A Multi-Camera Vehicle Tracking System based on City-Scale Vehicle Re-ID and Spatial-Temporal Information

Minghu Wu[1], Yeqiang Qian[2], Chunxiang Wang[1], and Ming Yang[1]
[1]Department of Automation, Shanghai JiaoTong University
[2]University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University
wuminghu@sjtu.edu.cn, qianyeqiang@sjtu.edu.cn, wangcx@sjtu.edu.cn, mingyang@sjtu.edu.cn

Abstract

With the demands of the intelligent city and city-scale traffic management, city-scale multi-camera vehicle tracking (MCVT) has become a vital problem. The MCVT is challenging due to frequent occlusion, similar vehicle models, significant feature variation by different lighting conditions, and viewing perspective in different cameras. This paper proposes an MCVT system composed of single-camera tracking (SCT), vehicle re-identification (Re-ID), and multi-camera tracks matching (MCTM). In the SCT phase, we designed a tracker update strategy and used the Re-ID model in advance. We also adopted a template matching method to re-associate the discontinuous tracklets. As for vehicle Re-ID, we implemented a spatial attention mechanism based on the background model. Then we fully leveraged the labels of synthetic data to train attributes Re-ID models as the attributes features extractor. Finally, we proposed an MCTM method to leverage tracklets representation and spatial-temporal information efficiently. Our system is evaluated both on the City-Scale Multi-Camera Vehicle Re-Identification task (Track 2) and City-Scale Multi-Camera Vehicle Tracking task (Track 3) at the AI City Challenge. Our vehicle Re-ID method has achieved 3rd place of Track 2, with an mAP score of 66.50%, and achieved state-of-the-art results on the VeRi776 dataset. Our MCVT system has achieved 3rd place, yielding 76.51% IDF_1 of Track 3. Experimental results demonstrate that our system has achieved competitive performance for city-scale traffic management.

1. Introduction

Predicting and analyzing large-scale traffic flow is necessary for improving city-level traffic management. Tracking vehicles means integrating spatial and temporal information of the traffic flow. Therefore, the city-scale multi-camera vehicle tracking (MCVT) system is attracting growing attention. As shown in Figure 1, the MCVT system extracts the vehicle trajectory going through a large area from the cameras at different locations. The main components of an MCVT system include single-camera tracking (SCT), vehicle re-identification (Re-ID), and multi-camera tracklets matching (MCTM). Different from classical single-camera multiple object tracking (MOT), MCVT matches the tracklets of an identical vehicle in different cameras, which may have overlapping or non-overlapping field of view (FOV), and generates one complete global trajectory. There are two major challenges the MCVT task is faced with. Firstly, frequent noise in detection and heavy occlusion make the tracking prone to lose or mismatch during SCT process. Secondly, similar vehicle models and appearance feature variations by illumination and perspective both disturb the Re-ID process. The instability in perception and the confusable figure of cars consequentially degrade the performance of the MCVT system.

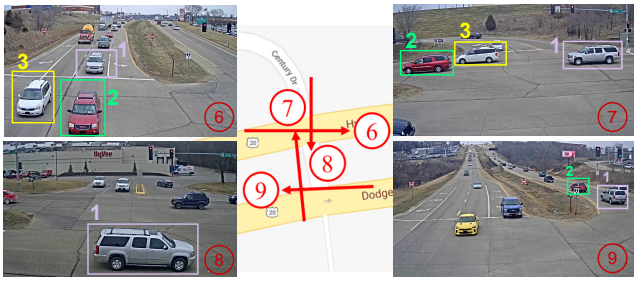


Figure 1. Overview of city-scale multi-camera vehicle tracking and Re-ID. The number near each bounding box denotes the vehicle ID, while the number inside the red circle denotes the camera ID.

The tracking-by-detection strategy with data association algorithms is widely applied to link detection outputs across frames for the SCT step [1, 46, 42]. However, the tracking performance of the existing methods shrinks when the speed or heading of the vehicles changes acutely, or mutual occlusion between vehicles happens. For vehicle Re-

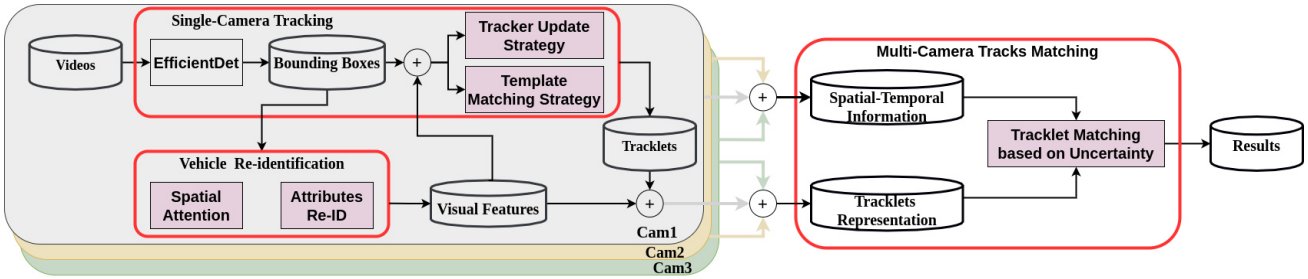


Figure 2. Framework of our MCVT system. The MCVT system is composed of three modules, a single-camera tracking (SCT), a vehicle re-identification(Re-ID), and a multi-camera tracklets matching (MCTM) .

