VERI-Wild: A Large Dataset and a New Method for Vehicle Re-Identification in the Wild

Yihang Lou[1,4] Yan Bai[1,5] Jun Liu[2] Shiqi Wang[3] Ling-Yu Duan[1,4,*]
[1]Peking University, Beijing, China [2]Nanyang Technological University, Singapore
[3]City University of Hong Kong, Hongkong, China [4]Peng Cheng Laboratory, Shenzhen, China
[5]Hulu, Beijing, China
{yihanglou, yanbai, lingyu}@pku.edu.cn, jliu029@ntu.edu.sg, shiqwang@cityu.edu.hk

Abstract

Vehicle Re-identification (ReID) is of great significance to the intelligent transportation and public security. However, many challenging issues of Vehicle ReID in real-world scenarios have not been fully investigated, e.g., the high viewpoint variations, extreme illumination conditions, complex backgrounds, and different camera sources. To promote the research of vehicle ReID in the wild, we collect a new dataset called VERI-Wild with the following distinct features: 1) The vehicle images are captured by a large surveillance system containing 174 cameras covering a large urban district (more than $200km^2$). 2) The camera network continuously captures vehicles for 24 hours in each day and lasts for 1 month. 3) It is the first vehicle ReID dataset that is collected from unconstrained conditions[1]. VERI-Wild contains more than 400 thousand images of 40 thousand vehicle IDs. In this paper, we also propose a new method for vehicle ReID, in which, the ReID model is coupled into a Feature Distance Adversarial Network (FDA-Net), and a novel feature distance adversary scheme is designed to online generate hard negative samples in feature space to facilitate ReID model training. The comprehensive results show the effectiveness of our method on the proposed dataset[2] and the other two existing datasets.

1. Introduction

Vehicle Re-Identification (ReID) aims to retrieve images of a query vehicle from a large-scale vehicle database, which is of great significance to the urban security and city management [9][31]. The straightforward method is to identify vehicles by the recognition of license plates [10][4].

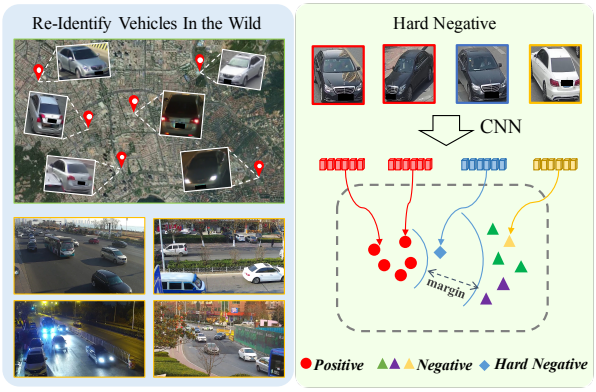


Figure 1. Left: Our dataset is collected with a large-scale real video surveillance system consisting of 174 cameras distributed in a urban district ($>200km^2$). Right: A hard negative generation method is proposed to boost the vehicle ReID performance.

However, in many circumstances, the license plates cannot be clearly captured, sometimes even removed, occluded, or faked. As a result, there is an exponential increase in the demand for the visual appearance based vehicle ReID techniques. The development of deep learning and existing annotated datasets have greatly facilitated the vehicle ReID research. However, the diversity in terms of viewpoint, background, and illumination variations present great challenges to the vehicle ReID models in real-world applications.

In vehicle ReID, the dataset is crucial to comprehensively and fairly evaluate the performance of the ReID methods. However, to the best of our knowledge, all of the existing vehicle ReID datasets [9][26][10] are captured under constrained conditions, and generally have limitations in the following aspects: 1) The number of vehicle identities and images are not large enough to the needs of practical application. 2) The limited camera numbers and covering areas do not involve complex and variant backgrounds in a variety of real-world scenarios. 3) The camera views are highly restricted. For most vehicle datasets, the samples are collected from checkpoint cameras that only capture the front and rear views, and the severe occlusion is also

*Ling-Yu Duan is the corresponding author.

[1]The unconstrained condition arises from the data collection in a real surveillance camera network of a city-scale district, covering huge diversity of viewpoints, resolutions, illuminations, camera sources, weathers, occlusions, backgrounds, vehicle models in the wild, etc.

[2]The dataset is available at https://github.com/PKU-IMRE/VERI-Wild.

not taken into consideration. 4) Most of current datasets are constructed from short-time surveillance videos without significant illumination and weather changes. These limitations may oversimplify the practical challenges of the ReID task, and the ReID models developed and evaluated on such datasets could be inevitably questioned regarding the generalization capability in the wild.

The above issues motivate us to create a new vehicle ReID dataset in the Wild (VERI-Wild) with the following distinctive features: 1) The dataset is captured via a large Closed Circuit Television (CCTV) system, which contains 174 surveillance cameras and covers a large urban district of more than $200km^2$. 2) The unconstrained capture conditions involve complex backgrounds, various viewpoints and occlusion in the wild. 3) The 174 cameras capture for 24h$\times$30days, such that various weathers and illumination conditions are considered. 4) Cleaning from 12 million vehicle images, VERI-Wild contains 416,314 images of 40,671 IDs. VERI-Wild is currently the most challenging dataset for vehicle ReID in real scenarios (see Fig. 1).

Due to the large quantity of vehicle IDs and images, the proposed VERI-Wild dataset poses significant challenges to vehicle ReID. One of the challenges is the similar vehicle problem, where many vehicles with different IDs can have very similar appearances, especially when these vehicles belong to the same vehicle model (see Fig. 1). As such, the remaining visual clues to differentiate such similar appearances are the local characteristic details such as the decorations and customized marks. To promote the capability of the model in capturing subtle differences, it is a wise choice to provide such hard negative pairs for training. Previous attempts [6, 28] focus on seeing more hard negatives by mining them from training set. However, selecting them from the whole training set leads to high computational cost. Moreover, the hard negative samples are limited, and iteratively training them may lead to over-fitting.

