Embedding Adversarial Learning for Vehicle Re-Identification

Yihang Lou, *Member, IEEE*, Yan Bai, *Member, IEEE*, Jun Liu, Shiqi Wang, *Member, IEEE*, and Ling-Yu Duan, *Member, IEEE*

***Abstract*— The high similarities of different real-world vehicles and great diversities of the acquisition views pose grand challenges to vehicle re-identification (ReID), which traditionally maps the vehicle images into a high-dimensional embedding space for distance optimization, vehicle discrimination, and identification. To improve the discriminative capability and robustness of the ReID algorithm, we propose a novel end-to-end embedding adversarial learning network (EALN) that is capable of generating samples localized in the embedding space. Instead of selecting abundant hard negatives from the training set, which is extremely difficult if not impossible, with our embedding adversarial learning scheme, the automatically generated hard negative samples in the specified embedding space can greatly improve the capability of the network for discriminating similar vehicles. Moreover, the more challenging cross-view vehicle ReID problem, which requires the ReID algorithm to be robust with different query views, can also benefit from such a scheme based on the artificially generated cross-view samples. We demonstrate the promise of EALN through extensive experiments and show the effectiveness of hard negative and cross-view generation in facilitating vehicle ReID based on the comparisons with the state-of-the-art schemes.**

***Index Terms*— Vehicle Re-Identification, generative adversarial network, embedding adversarial learning, hard negatives, cross-view.**

I. Introduction

VEHICLE Re-Identification (ReID) aims to identify all the images of the same vehicle ID as the given query vehicle from a large scale database. With the increasing demand of social public security, vehicle ReID has emerged as a crucial technique for both industry and academia realms. The vehicle ReID has extensive applications on intelligent surveillance to analyze the behaviors and tracks of vehicles from multiple non-overlapping surveillance cameras. Previous works [1], [2] focus on using license plate to perform vehicle ReID. However, the license plate recognition often fails due to variant imaging conditions, and the license plates are deliberately occluded, removed or even faked in some extreme cases. In view of this, a general model based on visual characteristics is highly desired [3]–[12].

A considerable number of methods [3]–[6] perform vehicle ReID in an embedding space. Specifically, they map vehicle images into a high dimensional embedding space, where the samples of the same vehicle ID are obligatorily pulled closer and in contrast the samples of different IDs are pushed far away. In particular, recent visual based methods seek to learn an embedding model [4], [6] among which the triplet network is the most representative one. The main idea of the triplet network is to map images into an embedding space in which samples belonging to the same vehicle ID are closer than those of different IDs. More specifically, let $< x, x^p, x^n >$ denote a triplet unit, where x is an anchor sample, x^p denotes the sample with the same vehicle ID as x, and x^n represents a sample with a different vehicle ID. The constraint among these three items are formulated as $d(x, x^p) + \alpha \leq d(x, x^n)$, where α is the minimum margin between positive samples and negative samples, and $d(\cdot, \cdot)$ is the L2 normalized distance in the embedding space. Given K triplet units, the triplet loss function [13] can be defined as: $L = \sum_{k=1}^{K} \frac{1}{2} \max\{d(x, x^p) + \alpha - d(x, x^n), 0\}$. When the margin constraint is not satisfied, the loss will be back propagated. The triplet unit that does not satisfy the distance margin constraint is regarded as a hard triplet unit, and the x^n in the triplet unit is termed as a hard negative for x.

However, there are two critical challenges in vehicle ReID. First, two vehicles of different IDs may belong to the same vehicle model. More specifically, the features of similar vehicles tend to be close in the embedding space [6], which makes vehicle ReID extremely difficult. In such extreme cases, the available cues for discrimination are only subtle differences such as the tissue boxes, pendants and other decoration marks, as shown in Fig. 1(a). Another general class of problems that could hinder the performance improvement of vehicle ReID is the viewpoint variations, as shown in Fig. 1(b). For example, identifying the corresponding rear viewpoint image given only the front view and vice versa cast challenges on the translation capability of the ReID methods.

Manuscript received August 2, 2018; revised January 4, 2019; accepted February 13, 2019. Date of publication February 27, 2019; date of current version June 13, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001501, in part by the National Natural Science Foundation of China under Grant U1611461 and Grant 61661146005, in part by the Shenzhen Municipal Science and Technology Program under Grant JCYJ20170818141146428, and in part by the National Research Foundation, Prime Minister's Office, Singapore, under the NRF-NSFC Grant NRF2016NRF-NSFC001-098. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun Fu. *(Corresponding author: Ling-Yu Duan.)*

Y. Lou is with the National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China (e-mail: yihanglou@pku.edu.cn).

Y. Bai is with the National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China, and also with Hulu LLC, Beijing 100102, China (e-mail: yanbai@pku.edu.cn).

J. Liu is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: jliu029@ntu.edu.sg).

S. Wang is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: shiqwang@cityu.edu.hk).

L.-Y. Duan is with the National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: lingyu@pku.edu.cn).

Digital Object Identifier 10.1109/TIP.2019.2902112

Fig. 1. Illustration of two crucial challenges in vehicle ReID. (a) High similarity between different vehicle IDs; (b) viewpoint variation within the same vehicle ID. The top two rows show different vehicle IDs of the same/similar model type, and their subtle differences are highlighted with the red box. The bottom two rows demonstrate the vehicle images with the same ID captured from different viewpoints.

Though substantial diversities are exhibited in the methodologies applied to address these arising vehicle ReID problems, some common characteristics are also shared. In particular, both the above two issues are rooted from the limited knowledge that can be obtained given the available samples for vehicle discrimination and identification. For example, a desired embedding model should be discriminative enough to scrupulously observe the subtle differences between the visually similar vehicles with different IDs, and increasing the number of hard negatives that break the distance constraints for triplet constitution has been proven to be pivotal for training a discriminative embedding model [13], [14]. As such, a typical paradigm is training the model by feeding more hard negative pairs that are very close in the embedding space [14], [15], and most of the existing works [10], [13] focus on selecting hard negatives from the training set. However, the limited capacity of the affordable selection often creates a large gap between the diversity of real-world images and the limited number of hard negatives in the training set, leading to poor ReID performance. Regarding the cross-view vehicle ReID, the dependencies of different viewpoints of an identical vehicle carry important information for vehicle ReID. As such, it is imperative to generate cross-view images, which provide valuable supplementary information to identify vehicles with different viewpoints.

The above issues motivate us to focus on the converging point regarding the generation of particularly demanded samples in the embedding space, which fundamentally desires highly versatile generative models. Adversarial networks, especially the Generative Adversarial Networks (GANs) [16], have shown great potential in image generation for ReID, such as artificially generating images with different styles [17] or poses [18], [19]. In this work, we propose an embedding adversarial learning network (EALN), in which a novel adversarial scheme on feature distance is designed within the embedding space. Specifically, for the feature distance between the input and the generated images, the generator attempts to minimize the feature distance while the discriminator tries to maximize it. As the ReID process is essentially performed in the embedding space, based on the proposed EALN, we can apply adversarial learning on feature distance to effectively generate meaningful samples and facilitate the vehicle ReID task. As such, the image generation process is manipulated in the specified embedding space, and substantial performance improvements for vehicle ReID are observed. The three main contributions of our work can be summarized as follows,

- We propose a novel generative scheme termed as EALN by imposing the embedding adversarial learning between the generator and discriminator. Moreover, the embedding discriminator in EALN serves as a representation network for feature extraction in ReID, such that an end-to-end optimization between image generation and feature representation can be achieved.
- We develop a novel online hard negative generation method based on the proposed EALN, which is able to continuously generate hard negatives to adapt existing embedding discriminator, leading to a more discriminative embedding discriminator.
- We further extend the EALN to improve the cross-view vehicle image generation, aiming to address the problem of viewpoint inconsistency in vehicle ReID. The given viewpoint is translated to multiple view points such that the comprehensive cross-view images are used to further enhance the performance of our ReID scheme.

II. Related Work

The ReID task has been widely studied and most recent efforts focus on person ReID [20]–[26]. Though there are great similarities between person and vehicle ReID, specific characteristics and challenges have also been particularly observed for vehicles. In this section, we first review the recent methods on vehicle ReID. We then briefly introduce the representative generative methods based on GAN. Finally, the GAN based ReID methods are discussed.

A. Vehicle Re-Identification

Several methods [5], [27] attempt to retrieve vehicles based on the attributes such as color, vehicle model and spatial-temporal characteristics. Feris *et al.* [27] proposed an end-to-end system for vehicle detection and retrieval based on semantic attributes. Liu *et al.* [5] proposed a vehicle ReID system that adopts a coarse-to-fine search strategy with additional spatial-temporal information. Shen *et al.* [11] proposed a two-stage framework that incorporates complex spatial-temporal information for effectively regularizing the ReID results.

Recent methods [6], [28]–[30] adopt deep metric networks (*i.e.*, siamese or triplet network) to learn a feature embedding space where "the samples belonging to the same vehicle ID are closer than those belonging to the different". Liu *et al.* [6] introduced a mixed difference network in which both the vehicle model and ID information are used as supervised signals for learning an embedding model. Bai *et al.* [28] proposed a group sensitive triplet embedding model, which focuses

on alleviating negative impacts of intra-class variances by incorporating group-wise variances into a structured triplet embedding model.