ID, the tendency is to focus on exploiting metric learning and generating discriminant visual features from convolutional neural networks (CNN) [39, 26, 34, 2]. However, these Re-ID models are likely to be confused by similar vehicle models and background, variety in vehicle orientation. To tackle MCTM task, Law *et al.* [21] and Kim *et al.* [19] combine video information with GPS location to match the tracklets from different cameras which have the overlapping FOV. For different cameras with non-overlapping FOV, some methods [29, 3] match tracklets by building the distance matrix using the vehicle appearance feature. These methods depend on a distance threshold to decide whether two tracklets are matching. However, due to the fact that the vehicle has significant appearance feature variation in different cameras, it is impossible to assume that the tracklets pairs with the smallest feature distances are true positives.

In this paper, we constructed a MCVT system based on city-scale vehicle re-identification and spatial-temporal information. The framework of our MCVT system is shown in Figure 2. In the SCT phase, following the tracking-by-detection paradigm, we adopted EfficientDet [38] as our vehicle detector for effectiveness. To overcome occlusion interference, we utilized our Re-ID model to extract the discriminative appearance feature of vehicles. We also designed a tracker update strategy and template matching strategy to overcome the challenges that the motion of vehicles frequently changes when close to camera or vehicle turns.

As for vehicle Re-ID, to improve the robustness to appearance feature variation of our Re-ID algorithm, we trained the vehicle attributes Re-ID models by the labels of synthetic data to extract the attributes features. Then we balanced the Re-ID data by data augmentation methods and implemented a spatial attention mechanism from the background model to get a more accurate vehicle representation.

In terms of tracklet matching, we presented an efficient tracklets representation strategy to overcome the appearance bias caused by occlusion or truncation. Given it is hard to match the tracklets only through the distance matrix, an uncertainty-based tracklets matching method is raised to measure the degree of match between tracklets. Finally, the spatial-temporal information of a local vehicle's trajectory is fully utilized to improve the robustness and accuracy of MCTM.

The main contributions of this paper are summarized as follows:

- We adopted the tracker update strategy and template matching method to enhance the SCT performance.
- We leveraged the synthetic data to train attributes Re-ID models as attributes feature extractors. We also proposed a spatial attention mechanism from the background model to improve the retrieval performance of the Re-ID model.
- We proposed an efficient tracklets representation strategy and an uncertainty-based tracklets matching method, which leverages the spatial-temporal information of a local vehicle's trajectory to improve the robustness and accuracy of our MCTM system.
- Our mehod achieved 3rd place both in Track 2 and Track 3 at the AI City Challenge. It also achieved state-of-the-art results on VeRi776 dataset.

2. Related Work

2.1. Single-Camera Tracking

Single-camera tracking (SCT) is a computer vision task that aims to analyze videos to identify and track objects belonging to one or more categories [4], such as vehicles without any prior knowledge about the appearance and number of targets. The current SCT algorithms could divide into two types. One follows the tracking-by-detection paradigm, while the other is jointing object detection with Re-ID in a single network [48] to accomplish the SCT task.

In recent years, a large number of object detection methods [3, 2, 50] have been proposed. With the improvement of object detection techniques, many SCT studies turn

to the tracking-by-detection paradigm. The tracking-by-detection paradigm uses a detection model for target localization and obtains the trajectories of targets by data association between adjacent frames. The Simple Online and Realtime Tracking (SORT) algorithm [1] leveraged Faster R-CNN [38] for the detection of targets. It used a relatively simple approach that consisted of predicting object motion using the Kalman filter [16] and then associating the detections together with the help of the Hungarian algorithm [20], using intersection-over-union (IoU) distances to compute the cost matrix. DeepSort [46] incorporated visual information extracted by a custom residual CNN to alleviate identity switches in SORT. The CNN provided a normalized vector with 128 features as output, and the cosine distance between those vectors was added to the affinity scores used in SORT. TrackletNet Tracker [42] is based on a tracklet graph model to generate tracklets by appearance similarity and spatial consistency.

JDE [19] and FairMOT [48] incorporate the appearance embedding model into a single-shot detector, so that the model can simultaneously output detections and the corresponding embeddings.

