PhD Learning: Learning with Pompeiu-hausdorff Distances for Video-based Vehicle Re-Identification

Jianan Zhao[1], Fengliang Qi[1], Guangyu Ren[2*], Lin Xu[1†]
[1]Shanghai Em-Data Technology Co., Ltd. [2]Imperial College London

Abstract

Vehicle re-identification (re-ID) is of great significance to urban operation, management, security and has gained more attention in recent years. However, two critical challenges in vehicle re-ID have primarily been underestimated, i.e., 1): how to make full use of raw data, and 2): how to learn a robust re-ID model with noisy data. In this paper, we first create a video vehicle re-ID evaluation benchmark called VVeRI-901 and verify the performance of video-based re-ID is far better than static image-based one. Then we propose a new Pompeiu-hausdorff distance (PhD) learning method for video-to-video matching. It can alleviate the data noise problem caused by the occlusion in videos and thus improve re-ID performance significantly. Extensive empirical results on video-based vehicle and person re-ID datasets, i.e., VVeRI-901, MARS and PRID2011, demonstrate the superiority of the proposed method. The source code of our proposed method is available at https://github.com/emdata-ailab/PhD-Learning.

1. Introduction

Vehicle re-identification (re-ID) aims to locate and recognize a vehicle of interest across multiple non-overlapping cameras in various traffic intersections. It is of great significance to urban operation, management, security [31, 72, 38], and has gained more attention in recent years [33, 65, 50, 57, 46]. The challenges are exponentially increasing for the visual appearance based vehicle re-ID tasks, such as tiny intra-class variations, multiple camera viewpoints, various illumination conditions, severe occlusions, and complex traffic conditions [51, 46, 64], *e.g.*, a car might glow different colors at varying viewing angles and light beams, while the vehicles of same model usually exhibit limited visual differences. Vehicle re-ID can be conducted on either images or videos. The existing vehicle re-ID has been extensively studied for still images via matching spatial appearance features [50, 38, 57, 46, 65]. However, static image-based approaches are intrinsically limited due to the visual ambiguities (*e.g.*, occlusions, viewpoints, and resolutions) and the lack of spatio-temporal information. Video sequences contain richer spatial and temporal clues and are beneficial for identifying a vehicle under complex surveillance conditions. Currently, making use of videos brings new challenges to vehicle re-ID. The diffi-

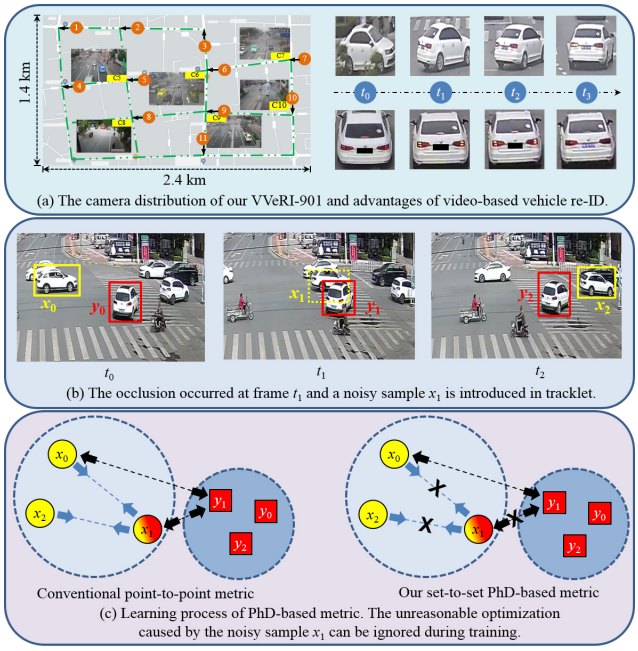


Figure 1. Schematic illustration of the advantage of video-based vehicle re-ID and the proposed Pompeiu-hausdorff distance (PhD) Learning method. (a): We create a video-based vehicle re-ID benchmark from complex traffic intersections. The rich spatial-temporal information in video can resist visual ambiguities. (b): Severe partial and full occlusions frequently occur in a traffic intersection surveillance video. It would introduce a large number of occlusion samples for recognition. (c): The proposed set-to-set PhD Learning method for video-to-video matching. The noisy sample (*e.g.*, x_1) can be eliminated automatically during the optimization process. Colors (*i.e.*, yellow and red) indicate the semantical visual appearance, while shapes (*i.e.*, squares and circles) represent the annotated label (*i.e.*, ground truth).

*Work done while an intern at Shanghai Em-Data Technology Co., Ltd.
†Contact Author (Email: lin.xu5470@gmail.com)

culties mainly come from the following two aspects: 1) An adequate quantity and high-quality video-based vehicle re-ID dataset is absent. To the best of our knowledge, most of the current vehicle re-ID datasets are constructed from sampled static images [31, 33, 50, 38, 57, 46, 65], where the consecutive spatial-temporal information are insufficient. Moreover, the diversity (*e.g.*, variations in viewpoint, occlusion, illumination, and resolution) of cameras' captured data are also oversimplified. These restrictions might make limited contributions to construct a reliable and robust appearance-based model. The right subfigure in Figure 1(a) presents a toy example to illustrate the advantages of the successive video data for re-ID. Two video tracklets with the same identity (ID) would be matched more accurately at frame t_3. 2) An effective video-based vehicle re-ID method of seeking discriminative features from the videos is also critically needed. Video-based re-ID benefits from rich spatial-temporal data to resist the aforementioned visual ambiguities. However, it also brings additional difficulties in accurately matching video sequences, especially the problem of matching frames from the videos with occlusion [41, 52]. As illustrated in Figure 1(b), two video tracklets $\{x_0, x_1, x_2\}$ and $\{y_0, y_1, y_2\}$ are labeled with IDs X and Y, respectively. At Frame t_1, the visual appearance of bounding box x_1 is heavily occluded by that of y_1. It leads to x_1 has a very similar visual feature with y_1, while still been labeled as ID X. These kinds of occlusion samples frequently occur in a surveillance video captured from crowded scenes (*e.g.*, traffic intersections). It will cause great difficulties for subsequent identification and deteriorate the recognition performance significantly.

