle with a route that does not connect to the main road (as shown in Fig. 5-b). We only keep the trajectory that has enter entry or exit-entry in the $zone_2$ or $zone_4$ (as shown in Fig. 5-a). However, there is a trajectory with two or more tracklets connections (as shown in Fig. 5-c); we break them into the correct route by using redefined $zone$ in each camera. After obtaining a higher accuracy trajectory by operating CW, the following Section illustrates the matching path from multi-camera.

3.5. Inter-Camera Matching

3.5.1 Adjacent Similarity Association

Adjacent Matching is used to fuse two trajectories of the neighboring camera and cluster between the zones of different cameras. We estimate the similarity between each trajectory between two bordering cameras. We compute trajectory features represented by averaged features of bounding boxes of all frames. We calculate the features by using the bounding boxes that size is bigger than ψ_b. The similarity of tracklets T_i and T_j can be computed using cosine similarity of average feature of trajectory $\bar{F}_i$ and $\bar{F}_j$:

$$s\left(T_i, T_j\right) = \frac{\bar{F}_i \times \bar{F}_j}{\left\|\bar{F}_i\right\| \times \left\|\bar{F}_j\right\|} \tag{6}$$

3.5.2 Adjacent Filter

After calculating feature similarity, we reduce the pressure of clustering in inter-camera association by filtering out the wrong matching of two adjacent tracklets in bordering cameras. First, to determine the connection of the two trajectories, we use the redefined zones, which are the entry and exit entrance of tracklets, to remove unconsiderable pairs of tracklets. For more detailt, the camera $zone_2$ is connected to the $zone_4$ of the next camera, and $zone_4$ is connected

Method	IDF1	IDP	IDR	Precision	Recall
Baseline	65.45	88.98	51.76	91.58	53.27
+RW	76.00	83.07	70.03	86.33	72.78
+CW	79.95	84.03	76.25	87.02	78.96
+IC	81.15	85.34	77.35	87.65	79.45

Table 3. The performance of Traffic-Aware method on the leaderboard.

Detector	IDF1	IDP	IDR	Precision	Recall
Mask R-CNN	79.05	86.71	72.64	89.50	74.98
YOLOv5	81.15	85.34	77.35	87.65	79.45
scaled-YOLOv4	80.78	86.70	75.61	89.12	77.72
YOLOR	80.75	86.93	75.38	89.71	77.79

Table 4. The performance of each Detector on the leaderboard.

to the $zone_2$ of the previous camera (as shown in Fig. 3). Second, by the time frame index, the period of the end time of the first tracklet with the start time of the second one is lower than ψ_t. We only keep the twos that satisfy the condition.

3.5.3 Inter Matching

Inter matching is used for clustering between all connected multi-cameras. It is used to cluster all tracklets in the camera. We use a hierarchical algorithm to produce the global IDs of vehicles for MTMC. For further detail, we use the agglomerative clustering algorithm for matching IDs, and there are different thresholds from $zone_2$ and $zone_4$ back and forth (as shown in Table 2).

4. Experiments

4.1. Implementation Details

The framework has been implemented and tested on RTX A6000 GPUs with 48GB memory and two-thread Intel i9-9900X 3.50GHz. In the detection process, we test YOLOv5 [15], scaled-YOLOv4 [40], and YOLOR [41] models pre-trained on COCO to perform vehicle detection. In the vehicle Re-ID process, we test the combination of several models with [27] to enhance the performance on both tracking and matching. ResNet [9], ResNext [45], ConvNeXt [20] as backbones. All backbone are pre-trained on ImageNet [5]. We use both modified SORT [1], DeepSORT [44], JDE [43], and FairMOT [47] to perform single-camera vehicle tracking in the single-camera tracking process and test which one gives a better result.

Tracker	IDF1	IDP	IDR	Precision	Recall
SORT	77.34	82.75	72.60	85.85	75.33
DeepSORT	80.08	84.30	76.27	86.66	78.40
JDE	81.09	86.96	75.96	89.32	78.01
FairMOT	81.15	85.34	77.35	87.65	79.45

Table 5. The performance of each Tracker on the leaderboard.