2.2. Vehicle Re-identification

Due to the wide application in intelligent transportation and the releases of large-scale annotated vehicle Re-ID datasets, vehicle Re-ID gains rapidly increasing attention in the past few years. Liu *et al.* [25] released a high-quality multi-viewed VeRi776 dataset. Tang *et al.* [40] proposed a city-scale traffic camera CityFlow dataset. Based on these datasets, numerous vehicle re-ID methods have been proposed recently.

Some methods adopted spatial-temporal information [34, 44] or vehicle attribute (e.g., color and type) [26, 6] to regularize the global vehicle representation learning. Some adopted the extra information, such as critical parts [7], viewpoint [29], or keypoint [17, 45] labels, which take advantage of local features to improve the representation ability. There are also some methods which use graphic engine [47, 39] or generative adversarial network [51, 27] to generate synthesize vehicle images with rich attributes to extend training datasets.

2.3. Multi-Camera Vehicle Tracking

Multi-camera vehicle tracking (MCVT) is a fundamental technique for traffic optimization. The recent progress in MCVT mainly benefits from single-camera tracking techniques and vehicle re-identification. Tang *et al.* [8] proposed a city-scale, well-annotated benchmark for MCVT, which also makes great advancement in this field. Recent approaches [32, 10] followed the processing pipeline of vehicle detection, SCT, Re-ID for visual feature extraction, and MCTM. For multi-cameras with overlapping FOVs, some methods [22, 12] combine temporal features from GPS location to match the tracklets from different cameras.

3. Methodology

3.1. Data Pre-Processing

As the training set only labels the vehicles that travel across multiple cameras, there are many unlabeled vehicles in the image. As shown in Figure 3 (a), there are three vehicles which weren't labeled. In order to make use of the training set to train the detection model and evaluate the performance of multi-object tracking, we have to rebuilt the training set to eliminate the interference of unlabeled vehicles.

Firstly, we used the Gaussian mixture model to generate a background model from video, as shown in Figure 3 (b). Then we cropped the bounding box of labeled vehicles and placed the cropped image on the background image. Therefore, we will get a training set in which all vehicles are labeled. As shown in Figure 3 (c).

(a) (b) (c)

Figure 3. Visualization of data pre-processing.

3.2. Vehicle Detection

Vehicle detection is the basis of multi-camera vehicle tracking. Therefore, its effectiveness directly affects the performance of the entire system. We evaluated five state-of-the-art detection algorithms, YOLOv4 [2], Mask R-CNN [8], Cascade RCNN [3], CenterNet [50], and EfficientDet [38]. The comparison result is shown in Table 1. Considering the effectiveness, we adopted EfficientDet for vehicle detection. EfficientDet is based on EfficientNets [37] and BiFPN. EfficientNets is an efficient backbone network, which utilizes a simple yet highly effective compound coefficient to uniformly scales all dimensions of depth/width/resolution. BiFPN is a bi-directional feature network that is enhanced with fast normalization, which enables easy and fast feature fusion.

3.3. Single-Camera Multi-target Tracking

Following the tracking-by-detection paradigm, DeepSort [46] was chosen as the baseline method for online multi-object tracking.

DeepSort is an online tracking algorithm which uses the the Kalman filter algorithm [4] to predict the position of the tracking target in the current frame and update the

Algorithm	mAP
YOLOv4 [2]	43.5
Mmdetection Mask R-CNN (X-101-32x8d-FPN) [8]	42.8
Mmdetection Cascade RCNN (X-101-64x4d-FPN) [3]	44.5
CenterNet(Hourglass-104) [50]	45.1
EfficientDet-D7X [38]	55.1

Table 1. Comparison of effectiveness on different detection algorithms on COCO2017 test-dev set [24].

tracker parameters. It also adds apparent feature information matching to improve tracking performance. This extension enables the algorithm to track the target within a longer period of occlusion, effectively reducing the number of ID transformations. However, in complex scenarios, DeepSort is still affected by occlusion, rapid change of vehicle velocity, and other unfavorable conditions, resulting in erroneous tracking results. In this regard, we have proposed corresponding methods as below to solve these problems.

3.3.1 Tracker Update Strategy

DeepSort depends on the Kalman filter algorithm to update the tracker parameters. The Kalman filter algorithm is based on constant velocity motion and a linear observation model. However, in pixel coordinates, the movement of the vehicle does not follow the constant velocity and linear observation model.

According to the Keyhole imaging principle [41], the pixel coordinates of vehicle position change rapidly when the vehicle is at a different distance from the camera. For instance, when the vehicle is closer to the camera, the position of vehicle in the pixel coordinates changes drastically between every two frames, which will cause the Kalman filter algorithm to predict the wrong position of the tracking target in the current frame.

