Tracklet-refined Multi-Camera Tracking based on Balanced Cross-Domain Re-Identification for Vehicles

Kai-Siang Yang* [1,2], Yu-Kai Chen* [1,2], Tsai-Shien Chen[1,2], Chih-Ting Liu[1,2], Shao-Yi Chien[1,2]
[1]NTU IoX Center, National Taiwan University
[2]Graduate Institute of Electronics Engineering, National Taiwan University
{siangyang, chenyukai, tschen, jackieliu}@media.ee.ntu.edu.tw
sychien@ntu.edu.tw

Abstract

Mutli-camera vehicle tracking and re-identification (re-ID) have gradually gained attention due to their applications in the intelligent transportation system. However, these problems are fundamentally challenging. Specifically, for vehicle tracking, we observe that the results generated from single camera tracking algorithm usually recognize tracklets with same identity as different vehicles when the tracklets are occluded. Hence, we propose a Tracklet Reconnection technique to refine tracking results with predefined zone areas and GPS information. The proposed method can efficiently filter invalid tracklet pairs and reconnect the split tracklets into complete ones, which is important for the afterwards multi-target multi-camera tracking. As for re-ID, we also find that when a large-scale auxiliary dataset is used to assist the learning of main dataset for better model capability and generalization, there is a performance drop caused by data imbalance when the full auxiliary dataset is applied. To tackle this problem, we introduce Balanced Cross-Domain Learning to avoid the overemphasis on larger auxiliary dataset by a newly introduced training data sampler and loss function. The extensive experiments validate the empirical effectiveness of our proposed components.

1. Introduction

Multi-camera vehicle tracking and re-identification (re-ID) aim to match and track the vehicles with same identity within cameras in a city-scale camera network. Recently, they gain increasing attention in both academia and industry due to several practical applications, such as the analysis and prediction of traffic flow and the implementation of intelligent transportation system.

For vehicle tracking, followed by common processing pipeline [23, 10, 15, 39], we split the whole algorithm into three parts: single-camera tracking (SCT), appearance feature re-ID and multi-target multi-camera tracking. For SCT, which tracks multiple detected objects under the same camera, an intuitive method is applying object detection model, such as Faster RCNN [27] or Mask RCNN [7], to detect vehicles and then utilize TrackletNet Tracker (TNT) [36] to generate trajectories of detected vehicles based on both temporal and appearance information. However, we find that following such straightforward baseline algorithm would end up in an unsatisfied results especially when the target objects are occluded. Hence, we propose a refinement module, named *Tracklet Reconnection* technique. Such module aims to refine the coarse outcomes from TNT by reconnecting the split tracklets which belong to same identity. In details, based on the generated zone areas, we define the completeness score of each tracklest. Then, only the tracklet pairs consist incomplete tracklet will be fed to our reconnection module. In addition, it utilizes the GPS coordinate to evaluate the orientation of each detected tracklet and can further pick out the potential split tracklets and enhance the performance of SCT.

As for vehicle re-ID, while lots of previous works proposed meticulously-designed efficient architectures, recently, there are increasing interests on multi-domain learning schemes to train a high-performance model which can leverage from more training samples, due to the release of several large-scale real world [20, 31] and synthetic [41] vehicle re-ID datasets. However, most literature only focused on the discrepancy of feature embedding space across different datasets without regard to the imbalance between main and auxiliary dataset which is exactly the case in 2021 AI City Challenge[†]. Specifically, training set of main dataset, CityFlowV2-ReID [31], only has 440 identities while the auxiliary one, Vehcile X [41], has up to 1,362

*denotes equal contribution

[†]https://www.aicitychallenge.org/

identities. He *et al*. [8] found that, with the full usage of auxiliary dataset, the imbalance would instead cause an unexpected performance drop. To solve such problem, in this paper, we propose a training scheme, named *Balanced Cross-Domain Learning* (BCDL). It contains a novel data sampler which ensures the training samples are evenly selected from the main and auxiliary datasets and also a new loss function that not only minimizes the domain gap between different datasets but also avoids the trivial training for identity recognition on the auxiliary dataset.

Extensive experiments prove that the empirical effectiveness of proposed components. We now highlight our contributions as follows:

- For single-camera tracking, we propose Tracklet Reconnection module to refine the mistakenly split tracklets by defining the zone areas and applying GPS information.
- For re-identification, we introduce a novel Balanced Cross-Domain Learning to better tackle the problem of data imbalance between main and auxiliary datasets.
- The extensive experiments shows the superiority of our proposed components.