In this paper, we have done the following two works to overcome the above limitations: 1) We firstly create a new Video-based Vehicle Re-Identification benchmark named *VVeRI-901* [1]. Some distinctive characteristics are summarized as: a) Unconstrained capture conditions involving multiple intersections motivate visual information diversity in viewpoint, resolution, and illumination, etc, as shown in Figure 2. b) Successive spatial and temporal information without any further down-sampling is contained to enhance the appearance-based model's robustness in tackling visual ambiguities. c) With the aid of rich information, more related research areas can be facilitated, like cross-resolution re-ID [29], cross-view matching [66], and multi-view synthesis [5]. 2) We then propose a set-to-set Pompeiu-hausdorff Distance (PhD) learning method for video-to-video matching. It can eliminate the occlusion samples automatically during the optimization process. Figure 1(c) illustrates the PhD learning method's optimization process. In the conventional metric learning method [23, 18, 1, 43], all the images within a mini-batch will be employed for optimizing the metric space, and the occlusion samples would play an adverse influence on the optimization, *e.g.*, the distance of positive pairs x_0 and x_1 (noise) with considerable visual discrepancy will be narrowed for the large ground distance, and the case is opposite for negative pairs (*e.g.*, x_1 and y_1). In contrast, as for the proposed PhD metric learning, the aforementioned detrimental positive pairs (*e.g.*, x_0 and x_1) could be excluded automatically due to the selected pairs to be optimized in PhD metric space can only be composed of samples from different video tracklets. Additionally, in terms of the negative with the highest similarity (*e.g.*, x_1 and y_1), it can also be eliminated in that the PhD measures the maximum mismatch between two point sets via the max-min optimization. In a nutshell, our main contributions are summarised as follows:

1. We create a new *VVeRI-901* benchmark for video-based vehicle re-ID. It is the first successive video-based vehicle re-ID benchmark captured from unconstrained real-world traffic intersections.

2. We propose a new PhD learning method for video-to-video matching in re-ID tasks. It can alleviate the occlusion problem in video-based re-ID and improve recognition performance significantly.

3. We verify the superiority of our proposed method on video-based re-ID tasks, including video-based vehicle re-ID and video-based person re-ID.

2. Related Work

2.1. Vehicle Re-ID Benchmarks

The pre-existing vehicle re-ID benchmarks are constructed from interval sampled static images. The spatial-temporal information and the diversity of the data might be insufficient. It will cause difficulties for subsequent identification. *VeRi-776* [34] contains 776 identities in $50,000$ images, which were collected from 20 surveillance cameras with different attributes and spatiotemporal labels. *PKU-VD* [58] has two large-scale sub-datasets VD1 and VD2, captured from two different cities with high-resolution cameras and surveillance cameras. *VehicleReID* [32] was collected from two non-overlapping surveillance cameras. It only captures the front and back viewpoints, and the occlusion is also not considered. *CityFlow* [50] is a city-scale traffic camera benchmark. It was collected from 40 cameras of 10 traffic intersections in an American mid-sized city. *CarsReid74k* [46] contains almost $74,000$ vehicle tracks with identity annotation. The 66 cameras captured the tracks on the bridges overlooking the same freeway. *VERI-Wild* [38] is a large-scale benchmark built in the wild, and over $400,000$ images were captured by an extensive surveillance system containing 174 cameras covering a large urban district more than 200km^2.

[1]Part of the *VVeRI-901* dataset is preliminarily released at https://gas.graviti.cn/dataset/hello-dataset/VVeRI901.

Benchmarks	IDs	Cameras	Boxes	N-Overlap	Multi-View	Multi-Reso	Multi-Illu	Occlusion	Video
VeRi-776 [34]	776	20	49,357	✗	✓	✓	✓	✗	✗
VehicleReID [32]	26,267	2	221,763	✓	✗	✓	✓	✗	✗
PKU-VD1 [58]	1,232	-	846,358	-	✗	✓	✓	✗	✗
PKU-VD2 [58]	1,112	-	807,260	-	✗	✓	✓	✗	✗
CityFlow [50]	666	40	229,680	✗	✓	✓	✓	✗	✗
VERI-Wild [38]	40,671	174	416,314	✓	✓	✓	✓	✓	✗
***VVeRI-901* (Ours)**	901	11	488,195	✓	✓	✓	✓	✓	✓

Table 1. Comparison with publicly available vehicle re-ID benchmarks. For each benchmark, the table illustrates the number of identities, bounding boxes, non-overlapping scenarios (N-Overlap), multiple viewing angles, multiple resolutions (Multi-Reso), multiple illumination conditions (Multi-Illu), occluded data (Occlusion), and sequential video-based (Video).

2.2. Video-based Re-ID Methods

The video-based re-ID can be regarded as an extension of single-shot image-based methods. It adopts an image sequence to further improve the matching accuracy [68]. McLaughlin *et al*. [41] first proposed a primary deep learning pipeline for video-based person re-ID. It selects discriminative frames is usually performed before feature extraction, aiming at selecting sufficient informative information while avoiding redundancy [3]. Typically adopted strategies are random sampling [3, 47, 56], restricted random sampling [53, 30, 28], snippet sampling [2], and periodical sampling [69], etc. Generating informative temporal representations is also a critical issue, and different sequential feature fusion methods have significant effects on the final performance of the models [13]. Among various methods, recurrent neural networks (RNNs) [41, 60], pooling (average or maximum) [70] and attention-based models [56, 36] are most widely applied. Besides, some proposed models incorporate the spatial and temporal feature extraction together via 3D CNN [26]. Almost all of the pre-existing works focus on the structure-related improvement, rarely on designing suitable metric learning methods for resolving video-based re-ID problems.

2.3. Distance Metric Learning

Distance Metric learning [23] aims to learn an optimal distance metric to measure the similarity among samples. Recently, deep metric learning [18, 1, 43, 55, 59, 48] shows a better ability to solve real-world problems and has attracted attention in various fields. Most existing supervised re-ID methods apply identification loss for identity classification (*e.g.*, cross-entropy loss) [49] and verification loss for metric learning (*e.g.*, triplet loss) [16]. Triplet loss [43] is designed initially for the face recognition problem, in which an anchor, a positive sample, and a negative sample are included. Quadruplet loss [4] is an improved version of triplet loss, which contains two different negative samples to learn a larger inter-class distance and a smaller intra-class distance compared to the triplet loss. However, almost all of the existing re-ID methods share the identical criterion of a point-to-point distance metric, and there is no valid set-to-set distance metric specially designed for sequential images in video-based re-ID.