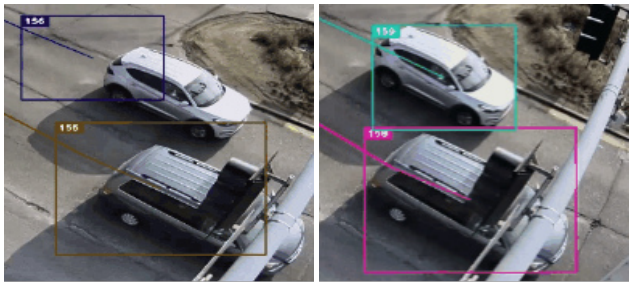
To deal with such a situation, for the matched target, instead of using the Kalman filter algorithm prediction's position $P_{KalmanFilter}$ to update its state, we used the associated detection position $P_{detection}$ to update its state. The current position in bounding box format of the target can be represented as:

$$P_{current} = \begin{cases} P_{detection} & \text{if } matched \\ P_{KalmanFilter} & \text{otherwise} \end{cases} \quad (1)$$

We only used the Kalman filter algorithm for the unmatched target. Figure 4 shows the visualization results without and with our tracker update strategy.

3.3.2 Robust Visual Features Extraction

DeepSort includes the deep visual features as association criteria. However, the Re-ID module which DeepSort uses has poor robustness and is difficult to cope with occlusion,



(a) (b)

Figure 4. Comparison of tracking results under two different tracker update strategies. (a) is the results that tracker update by Kalman filter algorithm prediction. (b) is the results that update by detection value.

vehicle turning, etc. Thence, we utilized our Re-ID model to extract the appearance feature of vehicles, as shown in Figure 2. The extracted features can also be used for the subsequent MCTM task. For the acquisition of the Re-ID model, we will elaborate on it in section 3.4.

3.3.3 Template Matching Strategy

From our observations, the tracks are prone to be switched when a vehicle turns with high speeds. Since the motion of the vehicle changes rapidly, the distance between prediction and detection will exceed the predefined threshold. As a result, the tracklet will move into an unmatched state. As shown in Figure 5 (a), the identity of a red vehicle is switched from 116 to 137 when it turns right.

When a tracklet moves into an unmatched state in the road, we assume there will be a real detection bounding box as matched detection hypothesis of the tracklet between 10 frames. Inspired by [23], we treated the visual features of the unmatched tracklet as a template, then we used the template to search a detection bounding box whose visual feature similarity with template is lower than the predefined threshold $matchThresh$. The search region is limited within a radius R from the last history position in the unmatched tracklet. The detection bounding box will be used to update the tracklet state space. Figure 5 (b) shows the tracking visualization results of using template matching strategy, the red vehicle maintain identity in 72 when it turns right.

3.4. Vehicle Re-identification

3.4.1 Network Structure

The baseline architecture of the whole Re-ID model consists of backbone, aggregation, and head. The three modules will be elaborated in below.

Backbone. ResNet [9] with IBN [31] structure and Non-local [43] module is used as the backbone to extract fea-

(a) (b)

Figure 5. Comparison of tracking results before and after using template matching strategy. (a) is the results that the tracks is switched when vehicle turns. (b) is the results of using the template matching strategy.

tures. It shows potentials to robustly extract appearance features for the presence of occlusions, noisy detection, different illumination conditions, and viewpoint changes.

Aggregation and Head. The generalized-mean (Gem) pooling and bnneck [28] are used as the aggregation layer and head, respectively. Compared to global average pooling and global max pooling, Gem pooling gets superior performance for image retrieval-based tasks, especially in vehicle Re-ID tasks.

3.4.2 Optimization Functions

We trained the whole vehicle Re-ID network with the fusion of circle loss [35] and triplet loss [11], with the equal-weighted. Triplet loss ensures that an image of a specific vehicle is closer to all other images of the same vehicle than to any images of other vehicles. While the circle loss with a more definite convergence target and the high flexibility in optimization will benefit deep feature learning. It has a unified formula for two elemental learning approaches, i.e., learning with class-level labels and learning with pair-wise labels [36].

3.4.3 Data Balance

We noticed that the data for Re-ID is imbalance, which different vehicle ID has different sample size. As a result, the Re-ID model will struggle to deal with imbalanced Re-ID data by focusing on minimizing the error rate for the majority vehicle ID while ignoring the minority vehicle ID. To overcome the shortness of data imbalance, the data augmentation methods are applied to increase the sizes of minority vehicle ID, until the sample of all vehicle IDs have the same sizes. The data augmentation methods include flipping, random erasing, random patch and auto-augment.

3.4.4 Spatial Attention

Since the dataset used in vehicle re-id is cropped by bounding box, which may introduce the other vehicles and extra background. To reduce the influence, we used a binary mask for the foreground vehicle as the spatial attention map.