Backbone	IDF1	IDP	IDR	Precision	Recall
ResNet	77.71	83.82	72.42	86.90	75.08
ResNeXt	81.09	86.85	76.04	89.24	78.12
ConvNeXt	80.96	86.80	75.85	89.15	77.90
Merge-all	81.29	87.04	76.26	89.37	78.30

Table 6. The performance of each Backbone on the leaderboard.

4.2. Datasets

Track 1 of the AI City Challenge 2022 uses the CityFlow [35] dataset for evaluation and ranking. CityFlow is one of the most prominent and figurative MTMC datasets captured in the actual scene of a United States city. For Track 1, the training set and validation set contain 3.25 hours of traffic video from 40 cameras, which locate at 10 intersections in a city, a length of about 2.5 kilometers. Furthermore, CityFlow contains various road traffic types, including intersections, road extensions, and highways. The test set includes six intersections for the competition and the detection samples from the committee. Moreover, we also use the data of CityFlowV2-ReID and Synthetic Vehicle X dataset [46] to train the Re-ID models.

4.3. Evaluation Metrics

The AI City challenge Track 1 [26] uses IDF1 [12, 31], IDP, and IDR as evaluation indicators, which estimate the trajectory consistency in the camera network, and calculate the ratio of correctly identified vehicles over the average number of ground-truth and predicted vehicles. More specifically, they count the IDF1 score by using the false negative ID (IDFN), false positive ID (IDFP), and true positive ID (IDTP):

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{7}$$

where IDFN, IDTN, and IDTP are defined as follows:

$$IDFN = \sum_{\vartheta} \sum_{t \in T_{\vartheta}} m\left(\vartheta, \varphi_m\left(\vartheta\right), t, \Delta\right) \tag{8}$$

$$IDFP = \sum_{\varphi} \sum_{t \in T_{\varphi}} m\left(\vartheta_m\left(\varphi\right), \varphi, t, \Delta\right) \tag{9}$$

(a) c041 (b) c042 (c) c043

(d) c044 (e) c045 (f) c046

Figure 6. Visualization of vehicle multi-camera tracking results.

$$IDTP = \sum_{\vartheta} len\left(\vartheta\right) - IDFN = \sum_{\varphi} len\left(\varphi\right) - IDFP \tag{10}$$

where the ground truth trajectory is denoted as ϑ, $\varphi_m\left(\vartheta\right)$ stages the best matches of the estimated trajectory for ϑ. φ is trajectory result. $\vartheta_m\left(\varphi\right)$ denotes the best matches of ground truth trajectory for φ. t states as the frame index. Δ is the IOU threshold that determines if computed bounding box matches the ground truth. $m(\cdot)$ represents a mismatch function which is set as 1 if there is a mismatch at t, otherwise, $m(\cdot)$ is set as 0.

4.4. Ablation Study

This section shows the result of the proposed method and the combination of models for each task. Our baseline uses ConvNeXt as the backbone for re-identification, FairMOT for tracking, and YOLOv5 for the detector.

Table 3 shows the effect of using each module separately on the results. We demonstrated the outcomes of the proposed RW, CW, and IC outcomes on performance. The RW increases the score by nearly 9%. CW and IC, respectively, further improve the performance of the model.

Furthermore, Table 4 illustrates our testing with both the given detection result from the committee (Mask R-CNN) and new results from the current state-of-the-art detectors (YOLOv5, scaled-YOLOv4, and YOLOR). As can be seen, all of the outcomes of new detectors have a higher score than the given result, and the YOLOv5 provides the highest IDF1.

Moreover, Table 5 describes the examination by replacing the tracker with SORT, DeepSORT, JDE, and FairMOT. The SORT with using only IoU gets the lowest score. The DeepSORT, JDE, and FairMOT scored 80.08, 81.09, and 81.15.

In addition, Table 6 verifies the influence of different backbone networks (ResNet, ResNeXt, ConvNeXt) on the model. Finally, the best performance was achieved by merging ResNet, ResNeXt, and ConvNeXt.

4.5. Quantitative Result

The final ranking result on the testing sequence is shown in Table 7, where our result is in bold. One week before the leaderboard finalizes, 50% of the test was used to evaluate, and we were on the top. However, last two days and after the finalization, our performance ranking dropped with a closing score of 81.29, which was lower than the first rank with 3.57%. Since it is improbable that our framework overfits the dataset, one potential answer for this reduction is the unbalanced separation of the experimental data.

