Viewpoint-aware Attentive Multi-view Inference for Vehicle Re-identification

Yi Zhou Ling Shao
Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE
School of Computing Sciences, University of East Anglia
y.zhou1@uea.ac.uk ling.shao@ieee.org

Abstract

Vehicle re-identification (re-ID) has the huge potential to contribute to the intelligent video surveillance. However, it suffers from challenges that different vehicle identities with a similar appearance have little inter-instance discrepancy while one vehicle usually has large intra-instance differences under viewpoint and illumination variations. Previous methods address vehicle re-ID by simply using visual features from originally captured views and usually exploit the spatial-temporal information of the vehicles to refine the results. In this paper, we propose a Viewpoint-aware Attentive Multi-view Inference (VAMI) model that only requires visual information to solve the multi-view vehicle re-ID problem. Given vehicle images of arbitrary viewpoints, the VAMI extracts the single-view feature for each input image and aims to transform the features into a global multi-view feature representation so that pairwise distance metric learning can be better optimized in such a viewpoint-invariant feature space. The VAMI adopts a viewpoint-aware attention model to select core regions at different viewpoints and implement effective multi-view feature inference by an adversarial training architecture. Extensive experiments validate the effectiveness of each proposed component and illustrate that our approach achieves consistent improvements over state-of-the-art vehicle re-ID methods on two public datasets: VeRi and VehicleID.

1. Introduction

Vehicle re-identification (re-ID) aims to spot a vehicle of interest from multiple non-overlapping cameras in surveillance systems. It can be applied to practical scenarios in intelligent transportation systems such as urban surveillance and security. However, compared with a similar topic called person re-ID [5, 19, 27, 37, 12, 35] which has been extensively explored, vehicle re-ID encounters more challenges.

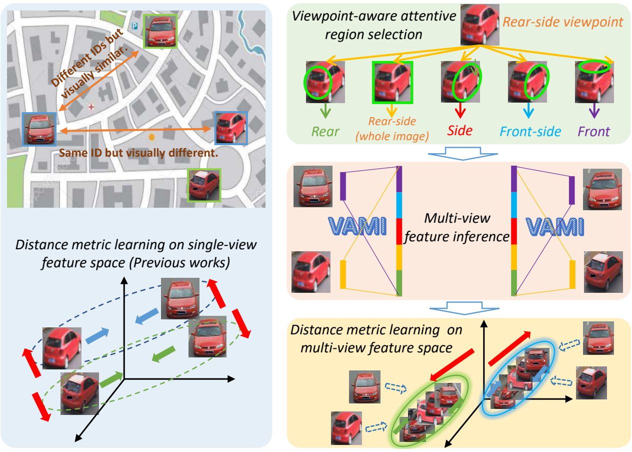

Different IDs but visually similar.
Same ID but visually different.
Distance metric learning on single-view feature space (Previous works)
Viewpoint-aware attentive region selection
Rear-side viewpoint
Rear
Rear-side (whole image)
Side
Front-side
Front
VAMI
Multi-view feature inference
VAMI
Distance metric learning on multi-view feature space

Figure 1. The left part shows the challenges of vehicle re-ID and a general framework of previous works. The right part illustrates the motivation of our proposed method that infers a multi-view feature representation from a single-view input, thus distance metrics can be optimized in the viewpoint-invariant multi-view feature space.

The top-left part of Fig. 1 reveals the two main obstacles of vehicle re-ID. One inherent difficulty is that a vehicle captured in different viewpoints usually has dramatically varied visual appearances. In contrast, two different vehicles of the same color and type have a similar appearance from the same viewpoint. The subtle inter-instance discrepancy between images of different vehicles and the large intra-instance difference between images of the same vehicle make the matching problem unsatisfactorily addressed by existing vision models. Recent re-ID methods are mainly proposed for persons, which can be categorized into three groups including feature learning [33, 10, 32, 40, 28], distance metric learning [30, 39, 24, 2] and subspace learning [38, 1, 31, 34]. All these methods utilize features of originally captured views to train models and compute distances between vehicle pairs, in which multi-view processing was not sufficiently considered. Directly deploying person re-ID models for vehicles does not achieve expected performance since features such as color and texture of clothes and pants can be used for humans even with large viewpoint variations but not for vehicles. Many vehicle re-ID researchers also noticed the challenges, thus preferred to make use of license plate or spatial-temporal information [15, 26, 23] to obtain promising results. However, license plate recognition requires high-resolution images in front or rear viewpoints, and the spatial-temporal information is usually hard to be obtained. Therefore, a general model only based on visual appearances is more desired for vehicle re-ID.