In this paper, we propose a Feature Distance Adversarial Network (FDA-Net), in which a novel adversary scheme on feature distance is designed in the embedding space. Within such scheme, a generator aims to online generate hard negative samples from both visual appearance and feature distance perspectives to cheat the embedding discriminator, while the embedding discriminator tries to discriminate them. A similarity constraint is imposed on the generator to make the generated hard negative to be visually similar to the real input, and meanwhile an extra attention regularization is further designed to enforce it to present subtle differences. Besides, the feature representation model (feature extractor) for vehicle ReID is seamlessly coupled into FDA-Net as the embedding discriminator, and end-to-end optimization can be achieved. As the adversary training proceeds, the generated hard negatives would become harder, which in turn promotes the discriminator to become more discriminative. The key idea of generating hard negative samples has significantly improved the state-of-the-art performance on vehicle ReID benchmarks.

Our main contributions are summarized as follows:

(1) A large-scale challenging dataset, VERI-Wild, is proposed for vehicle ReID evaluation in the wild. VERI-Wild is the first vehicle ReID dataset captured from an unconstrained large-scale real-world camera network.

(2) We design a FDA-Net to facilitate the ReID model learning by incorporating a novel feature distance adversary. In FDA-Net, the hard negatives are continuesly online generated to facilitate the learning of more discriminative embedding discriminator.

(3) The FDA-Net achieves superior performance over the state-of-the-art approaches on all the evaluated vehicle ReID datasets. The VERI-Wild dataset and the feature distance adversary scheme is expected to facilitate the large-scale vehicle ReID research from the perspective of figuring out the ReID performance bottleneck in the wild.

2. Related Work

Vehicle ReID Datasets. Recent vehicle ReID methods are mainly evaluated on two public datasets, VehicleID [9] and VeRI-776 [10]. Although impressive results have been achieved on these datasets, the vehicle ReID problem is still far from being addressed in the real-world scenarios. The practical challenging factors have not been fully considered in VehicleID [9] or its extension [26], since they both contain very limited viewpoints (only two views, namely, front and rear). Moreover, they do not contain complex background. Almost no occlusion or illumination changes are considered by them. The samples in VeRI-776 [10] are captured by 18 cameras in a circular road of 1.0 km^2 areas for a short time period (4:00 pm to 5:00 pm in only one day). Again, the limitations of VeRI-776 also lie in the small number of vehicle IDs, simple scenarios, low resolution, etc.

Vehicle Re-Identification. Vehicle ReID has attracted more research efforts in past two years. Liu *et al.* [10] proposed a "PROVID" ReID model that employed visual feature, license plate and spatial-temporal information to explore the ReID task. Shen *et al.* [18] proposed a two-stage framework that incorporates complex spatial-temporal information for effectively regularizing the ReID results. Recent methods [9][29][10] focus on learning an embedding model which map the samples into an embedding space where the samples of the same ID are closer than those of the different, and the similarities between vehicles are measured by the feature distances. Liu *et al.* [9] introduced a mixed difference network using vehicle model and ID information to strengthen the feature representation. Zhou *et al.* [31] designed a multi-view inference scheme to generate global-view feature representation to improve the vehicle ReID. Different from the above methods, our work aims

to explore generating hard negatives in the *feature space* to improve the discriminative capability of the ReID model.

GAN and GAN in ReID. GANs have achieved great success in many tasks, such as image generation [15][5] and translation [32][2][3]. Recent ReID methods also explore GAN both in vehicle and person ReID fields [30][22][13][12]. Zheng *et al.* [30] adopted the DCGAN [15] by using Gaussian noises to generate unlabeled person images before training. Wei *et al.* [22] proposed a PTGAN to transfer person images between different styles for reducing domain gap. Zhou *et al.* [25] designed a GAN model to generate cross-view vehicle images to improve cross-view ReID. Lou *et al.* [12] proposed to generate desired vehicle images from same-view and cross-view to facilitate ReID model training. Some other methods focus on image transferring between different datasets [22][3] or generating different human poses [13], but they are not suitable for vehicles.

Hard Example Learning. Learning from hard examples has always been a hot research topic [11][12][28]. Loshchilov *et al.* [11] proposed to online select hard examples according to loss in SGD optimization. Yuan *et al.* [28] proposed a hard-aware cascaded method to select hard examples for efficient training. However, the diversity of the hard examples in training set is insufficient compared to those in the real-world. Wang *et al.* [20] proposed to add mask to obtain hard positives for improving the robustness against occlusion in detection.

3. VERI-Wild Dataset

3.1. Description of VERI-Wild

We collect a large-scale vehicle ReID dataset in the wild (VERI-Wild), which is captured from an existing large CCTV camera system consisting of 174 cameras across one month ($30 \times 24h$) under unconstrained scenarios. The cameras are distributed in a large urban district of more than $200km^2$. The YOLO-v2 [16] is used to detect the bounding box of vehicles. Our raw vehicle image set contains 12 million vehicle images, and 11 volunteers are invited to clean the dataset for 1 month. After data cleaning and annotation, 416,314 vehicle images of 40,671 identities are collected. We present the statistics of VERI-Wild in Fig. 3, and the sample images from VERI-Wild are also compared in Fig. 2. For privacy consideration, the license plates are masked in our dataset. The distinctive features of VERI-Wild are summarized into the following aspects:

Unconstrained capture conditions in the wild. The VERI-Wild dataset is collected from a real CCTV camera system consisting of 174 surveillance cameras, in which the unconstrained capture conditions pose great challenges.