5. Conclusions

This paper illustrates a robust traffic-aware city-scale multi-camera Vehicle Tracking. The method has been

Rank	Team ID	IDF1
1	28	84.86
2	59	84.37
3	37	83.71
4	50	83.48
5	70	82.51
6	36	82.18
7	10	81.71
8	118	81.66
9	110	81.4
10	**94 (Ours)**	**81.29**

Table 7. **Leaderboard of City-Scale Multi-Camera Vehicle Tracking**. 0.8129 is the final score of Dataset A in 2022 AI City Challenge Track 1.

shown to successfully define the corrected movement direction, which goes through the multi-camera. The proposed method's performance shows its effectiveness and efficiency in determining a vehicle's route in different camera views. We submit results to the AI City 2022 Challenge on Track 1 MTMC tracking contest and get the score that competes against other participant teams on the 2022 challenge leaderboard. We will improve each component further for future work, both in time processing and robustness. For example, we manage typical detection errors, e.g., false positives due to vibrant or reflective background and deficiency of detection in large or small vehicles. Moreover, we will learn a better model for the route that can acclimate to the new scenes and weather circumstances. We will adopt the proposed method to handle online streaming of multiple traffic videos and performance optimization to run the pipeline on edge devices and embedded platforms.

6. Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT). (2020R1A2C3011286)

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, Phoenix, AZ, USA, Sept. 2016. IEEE. 2, 5, 6

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv*, Apr. 2020. 2

[3] Weihua Chen, Lijun Cao, Xiaotang Chen, and Kaiqi Huang. An Equalized Global Graph Model-Based Approach for Multicamera Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11):2367–2381, Nov. 2017. 3

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1320–1329, Honolulu, HI, July 2017. IEEE. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, June 2009. IEEE. 6

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *International Conference on Learning Representations*, 2020. 3

[7] Ross Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, Santiago, Chile, Dec. 2015. IEEE. 2

[8] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, Venice, Oct. 2017. IEEE. 2, 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. 6

[10] Shuting He, Hao Luo, Weihua Chen, Miao Zhang, Yuqi Zhang, Fan Wang, Hao Li, and Wei Jiang. Multi-Domain Learning and Identity Mining for Vehicle Re-Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2485–2493, Seattle, WA, USA, June 2020. IEEE. 3

[11] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based Object Re-Identification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14993–15002, Montreal, QC, Canada, Oct. 2021. IEEE. 3

[12] Hung-Min Hsu, Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang, and Kwang-Ju Kim. Multi-Target Multi-Camera Tracking of Vehicles Using Metadata-Aided Re-ID and Trajectory-Based Camera Link Model. *IEEE Transactions on Image Processing*, 30:5198–5210, 2021. 6

[13] Hung-Min Hsu, Tsung-Wei Huang, Gaoang Wang, Jiarui Cai, Zhichao Lei, and Jenq-Neng Hwang. Multi-Camera Tracking of Vehicles based on Deep Features Re-ID and Trajectory-Based Camera Link Models. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–424, Long Beach, CA, USA, June 2019. 3

[14] Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic Object Detection. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 918–927, Waikoloa, HI, USA, Jan. 2021. IEEE. 4

[15] Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Liu

Changyu, Jiacong Fang, Abhiram V, Laughing, Tkianai, YxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, Imyhxy, Lorenzo Mammana, AlexWang1900, Cristi Fati, Diego Montes, Jan Hajek, Laurentiu Diaconu, Mai Thanh Minh, Marc, Albinxavi, Fatih, Oleg, and Wanghaoyang0106. ultralytics/yolov5, Oct. 2021. 2, 4, 6

[16] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, Mar. 1960. 5

[17] Chong Liu, Xiaojun Chang, and Yi-Dong Shen. Unity Style Transfer for Person Re-Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6886–6895, Seattle, WA, USA, June 2020. IEEE. 3

[18] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4124–4132, Nashville, TN, USA, June 2021. IEEE. 3

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, volume 9905, pages 21–37. Springer International Publishing, Cham, 2016. 2