2. Related Work

The whole multi-camera vehicle tracking algorithm is commonly split into three steps: object detection and single-camera tracking (SCT), multi-camera vehicle re-identification (re-ID), and multi-target multi-camera tracking (MTMCT):

Detection and Single Camera Tracking respectively generate frame-level detections and associate the detected bounding boxes across frames into a tracklet. For the detection models, there are two common implementation: one-stage and two-stage frameworks. The former ones, such as SSD [19] and YOLOv3 [26], which combine detection and recognition into one integrated model while the latter ones, like Mask R-CNN [7] split it into region proposal network (RPN) and another classification model to improve the prediction of bounding boxes. In this paper, we use Mask R-CNN as object detection model to detect all vehicles in each video frame. As for SCT, regrading tracking as a template-matching problems, Deep-Sort [38] additionally integrates appearance feature with bounding box information to improve the performance, while combining the Kalman filter [11] and Hungarian algorithm [12]. To handle longtime occlusions, some works [29, 34] learn a relatively long-term appearance which is more compatible in various conditions. Most recently, TNT [36] simultaneously considers temporal and appearance information and solves tracking problem by graph-based model. However, we find that there are still some unsatisfying results from TNT especially when target objects are occluded. Hence, in this paper, we further propose a Tracklet Reconnetion technique which can benefit from the extra GPS information and refine the primary tracklet results into more robust one.

Deep Learning based Vehicle Re-identification aims to find the objects with same identity captured by a large-scaled camera network. Prior works usually proposed new re-ID frameworks, commonly spatial [37, 44, 21, 2] or channel-wise [1] attentive models; however, due to the release of large-scale vehicle re-ID datasets, such as real-world VERI-Wild [20] dataset with 416,314 images from 40,671 identities and synthetic VehicleX [41] dataset with 192,150 images (can simulate more if required) from 1,362 identities, there are emerging interests in multi-domain learning which is expected to enhance the model capability and generalization with the assistance of large-scale auxiliary dataset. While the previous literature mostly focused on the minimization of the image-level [43, 3] or feature-level [13, 17] domain discrepancy, He *et al*. [8] found the negative effect caused by the full adoption of large-scale auxiliary dataset. Such training sample imbalance issue encourages us to design a more robust training scheme to better apply the large-scale auxiliary dataset.

Multi-Target Multi-Camera Tracking tracks multiple detected objects across multiple cameras of overlapping or non-overlapping views. Recent approaches utilize multiple constraints, such as appearance similarity with tracklet-based features [16, 28, 4, 18], topology reasoning [10, 9] or transition time window [14, 32, 35, 40], to reduce the search space and find the potential matching pairs. For more details, Lee *et al*. [14] built a camera link model with bidirectional transition time distribution in an unsupervised manner, while Hsu *et al*. [10] introduced a more complex camera link model with a trajectory matching algorithm. In this paper, we also utilize a camera link model with spatial and temporal constraints for an effective MTMCT framework.

3. Proposed Method

The proposed vehicle camera tracking system contains three parts, shown in Fig. 1. First, as described in Sec. 3.1, given each video sequence, we conduct single-camera tracking (SCT) to locate the vehicles along frames and combine them to form several vehicle tracklets. However, the severe occlusion hinders the performance; hence, we propose *Tracklet Reconnection* technique to re-connect the original tracklets based on zone areas and GPS information in the sequence. Then, in order to match vehicles across cameras, we need a well-performed vehicle feature extractor, which is studied in the task of vehicle re-identification

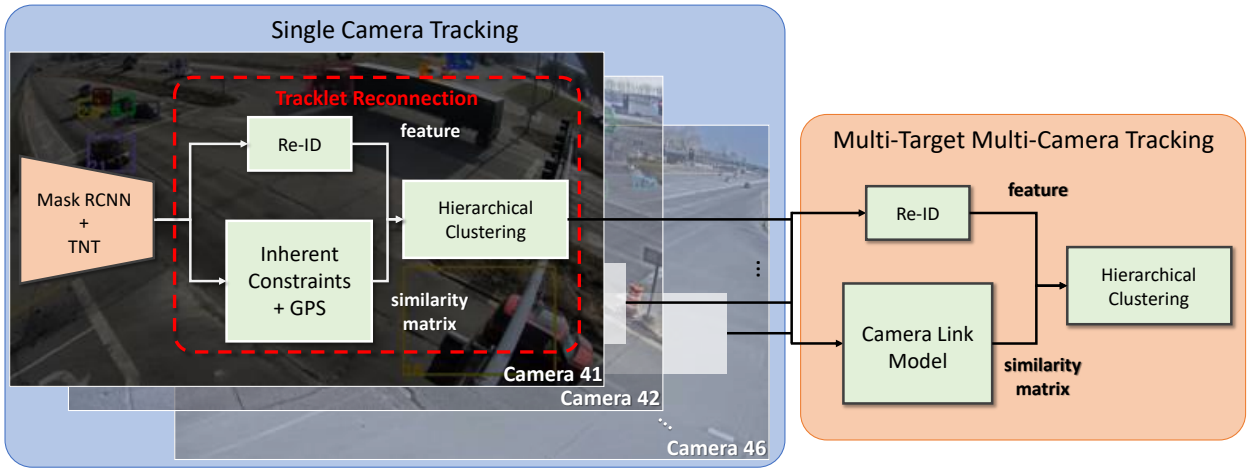

Figure 1: **The framework for vehicle camera tracking system.** Our pipeline system consists of two parts. First, in single camera tracking, vehicles are detected in frame level by Mask R-CNN [7] and associated into tracklets by TackletNet Tracker [36]. Then, the proposed Tracklet Reconnection technique refines the unsatisfied results by defining completeness, inherent constraints, and GPS information. Secondly, multi-target multi-camera tracking performs hierarchical clustering to match tracklets within different cameras, using the combination of re-ID appearance features and inherent constraints.