In particular, there are some efforts aiming to improve ReID performance from different aspects, such as hard examples selecting [10], [14], [15] and cross-view feature representation [4], [31]. Yuan *et al.* [10] proposed a hard-aware deeply cascaded embedding method that ensembles a set of models in a cascaded manner to select hard examples from training set for efficient learning. Moreover, cross-view feature representation is also an attractive method. Zhou and Shao [4] focused on multi-view feature representation, and proposed a viewpoint-aware attention model to select core regions at different viewpoints. Then these attentive regions from multi-view are combined by an adversarial learning scheme. Similarly, Zhou *et al.* [31] proposed a long short-term memory (LSTM) network to model appearance transformations across different viewpoints of vehicles. The cross-view vehicle ReID task is quite relevant to cross-view person ReID. Li *et al.* [32] proposed a Cross-view Dictionary Learning (CDL) method for the multi-view person feature learning, which learns a pair of dictionaries from two views with a projective learning strategy. Moreover, a multi-level representation scheme from image, horizontal part and patch level is further imposed on CDL to facilitate model learning. Alessandro *et al.* [33] proposed to mitigate the cross-view ambiguity by learning discriminative feature with a novel loss function, which aims at simultaneously mining effective intra-class and inter-class relationships in feature domain of the person identities. Li *et al.* [34] proposed a novel Discriminative Semi-coupled Projective Dictionary Learning (DSPDL) to enhance the feature matching between the cross-view person images.

B. Image Translation by GAN

Generative adversarial networks (GANs) have been repeatedly proved to be promising in image generation. Goodfellow *et al.* [16] introduced GAN in which they simultaneously train two models, a generative model that captures the data distribution, and a discriminative model that tries to distinguish a sample is from the training data or generated data. Recently, many variants of GAN [35]–[46] have been proposed to tackle different tasks, *e.g.*, image-to-image translation, natural style transfer, super-resolution, *etc*. In these tasks, image-to-image translation has attracted numerous attentions. Radford *et al.* [47] proposed a deep convolutional GAN architecture (DCGAN) with a set of conditional constraints to learn a mapping function from the input to output images. Zhu *et al.* [48] proposed CycleGAN which learns an unpaired image-to-image mapping between two different domains ($A \rightarrow B$ or $B \rightarrow A$) using cycle consistency. Choi *et al.* [49] proposed a scalable mapping approach, StarGAN, which uses a single model to implement one-to-many domain translation.

C. GANs Based Person ReID

Due to the powerful generation capability of GANs, it is natural to apply GANs in the visual analytics tasks. In particular, GANs have been successfully applied in person ReID [17]–[19], [50], [51]. These works can be summarized into two categories. The first category generates conditioned images as an offline data augmentation or translation method. For example, Zheng *et al.* [51] adopted DCGAN to generate unlabeled person images for augmenting training samples. Wei *et al.* [17] proposed a PTGAN based on CycleGAN to reduce dataset domain gap in person ReID, in which it imposed a mask constraint to CycleGAN for ensuring foreground unchanged during person image transferring. In analogous to [17], Deng *et al.* [50] used Cycle-GAN to reduce domain bias via an extra similarity constraint to change styles during person image transferring. The second group of works are proposed for specific tasks, such as pose generation and feature inference. Ma *et al.* [19] used person keypoints to transfer the human poses in a flexible way which disentangled the representation of foreground, background, and pose information from person images. Zhou and Shao [4], [31] proposed a feature adversarial training architecture for vehicle ReID, which implements multi-view feature inference from a single-view.

In this work, we propose an end-to-end embedding adversarial learning network by coupling embedding model optimization and sample generation for vehicle ReID. In our scheme, the performance of vehicle ReID is improved from different perspectives based on the generative model. A hard negative sample generation scheme is designed to improve the discriminative capability of embedding representation. To the best of our knowledge, this is the first adversarial model which considers generating hard negative samples in ReID field. Besides, we also propose a novel method to handle the view variations based on our embedding adversarial learning.

III. Embedding Adversarial Learning Based Vehicle ReID

In this section, we propose the EALN which is further incorporated into the vehicle ReID framework for hard negative and cross-view generation. Our proposed EALN based vehicle ReID framework is illustrated in Fig. 2. In particular, the EALN imposes an adversarial scheme between the generator and discriminator on feature distance in the embedding space. Given an input sample, such adversarial scheme enables the generator to generate a sample at a specific feature distance to the input. As such, the EALN serves as the basic building block for the proposed vehicle ReID methodology, which supports hard negative and cross-view generation with multiple adversarial learning strategies.

Overview: The first module is the hard negative adversarial learning between the generator G_{hard} and D_{hard}. The target of this scheme is generating a sample which is close to each input sample in the embedding space while of a negative identity. As the training proceeds, the negative samples generated by G_{hard} will be more difficult to differentiate. In turn, as being trained with these hard negatives, the D_{hard} will become more discriminative. The second module is the cross-view adversarial learning between G_{view} and D_{view}, which aims to generate a cross-view image of the given input by being enforced to be as close as possible, such that the vehicle identity can be maintained.

In summary, the hard negative adversarial learning serves to learn a more discriminative embedding discriminator D_{emb} for

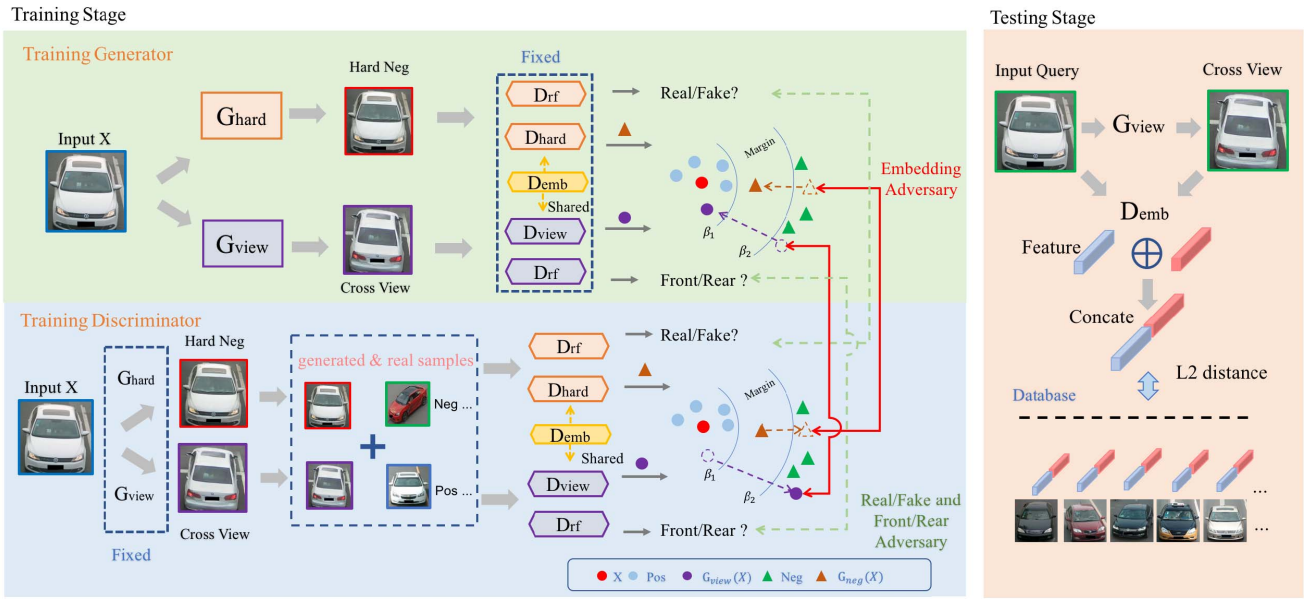


Fig. 2. Illustration of the proposed EALN for vehicle ReID. EALN contains embedding adversarial and real-fake adversarial scheme. Based on these two schemes, we design different distance constraints to achieve hard negative adversarial learning for hard negative generation and cross-view adversarial learning for cross-view sample generation. In the hard negative adversarial scheme, G_{hard} tries to generate a sample with close feature distance to each input sample, while D_{hard} tries to discriminate them. In cross-view adversarial scheme, G_{view} tries to generate a cross-view sample for each input image with as close feature distance as possible. The D_{rf} is applied in the above two schemes to determine whether the generated sample is real/fake or front/rear. We use generated samples by G_{view} and G_{hard} as well as training set to train a more discriminative embedding model D_{emb}. In the training stage, the generators and discriminators are alternatively optimized. In the testing stage, the embedding discriminator D_{emb} serves as the feature extractor. Additionally, for each input image, a generated cross-view image is used to achieve the fused feature representation.

feature representation, and the cross-view adversarial learning targets at obtaining a robust cross-view generator G_{view} to facilitate the cross-view ReID. It is worth mentioning that the embedding discriminators D_{hard} and D_{view} in Fig. 2 are parametrically shared, and we use D_{emb} to map the generated samples to identical embedding space. For better understanding of the adversarial scheme in hard negatives and cross-view generation, we differentiate D_{hard} and D_{view} in our description.