Complex capture conditions. The 174 surveillance cameras are distributed in an urban district over $200km^2$,

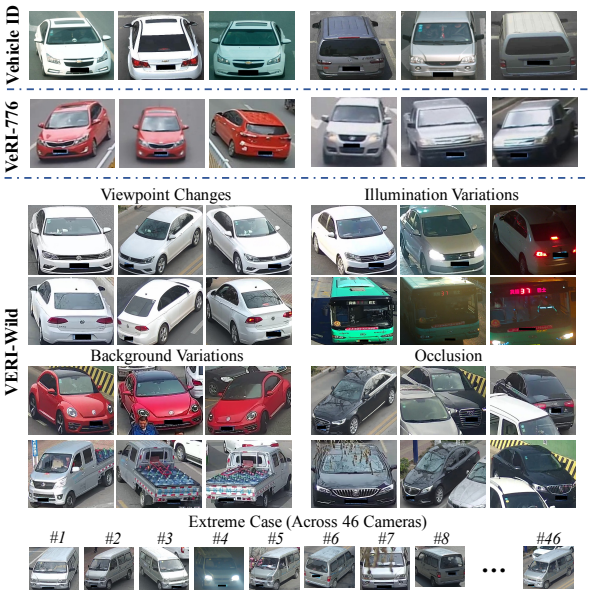


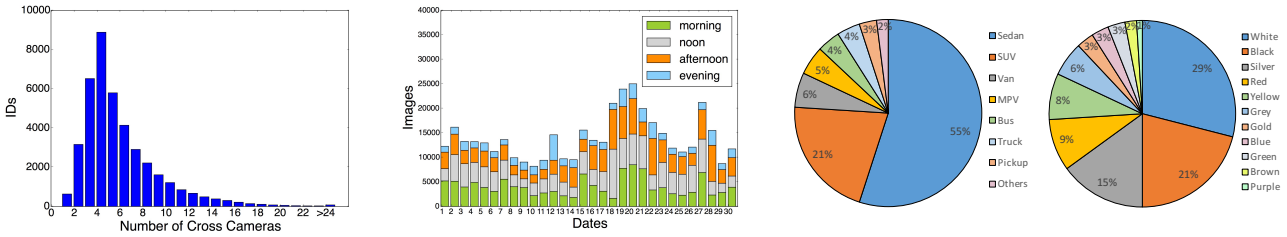
Figure 2. Comparison among the samples of VehicleID [9], VeRI-776 [10] and VERI-Wild datasets. Our collected VERI-Wild dataset poses many more practical challenges for vehicle ReID, e.g., significant viewpoint, illumination, and background variations, and severe occlusion. Another challenge in our dataset is that one vehicle may appear across numerous cameras, e.g., in an extreme case, the same vehicle appears in 46 surveillance cameras.

Table 1. Comparisons among the VehicleID [9], the VeRI-776 [10], and the created VERI-Wild datasets for vehicle ReID.

Dataset	VehicleID	VeRI-776	VERI-Wild
Images	221,763	49,360	**416,314**
Identities	26,267	776	**40,671**
Cameras	12	18	**174**
Capture Time	N/A	18h	**125,280h**
Views	2	6	**Unconstrained**
Spatio-temporal Relation Annotation	×	✓	✓
Tracks Across Cameras	×	×	✓
Camera ID	×	×	✓
Timestamp	×	×	✓
Occlusion	×	×	✓
Complex Background	×	×	✓
Morning	✓	×	✓
Afternoon	✓	✓	✓
Night	×	×	✓
Rainy Weather	×	×	✓
Foggy Weather	×	×	✓

presenting various backgrounds, resolutions, viewpoints, and occlusion in the wild, as shown in Fig. 2. In extreme cases, one vehicle even appears in more than 40 different cameras, which is very challenging for ReID algorithms.

Large time span involving severe illumination and weather changes. The VERI-Wild is collected from a duration of $174 \times 24 \times 30 = 125,280$ video hours. Fig. 3 (b) shows the vehicle distributions in 4 time slots of 24h, *i.e.*, morning, noon, afternoon, evening across 30 days. Also the VERI-Wild contains poor weather conditions, such as rainy,



(a) (b) (c) (d)

Figure 3. The statistics of VERI-Wild dataset. (a) The number of identities across different cameras, *i.e.*, 1-174 cameras; (b) The number of IDs captured in each day; (c) The distribution of vehicle types; (d) The distribution of vehicle colors.

foggy, etc, which are not contained in previous datasets.

Rich Context Information. We provide rich context information such as camera IDs, timestamp, tracks relation across cameras, which are potential to facilitate the research on behavior analysis in camera networks, like vehicle behavior modeling [14], cross-camera tracking [7] and graph-based retrieval [24].

3.2. Evaluation Protocol

The VERI-Wild is randomly divided into two parts for training and testing, as shown in Table 2. To better evaluate ReID methods, we further split the test set into three subsets, as shown in Table 3.

Table 2. The splitting for training and testing sets. (IDs/Images)

Dataset	Train	Probe	Gallery
VehicleID [9]	13,164/100,182	2,400/2,400	2,400/17,638
VeRI-776 [10]	576/37,778	200/1,678	200/11,579
VeRI-Wild	30,671/277,797	10,000/10,000	10,000/128,517

Table 3. Descriptions of the subset of the test set.