(re-ID). In Sec. 3.2, we introduce how to leverage the methods in re-ID and propose *Balanced Cross-Domain Learning* to make our feature extractor better trained with the assistance of large-scale auxiliary dataset. Last, with the feature extractor, in Sec. 3.3, we introduce our multi-target multi-camera tracking (MTMCT) and explain how we utilize the vehicle features and the inherent constraints (both spatially and temporally) in the videos to link and track the tracklets across different cameras.

3.1. Single-Camera Tracking

Detection and Single-Camera Tracking Algorithm. The first step in the multi-camera vehicle tracking is to detect and locate vehicles in each frame. Generating reliable detections is extremely important for the afterwards vehicle tracking; thus, we adopt a highly-performed instance segmentation framework, Mask R-CNN [7] as our detection model. It efficiently detects objects in a frame while simultaneously generating segmentation masks of objects.

After generating the detection results, we adopt the TrackletNet Tracker (TNT) [36] as our single camera tracking model, which combines temporal and appearance information together. Given the detection results in different frames under the same camera, the TNT would form a graph that consists of vertices and edges, which respectively represent the detection association (tracklets) and similarity between tracklets generated from TrackletNet. Finally, the graph partition is applied to cluster the vertices with the higher similarity into one group.

Tracklet Reconnection. From the SCT results generated by TNT, we observe that there are some split tracklets caused by heavy occlusion, which mostly happens while vehicles waiting for the traffic light. With the original tracking method, this issue would make it more difficult to recognize the occluded vehicles and lead to frequent ID switches. Therefore, inspired by [10, 15], we propose a refinement module called *Tracklet Reconnection* to deal with this problem by defining several complete trajectories in each camera, calculating GPS location of vehicles and then performing single camera re-ID with the model described in Sec. 3.2 to associate the split tracklets.

The complete trajectories in each camera will pass through several zones generated by the entry and exit information of each tracklet. Different from [10] that only using the zone list to describe a tracklet for cross-camera tracking, we found that it is useful for tracklets refinement inside a single camera. The first few steps for generating the ordered zone lists and assigning each tracklets to predefined ones are based on [10]. In brief, we first collect the coordinates of bounding boxes and perform clustering to decide the location of zones. Then based on the observation, we can define some ordered zone lists representing complete trajectories. Last, given any tracklet generated by TNT, we can assign it to the zone lists based on the overlapping area between the bounding boxes of tracklets and the zones. Fig. 2 demonstrates our results of zones generation in the testing videos.

After assigning each tracklet to one of the complete trajactories, we can easily define its completeness. With this score, under a single camera, we can further reconnect two

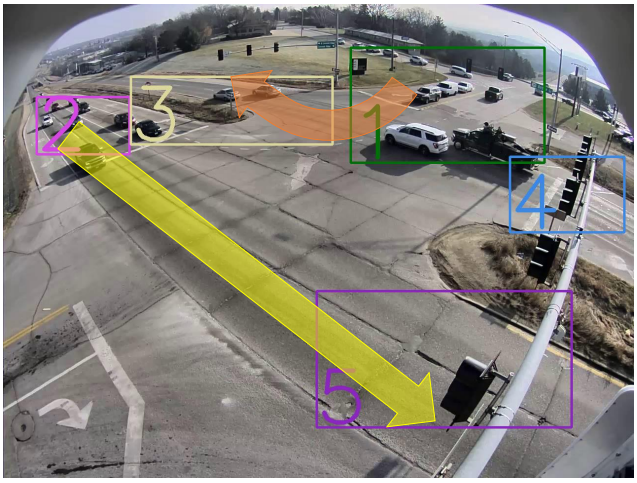

Figure 2: **Zones and Trajectories.** There are five zones detected in this video. The yellow trajectory can be described by zone list $(2, 5)$, and the orange one can be described by zone list $(1, 3)$.

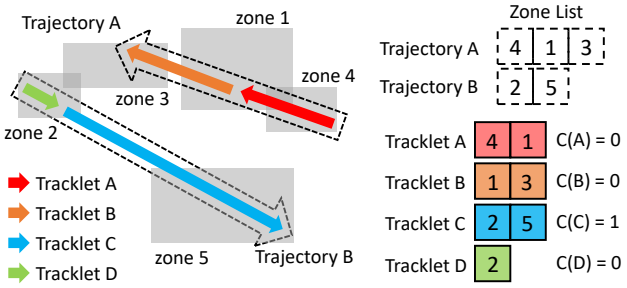

Figure 3: **Examples of completeness determination.** In this example, there are two predefined trajectories, A and B. Besides, four tracklets (A,B,C,and D) are respectively assigned to the zone list. Tracklet C is the only one complete tracklet with the first and the last zone in the zone list both existing in corresponding Trajectory B, while Tracklet A,C,and D are incomplete for mismatching with the first or the last zone to the corresponding trajectories.

tracklets that one or both of them are incomplete. Given a zone list of a tracklet $ZL^T = (Z_a, Z_b, ...)$ and its assigned zone list of a complete trajactory $ZL^c = (Z_i, Z_j, ...)$, where Z is the zone in the video, the completeness of a tracklet $(C(T))$ is formulated as:

$$C(T) = \begin{cases} 1, \ if \ (ZL_f^T = ZL_f^c) \wedge (ZL_l^T = ZL_l^c) \\ 0, \ otherwise \end{cases} \tag{1}$$

where f and l means the first and the last zone in the zone list. The tracklet with score value 1 means that it matches with the corresponding trajectory, which suggests that this tracklet is complete. While the tracklet with value 0 is defined incomplete, and may be selected as a candidate in the single camera reconnection.