A. Embedding Adversarial Learning

The design principle of the embedding adversarial learning is manipulating the generation of the samples in the embedding space. The visualization of the embedding adversarial framework with the generator and discriminator in the embedding space is shown in Fig. 3. For the generator, we constrain the generated sample $G(x)$ to be close to the given sample x within a relative distance constraint α in the embedding space. On the contrary, the discriminator D_{emb} tries to push the generated $G(x)$ away from x to ensure the relative distance to be larger than α. Therefore, an adversarial scheme on feature distance between the generator and discriminator is formed in the embedding space, which is formulated as follows,

$$\min_{G}\max_{D_{emb}} \mathbb{E}_{x\sim pdata(x)}\max\{d(x, G(x)) - \alpha), 0\}, \quad (1)$$

$$d(x, G(x)) = \|D_{emb}(x) - D_{emb}(G(x))\|_2^2, \quad (2)$$

where $D_{emb}(x)$ is the mapping of x in embedding space through discriminator D_{emb}. The generator and embedding discriminator are trained alternatively. When training the generator, we fix the D_{emb} and use it as a robust feature extractor.

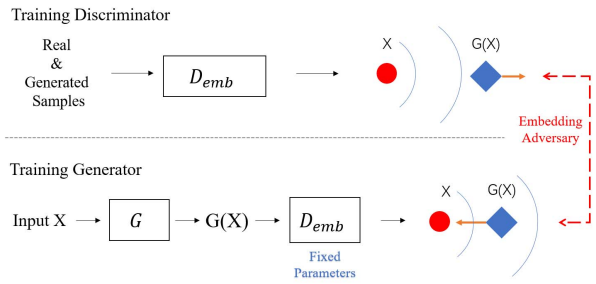


Fig. 3. Illustration of embedding adversarial learning between generator and discriminator. The generator G aims to generate samples close to the given sample in embedding space and the discriminator D_{emb} tries to push the generated samples far away.

Correspondingly, during the training of D_{emb}, we also fix the generator and use the generated samples to optimize D_{emb}.

More specifically, with the embedding adversarial mechanism, we can jointly optimize the generator and the discriminator in the embedding space. In particular, there are two advantages of embedding adversarial learning scheme. First, the embedding discriminator D_{emb} in GAN can be considered as the feature extractor for embedding representation in vehicle ReID. Rather than using multiple separated training stages as in [17]–[19], [50], this scheme achieves the end-to-end optimization. Second, through embedding adversarial scheme, we can continuously generate samples that satisfy the given feature distance constraints, and meanwhile these generated samples can be used to further optimize the embedding discriminator D_{emb}. Through different constraints on feature distance, the EALN can be flexibly applied for different purposes.

In the following section, we will describe how the embedding adversarial learning can be applied to hard negative and cross-view generation.

B. Hard Negative Adversarial Learning

The embedding adversarial learning enables us to generate samples in specified regions in the embedding space, which motivates us to generate hard negatives (*i.e.*, the sample pair whose feature distance are close yet the IDs are different) to train a more discriminative embedding model for ReID. To generate effective hard negatives for embedding model training, we extend embedding adversarial mechanism to hard negative adversarial learning, which consists of two parts, *i.e.*, embedding adversarial learning and real-fake adversarial learning.

Given an input vehicle image x, we aim to generate a hard negative sample $G_{hard}(x)$, which is closer to input x than real negatives in the embedding space and meanwhile of a negative identity. Thus, we constrain the generated $G_{hard}(x)$ located in the neighboring margin specified via β_1 and β_2,

$$\beta_1 \leq d(x, G_{hard}(x)) \leq d(x, x^n) - \beta_2, \tag{3}$$

where $d(\cdot, \cdot)$ is the L2 normalized distance in the embedding space and x^n is the real negative sample. The parameter β_1 is the minimum distance from x, which aims to constrain the $G_{hard}(x)$ to locate at a certain distance to x. Without this constraint, the G_{hard} will tend to generate a sample with the same appearance as x to satisfy the right part in Eq. (3). The β_2 is the margin gap from $G_{hard}(x)$ to the real negative samples, which aims to constrain the generated $G_{hard}(x)$ to be close to x and far away from negative samples in the training set. The combination of β_1 and β_2 can be interpreted as: the $G_{hard}(x)$ is enforced to be located in an annular belt around x with minimum distance of β_1 to x and width of β_2. With such constraints, the generated $G_{hard}(x)$ tends to break the distance constraint in the triplet loss. Therefore, these generated negatives cannot be easily discriminated from the positives. The embedding loss for G_{hard} can be formulated as:

$$\begin{aligned} L_{G\_hard} = {} & \mathbb{E}_{x\sim pdata(x)} \max\{\beta_1 - d(x, G_{hard}(x)), 0\} \\ & + \max\{d(x, G_{hard}(x)) + \beta_2 - d(x, x^n), 0\} \end{aligned} \tag{4}$$

As the opposite side in embedding adversarial learning, the embedding discriminator D_{hard} tries to push the $G_{hard}(x)$ away from x and maintains a certain margin gap, *i.e.*, $d(x, x^p) + \alpha \leq d(x, G_{hard}(x))$. Thus, the embedding loss for D_{hard} can be formulated as follows:

$$L_{D\_hard} = \mathbb{E}_{x\sim pdata(x)} \max\{d(x, x^p) + \alpha - d(x, G_{hard}(x)), 0\}. \tag{5}$$

Note that the generated $G_{hard}(x)$ are all treated as negative samples in embedding optimization. To learn a more effective embedding discriminator D_{emb}, not only the actual negative samples x^n in the training set but also the artificially generated hard negatives $G_{hard}(x)$ are used for training. Moreover, we adopt a multi-loss training strategy containing triplet loss and softmax loss, which has been demonstrated to be effective in [52], [53].

The other real-fake adversarial learning tries to differentiate the real sample x and the generated one $G_{hard}(x)$, and such adversarial scheme aims to make the generated images $G_{hard}(x)$ more realistic. Thus the loss of real-fake adversarial learning L_{real_fake} can be formulated as follows:

$$\begin{aligned} & L_{real\_fake} \\ & = \mathbb{E}_{x\sim pdata(x)}[log D_{rf}(x) + log(1 - D_{rf}(G_{hard}(x)))]. \end{aligned} \tag{6}$$

where the generator G_{hard} aims to minimize the objective while the real-fake discriminator D_{rf} tries to maximize it.

Hence, the final loss functions for hard negative adversarial learning with GAN include both the generator and discriminator, which can be represented as follows:

$$L_{G\_hard} = L_{real\_fake} + \lambda_{ed} L_{G\_hard}, \tag{7}$$

$$L_{D\_hard} = -L_{real\_fake} + \lambda_{ed} L_{D\_hard}. \tag{8}$$

where λ_{ed} is a hyper parameter to balance the adversarial schemes.

Compared to the offline generation methods [17], [50] which use a separate GAN to generate all the samples before training an embedding model, we couple the hard negative sample generation and embedding model optimization in the embedding adversarial learning. Therefore, we can continuously generate harder negatives to adapt existing embedding discriminator. With the proceed of the training, the generated hard negatives can consistently satisfy the distance constraint to train a better embedding discriminator. With the alternative embedding adversarial learning, the generation of hard negatives will become more suffering, and correspondingly, the embedding discriminator will become more discriminative.

C. Cross-View Adversarial Learning

Viewpoint variation is an another crucial factor that hinders the ReID performance. Based on embedding adversarial learning, we aim to address the issue of vehicle feature matching between different views via cross-view generation. More specifically, in ReID, an extra cross-view vehicle image is generated for each input vehicle image to achieve a fused feature representation. Different from hard negative generation, cross-view generation is a domain transfer problem (from viewpoint $A \rightarrow B$ or $B \rightarrow A$). Though CycleGAN [48] is a recent representative work on domain transfer, CycleGAN transfers viewpoint on set level, implying that it cannot ensure the identity maintenance. Hence, we utilize embedding adversarial learning to address the limitation of CycleGAN for instance-level viewpoint transfer.

We extend embedding adversarial scheme into cross-view adversarial learning to improve the quality of cross-view vehicle image generation, including embedding adversarial learning and front-rear adversarial learning. Given an input vehicle image x, the aim of cross-view generation is to produce a cross-view image $G_{view}(x)$ with the same identity as the given vehicle x. We consider two viewpoints, front A and rear B, to learn a mapping function between them, since most surveillance cameras capture the front and rear views of vehicles, such as vehicle ID dataset [6]. The training samples include $\{x_a\}_{a=1}^N$ and $\{x_b\}_{b=1}^N$, where $x_a \in A$ and $x_b \in B$.

We impose an extra embedding constraint based on the embedding adversarial learning that enforces the generated cross-view image $G_{view}(x)$ to be close to the input x within a distance constraint α in the embedding space. Given an image $x \in \{A, B\}$, the embedding distance constraint for G_{view} can be formulated as follows:

$$L_{G\_view} = \mathbb{E}_{x\sim pdata(x)}\max\{(d(x - G_{view}(x)) + \alpha - d(x, x^n), 0)\}, \quad (9)$$

By contrast, embedding discriminator D_{view} tries to push the $G_{view}(x)$ away from x at a distance of α. Thus, the embedding loss for D_{view} is given by,

$$L_{D\_view} = \mathbb{E}_{x\sim pdata(x)}\max\{(d(x, x^p) + \alpha - d(x, G_{view}(x)), 0)\}. \quad (10)$$

Another front-rear adversarial scheme is used to perform cross-view generation, which tries to determine whether the generated samples are real front or real rear. From view $A \rightarrow B$, the objective function can be expressed as:

$$L(G^B_{view}, D_B, A, B) = \mathbb{E}_{x_b\sim pdata(x_b)}[log D_B(x_b)] + \mathbb{E}_{x_a\sim pdata(x_a)}[log(1 - D_B(G^B_{view}(x_a)))], \quad (11)$$

where G^B_{view} aims to minimize the objective against an adversarial D_B that attempts to maximize it. The objective function $G^A_{view} : B \rightarrow A$ and its discriminators D_A are defined analogously. Hence, the adversarial loss function can be represented as follows:

$$L_{view} = L(G^B_{view}, D_B, A, B) + L(G^A_{view}, D_A, B, A). \quad (12)$$

Therefore, the final loss function which optimizes the generator G_{view} and discriminator D_{view} in cross-view adversarial learning can be formulated as follows:

$$L_{G\_view} = L_{view} + \lambda_{cyc} L_{cyc} + \lambda_{ed} L_{G\_view}, \quad (13)$$

$$L_{D\_view} = -L_{view} + \lambda_{ed} L_{D\_view}, \quad (14)$$

where L_{cyc} is the cycle consistency loss between G^B_{view} and G^A_{view} in [48] to improve the mapping stability. λ_{ed} and λ_{cyc} are the hyper parameters that balance the objectives.