In this paper, we propose a viewpoint-aware attentive multi-view inference (VAMI) model to infer multi-view features from single-view image inputs. Then, distance metrics can be learned on the generated viewpoint-invariant multi-view feature space. The right column of Fig. 1 illustrates the motivation of our proposed VAMI. The main contributions of the VAMI are highlighted as follows:

(1) A viewpoint-aware attention model is proposed to obtain attention maps from the input image. The high-scored region of each map shows the overlapped appearance between the input vehicle's view and a target viewpoint. For instance, to infer the side view feature from a front-side view input image, the VAMI only pays attention to the vehicle's side pattern while ignoring the front region.

(2) Given the attentive features of a single-view input, we design a conditional multi-view generative network to infer a global feature containing different viewpoints' information of the input vehicle. The adversarial training mechanism and auxiliary vehicle attribute classifiers are combined to achieve effective feature generation.

(3) In addition to inferring multi-view features, we embed pairwise distance metric learning in the network to place the same vehicle together and push different vehicles away. The distance metric can be more suitably optimized in the generated viewpoint-invariant feature space. Extensive ablation studies and comparison experiments conducted on the public VeRi and VehicleID datasets have demonstrated the effectiveness and superiority of our VAMI over state-of-the-art vehicle re-ID approaches.

2. Related Work

Vehicle Re-ID. Inspired by person re-ID, vehicle re-ID has attracted more attention in the past two years. Deep relative distance learning [13] is designed for learning the difference between similar vehicles based on a new VehicleID dataset which contains two viewpoints: front and rear. Liu *et al.* [14, 15] released the VeRi-776 dataset where vehicles have more available viewpoints. They employed visual feature, license plate and spatial-temporal information to explore the re-ID task. Zhou *et al.* [41] designed a conditional generative network to infer cross-view images from input view pairs and then combined the input and generated views to improve the re-ID performance. Moreover, Shen *et al.* [23] and Wang *et al.* [26] proposed the visual-spatio-temporal path proposals and orientation invariant feature embedding as well as spatial-temporal regularization, respectively, to focus on exploiting vehicles' spatial and temporal information to address the vehicle re-ID task.

Attention Mechanisms. Visual attention mechanisms aim to automatically focus on the core regions of image inputs and ignore the useless parts. The visual attention models' ability of selective feature extraction has been extensively explored in many applications including image classification [16, 25], fine-grained image recognition [4, 36], image captioning [18, 2] and VQA [42]. Existing attention mechanisms can be mainly categorized into two groups. One is the fully-differentiable soft attention model which aims to learn attention maps to weight different regions of an image. The other is the hard attention model which is a stochastic process sampling hidden states with probabilities. Hard attention models are not differentiable and usually learned by reinforcement learning [21].

Generative Adversarial Networks (GANs). GANs consist of a generative model G and a discriminative model D competing against each other in a two-player min-max game. It has achieved great success on many vision tasks such as image generation [6, 20] and image translation [9, 43]. In essence, the design of the adversarial learning leads to the success of GAN, mainly because it forces the generated samples to be indistinguishable from real data. Moreover, many works extend GAN to conditional frameworks such as InfoGAN [3], AC-GAN [17] and CycleGAN [43] to investigate better models for generation tasks. In this paper, we propose an attentive multi-view feature generative network by adversarial learning.

3. Proposed Methods

3.1. Problem Formulation

The overall target of vehicle re-ID is the same with that for person. Given a query vehicle image, a ranking list of candidates in the gallery set is desired, placing images of the query vehicle's identity in top positions and vice versa. Define a pair of images $(\mathbf{I}_i, \mathbf{I}_j)$ and their corresponding similarity label $\mathbf{l}_{ij}$. If $\mathbf{I}_i$ and $\mathbf{I}_j$ are two views from the same vehicle, then $\mathbf{l}_{ij} = 1$, while $\mathbf{l}_{ij} = 0$ if they are from different vehicles. For each single-view input image $\mathbf{I}$, we aim to map its feature to a multi-view representation $\mathbf{f}$ by the following function:

$$\mathbf{f} = T(\text{concat}(\{\mathbf{x}_v\}_{v=1}^{V})) = T(\text{concat}(F(\mathbf{I}) \cdot \{\alpha_v\}_{v=1}^{V})). \tag{1}$$

The operator $F(\cdot)$ is to extract the feature of the input image $\mathbf{I}$. $\{\alpha_v\}_{v=1}^{V}$ is obtained by the viewpoint-aware attention model to select overlapped regions between the input view and a target viewpoint v, where V is the defined number of viewpoints. Moreover, the operator $T(\cdot)$ denotes the transformation from the concatenated attentive single-view features $\{\mathbf{x}_v\}_{v=1}^{V}$ to the inferred multi-view features.