Before start matching and reconnecting two tracklets with single camera re-ID, we can filter out some invalid tracklet pairs with the completeness score to reduce the matching space. The valid candidate pair $P_{ij} = (ZL^{T_i}, ZL^{T_j})$ of tracklet T_i and T_j should satisfy the condition that one or both of the tracklets are incomplete. Thus, the pair that both of them are "complete" will not participate in our single camera reconnection. For example, in Fig. 3, (Tracklet A, Tracklet B) are a valid candidate pair for both being incomplete. Also, complete Tracklet C and incomplete Tracklet D can be selected as a candidate pair. To formulate it, we use C_{com} to denote the condition used in the reconnection afterwards:

$$C_{com} \ \Leftrightarrow \ \sim (C(T_i) \ \wedge \ C(T_j)) \tag{2}$$

In addition, we also apply some inherent constraints to further filter out the invalid pairs P. First, in the time domain, the pairs should meet the condition denoted as C_{time}, where the pairs with overlapping region or with unreasonable traveling time are removed. For the spatial constraint, the reconnection part between tracklets must be formed inside the same zone, because the zone covers the enrty and exit positions of tracklets. This inherent condition can be formulated as

$$C_{inher} \Leftrightarrow C_{time} \wedge ((ZL_f^{T_i} = ZL_l^{T_j}) \vee (ZL_l^{T_i} = ZL_f^{T_j})) \tag{3}$$

Last, besides the zone constraints, we apply the GPS information to further enhance the quality of candidate pairs based on [15]. Given a 3×3 homography matrix M provided by [24], we use the center point of the bounding box $(C_x, C_y) = (x + \frac{1}{2}w, \ y + \frac{1}{2}h)$ to calculate the GPS coordinates (G_x, G_y) by the following formulation:

$$M \cdot \begin{bmatrix} C_x \\ C_y \\ 1 \end{bmatrix} = \begin{bmatrix} G_x \\ G_y \\ 1 \end{bmatrix} \tag{4}$$

With GPS coordinates, the orientation vector of a vehicle tracklet i can be approximately calculated by $\vec{v_i} = (G_{x,l} - G_{x,f}, G_{y,l} - G_{y,f})$, where the subscript f and l means the first and the last point in the tracklet. Then we can obtain the similarity $sim_{i,j}$ of tracklet i and j by the cosine similarity:

$$sim_{i,j} = \frac{\vec{v_i} \cdot \vec{v_j}^T}{\|\vec{v_i}\|_2 \|\vec{v_j}\|_2} \tag{5}$$

Finally, with features generated from re-ID feature extractor, we can perform hierarchical clustering on candidate pairs under a single camera matching to those predefined conditions:

$$"P \ is \ a \ candidate" \ \Leftrightarrow \ C_{com} \wedge C_{inher} \wedge (sim_{i,j} \geq 0) \tag{6}$$