D. End-to-End Training and Testing

1) Training: In the training stage, the optimization for hard negative adversarial learning and cross-view adversarial learning are optimized simultaneously. In particular, the embedding discriminators D_{hard} and D_{view} in Fig. 2 are parametrically shared, which can be denoted by D_{emb} for brevity. Generators and discriminators are alternatively trained in the training stage. During the training of the generator, an input image from real training set is fed to G_{hard} and G_{view} to generate a hard negative sample and a cross-view sample, then they are fed to the fixed embedding discriminator D_{emb} to compute corresponding loss on feature distance constraint. During the training of the embedding discriminator, for each input sample, a hard negative sample and a cross-view sample will be generated by the fixed G_{hard} and G_{view}, which are further combined with real training samples, and their feature distances are optimized by embedding discriminator D_{emb} (*i.e.*, D_{hard} and D_{view}) for discrimination. These hard negative and cross-view samples are continuously generated with EALN, and meanwhile the embedding model D_{emb} is also consistently optimized. The whole training procedure is end-to-end. After adversarial learning, the D_{emb} serves as a feature extractor for the representation of vehicle images.

2) Testing: In the testing stage, each input vehicle image is fed into D_{emb} to obtain feature representation. In ReID procedure, a candidate list for each given query is returned from the database, which is sorted by the L2 normalized feature distances between the query and reference images. Moreover, for the ReID experiments involved cross-view generation, a cross-view image is generated by G_{view} for each input image. Then both the input image and generated cross-view image are fed to D_{emb} to produce features, which are then concatenated to obtain the final representation for ReID.

E. Implementation Details

1) Network Architecture: For the generator network, we adopt the architecture from Johnson *et al.* [54]. This network contains two convolution layers with the stride size of 2 for down-sampling, 9 residual blocks [55], and two convolutional layers with the stride size of $1/2$ for upsampling. The size of training vehicle images is 224×224. Instance normalization [56] is also applied in the generator. The discriminator network contains multiple subnetworks for different adversarial schemes. For the real-fake discriminator network, we use a PatchGAN of size 70×70 as in [48], [57]. VGG_CNN_M_1024 (*VGGM*) is selected as the base network for VeRI-776 and Resnet50 is selected for VehicleID. The embedding discriminator serves as a feature extractor, and the output of L2 normalization layer is treated as the feature representation for ReID.

2) Hyper Parameters in the Training Stage: Regarding parameters for distance constraint, we set $\alpha = 0.6$ in embedding discriminator, and $\beta_1 = 0.6$, $\beta_2 = 0.3$ in hard negative generator. Regarding the weights for loss functions, we set $\lambda_{ed} = 1$ and set $\lambda_{cyc} = 10$ as in [48]. Specially, in the last 10 epochs, we set $\lambda_{ed} = 0$ in Eq.14, which means we remove the embedding constraint of L_{D_view} in model training.

Learning rate starts from 0.001 for embedding discriminator, and starts from 0.0002 for other generators and discriminators. We maintain the same learning rate for the first 20 epochs and linearly decay the rate to zero over the next 30 epochs. The size of mini-batch, momentum and weight decay are set to 4, 0.9 and 0.0002, respectively.

IV. Experimental Results

To validate the proposed scheme, we conduct quantitative and qualitative experiments on VehicleID [6] and VeRI-776 [5] datasets. We first quantitatively compare the proposed scheme with the state-of-the-art methods. Subsequently, qualitative analyses are performed, including the visualization of the online generated hard negative and cross-view samples. Finally, more analyses regarding the distance distribution in the embedding space and response maps are provided, which facilitate better understanding of the proposed methodology.

To evaluate the proposed two adversarial learning schemes, we conduct experiments with following different structures (1) Embedding Network without any adversarial learning (EN); (2) EALN with Hard Negative Adversarial Learning (Hard-EALN); (3) EALN with Hard Negative Adversarial Learning and Cross-View Adversarial Learning (Hard-View-EALN):

- "EN": This structure is a traditional triplet embedding network. We adopt the multi-loss (triplet loss and softmax loss) as supervision signals, which is a widely used baseline model [6], [53].
- "Hard-EALN": This structure contains hard negative adversarial learning based on EALN.
- "Hard-View-EALN": This structure contains one more cross-view adversarial learning compared to Hard-EALN.

A. Datasets

We conduct experiments on VehicleID and VeRI-776 datasets, strictly following the evaluation protocols described in [5] and [6], respectively.

- VehicleID [6] dataset consists of the training set with 110 178 images of 13 134 vehicles and the testing set with 111 585 images of 13 133 vehicles, which are captured by different surveillance cameras in a city. The VehicleID dataset contains two viewpoints: front-view and rear-view. Following the configurations in [6], we use three test subsets of different sizes, *i.e.*, 7,332 images of 800 vehicles in small size, 12 995 images of 1,600 vehicles in medium size and 20,038 images of 2,400 vehicles in large size.
- VeRI-776 [5] dataset consists of the training set with 37 778 images of 576 vehicles and the testing set with 11,579 images of 200 vehicles, which are all captured by 20 cameras in an unconstrained traffic scenario. Each vehicle is captured by 2-18 cameras covering an 1 km^2 area. The vehicle images in VeRI-776 have multiple viewpoints *i.e.*, front, front-side, side, rear-side, rear. Additionally, this dataset provides spatial-temporal information of the captured vehicles.

B. Quantitative Results

In this subsection, the performance of the proposed scheme is compared with the state-of-the-art methods quantitatively both on vehicle ReID and VeRI-776 datasets.

1) Evaluation Metrics: We adopt two evaluation metrics, mean Average Precision (mAP) and Cumulative Match Curve (CMC) in our experiments.

a) Mean average precision: The mAP metric evaluates the overall performance for ReID. Average precision is calculated for each query image as follows:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times gt(k)}{N_{gt}}, \tag{15}$$

where k is the rank in the sequence of retrieved vehicles, n is the number of retrieved vehicles, and N_{gt} is the number of relevant vehicles. $P(k)$ is the precision at cut-off k in the recall list and $gt(k)$ indicates whether the k-th recall image is correct or not. Therefore, the mAP is defined as follows:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}, \tag{16}$$

where Q is the number of total query images. Moreover, Top K match rate is also reported in the experiments.

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

$$CMC@k = \frac{\sum_{q=1}^{Q} gt(q,k)}{Q}, \tag{17}$$

where $gt(q,k)$ equals 1 when the groundtruth of q image appears before rank k. In VehicleID and VeRI-776 datasets, we follow the CMC evaluation in [5] and [6], respectively.

2) Compared Methods: We compare the proposed method with the recently proposed Vehicle ReID methods which achieve excellent performance on these benchmarks.

a) Mixed Diff + CCL [6]: Liu *et al.* designed a mixed supervision consisting of vehicle model and vehicle ID, which is called coupled cluster triplet loss to train a representation network.

b) XVGAN [59]: Zhou *et al.* proposed to generate cross-view image from the input view of a vehicle, then utilized the generated views to compute distances.

c) HDC + Contrastive [10]: Yuan *et al.* leveraged a cascade scheme to find hard negatives in the training set to achieve efficient feature learning.

d) FACT + Plate-SNN + STR [5]: Liu *et al.* adopted a progressive vehicle ReID strategy consisting of FACT feature, license plate feature and spatial-temporal relationship to implement coarse-to-fine search.

e) SiameseVisual [11]: Shen *et al.* used a siamese-CNN + Path-LSTM model to retrieve vehicle using combined visual feature and spatial-temporal information.

f) OIFE [60]: Wang *et al.* used the key-point alignment to achieve an orientation invariant feature embedding module for vehicle feature representation.

g) VAMI [4]: Zhou *et al.* aimed to solve multi-view vehicle ReID problem by employing a viewpoint-aware multi-view inference to produce global-view feature representation from arbitrary viewpoints image.

3) Evaluation on VehicleID Dataset: The performance comparisons on VehicleID dataset are listed in the Table I. From the comparison between EN and Hard-EALN, we can find that hard negative adversary significantly improves the ReID performance from 67.5% mAP to 75.1% mAP which consistently proves the effectiveness of the generated hard negatives. Moreover, with additional cross-view adversary, we can obtain further 2.4 % mAP gains in Hard-View-EALN.