To train the Re-ID model for Track 2, we re-detected the foreground vehicle with the Mask R-CNN [8] for the dataset in Track 2. To train the Re-ID model for Track 3, as the Re-ID dataset in Track 3 is copped from videos and the static background can be extracted, we used the Gaussian mixture model to generate a robust background model from multi-frame image information. In the end, we can get the precise binary mask of vehicles. Combining the features of spatial attention images $F_{spatial}$ and the features of original images $F_{original}$, we can get a more accurate vehicle representation. The fusion features can be written as following:

$$F = F_{original} + \gamma F_{spatial} \tag{2}$$

Where γ is the weight parameter of $F_{spatial}$.

3.4.5 Attribute Re-ID

In addition to the real-world traffic data set, there is a well-annotated synthetic vehicle dataset provided by AI CITY CHALLENGE. The synthetic vehicle includes an orientation label, color label, and type label. These attributes are important for vehicle Re-ID. For example, the vehicle Re-ID model is likely to be confused by different candidates with similar orientation. The orientation attributes are helpful to overcome this drawback. Inspired by [52, 14], for each attribute, we treated each category of the attribute as an ID to practice the attribute Re-ID process. The attribute features can be extracted from attribute Re-ID model and output the attribute ID distance matrix. To increase the influence of color and type, while depressing the influence of similar orientation, we plus the color distance D_{color}, type distance D_{type}, and vehicle ID distance $D_{vehicle}$, then minus the orientation distance $D_{orientation}$. The fusion distance matrix could be formulated as:

$$D = D_{vehicle} + \alpha D_{color} + \beta D_{type} - \lambda D_{orientation} \tag{3}$$

Where α, β and λ are the weight parameter.

3.5. Multi-Camera Tracks Matching

3.5.1 Traffic Spatial-Temporal Information Reasoning

We proposed a tracklet matching strategy which is based on traffic spatial-temporal information, to narrow the matching candidates and get more stable tracklet synchronization matching results.

Figure 6 depicts the traffic geometry information of the six cameras. There is no overlapping FOV between different camera views. However, in the real-world traffic scene, limited by traffic topology and the traffic rules, the movements of vehicles are predictable. This allows us to model the traffic spatial-temporal information.

Figure 6. The traffic geometry information of cameras.

To assign a movement model for each given vehicle tracklet, inspired by [12], we used the pre-defined zones in the image. As is shown in Figure 7, each vehicle tracklet will pass through two of the zones. According to the zone list that the vehicle tracklet passes, we can extract the spatial information for the vehicle and finally assign a movement model for the tracklet.

According to the movement model of the tracklet and the distance between the cameras, we can estimate when the vehicle will appear on which camera. For example, if a vehicle is in camera 1 and driving towards camera 2. The distance between camera 1 and camera 2 is 1000m. Following the traffic rules, the vehicle can only appear in another camera 2 within 10 seconds to 200 seconds. Based on traffic spatial-temporal information reasoning, we can narrow the tracklet matching candidates to achieve more robust matching.

3.5.2 Tracklet Representation

According to our observations, the appearance features are ineffective in representing a tracklet when targets are occluded or truncated. Inspired by [13], we proposed an efficient tracklet representation approach to overcome the appearance bias.

Denote T_{vehicle} as the tracklet of each vehicle and T_{frame} as the track of vehicle in each frame. Considering vehicles may wait in front of traffic lights and produce repetitive T_{frame}, we selected the T_{frame} if the distance between it and all other T_{frame} in T_{vehicle} is larger than $disThresh$. Then we extracted the image-based appearance features of the T_{frame} base on Vehicle Re-ID model. The tracklet feature is updated to the mean feature of all seleted T_{frame}.

As the features of T_{frame} which are heavily occluded or truncated will be the outlier, we removed the T_{frame} whose distance between its features and tracklet features is larger than $featThresh$. Then we continued to update the tracklet feature until there are no outlier T_{frame}.

Finally, we let a set of selected T_{frame} to represent the tracklet of each vehicle.

3.5.3 Tracklet Matching based on Uncertainty

With the help of spatial-temporal information and tracklets representation, we can narrow the matching candidates and get synchronization results for the MCVT task. However, concerning the tracklet matching, due to the significant appearance feature variation caused by different viewing perspective, it is hard to find an accurate threshold to determine whether two tracklets are matching. For example, the feature distance of one negative tracklet pairs which have a similar viewing perspective may small, while the feature distance of the other positive tracklet pairs which have a different viewing perspective may be large.

To tackle this problem, we used uncertainty to measure the degree of match between Tracklets. For N T_{frame} of a query tracklet, we can get N matched gallery tracklets by the distance matrix. The N gallery tracklets come from K targets and $K < N$ as several gallery tracklets may share one target. Denote t is a weight parameter. When the number of matching gallery tracklets from a target is more than $t*N$, we treated it as a state of certainty and the query tracklet will be assigned to the target. We will remove the query tracklet in matching candidates if the query tracklet is in an uncertain state.