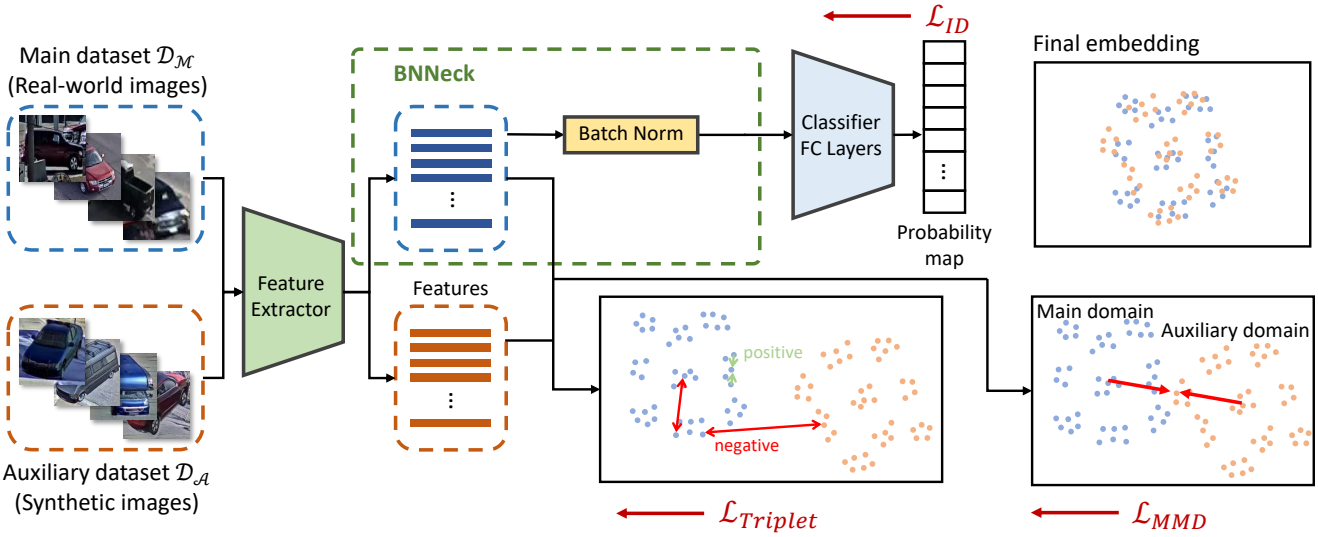

Figure 4: **Multi-Camera Vehicle Re-Identification.** To deal with the imbalance of training samples between main and auxiliary datasets, the training batch is constituted of fixed rate of identities from two datasets. After the features are extracted, they are used to compute the batch-hard triplet loss, $\mathcal{L}_{Triplet}$, along with domain loss, $\mathcal{L}_{MMD}$, for the distance metric learning. As for the identity learning, we only train the classification loss, $\mathcal{L}_{ID}$, for the samples from main dataset to avoid excessive trivial identities from auxiliary dataset.

3.2. Multi-Camera Vehicle Re-Identification

Balanced Cross-Domain Learning. To better train a re-ID model, in this section, we propose a new learning scheme, named Balanced Cross-Domain Learning (BCDL), to tackle the imbalance issue when we use large-scale auxiliary dataset to assist the training on relatively smaller main dataset. We illustrate the learning scheme in Fig. 4. For simplification, in the following paragraph, we denote main dataset as $\mathcal{D}_{\mathcal{M}}$ and auxiliary dataset as $\mathcal{D}_{\mathcal{A}}$. Our purpose is to train a model on both $\mathcal{D}_{\mathcal{M}}$ and $\mathcal{D}_{\mathcal{A}}$ to achieve better performance on $\mathcal{D}_{\mathcal{M}}$.

An intuitive approach of the utilization of $\mathcal{D}_{\mathcal{A}}$ is to directly merge the $\mathcal{D}_{\mathcal{A}}$ into $\mathcal{D}_{\mathcal{M}}$ and treat the identities from two datasets as different ones. However, based on the experiments conducted by He *et al*. [8], it can be observed that the performance would unexpectedly drop with the full usage of $\mathcal{D}_{\mathcal{A}}$. It can be inferred that the quantity of data in $\mathcal{D}_{\mathcal{A}}$ is much larger than those in $\mathcal{D}_{\mathcal{M}}$, so the training process would sample the data from $\mathcal{D}_{\mathcal{A}}$ with higher frequency and, therefore, overemphasize it. To tackle such problem, we propose a balanced cross-domain learning which constitutes every training batch by fixed rate of training data randomly sampled from both $\mathcal{D}_{\mathcal{M}}$ and $\mathcal{D}_{\mathcal{A}}$. Besides, in case the model is overly trained on large-scale $\mathcal{D}_{\mathcal{A}}$, for single epoch, the sampler would stop sampling excessive data from $\mathcal{D}_{\mathcal{A}}$ after all training data from $\mathcal{D}_{\mathcal{M}}$ is sampled. Based on this method, we can avoid training model partially fitting on $\mathcal{D}_{\mathcal{A}}$. Additionally, we also find the huge amount of identities from $\mathcal{D}_{\mathcal{A}}$ would bring the trivial identity recognition on $\mathcal{D}_{\mathcal{A}}$ and lead to performance drop. Therefore, we empirically train the classification loss (also called ID loss) $\mathcal{L}_{\text{ID}}$ only for the data sampled from $\mathcal{D}_{\mathcal{M}}$.

Finally, motivated by Liu *et al*. [17], we also apply Maximum Mean Discrepancy (MMD) measure [6], which is usually used to minimize the distance between distributions of two different domains, as our cross-domain loss. The MMD loss $\mathcal{L}_{\text{MMD}}$ can be formulated as follow with the notation of the distribution of x as p and the distribution of y as q:

$$\mathcal{L}_{\text{MMD}} = \| \frac{1}{n_m} \sum_{i=1}^{n_m} \phi(f_{c,i}^m) - \frac{1}{n_a} \sum_{j=1}^{n_a} \phi(f_{c,j}^a) \|_{\mathcal{H}}^2, \quad (7)$$

where ϕ is the mapping operation which projects the distribution into a Reproducing Kernel Hilbert Space $\mathcal{H}$ [5].

3.3. Multi-Target Multi-Camera Tracking

Since the movement of the vehicles are restricted by the road structures, we can easily recognize certain moving patterns and classify them into several trajectories. By assigning tracklets to define trajectories in 3.1, we can exploit the geographical relationship between cameras to remove invalid matching pairs. Besides the spatial constraint, we can further select the matching pairs with the temporal constraints, such as setting a time window or checking overlapping region in time domain.