From Table I, it is obvious that our Hard-View-EALN achieves the best performance. More specifically, the EN loss which combines softmax (treating identical ID samples as a category) and triplet loss is a very efficient method for vehicle ReID, which has also been proved in person ReID [52]. However, most of the compared methods focus on utilizing

TABLE I

PERFORMANCE COMPARISONS (%) WITH STATE-OF-THE-ART METHODS FOR VEHICLEID AND VERI-776 DATASET

VeRI-776				VehicleID								
Settings	Test Size= 11579			Settings	Query Number= 800			Query Number= 1600			Query Number= 2400		
Methods	mAP	r = 1	r = 5	Methods	mAP	r = 1	r = 5	mAP	r = 1	r = 5	mAP	r = 1	r = 5
LOMO [58]	9.78	23.87	39.14	LOMO [58]	-	19.76	32.01	-	18.85	29.18	-	15.32	25.29
GoogLeNet [8]	17.81	52.12	66.79	GoogLeNet [8]	-	47.88	67.18	-	43.40	63.86	-	38.27	59.39
XVGAN [59]	24.65	60.20	77.03	XVGAN [59]	-	52.87	80.83	-	49.55	71.39	-	44.89	66.65
FACT +Plate-SNN + STR [5]	27.77	61.44	78.78	CCL [6]	49.2	43.62	64.84	44.8	39.94	62.98	38.6	35.68	56.24
SiameseVisual [11]	29.48	41.12	60.31	Mixed Diff+CCL [6]	54.6	48.93	75.65	48.1	45.05	68.85	45.5	41.05	63.38
OIFE [60]	48.00	65.92	87.66	HDC + Contrastive [10]	65.5	-	-	63.1	-	-	57.5	-	-
VAMI [4]	50.13	77.03	90.82	VAMI [4]	-	63.12	83.25	-	52.87	75.12	-	47.34	70.29
Defense Triplet [61]	49.95	80.86	90.23	Defense Triplet [61]	68.9	65.2	77.93	65.37	62.02	76.69	61.37	57.20	71.91
EN (ours)	47.85	79.67	89.45	EN (ours)	67.5	64.23	76.14	65.4	61.45	75.65	62.0	57.27	72.41
View-EALN (ours)	50.32	81.34	90.88	View-EALN (ours)	70.2	67.19	78.20	67.45	63.23	77.12	63.88	59.98	74.20
Hard-EALN (ours)	55.36	84.26	92.25	Hard-EALN (ours)	75.1	71.79	86.20	72.0	68.76	80.73	69.5	65.88	79.98
Hard-View-EALN (ours)	**57.44**	**84.39**	**94.05**	Hard-View-EALN (ours)	**77.5**	**75.11**	**88.09**	**74.2**	**71.78**	**83.94**	**71.0**	**69.30**	**81.42**

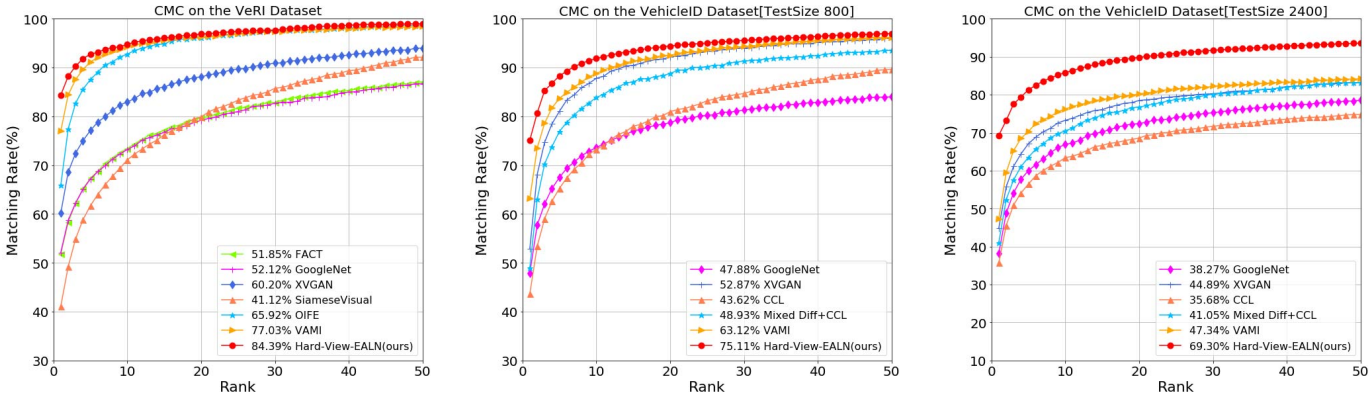


Fig. 4. The CMC comparisons of different methods on VeRI-776 and VehicleID datasets.

extra semantic information such as vehicle model or vehicle viewpoint, ignoring the importance of vehicle ID classification. HDC+Contrastive [10] adopts a similar strategy, but they focus on extracting hard negatives in the dataset for better training. Comparing Hard-EALN and HDC+Contrastive, we can find our online hard negative generation scheme can achieve more significant performance improvements. Besides, HDC+Contrastive cascades a set of GoogleNet which is a more complex method than a single network in our model, which further demonstrates the effectiveness of our method. Regarding the method VAMI [4], though the motivation of it is similar to our cross-view methodology, they aim to infer viewpoint aware attentive regions for multi-view representation. In three different scale test sets, our Hard-View-EALN can consistently outperform it by an obvious performance margin.

We plot the CMC curve of our Hard-EALN and comparison methods on Vehicle ID dataset, as shown in Fig. 4. The results both on small scale (Query Number = 800) and large scale (Query Number = 2400) are provided. In particular, we achieve much higher Rank 1 value (Top 1 Accuracy), implying that our method can better differentiate subtle differences benefiting from hard negative adversary. In addition, it is worth noting that in Fig. 4 our method has higher superiorities in large scale test set (TestSize = 2400) than in small scale one (TestSize = 800), which demonstrates stronger discriminative capabilities when the number of similar vehicles is increased.

4) Evaluation on VeRI-776 Dataset: The ReID performance comparisons on VeRI-776 dataset are shown in Table I. As for the incremental results of structure "EN", "Hard-EALN" and "Hard-View-EALN", we can find the hard negative adversary contributes significant performance boosting, and it brings about 7.51% mAP performance gains in "Hard-EALN" (from 47.85% mAP to 55.36% mAP). Moreover, additional cross-view adversary further promotes 2.08% mAP gains in "Hard-View-EALN". In both VehicleID and VeRI-776 datasets, we can find that significant performance improvements can be achieved by the combination of hard negative and cross-view adversary, which proves the effectiveness of embedding adversarial learning. Regarding the contributions of the performance gain, hard negative adversary plays the dominant role, indicating that the online generated hard negatives are of great potential to improve discriminative capability of the discriminator network.

Regarding the comparisons with state-of-the-art methods in VeRI-776 dataset, our Hard-EALN outperforms the state-of-the-art method VAMI [4] by 5.23% mAP. With the cross-view scheme, an additional 2.08% mAP gains can be achieved.

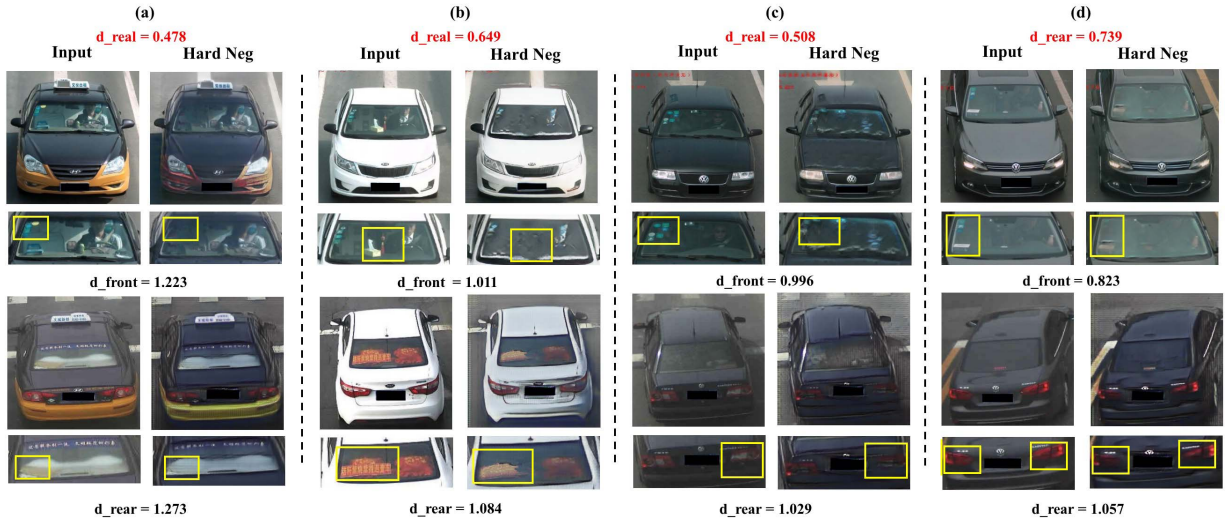


Fig. 5. Visualization of Hard Negative generation on Vehicle ID dataset. The input samples contain front view and rear view, and their corresponding hard negatives are provided. We crop and zoom in the regions with subtle differences compared to the input. The notations d_front and d_rear are feature distances between input and the generated hard negative of front and rear view in D_{emb}. In particular, d_real denotes the average distance of the input and its real positive instances. From the distance comparison between d_front, d_rear and d_real (d_rear and $d\_front >> d\_real$), we can find that though the hard negative generator is able to generate very hard negatives, the D_{emb} can still discriminate them in the embedding space.