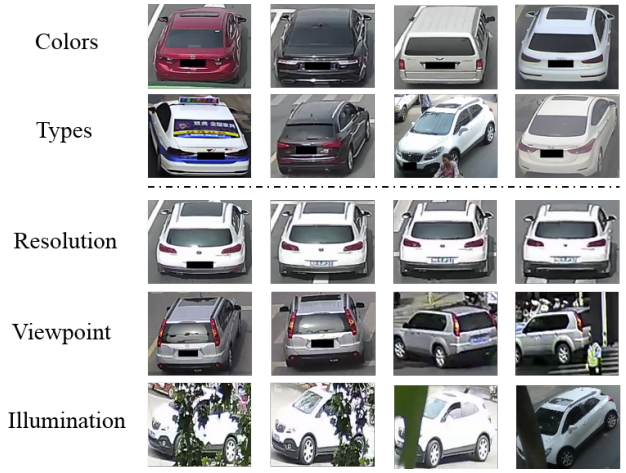
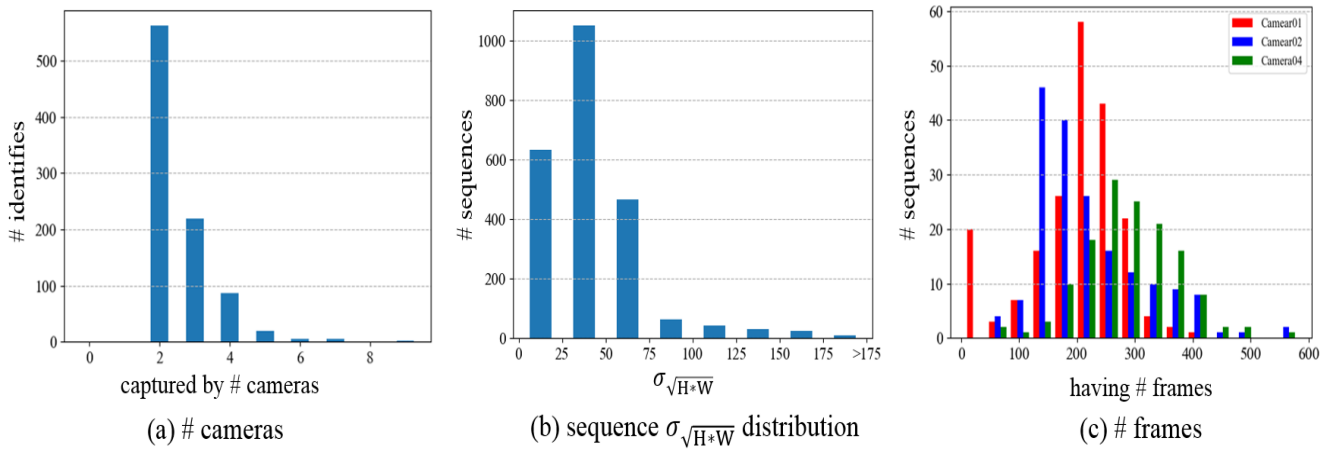


Figure 2. Some image and tracklet instances sampled from our *VVeRI-901* dataset. Within a tracklet, more visual information are included, and temporal information can be explored with appearance-based models for vehicle re-ID task.

3. *VVeRI-901* Benchmark

The raw video data of the *VVeRI-901* are captured from a mid-sized city of China with an area of 1.1km $\times$ 2.4km. In this region, 11 non-overlapping surveillance cameras are deployed at different traffic intersections. We tailor the raw videos captured by all the cameras from $7:00$ am to $10:00$ am when traffic is massive, and more vehicle IDs can be recorded accordingly. We mask the license plates in the *VVeRI-901* for privacy consideration and make the model more focus on the vehicle's visual appearance.

Data Annotations: To acquire the tracklets of each vehicle in every single camera, we manually annotate the bounding box of each tracklet with the help of the Computer Vision Annotation Tool (CVAT) [44]. After getting all the tracklets in each camera, we associate the same vehicle that appears at different cameras with the auxiliary spatial-temporal cues. It is worth noting that we find some vehicles stopping and keeping still at zebra crossings in some



(a) # cameras (b) sequence $\sigma_{\sqrt{H*W}}$ distribution (c) # frames

Figure 3. Statistic of the *VVeRI-901* dataset. (a): The number of identities across different cameras. (b): The distribution of the variance of resolution within a tracklet. (c): The distribution of the length of tracklet at traffic intersections.

raw videos. These vehicles provide the same appearance in each frame and poss limited variations in a tracklet, leading to severe information redundancy in the dataset. In order to avoid this redundancy issue, we define a region of interest (ROI) for each scenario and retain the clips within the ROI.

Data Statistics: The proposed dataset contains 901 IDs (*i.e.*, 451 IDs for training and 450 IDs for testing), 2,320 tracklets, and 488,195 bounding boxes. As sketched in Figure 3(a), all the IDs are captured by at least two cameras, and most of the IDs are captured by 2-4 cameras, indicating that *VVeRI-901* is an ideal benchmark for algorithms to explore multiple queries or re-ranking methods. In Figure 3(b), we calculate the distribution of the standard deviation of $\sqrt{\mathrm{H} * \mathrm{W}}$ for the tracklets, herein and after, H and W represent the height and width of each frame in a tracklet, respectively. Most of the tracklets render more or less scale changes due to the vehicle's high moving speed. Figure 3(c) shows the distribution of the sequence length at traffic intersections. The sequence length presents a distinct distribution at different crossroads due to the discrepancy between various scenes. Generally, most of the sequences from all cameras contain 100-400 frames, guaranteeing information diversity in the proposed benchmark dataset.

Data Characteristics: The leading contribution of the proposed benchmark is providing challenging factors to facilitate the development of vehicle re-ID methods in realistic scenarios. As shown in Table 1, the main practical factors in vehicle re-ID tasks, including occlusions and viewpoints, are fully considered only in the VERI-Wild [38] and our benchmark, *i.e.*, *VVeRI-901*. However, compared with *VVeRI-901*, vehicle images from the side view and severe occlusions can hardly be found in VERI-Wild. Only images with limited viewpoints and minor occlusions are provided. Furthermore, the *VVeRI-901* is the only existing video-based vehicle re-ID benchmark. It benefits from the sequential video information, and some small but informative clues can be preserved for matching and identification. The video is closer to the raw data captured from the surveillance system from the real-world application perspective, making the *VVeRI-901* more practical and challenging. More detailed information about our *VVeRI-901* benchmark can be found in our supplementary materials.