With re-ID feature extractor, we can generate the tracklet-based features by simply averaging all features of

each frame and construct a Euclidean distance matrix of each tracklet pairs. Given this matrix, hierarchical clustering is performed under the following constraints:

1. Select the tracklet pairs which are possibly connected under given geographical relationship as a candidate .

2. Remove the tracklet pairs from candidates if there are overlapping regions in the time domain.

3. Use a time window to further filter the invalid pairs whose traveling time across cameras is unreasonable.

Constraint 1. is our **spatial constraints**. To associate two tracklets within different cameras, we use the geographical relationship between cameras to develop a camera link model, which is based on [10]. This model consists of the trajectory pairs which contain potential connected trajectories from different cameras. To simplify the network, only the trajectories within adjacent cameras are selected as a pair. Constraints 2.&3. are **temporal constraints**. Besides spatial constraints, temporal constraints are adopted for a better MTMCT quality. In non-overlapping camera scenarios, we remove the tracklet pairs with overlapping in the time domain from candidates. Furthermore, a time window is set for each trajectory pair to filter unreasonable matching pairs whose traveling time across cameras is out of range.

4. Experiments

4.1. Datasets and Evaluation Metrics

In this section, we will introduce the datasets released by the official of 2021 AI City Challenge, mainly for Track 2 (City-Scale Multi-Camera Vehicle Re-Identification) and Track 3 (City-Scale Multi-Camera Vehicle Tracking). Note that the usage of external datasets is prohibited in the challenge, so we do not use other datasets.

Multi-camera Vehicle Tracking. The dataset for Track 3, named CityFlowV2 [33, 22, 25], includes 46 camera views within 6 different scenarios. Training set consists of 3 scenarios with 36 cameras, and validation set consists of 2 scenarios with 23 cameras among which 19 cameras have existed in training set. Remaining 6 cameras form a scenario used as testing set. The camera views in testing set are non-overlapping, while the scale of detections are much smaller than training and validation set which makes this dataset more challenging. We use the official evaluation matrix, namely IDF1, to evaluate the experiment results.

Vehicle Re-Identification The datasets for Track 2 contains one real-world dataset, along with one synthetic dataset. As the real-world one, CityFlowV2-ReID [33, 22, 25] (CFV2-ReID) contains 85,058 images of 880 vehicles captured by 46 cameras which is split into 440 vehicles with 52,717 images for training and other 440 vehicles with 32,341 images of for testing. For the testing set, 1,103 images are for queries and 31,238 images are for galleries. Due to the limitation of submissions, we manually split the training set of CFV2-ReID into training and validation set which respectively includes 21,760 images of 340 vehicles and 10,581 images of 100 vehicles, denoted as Split-train and Split-test, to evaluate the performance of each method. As for the synthetic one, Vehicle X [42, 30] contains 192,150 images of 1,362 vehicles for usage. Besides to identity, the images are also labeled with colors, vehicle types, orientation, etc. As in previous vehicle re-ID works, we employ the standard metrics, namely the cumulative matching curve (CMC) and the mean average precision (mAP) to evaluate the results.

4.2. Implementation Details

Multi-camera Vehicle Tracking. As described in Sec. 3, we perform hierarchical clustering twice respectively in Tracklet Reconnection for both single camera tracking (SCT) and multi-target multi-camera tracking (MTMCT). The threshold used to restrict feature distance and the max iteration of clustering are respectively set as 10 and 50 in SCT, while those in MTMCT are defined as 10 and 200 due to more candidate pairs across cameras than within a single camera.

Vehicle Re-Identification Our baseline model is modified from the code released by He *et al*. [8]. In the training stage, for each training batch, we will randomly sample 12 identities and 6 instances for every vehicle. Then, the training images would be resized into 352×352 and extracted as representative features by ResNet101_IBN_a as backbone. Finally, the batch-hard triplet loss $\mathcal{L}_{Triplet}$ and classification loss (cross-entropy Loss) $\mathcal{L}_{ID}$ would jointly be computed to train the backbone model. As for the utilization of auxiliary dataset $\mathcal{D}_{\mathcal{A}}$, followed by our balanced cross-domain learning as illustrated in Fig. 4, we randomly sample 8 and 4 identities from $\mathcal{D}_{\mathcal{M}}$ (CFV2-ReID) and $\mathcal{D}_{\mathcal{A}}$ (Vehicle X) for every single batch. And, the features from both $\mathcal{D}_{\mathcal{M}}$ and $\mathcal{D}_{\mathcal{A}}$ will be additionally used to compute $\mathcal{L}_{MMD}$ achieve domain generalization.