Test Size	Small	Medium	Large
Identities	3,000	5,000	10,000
Images	41,816	69,389	138,517

In the ReID process, for each given query, a candidate list sorted by the feature distances between the query and reference images is returned from the database. The mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) are used as performance metrics.

Mean Average Precision: The mAP evaluates the overall performance for ReID, and is defined as follows:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{gt}}, \quad mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{1}$$

where k is the rank in the recall list of size n, and N_{gt} is the number of relevant vehicles. $P(k)$ is the precision at cut-off k and $gt(k)$ indicates whether the k-th recall is correct or not. Q is the number of total query images. Moreover, Top K match rate is also reported in the experiments.

Cumulative Match Characteristics: The CMC curve shows the probability that a query identity appears in different-sized candidate lists. The cumulative match characteristics at rank k can be calculated as:

$$CMC@k = \frac{\sum_{q=1}^{Q} gt(q,k)}{Q}, \tag{2}$$

Figure 4. An example of the real hard negative pair. The two vehicles look very similar, and only subtle differences can be observed, e.g., the details behind the windscreen.

where $gt(q,k)$ equals 1 when the groundtruth of q image appears before rank k.

4. Proposed Method

In learning an embedding model, the hard negative samples play the predominant roles in facilitating the embedding model's discriminative capability [6][20]. The similarity metrics in the embedding space is represented by the feature distance. As shown in Fig. 4, in general, two samples in each real hard negative pair are often similar (with only subtle differences observed). Inspired by this, we design a novel feature distance adversary scheme to generate hard negative samples for enhancing the vehicle ReID model, which consists of two parts, *i.e.*, a similarity constraint and an attention regularization. Given an input vehicle, the similarity constraint is designed to enforce the generated hard negative to be visually similar to the input. To further improve the manipulating capability on subtle differences, an attention regularization is proposed to constrain the attentive regions of the input vehicle and the generated hard negative to be dissimilar. In this manner, the generated hard negatives tend to present visually similar but with subtle differences to the input. As the opposite of the adversary scheme, the discriminator is promoted to be more discriminative with more available hard negatives. Accordingly, a Feature Distance Adversary Network (FDA-Net) is designed in this paper, which includes a hard negative generator G and an embedding discriminator D.

4.1. Hard Negative Generator

Similarity Constraint. To get a visually similar negative to the input, we aim to constrain the generated hard negative closer to its positive than its real sampled negative.

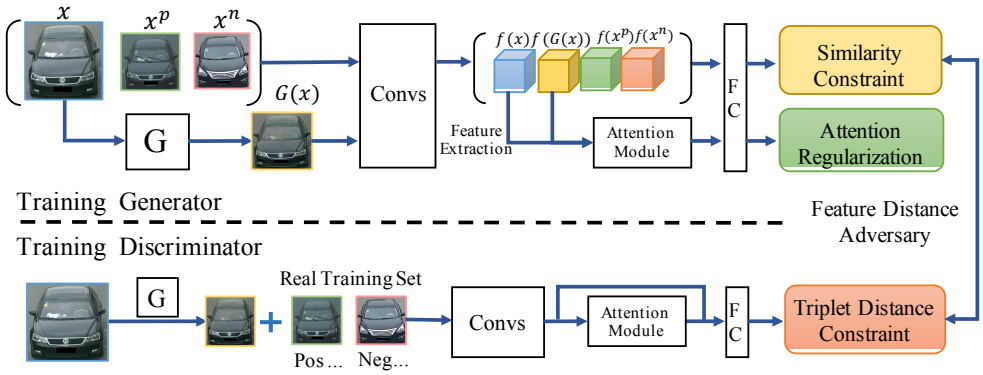


Figure 5. Illustration of the proposed FDA-Net. The feature distance adversary scheme is imposed between the generator G and the embedding discriminator D. The G tries to generate a hard negative sample under similarity constraint and attention regularization, while the D tries to discriminate them. The generator and discriminator are alternatively optimized. We use the generated hard negative $G(x)$ as well as training set to train a more discriminative embedding model D.

Besides, the generated samples should be different from the input. Given a real input vehicle image x, to generate such hard negative sample $G(x)$, the similarity constraint for G can be formulated as follows:

$$\begin{aligned}\|H(x), H(x^p)\|_2^2 + \beta \leq \|H(x), H(G(x))\|_2^2 \\ \leq \|H(x), H(x^n)\|_2^2 - \beta,\end{aligned} \tag{3}$$

where $H(\cdot)$ denotes the feature representation in the embedding space. x^p and x^n represent the real positive and negative of input x, respectively. Such a constraint can be explained as follows: the generated $G(x)$ is constrained to be away from x at a minimum margin gap β, and meanwhile it is also constrained to be away from the real negative x^n at a minimum margin gap β, as shown in Fig. 6. The right part in Eq (3) enforces $G(x)$ to be more similar to x, compared to the sampled real negative, while the left part in Eq (3) constrains $G(x)$ to avoid present the same appearance as x.

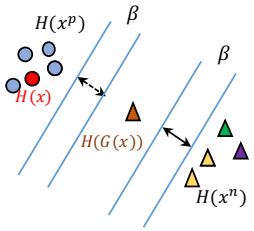


Figure 6. Illustration of the similarity constraint.