Evaluation Protocols: To evaluate the performance of vehicle re-ID tasks, there are two widely used performance metrics, namely, mean Average Precision (mAP) and Cumulative Matching Curve (CMC). More specifically, the mAP indicates the overall performance of re-identification:

$$AP = \frac{\sum_{l=1}^{n} S(k) \times gt(k)}{M_{gt}}, \tag{1}$$

where k is the rank in the order of vehicles of size n, M_{gt} is the number of relevant vehicles. $S(k)$ is the precision at cut-off k and $gt(k)$ indicates whether the k-th recall is correct. Therefore, the mAP is defined as:

$$mAP = \frac{\sum_{u=1}^{V} AP(u)}{V}, \tag{2}$$

where V represents the number of total query images. The CMC shows the ranking capabilities of re-ID models through quantitative evaluation. The CMC value at rank k can be calculated as:

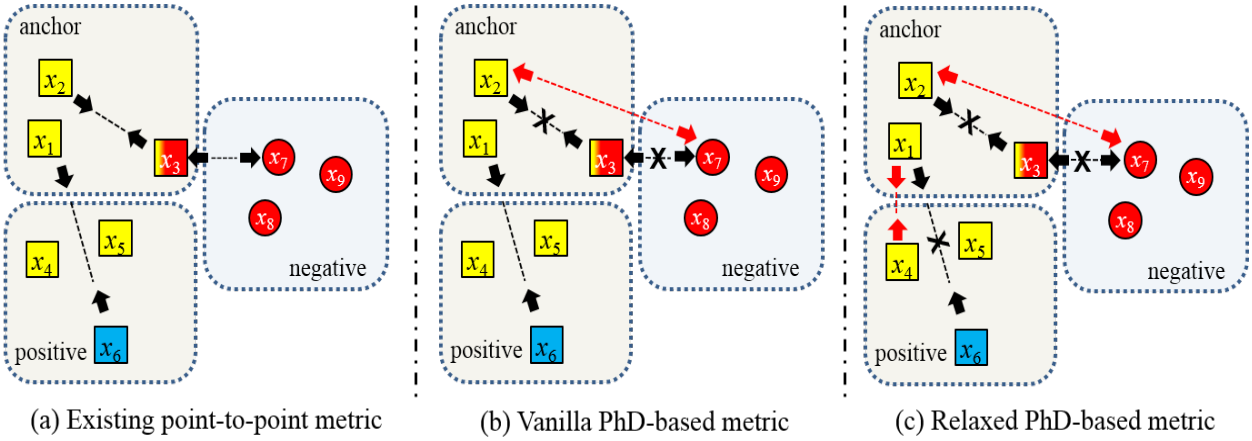
$$CMC@k = \frac{\sum_{u=1}^{V} gt(u,k)}{V}, \tag{3}$$

where $gt(u,k)$ is equal to 1 when the ground truth of u image appears before rank k.

4. Our Method

4.1. Vanilla Pompeiu-hausdorff Distance Metric

The pompeiu-hausdorff distance is widely used to measure the similarity between two sets of points [21, 8]. Let $\mathbb{S}_1$ and $\mathbb{S}_2$ be two non-empty subsets of the Euclidean



(a) Existing point-to-point metric (b) Vanilla PhD-based metric (c) Relaxed PhD-based metric

Figure 4. **The basic principle of Pompeiu-hausdorff distance (PhD) metric learning in solving the occlusion problem.** Shapes (*i.e.*, squares and circles) represent the annotated label while filling colours (*i.e.*, yellow, blue, and red) represent the semantic visual appearance. (a): Three kinds of harmful metric pairs due to the outlier (*i.e.*, x_3) and label noise (*i.e.*, x_6) when applying the existing point-to-point metric. (b): Without extra modification, the *vanilla* PhD metric can alleviate the impacts of an outlier but failed in resisting the label noise. (c): The *relaxed* PhD metric can alleviate the impacts of the outlier and label noise simultaneously.

metric space (G, d_{E}). Their pompeiu-hausdorff distance $d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2)$ denotes that any point in set $\mathbb{S}_1$ is at most at distance $d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2)$ to some point in set $\mathbb{S}_2$. Specifically, $d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2)$ can be defined as:

$$d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2) = \max\Big\{ \sup_{p\in\mathbb{S}_1} \inf_{q\in\mathbb{S}_2} d_{\mathrm{E}}(p,q), \sup_{q\in\mathbb{S}_2} \inf_{p\in\mathbb{S}_1} d_{\mathrm{E}}(p,q)\Big\}, \tag{4}$$

where sup and inf denote supremum and infimum, respectively. Equivalent form of Equation (4) [42] is:

$$\begin{aligned} d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2) &= \inf\left\{\varepsilon \geq 0; \mathbb{S}_1 \subseteq \mathbb{S}_2^{\varepsilon} \text{ and } \mathbb{S}_2 \subseteq \mathbb{S}_1^{\varepsilon}\right\}, \\ \mathbb{S}_1^{\varepsilon} &= \bigcup_{p\in\mathbb{S}_1} \left\{g \in G; d_{\mathrm{E}}(g,p) \leq \varepsilon\right\}, \end{aligned} \tag{5}$$

Equation (5) represents the set of all points within ϵ of the set $\mathbb{S}_1$ and sometimes called the ϵ-fattening of $\mathbb{S}_1$ or a generalized ball of radius ϵ around $\mathbb{S}_1$.

In video-based re-ID, a whole tracklet (*i.e.*, image sequence) can be regarded as a point set. Supposing $\mathbb{S}_1, \mathbb{S}_2 \subset \mathbb{R}^N$ are two sets composed of M and N points, respectively. The Pompeiu-hausdorff distance $d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2)$ between two sets of points $\mathbb{S}_1$ and $\mathbb{S}_2$ can be defined as:

$$\begin{aligned} d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2) &= \max\{d_{\mathrm{U}}(\mathbb{S}_1, \mathbb{S}_2), d_{\mathrm{U}}(\mathbb{S}_2, \mathbb{S}_1)\}, \\ d_{\mathrm{U}}(\mathbb{S}_1, \mathbb{S}_2) &= \max_{p\in\mathbb{S}_1} \min_{q\in\mathbb{S}_2} d_{\mathrm{E}}(p,q), \\ d_{\mathrm{U}}(\mathbb{S}_2, \mathbb{S}_1) &= \max_{q\in\mathbb{S}_2} \min_{p\in\mathbb{S}_1} d_{\mathrm{E}}(q,p), \end{aligned} \tag{6}$$

where $d_{\mathrm{H}}(\mathbb{S}_1, \mathbb{S}_2)$ is also called bi-directional Pompeiu-hausdorff distance. The $d_{\mathrm{U}}(\mathbb{S}_1, \mathbb{S}_2)$ and $d_{\mathrm{U}}(\mathbb{S}_2, \mathbb{S}_1)$ are two uni-directional Pompeiu-hausdorff distances between the two sets, which is the maximum value measuring the distance from points in one set to its nearest neighbor point in the other set. Furthermore, the bi-directional Pompeiu-hausdorff distance is the larger one of the two uni-directional Pompeiu-hausdorff distances.