4.3. Analysis of Tracklet Reconnection in Multiple Camera Vehicle Tracking

Table 1 demonstrates the effectiveness of our Tracklet Reconnection module. The first row is our baseline model. A camera link model [10] is built to reduce the search space in hierarchical clustering. When adding the constraints of completeness (C_{com}) and inherent characteristics (C_{inher}), the performance is improved by 12% in terms of IDF1, as

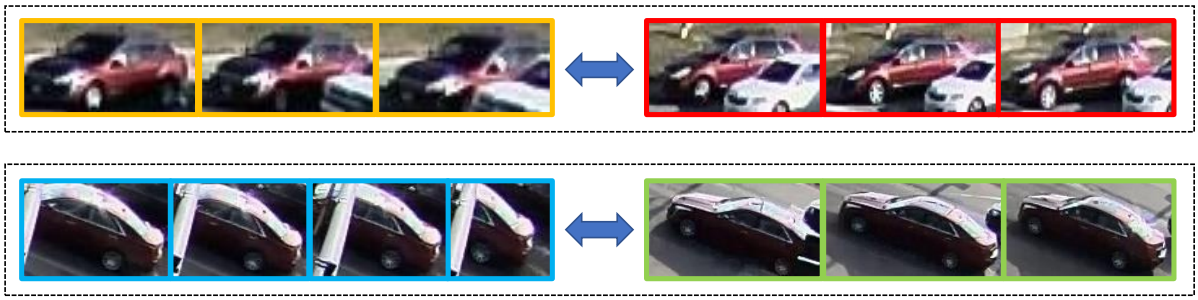

Figure 5: **Visualization of Tracklet Reconnection Technique.** In the top and bottom rows, we respectively show the two failure pairs of split tracklets generated by TrackerNet Tracker [36] due to the occlusion of target object. With the help of Tracklet Reconnection, they are correctly associated into single complete tracklet.

Table 1: **Analysis of Tracklet Reconnection on CityFlow V2.** TR : Tracklet Reconnection.

Method	TR	IDF1
Baseline [10]	✗	0.3805
	w/o GPS	0.5096
	w/ GPS	**0.5458**

shown in the second row. Furthermore, in the last row, when using the GPS information to clearly determine the orientation similarity between tracklets, it can finally improve 4% and achieve 0.5458 in terms of IDF1. Fig. 5 is the visualization results of our method.

4.4. Analysis of Balanced Cross-Domain Learning for Vehicle Re-Identification

We evaluate the effectiveness of our proposed method, Balanced Cross-Domain Learning (BCDL), on both Split-test and the official test set in CityFlowV2-ReID. Table 2 demonstrates the results on our Split-test validation set. We can see that in the second row of the table, if we add all the training data from the auxiliary dataset $\mathcal{D}_\mathcal{A}$, the bias of $\mathcal{D}_\mathcal{A}$ will crush the model performance, which performs even worse than not using $\mathcal{D}_\mathcal{A}$. Our porposed BCDL is trying to solve this problem. We sample 8 IDs from $\mathcal{D}_\mathcal{M}$ and 4 IDs from $\mathcal{D}_\mathcal{A}$ to form a batch and add $\mathcal{L}_{MMD}$ making two domains closer to achieve the better performance. The two results are shown in the third and fourth row in the table, respectively, and both methods successfully enhance the performance. Last, we show the performance of our method with BCDL and $\mathcal{L}_{MMD}$ on the test set of CityFlowV2-ReID in Table 3. It shows with the two methods, it can promisingly improve the mAP score from 29.96% to 37.97%.

Table 2: **Analysis of Balanced Cross-Domain Learning on Split-test set.** *All*: using the whole dataset of Vehicle X for training; *BCDL*: our proposed sampling strategy.

Method	Trained with Vehicle X	$\mathcal{L}_{MMD}$	Split-test mAP	Split-test rank-1
Baseline [8]	✗	✗	32.74	45.42
	All	✗	27.68	34.52
	BCDL	✗	40.12	51.93
	BCDL	✓	**41.25**	**53.23**

Table 3: **Analysis of Balanced Cross-Domain Learning on CityFlowV2-ReID.**

Method	Trained with Vehicle X	$\mathcal{L}_{MMD}$	CFV2-ReID mAP	CFV2-ReID rank-1
Baseline [8]	✗	✗	29.96	41.43
	BCDL	✗	37.11	47.73
	BCDL	✓	**37.97**	**48.23**

4.5. Competition Results

Multi-camera Vehicle Tracking. For track3, we adopt the method proposed by [10] as a baseline model. After adding the constraints of completeness judgement and inherent characteristics, we improve the IDF1 score by 12%. Additionally, with the help of GPS information, the performance is further improved by 4%. The final IDF1 scores is presented in Table 4. Our work has achieved 9^{th} place with **0.5458** in terms of IDF1.

Vehicle Re-Identification In the final leaderboard of track2, with the model which ensembles with different predictions from multiple training configurations, our team, team ID 79, has achieved 0.4240 mAP score on AI City Challenge 2021 Track 2. Table 5 shows the ranking, our performance ranks in the 21^{st} place.