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *arXiv*, Mar. 2022. 6

[21] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Lingyu Duan. VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3230–3238, Long Beach, CA, USA, June 2019. IEEE. 3

[22] Hao Luo, Weihua Chen, Xianzhe Xu, Jianyang Gu, Yuqi Zhang, Chong Liu, Yiqi Jiang, Shuting He, Fan Wang, and Hao Li. An Empirical Study of Vehicle Re-Identification on the AI City Challenge. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2021. 3, 4

[23] 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*, Long Beach, CA, USA, June 2019. 4

[24] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, pages 1–1, 2019. 3, 4

[25] Brendan Tran Morris, Cuong Tran, George Scora, Mohan Manubhai Trivedi, and Matthew J. Barth. Real-time video-based traffic measurement and visualization system for energy/emissions. 13(4):1667–1678. 2

[26] Milind Naphade, Shuo Wang, David C Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Sharifur Rahman, et al. The 6th ai city challenge. In *CVPR Workshop*, New Orleans, LA, USA, 2022. 6

[27] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In *Computer Vision – ECCV 2018*, pages 484–500. 2018. 6

[28] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G. Hauptmann. ELECTRICITY: An Efficient Multi-camera Vehicle Tracking System for Intelligent City. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2511–2519, Seattle, WA, USA, June 2020. IEEE. 3

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv*, May 2016. 2

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv*, Jan. 2016. 2

[31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In *Computer Vision – ECCV 2016 Workshops*, volume 9914, pages 17–35. Springer International Publishing, 2016. 6

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *Computer Vision – ECCV 2018*, volume 11208, pages 501–518. Springer International Publishing, Cham, 2018. 3

[33] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10778–10787, Seattle, WA, USA, June 2020. IEEE. 2

[34] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5033–5041, Boston, MA, USA, June 2015. IEEE. 2

[35] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4, 6

[36] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-Camera and Inter-Camera Vehicle Tracking and 3D Speed Estimation Based on Fusion of Visual and Semantic Features. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 108–1087, Salt Lake City, UT, USA, June 2018. IEEE. 3

[37] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4082–4089, Nashville, TN, USA, June 2021. IEEE. 4

[38] Duong Nguyen-Ngoc Tran, Long Hoang Pham, Ha Manh Tran, and Synh Viet-Uyen Ha. Scene recognition in traf-

fic surveillance system using Neural Network and probabilistic model. In *2017 International Conference on System Science and Engineering (ICSSE)*, pages 226–230, Ho Chi Minh City, Vietnam, July 2017. IEEE. 2

[39] Synh Viet-Uyen Ha, Duong Nguyen-Ngoc Tran, Tien Phuoc Nguyen, and Son Vu-Truong Dao. High variation removal for background subtraction in traffic surveillance systems. *IET Computer Vision*, 12(8):1163–1170, Dec. 2018. 2

[40] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling Cross Stage Partial Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13029–13038, June 2021. 2, 4, 6

[41] Chien-Yao Wang, I.-Hau Yeh, and Hong-Yuan Mark Liao. You Only Learn One Representation: Unified Network for Multiple Tasks. *arXiv*, May 2021. 2, 4, 6

[42] Xinchao Wang, Engin Turetken, Francois Fleuret, and Pascal Fua. Tracking interacting objects using intertwined flows. 38(11):2312–2326. 2

[43] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards Real-Time Multi-Object Tracking. In *Computer Vision – ECCV 2020*, volume 12356, pages 107–122. Springer International Publishing, Cham, 2020. 2, 5, 6

[44] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649, Beijing, Sept. 2017. IEEE. 2, 5, 6

[45] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995. IEEE, July 2017. 6

[46] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *ECCV*, 2020. 4, 6

[47] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *arXiv*, Sept. 2020. 2, 5, 6

[48] Zhedong Zheng, Minyue Jiang, Zhigang Wang, Jian Wang, Zechen Bai, Xuanmeng Zhang, Xin Yu, Xiao Tan, Yi Yang, Shilei Wen, and Errui Ding. Going Beyond Real Data: A Robust Visual Representation for Vehicle Re-identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2550–2558, Seattle, WA, USA, June 2020. IEEE. 3