4.2. Noises Elimination with Relaxed PhD Metric

We clarify the principle of PhD learning for resisting the noisy samples caused by the occlusion problem in Figure 4. Three video tracklets with two identities are exemplified here, and in each tracklet, three samples are included. Specifically, inspired by [67], two types of noisy samples, *outlier* (*i.e.*, x_3) and *label noise* (*i.e.*, x_6), are investigated here in details, which result from partial and full occlusion, respectively. As shown in Figure 4(a), in the existing point-to-point metric learning method, three kinds of metric pairs (*i.e.*, $\{x_1, x_6\}$, $\{x_2, x_3\}$ and $\{x_3, x_7\}$) caused by the noisy samples (*i.e.*, x_3 and x_6) are introduced. It would play an adverse influence on the optimization, *e.g.*, the distance of positive pairs x_2 and x_3 (x_6) with considerable visual discrepancy will be narrowed for the large ground distance and the case is opposite for negative pairs (*e.g.*,x_3 and x_7). The problems essentially originate from that in the conventional point-to-point method, each sample is treated independently, and the structure of the tracklet is not considered. In contrast, for the PhD metric shown in Figure 4(b), only the distance between two sets is taken into account, and the metric pair within the same tracklet (*e.g.*, $\{x_2, x_3\}$) can be discarded automatically. At the same time, according to definition of uni-directional Pompeiu-hausdorff distance $d_{\mathrm{U}}(\mathbb{S}_1, \mathbb{S}_2)$ ($d_{\mathrm{U}}(\mathbb{S}_2, \mathbb{S}_1)$) in the equation 6, the sample pairs between two sets with minimum distance are discarded, and the metric pair $\{x_3, x_7\}$ can also be avoided accordingly.

Algorithm 1: Pseudocode of PhD Learning.

Input: network F parameterized by θ, dataset $\mathcal{D}$ and hyperparameter k, τ.
while not MaxEpoch **do**
  /* Sample a mini-batch */
  $\mathcal{B} \leftarrow \{(\mathbb{S}_i, y_i) \sim \mathcal{D}\}_{i=1}^{|\mathcal{B}|}$
  /* Compute *relaxed* PhD metric d_H^k */
  for $\mathbb{S}_i = \{x_{i,1}, \cdots, x_{i,p}, \cdots, x_{i,S_1}\} \in \mathcal{B}$ **do**
    for $\mathbb{S}_j = \{x_{j,1}, \cdots, x_{j,q}, \cdots, x_{j,S_2}\} \in \mathcal{B}(i \neq j)$ **do**
      $d_{\mathrm{U}}^k(\mathbb{S}_i, \mathbb{S}_j) \leftarrow k\text{th max}\{\min\{F(x_{i,p})^\top F(x_j)\}\}$
      $d_{\mathrm{U}}^k(\mathbb{S}_j, \mathbb{S}_i) \leftarrow k\text{th max}\{\min\{F(x_{j,q})^\top F(x_i)\}\}$
      $d_{\mathrm{H}}^k(\mathbb{S}_i, \mathbb{S}_j) \leftarrow \max\{d_{\mathrm{U}}^k(\mathbb{S}_i, \mathbb{S}_j), d_{\mathrm{U}}^k(\mathbb{S}_j, \mathbb{S}_i)\}$
    end
  end
  /* Compute $\mathcal{L}_{\mathrm{PhD}}$ and the gradient */
  $\mathcal{L}_{\mathrm{PhD}} \leftarrow$ torch.nn.MarginRankingLoss$(d_{\mathrm{H}}^k, y, \tau)$
  $\delta\theta \leftarrow \partial_\theta \mathcal{L}_{\mathrm{PhD}}$
end
Output: network F

However, the *vanilla* PhD metric is sensitive to the label noise x_6 due to the maximum operation in the calculation of uni-directional Pompeiu-hausdorff distance. To further circumvent this difficulty, we propose a *relaxed* PhD metric learning by relaxing the maximum constraint of uni-directional distance. Denoting the distance as:

$$d_{\mathrm{H}}^k(\mathbb{S}_1, \mathbb{S}_2) = \max\Big\{ \underset{p\in\mathbb{S}_1}{k\text{th-max}}\{\min_{q\in\mathbb{S}_2} d_{\mathrm{E}}(p,q)\}, \underset{q\in\mathbb{S}_2}{k\text{th-max}}\{\min_{p\in\mathbb{S}_1} d_{\mathrm{E}}(q,p)\}\Big\}, \tag{7}$$

where $k\text{th-max}_{p\in\mathbb{S}_1}\{\min_{q\in\mathbb{S}_2} d_{\mathrm{E}}(p,q)\}$ means selecting kth maximum value in set $\mathbb{D}_1 = \{\min_{q\in\mathbb{S}_2} d_{\mathrm{E}}(p_i,q), p_i \in \mathbb{S}_1\}$ and vice versa. The *relaxed* PhD metric can automatically select the k best matching points of $\mathbb{S}_1$ ($\mathbb{S}_2$) because it identifies the subset of the whole set of size k that minimizes the uni-directional Hausdorff distance and the metric pair caused by the label noise x_6 can be avoided attributed to the largest distance as shown in Figure 4 (c).

4.3. Loss Objective of Set-based PhD learning

We propose the PhD loss objective to adopt a set-based sample selection strategy for the video-based re-ID task. The PhD loss is developed based on Batch Hard (BH) triplet loss [16]. Specifically, let f be the embedding feature of a sample x learned with a function, and $\mathbb{S}$ represents the tracklet of one object. Following the sampling strategy in [16], the batches are formed by randomly sampling P classes (object identities) and then randomly sampling K sequences of each class. For each sequence, S samples are sampled from the whole tracklet. Now, the BH triplet loss function for video-based re-ID can be re-defined as follow:

$$\mathcal{L}_{\text{BH-Tri}} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K*S}}^{\text{all anchors}} \Big[\tau + \overbrace{\max_{p=1\ldots K*S} d_{\mathrm{E}}(f_a^i, f_p^i)}^{\text{hardest positive}} - \underbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K*S \\ j\neq i}} d_{\mathrm{E}}(f_a^i, f_n^j)}_{\text{hardest negative}} \Big]_+ \tag{8}$$

where τ is the margin. The footnotes a, p and n represent the *anchor*, *positive* and *negative*, respectively.

To evaluate the video-based distance in a set-to-set manner, we revise the triplet strategy based on the proposed *relaxed* pompeiu-hausdorff distance and design the PhD loss objective. For a specific anchor $\mathbb{S}_a$, the PhD loss can be formulated as follow:

$$\mathcal{L}_{\text{PhD}} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}} \Big[\tau + \overbrace{\max_{p=1\ldots K} d_{\mathrm{H}}^k(\mathbb{S}_a^i, \mathbb{S}_p^i)}^{\text{hardest positive}} - \underbrace{\min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} d_{\mathrm{H}}^k(\mathbb{S}_a^i, \mathbb{S}_n^j)}_{\text{hardest negative}} \Big]_+ \tag{9}$$

where d_{H}^k represents the *relaxed* hausdorff distance. It should be noted that the sampling strategies in *vanilla* (k=1) and *relaxed* (k-th max) PhD can be regarded as batch-hard and semi-hard methods in [24] as mentioned. The pseudo-code of the PhD learning is described in the Algorithm 1.