TABLE II

THE REID PERFORMANCE (MAP) OF IDENTICAL-VIEW AND CROSS-VIEW ON THE SMALL SCALE (TESTSIZE = 800) TEST SET IN VEHICLEID

Methods	ALL	Same View	Cross View
Hard-EALN	75.1	89.6	37.8
Hard-View-EALN	77.5	89.2	42.6

TABLE III

THE REID PERFORMANCE (MAP) WITH CROSS-VIEW SCHEME ON THE VERI-776 DATASET

Methods	mAP	Rank 1	Rank 5
EN	47.85	79.67	89.45
View-EALN (ours)	50.32	81.34	90.88
EN-CycleGAN [49]	41.21	75.51	86.35
EN-StarGAN [50]	30.59	53.28	71.75

It is worth mentioning that VAMI [4] also concentrates on multi-view feature representation to improve the ReID performance, while our method pays more attention to improving subtle differences representation capability from hard negative adversary. Compared with OIFE [60], which used key-point alignment in vehicle feature representation, the proposed Hard-EALN achieves better performance with 7.36% mAP superiority. The CMC curves on VeRI-776 dataset are also provided in Fig. 4. It can be observed that our Hard-EALN has also achieved superior performances, especially in the higher ranking matching rate (our Hard-EALN 84.26% v.s. state-of-the-art 77.03% VAMI).

5) Analysis of Cross-View: To further investigate the performance gains in cross-view vehicle ReID, we list the ReID performance variations in the same/cross-view in Table II.

Given a query vehicle, for separate cross-view vehicle ReID, we treat the reference vehicle belonging to this vehicle but with the same view as the junk samples, which are not involved in mAP computation [62]. Our Hard-EALN model achieves 37.8% mAP in cross-view ReID and 89.6% mAP in same-view ReID, which demonstrates that the viewpoint variances significantly influence the ReID performance. Fortunately, our Hard-View-EALN artificially generates another view and fuses the two-view vehicle features to improve cross-view retrieval. It can be seen that 4.8 % mAP gains in cross-view ReID can be achieved by cross-view adversary in Hard-View-EALN. Due to the feature fusion strategy, the performance of the same-view ReID drops 0.4% in terms of mAP. Finally, in "ALL" mAP computation, 2.4 % mAP improvements can be obtained. Analogously, regarding the VeRI-776 dataset, based on the generation of vehicle images, 2.08% mAP performance improvement from 55.36% to 57.44% can be observed.

Moreover, in Table III, we present numerical results of cross-view scheme with other domain transfer approaches such as CycleGAN [48] and StarGAN [49]. From the comparison results, our View-EALN has significantly outperformed the CycleGAN as well as StarGAN. Compared with baseline method EN, it is very clear that the ReID performances based on StarGAN and CycleGAN drop significantly. Since these two methods only focus on the transfer of different domains and do not constrain the feature distances between the generated cross-view samples and the given inputs, the vehicle identity consistence cannot be ensured.

C. Qualitative Study

1) Vehicle Generation:

a) Hard negative generation: The input and generated hard negatives are pair-wisely shown in Fig. 5, from which we can observe that the hard negatives have very similar appearance with the inputs, and minor modifications have harmoniously been made for distinctions. For example, in the front-view pair, not only the annual inspection marks and the other decorations are erased, but also the vehicle windows are

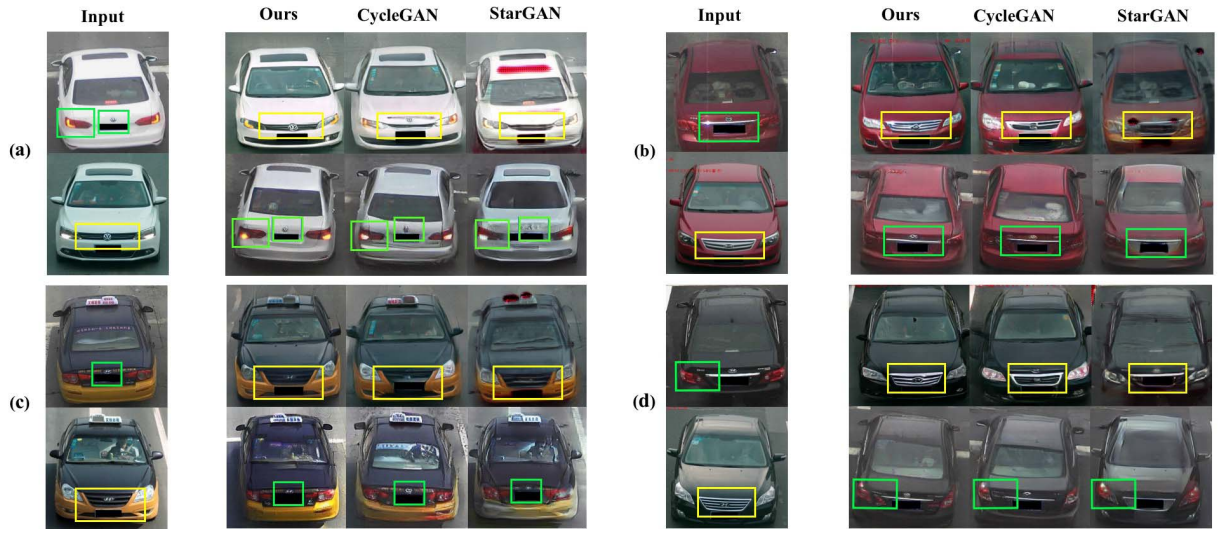


Fig. 6. Comparisons of different methods for cross-view generation. From left to right: input and the samples from our EALN, CycleGAN and StarGAN methods. Each input contains a front (upper) and rear (bottom) view vehicle, and their corresponding cross-view results are given. As such, the generated results can be compared with the ground-truth cross-view input.

blurred and distorted. In rear-view pair, the logos and lights are also occluded and distorted. In addition, We further demonstrate distance distributions in the embedding space including the distances between the inputs and their corresponding hard negatives from front view and rear view (d_front and d_rear), as well as the average distances between the input and its positives belonging to the same vehicle ID ($d(real)$). In Fig. 5 (a), both d_front and d_rear are much larger than d_real, implying that the discriminator D_{emb} is able to discriminate their subtle differences in the embedding space. Similar results can also be found in Fig. 5 (b)∼(d). Such evidences can prove that the hard negative adversarial learning can promote embedding discriminator to yield stronger discriminative capability.

To better illustrate the hard negatives generation procedure, we visualize the hard negatives every 10 epochs in Fig. 7. It can be seen that the generation of hard negatives undergoes 5 distinct stages. In the first two stages (0∼20 epochs), the hard negative generator attempts to apply contrast reduction and color jittering to cheat the hard negative discriminator. Along with the training process, hard negative adversarial learning would make the generator and discriminator stronger. In the following two stages (20∼40 epoch), we can find the generator attempts to modify the details of the input vehicle such as erasing decorations marks on window or changing objects behind windshields. However, at this stage the generation is not perfect enough since there are some fuzzy effects (from significant to inconspicuous along with training). Over 50 epochs, the hard negatives only show subtle differences compared to the input, which are extremely difficult to discriminate from both visual appearances and feature distances. From another perspective of hard negative adversary, the hard negative discriminator has already been a much stronger one.

b) Cross-view vehicle generation: The input and generated cross-views are shown in Fig. 6, from which we can find that generation from front-to-rear works better than rear-to-front. This phenomenon can be explained by the fact there

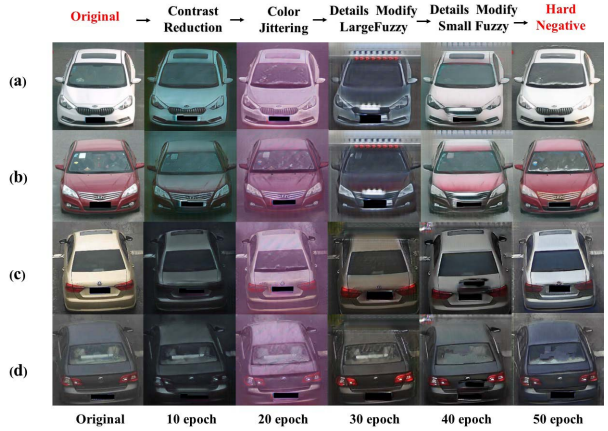


Fig. 7. Illustration of the generated hard negative samples for every 10 epochs.

are more complicated details existing on front view than rear view. We also compare the proposed scheme with the state-of-the-art methods including CycleGAN [48] and StarGAN [49], which are both representative methods in domain transfer. Different from style and facial expression transfer, the vehicle viewpoint transfer requires to change the whole body parts. As shown in Fig. 6, our method demonstrates an advantage in generating more realistic details with less artifacts. In particular, the details of the generated samples on lights and car inlet grilles are more visually reasonable and similar to real front vehicle. By contrast, the results of CycleGAN and StarGAN both show serious distortion and blur in cross-view generation. Moreover, the generated vehicle appearance (model type) has an obvious bias compared with the input vehicle, as shown in Fig. 6(a). This is owing to their objectives which do not consider instance identity guarantee mechanism. These visualizations also prove that the proposed embedding adversary can effectively improve the capability of the generator in terms of sample identity consistency and details representation.