When a gallery tracklet is assigned to several query tracklets, we will select the best matching query tracklet by distance matrix, and remove the unmatched query tracklets.

4. Experiments

We participated in the track 2: City-Scale Multi-Camera Vehicle Re-Identification, and track 3: City-Scale Multi-Camera Vehicle Tracking in the AI City Challenge.

We used ImageNet [5] and MSCOCO [24] to pre-train our detection model EfficientDet [38] and backbone networks (ResNet [9]) of Re-ID model. We also evaluated our Re-ID method on VeRi776 dataset for comparison with state-of-the-art methods, and the VeRi776 dataset was not used in AI City Challenge.

4.1. Datasets and Evaluation Metric

In the dataset for Track 2, there are 85,058 images, of which 52,717 come from 440 vehicle identities in the training set and 31,238 from the other 440 vehicle identities in the test set. An additional 1,103 images are used as queries.

In the dataset for Track 3, there are six scenarios with 46 cameras which include 16 intersections. Three of the scenarios with 36 cameras are used for training, two scenes of the scenarios with 23 cameras are for validation, and the

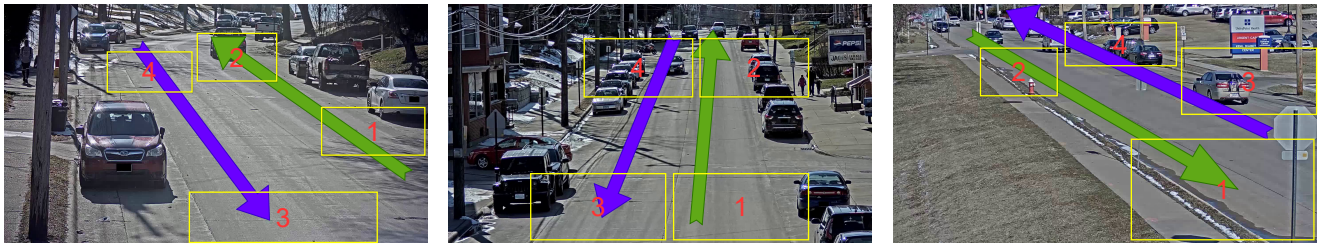


Figure 7. Examples of transitions for tracklet pairs with non-overlapping FOV between different camera views. The movement (blue) can be described by zone list [3, 4], while the other movement (green) can be described by zone list [1, 2]. The movement with the same color under different cameras will be the matching candidates.

remaining one is for testing. It contains 313931 bounding boxes for 880 annotated vehicle identities.

The mean Average Precision mAP [49] is used for Re-ID model performance evaluation. We also used $Rank_1$ to measure the percentage of the queries that get the true positive result ranked in the top 1 position.

We adopted IDF_1 [33] as the tracking evaluation metric. Denote $IDTP$ as the count of true positive IDs, $IDTN$ as the count of true negative IDs, $IDFP$ as the count of false positive IDs, and $IDFN$ as the count of false negative IDs. The IDF_1 could be calculated as:

$$IDF_1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{4}$$

4.2. Implementation Details

Our system is implemented under the Pytorch framework and NVIDIA 1080Ti GPU platform. We trained the Re-ID model via the SGD optimizer with a momentum of 0.9. The images are resized to 384×384, and batch size is set to 64 with 4 images per vehicle. As for the learning rate, we adopted the learning rate warm-up strategy and cosine rule to avoid premature over-fitting. In Table 2, we provided all the values of the tuned parameters in our experiment.

Parameter	Value
γ	0.9
α	0.3
β	0.05
λ	0.08
$disThresh$	20
$featThresh$	0.7
$matchThresh$	0.2
t	0.8
R	200

Table 2. The tuned parameters set in our experiment.

4.3. Ablation Study on Single-Camera Tracking (SCT)

We evaluated the effects of our tracking strategies in this place. We selected S01 from the training set as our validation set, in which unlabeled vehicles are removed by data pre-processing.

Method	Performance			
With TU		✓	✓	✓
With VFE			✓	✓
With TMS				✓
$IDF_1(\%)$	72.57	82.96	88.40	90.34

Table 3. Comparison with different strategies in online multi-object tracking.

TU corresponds to using our tracker update strategy, VFE corresponds to using our Re-ID module for robust visual features extraction, and TMS corresponds to using template matching strategy. As shown in Table 3, the tracker update strategy (TU) reduces the identity switches substantially and increases the IDF_1 by more than 10%, from 72.57% to 82.96%. The visualization results of using the tracker update strategy are shown in Figure 4, we can get a more accurate position of the tracking target by our tracker update strategy. The template matching strategy (TMS) and using our Re-ID module as visual features extractor (VFE) also achieved an impressive performance that effectively improves the IDF_1 of SCT from 82.96% to 90.34%. As shown in Figure 5 (b), we can re-associate the discontinuous tracklets after using the template matching strategy.