5. Experiments

5.1. Experiment Settings

Architecture: Standard ResNet-50 [15] and OSNet-AIN [71] pretrained on ImageNet are applied as the backbone for video-based vehicle and person re-ID, respectively. We adopt cross-entropy as identification loss (*i.e.*, ID loss [49]) for identity classification, triplet loss [16] and our PhD loss as verification loss for metric learning. As suggested by Luo *et al.* [40], we apply the structure of the BNNeck layer, which is positioned after features extracted by the backbone and before classifier fully connected layers. Furthermore, both the BNNeck and fully connected layers are initialized through Kaiming's initialization [14].

Datasets: Each image in a image tracklet [2] is prepossessed and resized to the same resolution (256×256 for *VVeRI-901*, 256×128 for *MARS* [70], and *PRID2011* [17]) in pixels

[2] According to the "probe-to-gallery" pattern [70], the queries and galleries are both image tracklets for supporting video-to-video re-ID mode.

and synchronously augmented with a random horizontal flip operation for each image tracklet.

Parameters: Adam [22] optimizer is selected with an initial learning rate of 3.5×10^{-4}. The commonly adopted warm-up strategy [9] is applied to bootstrap the network for better performance. In practice, the network is optimized for 120 epochs. We spend 10 epochs linearly increasing the learning rate from 3×10^{-6} to 3×10^{-4}, and it then decays by 10 at 40th epoch and 70th epoch respectively. In each

Strategies		mAP	R1	R5	R10	R20
Aggregation Strategies ($S = 6$)	Avg. Pooling	41.8	41.4	61.2	70.2	76.0
	Max Pooling	40.5	39.6	59.1	68.4	77.3
	Attention	39.8	39.2	61.4	66.1	74.7
	LSTM	37.2	36.5	54.3	61.8	69.8
Tracklets Lengths (Avg.)	$S = 4$	39.7	38.4	58.6	66.7	72.4
	$S = 6$	41.8	41.4	61.2	70.2	76.0
	$S = 8$	42.0	41.2	61.2	68.8	75.3
Sampling Strategies ($S = 6$, Avg.)	Evenly	41.8	41.4	61.2	70.2	76.0
	Random	45.7	44.3	64.3	73.1	80.0
	RandomSeq.	45.2	43.7	63.9	73.6	80.7

Table 2. Ablation study results of PhD learning on the *VVeRI-901*.

Methods	mAP	R1	R5	R10	R20
GoogLeNet [63]	41.4	40.8	59.6	65.3	72.0
ID Loss [35]	36.5	37.2	52.4	60.5	68.6
TCLNET-tri* [19]	44.0	45.5	58.0	67.1	72.9
MGH [61]	44.5	44.3	61.8	67.8	74.1
Triplet+ID(I-I) [16]	26.1	27.8	41.5	47.8	51.5
Triplet+ID(I-V) [16]	33.7	35.6	51.2	54.3	59.7
Triplet+ID(V-I) [16]	33.3	36.8	50.3	55.6	61.4
Triplet+ID(V-V) [16]	41.8	41.4	61.2	70.2	76.0
PhD+ID ($S = 6$, Avg., Ran.)	47.2	47.1	67.6	74.7	80.4

Table 3. Comparison results with other methods on the *VVeRI-901*.

training batch, we sample 8 identities (*i.e.*, $P = 8$), each with 4 tracklets (*i.e.*, $K = 4$), and the sequence length of each tracklet is 6 (*i.e.*, $S = 6$). Besides, the margin parameter τ defined in the loss objectives is set to 0.3.

5.2. Evaluation on Video-based Vehicle Re-ID

Aggregation Strategies Evaluation: Some commonly applied aggregation methods including temporal pooling [13], temporal attention [7] and RNN [41] are evaluated on the *VVeRI-901*. In the temporal pooling model, we consider two modes, *i.e.*,, max pooling and average pooling. In the temporal attention model, we apply an attention weighted average on the sequence of image features as suggested in [13]. For RNN fusion, we test the Long Short-Term Memory (LSTM). It can be found from the evaluation results in the top part of Table 2 that temporal pooling methods outperform the other methods, while the LSTM method is inferior in terms of mAP and CMC on *VVeRI-901*. Furthermore, the average pooling method shows superior performance compared with the max pooling.

Sequence Lengths Evaluation: Then, we explore the effects of different sequence lengths (*i.e.*, $S = 4$, $S = 6$, and $S = 8$) with the evenly sampling strategy. It can be seen in Table 2 that the mAP and CMC increase with the sequence length, while the performance of $S = 6$ and $S = 8$ are similar, indicating improvement can be limited when further increasing the sequence length.

Sampling Strategies Evaluation: We compare three commonly used sampling strategies: 1) Evenly sampling strategy divides the whole tracklet into S clips with the same length first and then select an image in each clip to construct the sequence. 2) Random sampling strategy randomly samples S frames in each tracklet. 3) Random sequence sampling strategy randomly choose a consecutive clip with a length of S. The comparison results are listed in Table 2. The random sampling strategy performs better than others cause of suitable randomness within the tracklets.

Comparison with The Existing Methods: We evaluate the PhD learning performance with some commonly used methods on the *VVeRI-901*. Four probe-to-gallery patterns [70] (*i.e.*, image-to-image (I-I), image-to-video (I-V), video-to-image (V-I), and video-to-video (V-V) pattern) are also considered. The comparison results in Table 3 demonstrate that the PhD combined with ID loss outperforms the other methods by a large margin.

5.3. Evaluation on Video-based Person Re-ID

Noise Robustness Evaluation: We evaluate the noise resistance ability of the proposed PhD learning method. The *vanilla* (*i.e.*, k=1) and *relaxed* (*i.e.*, k=S/2, where S is the tracklet length) PhD metric learning methods are compared in Table 4. Since the noise intensity correlates with the length of image tracklets, we also investigate the method by sampling different sequence length (*i.e.*, $S = 4$, $S = 6$, and $S = 8$). Empirical results on *MARS* show the *relaxed* (*e.g.*, k=S/2,S=8) PhD metric is more robust than the *vanilla* PhD. In contrast, the *vanilla* PhD loss shows superior performance on *PRID2011*. The reason lies in that the *MARS* uses DPM detector [11] and GMMCP tracker [6] for pedestrian detection and tracking, respectively. It brings in a number of false detection and tracking annotations [70]. The *relaxed* PhD loss with a larger value of k (and the corresponding S) is more effective to reduce these outliers. Compared with the *MARS*, the *PRID2011* dataset has fewer outliers due to manual annotations. Thus, the *relaxed* PhD loss may regard some hard positive samples as outliers and filter them, which result in insufficient optimization for the model parameters and degraded the performance compared with the *vanilla* PhD loss objective.