The desired $G(x)$ is located in an annular belt region around x. Such characteristics allow the generated $G(x)$ to be more likely to break the distance constraint when optimizing the embedding model, which is also the crucial role of the real hard negatives in training. Finally, the loss for similarity constraint on G can be formulated as:

$$\begin{aligned}L_{G\_sim} = \\ \mathbb{E}_x[\max\{\|H(x), H(x^p)\|_2^2 - \|H(x), H(G(x))\|_2^2 + \beta, 0\}] \\ + \mathbb{E}_x[\max\{\|H(x), H(G(x))\|_2^2 - \|H(x), H(x^n)\|_2^2 + \beta, 0\}].\end{aligned} \tag{4}$$

Attention Regularization. To strengthen the manipulating capability on local subtle differences (see Fig. 4), an attention regularization is further imposed. To promote differences in local regions, the attentive regions between x and $G(x)$ should be relatively far away in feature distance. We design an attention module ATT to perform feature response selection for the intermediate feature $f(x) \in \mathbb{R}^{h\times w\times c}$, then the output attention map is formulated as:

$$A(x) = ATT(f(x); \theta_{att}),\ A \in \mathbb{R}^{h\times w\times 1}. \tag{5}$$

Each patch $a_{i,j}(x)$ in $A(x)$ indicates the attention value (importance score) at (i, j) for $f(x)$. $A(x)$ is normalized with softmax function to be non-negative. The $f(x)$ and $f(G(x))$ are then weighted by $A(x)$ and fed into Fully Connected (FC) layers to get attentive feature representation as:

$$\begin{aligned}F(x) = FC(f(x) \odot A(x)), \\ F(G(x)) = FC(f(G(x)) \odot A(x)).\end{aligned} \tag{6}$$

To ensure the attentive regions are consistent on x and $G(x)$, both $f(x)$ and $f(G(x))$ are weighted by $A(x)$. The regularization enforce the $F(x)$ and $F(G(x))$ to be larger than a minimum margin γ as follows:

$$L_{G\_reg} = \mathbb{E}_x[\max\{\gamma - \|F(x), F(G(x))\|_2^2, 0\}]. \tag{7}$$

For clarity, we use $H(\cdot)$ and $F(\cdot)$ to denote the obtained feature with/without attention regularization. By using the attention regularization, the generator is explicitly promoted to make the subtle differences in some local regions.

4.2. Embedding Discriminator

The embedding discriminator D is fixed to compute feature distance for similarity measurement in training the generator. In contrast, during training the discriminator, G is fixed to generate hard negatives for D. Therefore, D aims to distinguish the hard negative $G(x)$ from x by enlarging their distances via triplet distance constraint, and the loss for L_{D_emb} can be formulated as:

$$\begin{aligned}L_{D\_emb} = \mathbb{E}_x[\max\{\|(H(x), H(x^p)\|_2^2 - \{\|H(x), H(G(x))\|_2^2 \\ + \alpha, 0\}],\ \alpha \geq 2\beta,\end{aligned} \tag{8}$$

where α is the minimum margin gap. In order to make the training more efficient and stable, we mix the generated

$G(x)$ and real negative x^n from the training set to optimize the embedding discriminator D. In addition, we incorporate the extra softmax loss to our training procedure, which is widely used in ReID. Thus, the overall loss L_{D_emb} for D in Eq. (8) is given by:

$$\begin{aligned} L_{D\_emb} = \mathbb{E}_x\left[-\log D_{cls}(I|x)\right] + \\ \mathbb{E}_x\left[\max\{\|H(x), H(x^p)\|_2^2 - \|H(x), H(z)\|_2^2 \right. \\ \left. +\alpha, 0\},\ z \in \{G(x) \cup x^n\}, \end{aligned} \tag{9}$$

where I is the ID label of real input sample x and D_{cls} is another classification objective of D. During the training of D, the sample z is alternatively selected from the union set of $G(x)$ and x^n. We also apply L_{D_emb} to the $F(x)$ to train the attention module. The embedding discriminator can be considered as a feature extractor for vehicle ReID.

4.3. Real/Fake Adversary

Besides satisfying the distance constraints, the generated hard negatives should appear as realistic vehicles. Thus, the real/fake adversarial scheme is further imposed. The output of real/fake discriminator $D_{rf}(x)$ indicates the probability of an image x to be a real one. The loss of $D_{rf}(x)$ can be formulated as a standard cross-entropy loss as:

$$L_{rf} = \mathbb{E}_x[logD_{rf}(x) + log(1 - D_{rf}(G(x)))]. \tag{10}$$

4.4. Overall Loss Function

Finally, the overall loss functions for optimizing the generator and discriminator can be represented as follows:

$$\begin{aligned} L_G &= \lambda L_{rf} + L_{G\_sim} + L_{G\_reg} \\ L_D &= -\lambda L_{rf} + L_{D\_emb}. \end{aligned} \tag{11}$$

where the λ is a hyper parameter to balance the two adversarial schemes. For the whole network, the G and D are alternatively optimized in an adversarial way.

4.5. Training and Testing Details

The feature distance adversarial learning allows us to couple the ReID model and discriminator as an embedding discriminator in FDA-Net. During updating G, the G is optimized to generate a negative sample that satisfies the hard negative's distance constraint with input in the embedding space. When updating D, for each input sample, a hard negative sample is generated by G, which is further combined with real training samples, then their feature distances are optimized by D for discrimination. So the generator is able to continuously generate hard negatives to adapt the iteratively updated embedding discriminator during training, and meanwhile these generated hard negatives can be used to further facilitate the training of D.

Therefore, in testing stage, the vehicle samples are fed into embedding discriminator D to get vehicle feature representation $F(\cdot)$ to perform feature matching in ReID.