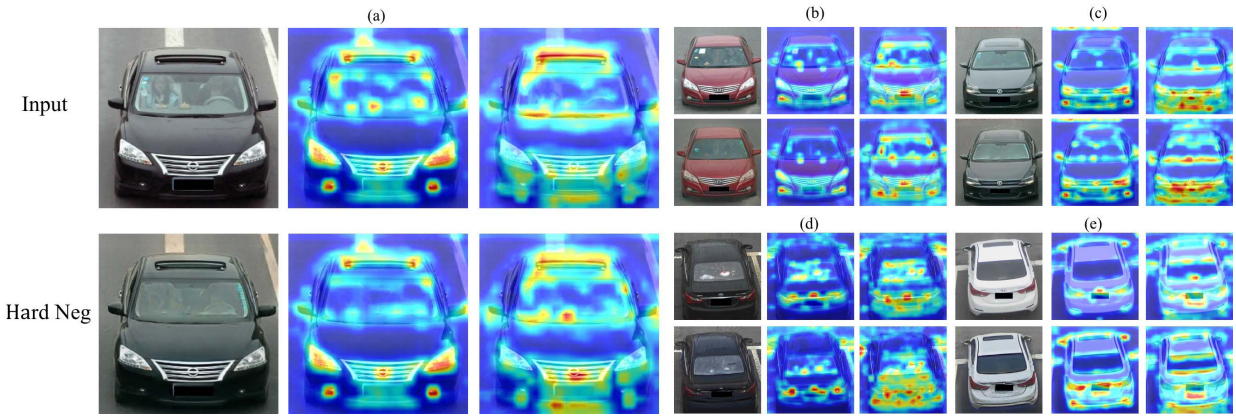


Fig. 8. Response maps of input and generated samples with different methods. Each comparison consists of vehicles of the original input (upper) and its hard negative generated by G_{hard} (lower) from the VehicleID dataset. These two columns responses maps (from left to right) represent Hard-EALN and EN, respectively.

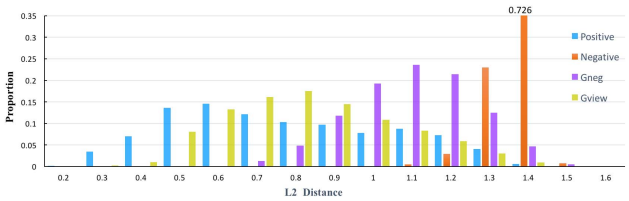


Fig. 9. Distance distribution on the vehicleID test set. The histogram represents the proportion of average distances distribution in positives, negatives, and our generated G_{neg} and G_{view}, respectively.

As vehicle cross-view generation is a very challenging task, we also have some bad cases. In particular, in rear-to-front generation, the vehicle inlet grille are sometimes slightly distorted. We also find that when the shapes of front and rear views have obvious differences (*e.g.* trunk), or the viewpoint is enormously changed, the results would be less satisfactory. More results are provided in the supplementary materials.

2) Response Maps Comparisons: In order to better understand our methods, we visualize the responses maps of input X and $G_{neg}(X)$ with EN and Hard-EALN methods, as shown in Fig. 8. The response map is extracted from the last convolutional layers (pool5) which undergoes the embedding layer to generate final feature representation. It is interesting to find that our Hard-EALN can better focus on the characteristics details compared to EN. For example, in Fig. 8 (a), the responses of our Hard-EALN are more concentrated which can be found on the headlight, foglight, vehicle logo as well as windshield. In particular, for the hard negative samples $G_{neg}(X)$, the signs and marks sticking left top corner and decorations at the middle of bottom are erased. Compared with EN, our Hard-EALN is more sensitive and has higher capability to capture subtle details on vehicle appearances. Similar results can also be found in Fig. 8 (b)$\sim$(e).

3) Distances Distribution: To demonstrate the effectiveness of generated hard negatives, we plot the distances distribution in embedding space of Vehicle ID test dataset in Fig. 9. We can find that average distances of our generated $G_{neg}(X)$ are much smaller compared to the negatives in the test set, and there are large distributions overlap with positive samples. Such evidences show that our generated hard negatives are much closer to the positives in the embedding space, also implying these generated hard negatives are more difficult than existing negatives in training set for the discriminator to differentiate from positives. This distribution can evidently illustrates the effectiveness of our proposed hard negative adversary scheme. Moreover, our generated cross-view samples distributions are quite similar and close to positives, which further proves our embedding adversarial learning is fairly effective and flexible.

V. Discussion

In this subsection, we will briefly discuss some concerned points regarding our EALN.

A. The Reasonability of Cross-View Generation

The motivation of cross-view generation is to get a cross-view vehicle image of the same model type of given query vehicle image, since some characteristics details on cross-view are impossible to be recovered in principle, especially for artificially placed decorated signs or marks. The proposed method aims to predict the shape and layout of the vehicle from another view, which provides an important clue for cross-view vehicle ReID.

B. The Choice of β_1, β_2 and Their Sensitivity

The value of β_1 and β_2 are decided according to the real sample distribution from the training set. As shown in Fig. 9, the distances between real positive samples mainly distribute close to 0.5, and the distances between real negative samples mainly distribute near 1.3. We expect to obtain hard negatives which mainly distribute larger than β_1, and meanwhile maintain a $\beta_2 + d_{pos}$ margin gap from the real negatives.

Moreover, from Eq (3), the value of β_1 and β_2 should satisfy $d_{pos} < \beta_1 < d_{neg} - \beta_2$. We then plot the performance changes with varying β_1 and β_2 in Fig. 10. It can be observed that β_1 is crucial to the performance improvement, since β_1 indicates the upper bound of the "hard degree" of the generated negative samples. With the increase of β_1, the generated negative samples are close to common negatives, and the performance

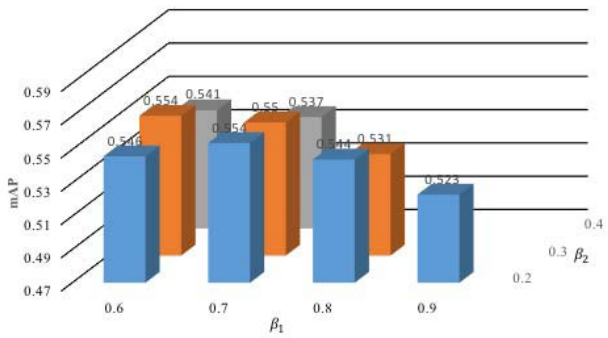


Fig. 10. The performance changes by varying β_1 and β_2 on VeRI dataset. According to the real distribution of d_{pos} and d_{neg}, we report the performance of reasonable value choosing of β_1 and β_2 under the above mentioned constraint $d_{pos} < \beta_1 < d_{neg} - \beta_2$.

gains tend to degrade a little because of less available hard negatives. Overall, even when the β_1 is larger than 0.9, we still achieve obvious performance gains from 48% to 52% in terms of mAP. As the β_2 is the margin gap from the selected real negatives, the different margin setting remains generally stable performance changes. Hence, in the optimization stage, β_1 which is the minimum relative distance is set to 0.6, and the value of β_2 which indicates the margin gap is set to 0.3.

C. The Complexity of EALN

In the training stage, each input vehicle is fed to the generator to generate an additional negative sample and a cross-view sample. Since the embedding model is incorporated into EALN as an embedding discriminator, the end-to-end optimization also avoids the multi-stage training pipeline, which leads to less optimization complexity. In the testing stage, the additional cost is caused by generating a cross-view sample when performing the ReID procedure. Therefore, in general, compared with the baseline embedding model optimization, our method introduces extra $O(n)$ complexity.

D. The Convergence Behavior of EALN

In the embedding adversarial scheme, the optimization objective of the embedding discriminator is identical with the common vehicle embedding models, and the generator in EALN provides extra hard negative samples. The convergence of the embedding discriminator can be ensured since theoretically the additional generated negative samples will not degrade the performance of the embedding model. We plot the loss curves of comparison EN and the embedding discriminator in Hard-EALN model in Fig. 11. It is clear that compared with the common embedding model, the Hard-EALN with more generated hard negatives can converge faster, especially in the early stage. As the training continues, the loss of Hard-EALN becomes more stable and lower than the baseline EN model, which also proves the effectiveness of hard negative generation scheme which aims to facilitate the embedding model training.

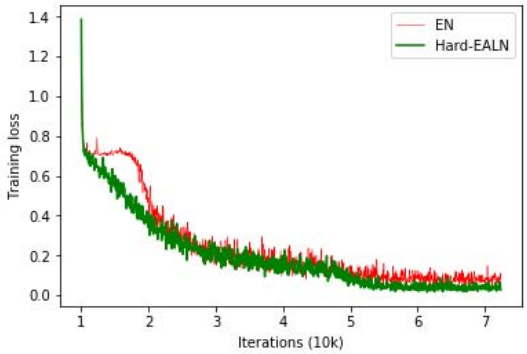


Fig. 11. Loss curves of EN and Hard-EALN in the training stage.

E. Limitations

Our approach performs fairly well in promoting the Vehicle ReID performance, but there are still several limitations to address:

- Our EALN focuses on using generation scheme to promote the discriminative capability of embedding model. However, compared with simply training an embedding model, our EALN involves the alternative training strategy between the generator and embedding discriminator, which consumes more time in the optimization stage.
- When training the hard negative generator G_{hard}, we randomly sample the real negative samples for computing the constraint in Eq(4). In fact, our motivation is to generate a negative sample which is close to the positives in the embedding space, such that the sampling of real negatives is also crucial and important. The selection scheme can be more flexible, such as using bootstrapping strategy or attributes selection to further promote the generation effects. This will be explored in future work.
- The hard negative scheme is only imposed in the training stage, while the cross-view scheme performs in the testing stage. Additional cross-view generation imports extra model storage and computation cost. Our future work will also explore feature adversarial scheme between different views in the training stage to improve cross-view feature representation, as well as avoid introducing additional computation in the testing stage.