The *MARS* also provides optional distractors (labelled

Tracklet Lengths	*MARS* [70]						*PRID2011* [17]			
	$k=1$ (*vanilla*)			$k=S/2$ (*relaxed*)			$k=1$ (*vanilla*)		$k=S/2$ (*relaxed*)	
	mAP	R1	R5	mAP	R1	R5	R1	R5	R1	R5
$S=4$	81.5	85.8	95.8	**82.5**	86.7	96.4	**85.4**	94.4	83.1	94.4
$S=6$	82.7	87.0	96.5	**83.6**	87.7	96.2	**92.1**	97.8	88.8	96.6
$S=8$	82.0	86.0	96.2	**84.1**	88.0	96.6	**89.9**	97.8	89.9	95.5

Table 4. Ablation study results on the *relaxed* parameter k and tracklet length S.

Methods (S=6)	*MARS*			
	mAP	R1	R5	R10
BH triplet (w/o noise)	82.2	86.5	95.9	97.4
BH triplet (w noise)	70.0	79.5	91.3	93.8
PhD (w/o noise)	86.2	88.9	97.0	97.9
PhD (w noise)	80.4	84.2	94.7	96.3

Table 5. Noise robustness evaluation results.

Loss Objectives	*MARS* [70]			*PRID2011* [17]	
	mAP	R1	R5	R1	R5
ID Loss [49]	79.1	84.4	94.7	85.4	95.5
Quadruplet+ID [4]	79.4	85.4	94.2	85.4	96.6
Triplet+ID [16]	82.2	86.5	95.9	87.6	96.9
PhD+ID	**83.6**	**87.7**	**96.2**	**92.1**	**97.8**

Table 6. Evaluation results of several loss objectives.

Methods	*MARS* [70]				*PRID2011* [17]		
	mAP	R1	R5	R20	R1	R5	R20
RQEN [45]	51.7	73.7	84.9	91.6	91.8	98.4	99.8
RRU+STIM [37]	72.7	84.4	93.2	96.3	92.7	98.8	99.8
GLTR [25]	78.5	87.0	95.8	98.2	95.5	**100.0**	100.0
STA [12]	80.8	86.3	95.7	98.1	-	-	-
AdaptiveGraph [54]	81.9	89.5	96.6	97.8	94.6	99.1	100.0
VRSTC [20]	82.3	88.5	96.5	-	-	-	-
NVAN [30]	82.8	**90.0**	-	-	-	-	-
STGCN [62]	83.7	89.9	96.4	98.3	-	-	-
TACAN [27]	84.0	89.1	96.1	98.0	95.3	-	-
TCLNET-tri* [19]	85.1	89.8	-	-	-	-	-
MGH [61]	85.8	**90.0**	96.7	98.5	94.8	99.3	100.0
SSDML [73]	65.7	74.4	89.4	95.0	86.5	98.9	100.0
MMDML [39]	81.6	86.3	95.7	98.1	85.4	95.5	100.0
SA Triplet [10]	81.8	85.3	95.4	98.2	90.2	99.6	100.0
PhD (Ours)	**86.2**	88.9	**97.0**	**98.6**	**96.6**	97.8	100.0

Table 7. Performance comparison with the state-of-the-art methods. Among them, the middle group of results represents some recent set-based metric learning methods.

as 'ID=0') in the gallery to evaluate the methods' robustness. We randomly insert a distractor in the image tracklet to mimic the label noise. Table 5 shows the performance degraded significantly for the batch-hard triplet, while the *relaxed* PhD is robust. The mAP of PhD dominates the batch-hard triplet by 4.0%. This performance gap could be further expanded to 10.4% after inserting additional noises.

Loss Objectives Evaluation: We compare some commonly used loss objectives with the *relaxed* PhD loss on *MARS* and *PRID2011* datasets. We train the *relaxed* PhD loss with a sequence length of $S=6$, and an *evenly sampling* technique is used for all the methods for a fair comparison. Comparison results in Table 6 show the *relaxed* PhD combined with ID loss outperforms the other losses.

Comparison with The State-of-the-art Methods: We compare our method with the state-of-the-art methods on the *MARS* and *PRID2011* in Table 7. To further evaluate the set-based PhD metric, we also make comparison with some exiting set-to-set metric learning methods (*e.g.*, SSDML [73], MMDML [39], and SA Triplet [10]). It also should be noted that we use different sampling strategies in ablation studies (*i.e.*, Table 4 and 6) and the "SOTA" contest (*i.e.*, Table 5 and 7). In ablation studies, we *evenly* sample S frames from tracklets for both training and querying (fixed S=6) for eliminating randomness. When compared with the SOTAs, we *random* sample S=6 frames for training, and use whole tracklets for querying as [61] to boost the performance. The results in Table 7 shows that the PhD method achieves 86.2% mAP without re-ranking on the *MARS*. It is a new SOTA result on this large-scale dataset. The 96.6% in rank-1 on the *PRID2011* also dominates other methods. The PhD metric outperforms other set-based metric learning methods by at least 4.4% in mAP on the *MARS* and 6.4% in rank-1 on the *PRID2011*. The comparison results demonstrate the superiority of our PhD learning method for both video-based vehicle and person re-ID tasks.

6. Conclusion

In this paper, we created a new video-based vehicle re-ID benchmark *VVeRI-901* to promote the research of vehicle re-ID. It is the first video-based vehicle re-ID benchmark captured from unconstrained intersections. The spatial-temporal clues in videos are rich, diverse, and change successively. It is beneficial for identifying a vehicle under complex surveillance conditions in the wild. Although more information can be obtained from vehicle videos, more challenges come along. *E.g.*, the noises induced by partial and full occlusions in vehicle videos are more severe. Thus, we further propose the video-to-video *relaxed* PhD learning method. The delicate matching and mining strategies in PhD metric learning can resist the noise and improve recognition significantly. Our current study mainly focuses on correctly stressing noises caused by severe occlusions while fully exploiting the abundant visual sequential information in collected vehicle videos. Video-based vehicle re-ID task on the created *VVeRI-901* is still challenging as the PhD merely achieves 47.2% mAP. The more vital spatial-temporal information, *e.g.*, timestamps and camera coordinates, is also pre-annotated in *VVeRI-901*. Our future work will focus more on modelling the prior knowledge and gain deeper insights on the video vehicle re-ID task.