Table 4: **Competition results of AICITY21 Track3.**

Rank	Team ID	Team Name	IDF1
1	75	mcmt	0.8095
2	29	fivefive	0.7787
3	7	CyberHu	0.7651
4	85	FraunhoferIOSB	0.6910
5	42	DAMO	0.6238
6	27	Janus Wars	0.5763
7	15	aiforward	0.5654
8	48	BUPT-MCPRL2	0.5534
9	**79**	**Ours**	**0.5458**
10	112	Dukbaegi	0.5452

Table 5: **Competition results of AICITY21 Track2.**

Rank	Team ID	Team Name	mAP
1	47	DMT	0.7445
2	9	NewGeneration	0.7151
3	7	CyberHu	0.6650
4	35	For Azeroth	0.6555
5	125	IDo	0.6373
21	**79**	**Ours**	**0.4240**

5. Conclusion

In this paper, we propose an efficient multi-camera vehicle tracking system which mainly contains two novel components. First, to refine the primary single-camera tracking results, the Tracklet Reconnection technique is introduced to associate multiple mistakenly split tracklets due to the occlusion of target objects. Second, when we use large-scale auxiliary dataset to assist the training on main dataset, the training sample imbalance problem would lead to unexpected performance drop. Hence, we further propose the Balanced Cross-Domain Learning with a new training data sampler and loss function to avoid overemphasizing on the auxiliary dataset. We conduct extensive experiments and show the empirical effectiveness of our proposed components.

Acknowledgment

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 109-2218-E-002-026), National Taiwan University (NTU-108L104039), Intel Corporation, Delta Electronics and Compal Electronics.

References

[1] Tsai-Shien Chen, Man-Yu Lee, Chih-Ting Liu, and Shao-Yi Chien. Viewpoint-aware channel-wise attentive network for vehicle re-identification. In *Proc. CVPR Workshops, Seattle, WA, USA*, 2020.

[2] Tsai-Shien Chen, Chih-Ting Liu, Chih-Wei Wu, and Shao-Yi Chien. Orientation-aware vehicle re-identification with semantics-guided part attention network. In *European Conference on Computer Vision*, pages 330–346. Springer, 2020.

[3] Viktor Eckstein, Arne Schumann, and Andreas Specker. Large scale vehicle re-identification by knowledge transfer from simulated data and temporal attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 616–617, 2020.

[4] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104*, 2018.

[5] Arthur Gretton, Karsten Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*, 2008.

[6] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pages 673–681, 2009.

[7] 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.

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 576–577, 2020.

[10] 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.

[11] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

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

[13] Sangrok Lee, Eunsoo Park, Hongsuk Yi, and Sang Hun Lee. Strdan: Synthetic-to-real domain adaptation network for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 608–609, 2020.

[14] Y. Lee, J. Hwang, and Z. Fang. Combined estimation of camera link models for human tracking across nonoverlapping cameras. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2254–2258, 2015.

[15] 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.

[16] Peng Li, Jiabin Zhang, Zheng Zhu, Yanwei Li, Lu Jiang, and Guan Huang. State-aware re-identification feature for multi-target multi-camera tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[17] Chih-Ting Liu, Man-Yu Lee, Chih-Wei Wu, Bo-Ying Chen, Tsai-Shien Chen, Yao-Ting Hsu, Shao-Yi Chien, and NTU IoX Center. Supervised joint domain learning for vehicle re-identification. In *CVPR Workshops*, pages 45–52, 2019.

[18] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[20] Yihang Lou, Yan Bai, Jun Liu, Shiqi Wang, and Ling-Yu Duan. Veri-wild: A large dataset and a new method for vehicle re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3235–3243, 2019.

[21] 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.

[22] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C. Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, and Siwei Lyu. The 2019 AI City Challenge. In *Proc. CVPR Workshops*, pages 452–460, Long Beach, CA, USA, 2019.

[23] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 2665–2674, June 2020.

[24] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th ai city challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, page 2665–2674, June 2020.

[25] Milind Naphade, Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Xiaodong Yang, Liang Zheng, Anuj Sharma, Rama Chellappa, and Pranamesh Chakraborty. The 4th AI City Challenge. In *Proc. CVPR Workshops*, Virtual, 2020.

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

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

[28] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018.

[29] Zheng Tang and Jenq-Neng Hwang. Moana: An online learned adaptive appearance model for robust multiple object tracking in 3d. *IEEE Access*, 7:31934–31945, 2019.

[30] 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 *ICCV*, 2019.

[31] 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.

[32] 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.

[33] 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 *Proc. CVPR*, pages 8797–8806, Long Beach, CA, USA, 2019.

[34] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 108–115, 2018.

[35] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 108–115, 2018.

[36] 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.

[37] 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.

[38] 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. IEEE, 2017.

[39] Chih-Wei Wu, Chih-Ting Liu, Cheng-En Chiang, Wei-Chih Tu, and Shao-Yi Chien. Vehicle re-identification with the space-time prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 121–128, 2018.

[40] Chih-Wei Wu, Chih-Ting Liu, Wei-Chih Tu, Yu Tsao, Yu-Chiang Frank Wang, and Shao-Yi Chien. Space-time guided association learning for unsupervised person re-identification. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2261–2265. IEEE, 2020.

[41] 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.

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

[43] Zhedong Zheng, Minyue Jiang, Zhigang Wang, Jian Wang, Zechen Bai, Xuanmeng Zhang, Xin Yu, Xiao Tan, Yi Yang, Shilei Wen, et al. Going beyond real data: A robust visual representation for vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 598–599, 2020.

[44] Yi Zhou and Ling Shao. Viewpoint-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.