Method	Performance					
With data balance		✓	✓	✓	✓	✓
With spatial attention			✓		✓	✓
With synthetic data				✓	✓	✓
With attribute Re-ID						✓
$mAP(\%)$	43.95	47.41	52.95	52.77	57.62	60.08
$Rank_1(\%)$	82.82	85.98	87.59	89.01	90.43	90.89

Table 4. Ablation study on vehicles Re-ID.

4.4. Ablation Study on vehicles Re-ID

The training set of Track 2 includes 440 vehicles. To conduct the ablation studies of the important components, we split 40 vehicle ID as the validation set from the training set, while left out the 400 vehicle IDs for training. we also evaluated the performance of the synthetic data on the

Method	$mAP(\%)$	$Rank_1(\%)$
UMTS [15]	75.9	95.8
PVEN [30]	79.5	95.6
SAVER [18]	79.6	96.4
VOC-ReID [52]	82.8	97.6
Our baseline model	80.0	96.6
Our baseline model (+spatial attention)	**83.4**	**97.8**

Table 5. Comparison with state-of-the art methods on VeRi776.

Rank	Team ID	Team Name	mAP
1	49	DMT	0.7445
2	9	NewGeneration	0.7151
3	**7**	**CyberHu (Ours)**	**0.6650**
4	35	For Azeroth	0.6555
5	125	IDo	0.6373

Table 6. The public leader board of Track 2.

Rank	Team ID	Team Name	IDF_1
1	75	mcmt	0.8095
2	29	fivefive	0.7787
3	**7**	**CyberHu (Ours)**	**0.7651**
4	85	FraunhoferIOSB	0.6910
5	42	DAMO	0.6238

Table 7. The public leader board of Track 3.

validation set. As shown in Table 4, by balancing the Re-ID data, we improved the performance to 47.41% mAP. Training the Re-ID model with synthetic data and implementing the spatial attention mechanism produced a similar performance that brings almost 5.5% mAP gains. Exploiting the three methods above together further improved the performance to 57.62% mAP. Utilizing the attributes features by attributes Re-ID models further get 2.46% mAP increasing, from 57.62% mAP to 60.08% mAP.

4.5. Re-ID Performance on VeRi776 Dataset

We also performed our Re-ID method on VeRi776 dataset and compared it with state-of-the-art methods in Table 5. Our baseline Re-ID model outperforms other state-of-the art methods, and applying spatial attention can further boost the performance. We only utilized the training set in VeRi776 for training and did not adopt other augmentation strategies such as using vehicle attribute (e.g., color and type) bias.

4.6. Performance on the CVPR AI City Challenge 2021

We (team ID 7) both participate in the City-Scale Multi-Camera Vehicle Re-Identification task (Track2) and City-Scale Multi-Camera Vehicle Tracking task (Track3) .

In Track2, the result of our submission on the private test set is shown in Table 6. We rank third place on the final public leaderboard with an mAP score of 66.50%. In addition, our proposed MCVT system has achieved third place and yielded 76.51% IDF_1 for Track3, as shown in Table 7.

5. Conclusion

In this paper, we proposed a novel MCVT system that has achieved top-3 ranking both in Track 2 and Track 3 of the AI City Challenge. Our system includes single-camera tracking (SCT), vehicle re-identification(Re-ID), and multi-camera tracklets matching (MCTM). We adjusted the structure of the MCVT system. The coarse visual features extraction step for SCT is removed now. The SCT and the MCTM directly take the precise visual features from the Re-ID module as input, which saves the system from repeated calculation. For the SCT module, we proposed a tracker update strategy and template matching strategy to improve its precision. For the Re-ID module, we designed a spatial attention method based on background modeling and a vehicle attributes Re-ID method to get discriminant visual features for the SCT and the MCTM. For the MCTM module, we fully utilized the spatial-temporal information to narrow the matching candidates, and formulated the uncertainty of tracklets matching to get synchronization results for the MCVT task.

These improvements contribute to the final accuracy of 66.50% mAP for Track 2 and 76.51% IDF_1 for Track 3. We also achieved state-of-the-art results on the VeRi776 dataset. Experimental results show the robustness of our system and it is easy to be upgraded by improving each component further.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (U1764264/61873165), and Shanghai Automotive Industry Science and Technology Development Foundation (1807).