Acknowledgment: This work was supported by the Shanghai Rising-Star Program (20QB1405500), and in part by the Open Project of Key Laboratory of Ministry of Public Security for Road Traffic Safety (2020ZDSYSKFKT03-1).

References

[1] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)*, 34(4):98, 2015. 2, 3

[2] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018. 3

[3] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 28(9):4192–4205, 2019. 3

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017. 3, 8

[5] Eddie Cooke, Peter Kauff, and Thomas Sikora. Multi-view synthesis: A novel view creation approach for free viewpoint video. *Signal Processing: Image Communication*, 21(6):476–492, 2006. 2

[6] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, June 2015. 7

[7] W. Ding, X. Wei, X. Hong, and Y. Gong. Complex spatial-temporal attention aggregation for video person re-identification. In *ICIP*, pages 2441–2445, 2020. 7

[8] Thomas Eiter and Heikki Mannila. Distance measures for point sets and their computation. *Acta Informatica*, 34(2):109–133, 1997. 4

[9] Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, 2019. 7

[10] Pengfei Fang, Pan Ji, Lars Petersson, and Mehrtash Harandi. Set augmented triplet loss for video person re-identification, 2020. 8

[11] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 7

[12] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, volume 33, pages 8287–8294, 2019. 8

[13] Jiyang Gao and Ram Nevatia. Revisiting temporal modeling for video-based person reid. *arXiv:1805.02104*, 2018. 3, 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 3, 6, 7, 8

[17] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 6, 8

[18] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 2, 3

[19] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, 2020. 7, 8

[20] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019. 8

[21] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. 4

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 7

[23] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013. 2, 3

[24] Ratnesh Kumar, Edwin Weill, Farzin Aghdasi, and Parthsarathy Sriram. Vehicle re-ID: an efficient baseline using triplet embedding. In *IJCNN19'*, 2019. 6

[25] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, 2019. 8

[26] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, volume 33, pages 8618–8625, 2019. 3

[27] Mengliu Li, Han Xu, Jinjun Wang, Wenpeng Li, and Yongli Sun. Temporal aggregation with clip-level attention for video-based person re-identification. In *WACV*, 2020. 8

[28] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018. 3

[29] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *CVPR*, pages 8090–8099, 2019. 2

[30] 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. In *arXiv:1908.01683*, 2019. 3, 8

[31] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016. 1, 2

[32] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016. 2, 3

[33] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-

identification for urban surveillance. In *ECCV*, pages 869–884. Springer, 2016. 1, 2

[34] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, pages 869–884. Springer, 2016. 2, 3

[35] X. Liu, L. Wu, M. Tao, and H. Ma. *A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance*. Computer Vision – ECCV 2016, 2016. 7

[36] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, pages 5790–5799, 2017. 3

[37] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, volume 33, pages 8786–8793, 2019. 8

[38] 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 *CVPR*, June 2019. 1, 2, 3, 4

[39] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1137–1145, 2015. 8

[40] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019. 6

[41] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 2, 3, 7

[42] J.R. Munkres. *Topology*. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000. 5

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2, 3

[44] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020. 3

[45] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018. 8

[46] Jakub Spanhel, Jakub Sochor, Roman Juránek, Petr Dobes, Vojtech Bartl, and Adam Herout. Learning feature aggregation in temporal domain for re-identification. *arXiv:1903.05244*, 2019. 1, 2

[47] Xinxing Su, Xiaoye Qu, Zhikang Zou, Pan Zhou, Wei Wei, Shiping Wen, and Menglan Hu. k-reciprocal harmonious attention network for video-based person re-identification. *IEEE Access*, 7:22457–22470, 2019. 3

[48] Han Sun, Zhiyuan Chen, Shiyang Yan, and Lin Xu. Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification. In *ICCV*, pages 6737–6747, 2019. 3

[49] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, pages 3800–3808, 2017. 3, 6, 8

[50] 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 *CVPR*, pages 8797–8806, 2019. 1, 2, 3

[51] Peng Wang, Bingliang Jiao, Lu Yang, Yifei Yang, Shizhou Zhang, Wei Wei, and Yanning Zhang. Vehicle re-identification in aerial imagery: Dataset and approach. In *ICCV*, pages 460–469, 2019. 1

[52] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, 21(6):1412–1424, 2018. 2

[53] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, and Qi Tian. Adaptive graph representation learning for video person re-identification. *arXiv:1909.02240*, 2019. 3

[54] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, and Qi Tian. Adaptive graph representation learning for video person re-identification. *arXiv:1909.02240*, 2019. 8

[55] Lin Xu, Han Sun, and Yuai Liu. Learning with batch-wise optimal transport loss for 3d shape recognition. In *CVPR*, pages 3333–3342, 2019. 3

[56] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, pages 4733–4742, 2017. 3

[57] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *ICCV*, pages 562–570, 2017. 1, 2

[58] Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *ICCV*, pages 562–570, 2017. 2, 3

[59] Shiyang Yan, Jun Xu, Yuai Liu, and Lin Xu. Hornet: a hierarchical offshoot recurrent network for improving person re-id via image captioning. *arXiv:1908.04915*, 2019. 3

[60] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016. 3

[61] Yichao Yan, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, Ying Tai, and Ling Shao. Learning multi-granular hypergraphs for video-based person re-identification. In *CVPR*, 2020. 7, 8

[62] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, 2020. 8

[63] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. pages 3973–3981, 06 2015. 7

[64] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv:2001.04193*, 2020. 1

[65] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020. 1, 2

[66] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017. 2

[67] Tianyuan Yu, Da Li, Yongxin Yang, Timothy M Hospedales, and Tao Xiang. Robust person re-identification by modelling feature uncertainty. In *ICCV*, pages 552–561, 2019. 5

[68] Ruimao Zhang, Jingyu Li, Hongbin Sun, Yuying Ge, Ping Luo, Xiaogang Wang, and Liang Lin. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, 28(10):4870–4882, 2019. 3

[69] Wei Zhang, Shengnan Hu, and Kan Liu. Learning compact appearance representation for video-based person re-identification. *arXiv:1702.06294*, 2017. 3

[70] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016. 3, 6, 7, 8

[71] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Learning generalisable omni-scale representations for person re-identification, 2020. 6

[72] 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. 1

[73] P. Zhu, L. Zhang, W. Zuo, and D. Zhang. From point to set: Extend the learning of distance metrics. In *2013 IEEE International Conference on Computer Vision*, pages 2664–2671, 2013. 8