4.6. Implementation Details

Network Architecture. For the generator network, two stride-2 convolutions, 9 residual blocks, and two stride 1/2 deconvolutions are used. The size of training images is 224×224. There are two subnetworks for discriminator network. For the real-fake discriminator, we use a PatchGAN of size 70×70 as in [32]. For the embedding discriminator, we use the $VGG\_CNN\_M\_1024$ (*VGGM*) as the base network for fair comparison, which is also adopted in [9]. The attention module is constructed with a 2-layer CNN with 1×1 filters and ReLU activation at the top.

Hyper Parameters. For the discriminator, α in triplet margin constraint is set to 0.6. For the generator, the β in similarity constraint is set to 0.3 and the γ in attention regularization is set to 0.7. The loss weight λ is set to 1. Learning rate starts from 0.001 for embedding discriminator, and starts from 0.0002 for other discriminator and generator. The learning rate is kept constant for the first 50 epochs and decay to zero in the next 50 epochs.

5. Experimental Results

5.1. Experiment setup

We conduct experiments on our proposed VERI-Wild dataset and two existing datasets, VehicleID and VeRI-776, by following the evaluation protocols in [10] and [9], respectively. For match rate computation in VERI-Wild, we follow the standard CMC protocol that all the references of a given query are in gallery. But in VehicleID [9] dataset, there are only one reference of a given query in gallery. In VeRI-776 [10], only cross-camera search is performed. For better evaluation, we perform comparisons as follows:

1) EN: This network is a conventional embedding network with triplet and softmax loss.

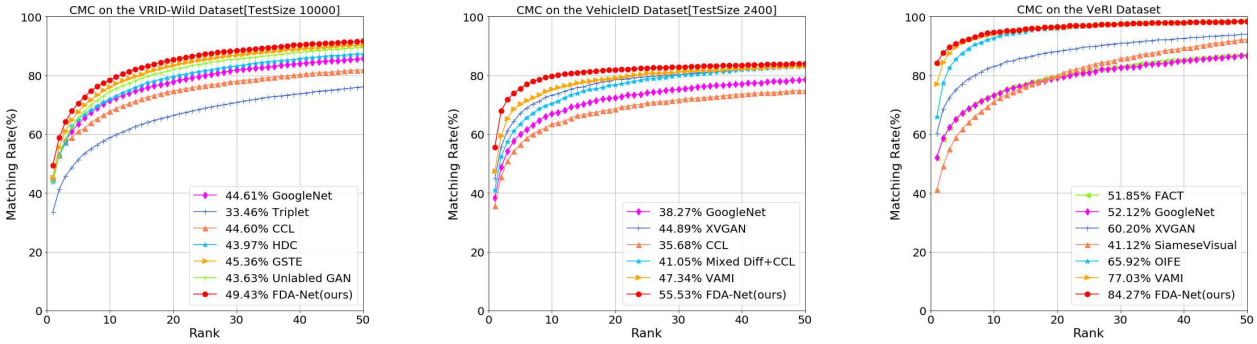
2) FDA-Net: This is the proposed FDA-Net.

3) FDA-Net $\ominus$ Att: This structure is similar to FDA-Net but without attention regularization.

5.2. Quantitative Results

5.2.1 Evaluation on VERI-Wild

To verify the proposed VERI-Wild is challenging and close to the real scenarios, we test the existing methods published in recent 2 years, and use the codes or models provided by their authors to evaluate the performance. In vehicle ReID, after the release of VehicleID and VeRI-776 datasets, the VGG_M usually serves as a baseline model for comparison, the performance of which is also provided. We present the experimental results in Tables 4 and 5. Clearly, the performances of these methods all dramatically drop on VERI-Wild compared to their results on existing VehicleID and VeRI datasets. For example, HDC [28] achieves mAP of 63.1 %, which is the best result on VehicleID dataset.



(a) (b) (c)

Figure 7. The CMC comparisons on VERI-Wild (TestSize = 10000), VeRI-776 and VehicleID (TestSize = 2400) test sets.

Table 4. The mAP performance on the VERI-Wild dataset.

Settings	Small	Medium	Large
GoogLeNet [27]	24.27	24.15	21.53
Triplet [17]	15.69	13.34	9.93
Softmax [10]	26.41	22.66	17.62
CCL [9]	22.50	19.28	14.81
HDC [28]	29.14	24.76	18.30
GSTE [1]	31.42	26.18	19.50
Unlabled GAN [32]	29.86	24.71	18.23
EN	28.77	24.63	19.48
FDA-Net $\ominus$ Att	32.40	27.10	21.13
FDA-Net	**35.11**	**29.80**	**22.78**

However, it only achieves the 29.14% mAP on VERI-Wild, significantly lower than its results on VehicleID. Such performance change indicates that VERI-Wild is a challenging dataset and valuable for the vehicle ReID research. Note that the match rate protocol on VERI-Wild and VehicleID are different, it can't be fairly compared.

Our proposed FDA-Net outperforms the other comparison methods. Compared to the baseline EN, the incremental improvements on FDA-Net $\ominus$ Att and FDA-Net demonstrate the feature distance adversary scheme can significantly facilitate the discriminative capability of the embedding model. More specifically, the FDA-Net outperforms FDA-Net $\ominus$ Att, indicating the effectiveness of the attention regularization. Compared to the Unlabled GAN [30] that uses Gaussian noises to randomly generate negatives, our FDA-Net presents much more improvements by exploring the potential of the hard negatives. The CMC curves on VERI-Wild large test set are shown in Fig. 7(a). Our method achieves higher rank 1 values than the compared methods.

Table 5. Match rate on the VERI-Wild dataset.