References

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing*, pages 3464–3468, 2016. 1, 3

[2] 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, 3, 4

[3] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018. 2, 3, 4

[4] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. Deep learning in video multi-object tracking: A survey. *Neurocomputing*, 381:61–88, 2020. 2, 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, 2009. 6

[6] Haiyun Guo, Kuan Zhu, Ming Tang, and Jinqiao Wang. Two-level attention network with multi-grain ranking loss for vehicle re-identification. *IEEE Transactions on Image Processing*, 28(9):4328–4338, 2019. 3

[7] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. Part-regularized near-duplicate vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3997–4005, 2019. 3

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 3, 4, 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 6

[10] Yuhang He, Jie Han, Wentao Yu, Xiaopeng Hong, Xing Wei, and Yihong Gong. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In *CVPR Workshops*, pages 576–577, 2020. 3

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[12] 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 *CVPR Workshops*, pages 416–424, 2019. 3, 6

[13] Hung-Min Hsu, Yizhou Wang, and Jenq-Neng Hwang. Traffic-aware multi-camera tracking of vehicles based on reid and camera link model. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 964–972, 2020. 6

[14] Tsung-Wei Huang, Jiarui Cai, Hao Yang, Hung-Min Hsu, and Jenq-Neng Hwang. Multi-view vehicle re-identification using temporal attention model and metadata re-ranking. In *CVPR Workshops*, 2019. 5

[15] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Uncertainty-aware multi-shot knowledge distillation for image-based object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11165–11172, 2020. 8

[16] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Broadcast Engineering*, pages 35–45, 1960. 3

[17] Pirazh Khorramshahi, Amit Kumar, Neehar Peri, Sai Saketh Rambhatla, Jun-Cheng Chen, and Rama Chellappa. A dual-path model with adaptive attention for vehicle re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6141, 2019. 3

[18] Pirazh Khorramshahi, Neehar Peri, Jun-cheng Chen, and Rama Chellappa. The devil is in the details: Self-supervised attention for vehicle re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 369–386, 2020. 8

[19] Kyung Hun Kim, Jun Ho Heo, and Suk-Ju Kang. Towards real-time multi-object tracking in cpu environment. *Journal of Broadcast Engineering*, 25(2):192–199, 2020. 2, 3

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3

[21] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision*, pages 734–750, 2018. 2

[22] Peilun Li, Guozhen Li, Zhangxi Yan, Youzeng Li, Meiqi Lu, Pengfei Xu, Yang Gu, Bing Bai, Yifei Zhang, and DiDi Chuxing. Spatio-temporal consistency and hierarchical matching for multi-target multi-camera vehicle tracking. In *CVPR Workshops*, pages 222–230, 2019. 3

[23] Zhenbang Li, Qiang Wang, Jin Gao, Bing Li, and Weiming Hu. End-to-end temporal feature aggregation for siamese trackers. In *2020 IEEE International Conference on Image Processing*, pages 2056–2060, 2020. 4

[24] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014. 4, 6

[25] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016. 3

[26] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2017. 2, 3

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019. 3

[28] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019. 5

[29] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 2, 3

[30] Dechao Meng, Liang Li, Xuejing Liu, Yadong Li, Shijie Yang, Zheng-Jun Zha, Xingyu Gao, Shuhui Wang, and Qingming Huang. Parsing-based view-aware embedding network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2020. 8

[31] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision*, pages 464–479, 2018. 4

[32] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *CVPR Workshops*, pages 588–589, 2020. 3

[33] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision*, pages 17–35, 2016. 7

[34] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1900–1909, 2017. 2, 3

[35] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 5

[36] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 5

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 3

[38] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2020. 2, 3, 4, 6

[39] Zheng Tang, Milind Naphade, Stan Birchfield, Jonathan Tremblay, William Hodge, Ratnesh Kumar, Shuo Wang, and Xiaodong Yang. Pamtri: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 211–220, 2019. 2, 3

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019. 3

[41] Joop J Van Vaals, Marijn E Brummer, W Thomas Dixon, Hans H Tuithof, Hans Engels, Rendon C Nelson, Brigid M Gerety, Judith L Chezmar, and Jacques A Den Boer. "keyhole" method for accelerating imaging of contrast agent uptake. *Journal of Magnetic Resonance Imaging*, 3(4):671–675, 1993. 4

[42] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 482–490, 2019. 1, 3

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 4

[44] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 379–387, 2017. 3

[45] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 379–387, 2017. 3

[46] 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, 2017. 1, 3

[47] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. *arXiv preprint arXiv:1912.08855*, 2019. 3

[48] 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 preprint arXiv:2004.01888*, 2020. 2, 3

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 7

[50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3, 4

[51] Yi Zhou and Ling Shao. Aware attentive multi-view inference for vehicle re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2018. 3

[52] Xiangyu Zhu, Zhenbo Luo, Pei Fu, and Xiang Ji. Voc-reid: Vehicle re-identification based on vehicle-orientation-camera. In *CVPR Workshops*, pages 602–603, 2020. 5, 8