VI. Conclusions

We present a novel embedding adversarial learning network for vehicle ReID, and the principle behind the design is to generate samples located in the specified vicinity areas in the embedding space. Such a unified end-to-end solution lay the groundwork for the subsequent improvement of vehicle ReID in terms of hard negative and cross-view sample generation. To further demonstrate the effectiveness of the proposed scheme, extensive qualitative and quantitive results are presented, which show that significant ReID accuracy improvement can be achieved.

There remain several open issues to be resolved in the future work. In particular, there is still substantial room for improving the performance for cross-view vehicle retrieval, and it is an extremely difficult task, even for humans, to infer

the details of a cross-view sample given a certain view, especially for artificially placed decorated signs or marks. In practice, a worthwhile direction is to model the vehicle with the 3D information, which can be further exploited for cross-view generation. Moreover, the proposed EALN can be extended in many ways to facilitate other tasks (*e.g.*, fine-grained image generation, image enhancement), which are worth investigating in our future work.

REFERENCES

[1] O. Bulan, V. Kozitsky, P. Ramesh, and M. Shreve, "Segmentation- and annotation-free license plate recognition with deep localization and failure identification," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2351–2363, Sep. 2017.

[2] C. Gou, K. Wang, Y. Yao, and Z. Li, "Vehicle license plate recognition based on extremal regions and restricted Boltzmann machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1096–1107, Apr. 2016.

[3] X. Liu, W. Liu, H. Ma, and H. Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[4] Y. Zhou and L. Shao, "Viewpoint-aware attentive multi-view inference for vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6489–6498.

[5] X. Liu, W. Liu, T. Mei, and H. Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 869–884.

[6] H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, "Deep relative distance learning: Tell the difference between similar vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2167–2175.

[7] J. M. Ernst, J. V. Krogmeier, and D. M. Bullock, "Estimating required probe vehicle re-identification requirements for characterizing link travel times," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 1, pp. 50–58, Jan. 2014.

[8] L. Yang, P. Luo, C. C. Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3973–3981.

[9] D. Zapletal and A. Herout, "Vehicle re-identification for automatic video traffic surveillance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun./Jul. 2016, pp. 25–31.

[10] Y. Yuan, K. Yang, and C. Zhang. (2016). "Hard-aware deeply cascaded embedding." [Online]. Available: https://arxiv.org/abs/1611.05720

[11] 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. IEEE ICCV*, Jun. 2017, pp. 1900–1909.

[12] Y. Tian, H.-H. Dong, L.-M. Jia, and S.-Y. Li, "A vehicle re-identification algorithm based on multi-sensor correlation," *J. Zhejiang Univ. Sci. C*, vol. 15, no. 5, pp. 372–382, 2014.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[14] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 3–20.

[15] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 761–769.

[16] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[17] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[18] X. Ma *et al.*, "Person re-identification by unsupervised video matching," *Pattern Recognit.*, vol. 65, pp. 197–210, May 2017.

[19] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. (2017). "Disentangled person image generation." [Online]. Available: https://arxiv.org/abs/1712.02621

[20] W. Lin *et al.*, "Learning correspondence structures for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2438–2453, May 2017.

[21] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[22] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4806–4817, Oct. 2017.

[23] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.

[24] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, Jun. 2017.

[25] J. García, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni, "Discriminant context information analysis for post-ranking person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1650–1665, Apr. 2017.

[26] Y.-J. Cho and K.-J. Yoon, "PaMM: Pose-aware multi-shot matching for improving person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3739–3752, Aug. 2018.

[27] R. S. Feris *et al.*, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 28–42, Feb. 2012.

[28] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2385–2399, Sep. 2018.

[29] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. (2015). "Embedding label structures for fine-grained feature representation." [Online]. Available: https://arxiv.org/abs/1512.02895

[30] J. Wang *et al.*, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.

[31] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3287, Jul. 2018.

[32] S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2963–2977, Dec. 2018.

[33] A. Borgia, Y. Hua, E. Kodirov, and N. M. Robertson, "Cross-view discriminative feature learning for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5338–5349, Nov. 2018.

[34] K. Li, Z. Ding, S. Li, and Y. Fu, "Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification," in *Proc. AAAI*, 2018, pp. 1–8.

[35] M. Mirza and S. Osindero. (2014). "Conditional generative adversarial nets." [Online]. Available: https://arxiv.org/abs/1411.1784

[36] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, "Pixel-level domain transfer," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 517–532.

[37] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 776–791.

[38] C. Ledig *et al.* (2017). "Photo-realistic single image super-resolution using a generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.04802

[39] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[40] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 469–477.

[41] Z. Yi, H. Zhang, P. Tan, and M. Gong. (2017). "DualGAN: Unsupervised dual learning for image-to-image translation." [Online]. Available: https://arxiv.org/abs/1704.02510

[42] J. Zhao, M. Mathieu, and Y. LeCun. (2016). "Energy-based generative adversarial network." [Online]. Available: https://arxiv.org/abs/1609.03126

[43] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[44] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[45] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1486–1494.

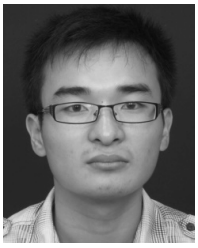
[46] H. Zhang *et al.* (2017). "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1612.03242

[47] A. Radford, L. Metz, and S. Chintala. (2015). "Unsupervised representation learning with deep convolutional generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1511.06434

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks." [Online]. Available: https://arxiv.org/abs/1703.10593
[49] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. (2017). "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation." [Online]. Available: https://arxiv.org/abs/1711.09020
[50] 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. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.
[51] Z. Zheng, L. Zheng, and Y. Yang. (2017). "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*." [Online]. Available: https://arxiv.org/abs/1701.07717
[52] Y. Wen, K. Zhang, and Z. Li, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, 2016, pp. 499–515.
[53] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1114–1123.
[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 694–711.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[56] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. (2016). "Instance normalization: The missing ingredient for fast stylization." [Online]. Available: https://arxiv.org/abs/1607.08022
[57] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. (2017). "Image-to-image translation with conditional adversarial networks." https://arxiv.org/abs/1611.07004
[58] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.
[59] Y. Zhou and L. Shao, "Cross-view GAN based vehicle generation for re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–12.
[60] Z. Wang *et al.*, "Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 379–387.
[61] A. Hermans, L. Beyer, and B. Leibe. (2017). "In defense of the triplet loss for person re-identification." [Online]. Available: https://arxiv.org/abs/1703.07737
[62] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

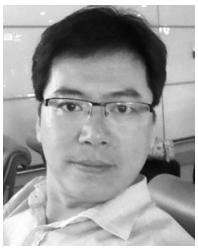
Yihang Lou (M'18) received the B.S. degree in software engineering from the Dalian University of Technology, Liaoning, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Peking University, Beijing, China. His current interests include large-scale image retrieval and video content analysis.

Yan Bai (M'18) received the M.S. degree from the School of Electrical Engineering and Computer Science, Peking University, China, in 2018. She is currently with HuLu LLC. Her research interests include large-scale video retrieval and fine-grained visual recognition.



Jun Liu received the B.Eng. degree from Central South University, China, in 2011, and the M.Sc. degree from Fudan University, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include computer vision and machine learning.

Shiqi Wang (M'15) received the B.S. degree in computer science from the Harbin Institute of Technology in 2008, and the Ph.D. degree in computer application technology from Peking University in 2014. From 2014 to 2016, he was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. From 2016 to 2017, he was with the Rapid-Rich Object Search Laboratory, Nanyang Technological University, Singapore, as a Research Fellow. He is currently an Assistant Professor with the Department of Computer Science, City University of Hong Kong. He has given over 40 technical proposals to ISO/MPEG, ITU-T, and AVS standards. His research interests include video compression, image/video quality assessment, and image/video search and analysis.



Ling-Yu Duan (M'06) received the Ph.D. degree in information technology from The University of Newcastle, Callaghan, Australia, in 2008. He has been serving as an Associate Director of the Rapid-Rich Object Search Laboratory, a joint lab between Nanyang Technological University, Singapore, and Peking University (PKU), China, since 2012. He is currently a Full Professor with the National Engineering Laboratory of Video Technology, School of Electronics Engineering and Computer Science, PKU. His research interests include multimedia indexing, search, and retrieval, mobile visual search, visual feature coding, and video analytics. He received the *EURASIP Journal on Image and Video Processing* Best Paper Award in 2015, the Ministry of Education Technology Invention Award (first prize) in 2016, the National Technology Invention Award (second prize) in 2017, the China Patent Award for Excellence in 2017, and the National Information Technology Standardization Technical Committee Standardization Work Outstanding Person Award in 2015. He is serving as the Co-Chair for the MPEG Compact Descriptor for Video Analytics. He was a Co-Editor of the MPEG Compact Descriptor for Visual Search Standard (ISO/IEC 15938-13). He is currently an Associate Editor of the *ACM Transactions on Intelligent Systems and Technology* and the *ACM Transactions on Multimedia Computing, Communications, and Applications*.