Settings	Small		Medium		Large	
Methods	R = 1	R = 5	R = 1	R = 5	R = 1	R = 5
GoogLeNet [27]	57.16	75.13	53.16	71.1	44.61	63.55
Triplet [17]	44.67	63.33	40.34	58.98	33.46	51.36
Softmax [10]	53.4	75.03	46.16	69.88	37.94	59.89
CCL [9]	56.96	75.0	51.92	70.98	44.6	60.95
HDC [28]	57.1	78.93	49.64	72.28	43.97	64.89
GSTE [1]	60.46	80.13	52.12	74.92	45.36	66.5
Unlabled GAN [32]	58.06	79.6	51.58	74.42	43.63	65.52
EN	57.13	77.33	52.86	73.18	43.02	66.3
FDA-Net $\ominus$ Att	61.93	80.48	55.62	75.64	46.48	68.36
FDA-Net	**64.03**	**82.8**	**57.82**	**78.34**	**49.43**	**70.48**

* All the references of any given query are in gallery.

Table 6. Performance on the VehicleID dataset.

Settings	Test Size=1600			Test Size=2400		
Methods	mAP	R=1	R=5	mAP	R=1	R=5
LOMO [8]	-	18.85	29.18	-	15.32	25.29
DGD [23]	-	40.25	65.31	-	37.33	57.82
GoogLeNet [27]	42.85	43.40	63.86	40.39	38.27	59.39
FACT [10]	-	44.59	64.57	-	39.92	60.32
XVGAN [25]	-	49.55	71.39	-	44.89	66.65
CCL [9]	44.8	39.94	62.98	38.6	35.68	56.24
Mixed Diff [9]	48.1	45.05	68.85	45.5	41.05	63.38
HDC [28]	63.1	-	-	57.5	-	-
VAMI [31]	-	52.87	75.12	-	47.34	70.29
EN	55.78	51.73	73.08	52.20	47.62	67.81
FDA-Net $\ominus$ Att	62.71	57.23	76.14	59.26	52.06	72.41
FDA-Net	**65.33**	**59.84**	**77.09**	**61.84**	**55.53**	**74.65**

* For CMC only one reference of any given query is in gallery.

5.2.2 Evaluation on VehicleID

The results on VehicleID are shown in Table 6. Our FDA-Net consistently obtains better performance over other approaches in both 1600 and 2400 test size. Both VAMI [31] and XVGAN [25] involve GANs, and they focus on generating cross-view vehicle images from the input view of a vehicle for improving cross-view ReID. However, our approach achieves superior performance from the perspective of hard negatives generation. HDC [28] also pays attention to the hard negatives, while they focus on mining hard negatives in the training set. Compared to the HDC, our generation scheme achieves more superior performance. It is worth mentioning that HDC cascades a set of GoogleNet which is a much deeper network than VGG_M used in our model, which further proves the effectiveness of our method. The CMC curves comparison on VehicleID dataset are shown in Fig. 7(b).

Other GAN relevant ReID methods focus on person image style transferring [22][3] or pose generation [13], which are not suitable for fair comparison or applied to vehicles.

5.2.3 Evaluation on VeRI

Table 7 shows the results on the VeRI-776 dataset. The proposed method outperforms the state-of-the-art method VAMI [31] by 5.36% mAP. In particular, VAMI used GAN

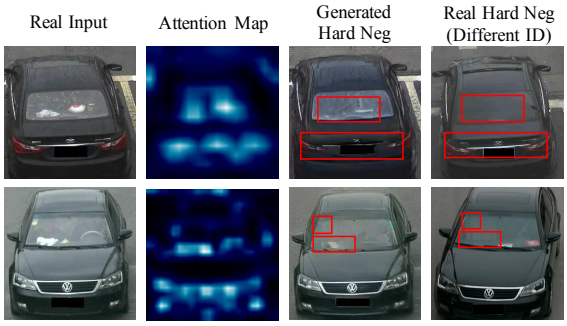


Figure 8. Examples of hard negative generation results. The generated hard negatives (3rd column) are similar to the real inputs (1st column), yet with differences in certain regions. These regions with differences are learned from the attention module with large responses in the corresponding attention map (2nd column).

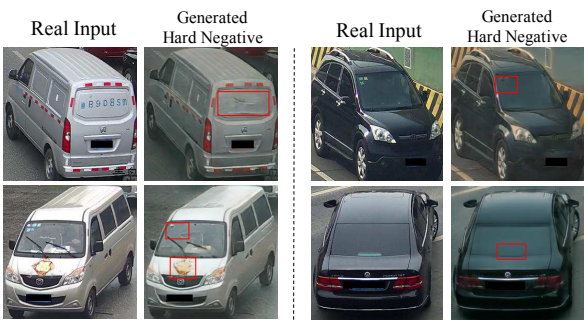


Figure 9. Comparison examples of generated hard negative pairs.

to infer feature for multi-view feature representation, while we focus on improving the capability of representing subtle details via hard negative scheme. Compared with OIFE [21], which used key-point alignment in vehicle feature representation, the proposed FDA-Net achieves much better performance. The CMC curves on VeRI-776 dataset are also provided in Fig. 7(c). Our method has also achieved much superior performance, especially in the rank 1 match rate (84.27% v.s. state-of-the-art 77.03% VAMI), meaning that the adversary scheme can significantly improve the discrimination capability of ReID model on subtle differences.

Table 7. Performance comparisons on the VeRI-776 dataset.

Methods	mAP	r = 1	r = 5
LOMO [8]	9.64	25.33	46.48
DGD [23]	17.92	50.70	67.52
GoogLeNet [27]	17.81	52.12	66.79
FACT [10]	18.73	51.85	67.16
XVGAN [25]	24.65	60.20	77.03
OIFE [21]	48.00	65.92	87.66
SiameseVisual [18]	29.48	41.12	60.31
FACT +Plate + STR [10]	27.77	61.44	78.78
VAMI [31]	50.13	77.03	90.82
EN	47.85	79.67	89.45
FDA-Net $\ominus$ Att	53.46	83.97	91.59
FDA-Net	**55.49**	**84.27**	**92.43**

5.3. Qualitative Results

Visualization of Hard Negatives. In Fig. 8 and Fig. 9, the input and generated hard negatives are pair-wisely shown. It can be observed that the hard negatives have very similar appearance with the input vehicles, and minor modifications have been harmoniously made for distinction. Compared with real hard negatives, our generated hard negatives also present challenging discrimination difficulties.

6. Discussion

The significance of real/fake loss. The "real/fake" loss allows the generated images to look more like a real one, also used in [22][30][3][13]. Without real/fake loss, FDA-Net would still introduce blur, deformation after complete training. Such vehicles do not exist in the real-world, when the discriminator is strong enough, such samples are of less significance to facilitate model training.

The choice of α, β and γ. We set $\alpha = 0.6$ in triplet margin which is a widely used setting in ReID. We also experiment with α from 0.4 to 0.7, and the performances are close. According to Eq(3) and Eq(8), $\|H(x), H(x^p)\|_2^2 + 2\beta \leq \|H(x), H(x^n)\|_2^2$ and $\|H(x), H(x^p)\|_2^2 + \alpha \leq \|H(x), H(x^n)\|_2^2$, it is known that $2\beta \leq \alpha$, thus we set $\beta = 0.3$. Different from margin constraint like α and β, the constraint of γ is similar to verification loss [19], we find that $\gamma = 0.7$ works well in our experiments.

7. Conclusion

In this work, we contribute a large-scale VERI-Wild dataset, which presents rich variants on backgrounds, illumination, occlusion and viewpoints, etc. The VERI-Wild is expected to facilitate the development and evaluation of the ReID methods in realistic scenarios.

In particular, we present a novel FDA-Net for vehicle ReID, which is able to elegantly generate hard negative samples in the embedding space for training a more discriminative ReID model. The ReID model coupled into FDA-Net, can be end-to-end optimized with feature distance adversarial scheme. With more available hard negatives, the embedding model's discriminative capability can be further facilitated, which has been well validated.

Acknowledgement: This work was supported by the National Natural Science Foundation of China under Grant 61661146005 and Grant U1611461, and in part by the National Basic Research Program of China under Grant 2015CB351806, and in part by Hong Kong RGC Early Career Scheme under Grant 9048122 (CityU 21211018).

References

[1] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan. Group sensitive triplet embedding for vehicle re-identification. *IEEE Transactions on Multimedia*, 2018.

[2] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-

domain image-to-image translation. *arXiv preprint*, 1711, 2017.

[3] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proc. of Computer Vision and Pattern Recognition*, 2018.

[4] S. Du, M. Ibrahim, M. Shehata, and W. Badawy. Automatic license plate recognition (alpr): A state-of-the-art review. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2):311–325, 2013.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. of International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

[6] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. of European Conference on Computer Vision*, pages 241–257, 2016.

[7] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, 2008.

[8] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[9] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proc. of the Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.

[10] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *Proc. of European Conference on Computer Vision*, pages 869–884. Springer, 2016.

[11] I. Loshchilov and F. Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

[12] Y. Lou, Y. Bai, J. Liu, S. Wang, and L.-Y. Duan. Embedding adversarial learning for vehicle re-identification. *IEEE Transactions on Image Processing*, 2019.

[13] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *Proc. of Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[14] B. T. Morris and M. M. Trivedi. Learning, modeling, and classification of vehicle track patterns from live video. *IEEE Transactions on Intelligent Transportation Systems*, 9(3):425–437, 2008.

[15] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Computer Science*, 2015.

[16] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.

[17] F. Schroff, K. Dmitry, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[18] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *Proc. of IEEE International Conference on Computer Vision*, pages 1918–1927. IEEE, 2017.

[19] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Proc. of Advances in neural information processing systems*, pages 1988–1996, 2014.

[20] X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Computer Vision and Pattern Recognition*, 2017.

[21] Z. Wang, L. Tang, X. Liu, Z. Yao, S. Yi, J. Shao, J. Yan, S. Wang, H. Li, and X. Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *Proc. of Computer Vision and Pattern Recognition*, pages 379–387, 2017.

[22] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proc. of Computer Vision and Pattern Recognition*, 2018.

[23] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proc. of Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

[24] J. Xu, V. Jagadeesh, Z. Ni, S. Sunderrajan, and B. Manjunath. Graph-based topic-focused retrieval in distributed camera network. *IEEE Transactions on Multimedia*, 15(8):2046–2057, 2013.

[25] Z. Y. and S. L. Cross-view gan based vehicle generation for re-identification. In *Proc. of British Machine Vision Conference (BMVC)*, 2017.

[26] K. Yan, Y. Tian, Y. Wang, W. Zeng, and T. Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *Proc. of Conference on Computer Vision*, pages 562–570, 2017.

[27] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proc. of Computer Vision and Pattern Recognition*, pages 3973–3981, 2015.

[28] Y. Yuan, K. Yang, and C. Zhang. Hard-aware deeply cascaded embedding. *arXiv preprint arXiv:1611.05720*, 2016.

[29] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. *arXiv preprint arXiv:1512.02895*, 2015.

[30] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 3, 2017.

[31] Y. Zhou and L. Shao. Viewpoint aware attentive multi-view inference for vehicle re-identification. In *Computer Vision and Pattern Recognition*, pages 6489–6498, 2018.

[32] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.