After modeling $\mathbf{f}$, we simultaneously aim to optimize the network minimizing a loss function $\mathcal{L}_{reid}$ to shorten the dis-

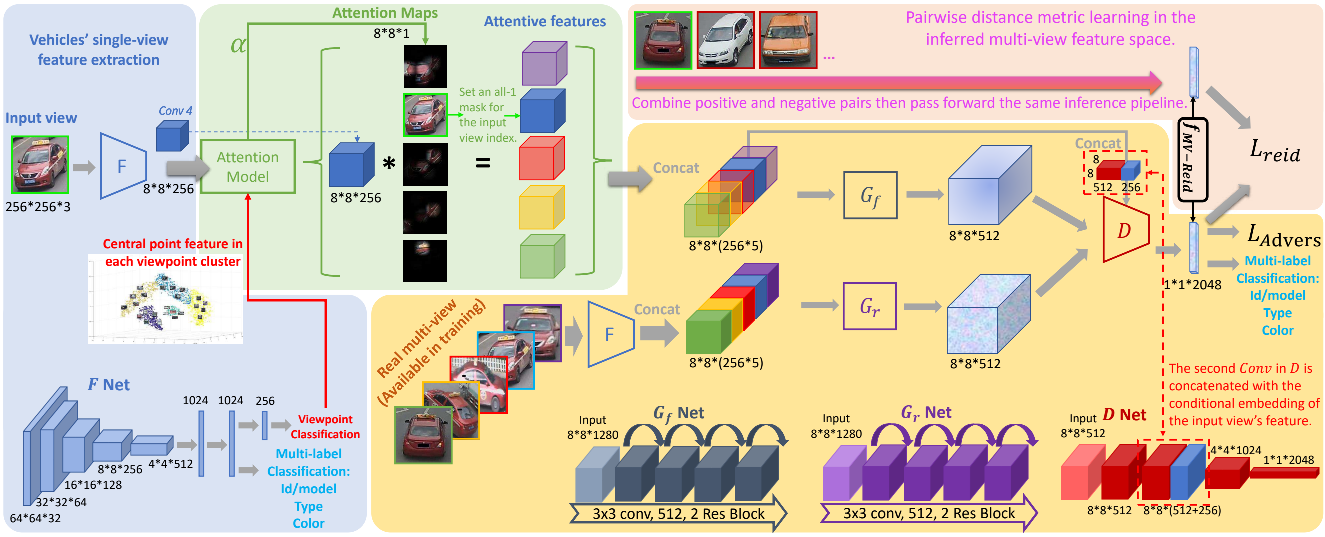

Vehicles' single-view feature extraction
Input view
256*256*3
F
Conv 4
8*8*256
Attention Model
α
Attention Maps
8*8*1
8*8*256
*
=
Set an all-1 mask for the input view index.
Attentive features
Central point feature in each viewpoint cluster
F Net
1024
1024
256
Viewpoint Classification
Multi-label Classification: Id/model Type Color
8*8*256
4*4*512
16*16*128
32*32*64
64*64*32
Pairwise distance metric learning in the inferred multi-view feature space.
...
Combine positive and negative pairs then pass forward the same inference pipeline.
Concat
8*8*(256*5)
G_f
8*8*512
Concat
8
8
512
256
D
$f_{MV-Reid}$
L_{reid}
L_{Advers}
1*1*2048
Multi-label Classification: Id/model Type Color
Real multi-view (Available in training)
F
Concat
8*8*(256*5)
G_r
8*8*512
The second Conv in D is concatenated with the conditional embedding of the input view's feature.
Input 8*8*1280
G_f Net
3x3 conv, 512, 2 Res Block
Input 8*8*1280
G_r Net
3x3 conv, 512, 2 Res Block
Input 8*8*512
D Net
4*4*1024
1*1*2048
8*8*512
8*8*(512+256)

Figure 2. An overview of the architecture of VAMI. The F Net is for learning single-view features containing vehicles' intrinsic information such as model, color and type. Moreover, viewpoint features can be also learned so that the central point feature of each viewpoint cluster over the whole training set can be obtained and used for attention learning. The attention model aims to output viewpoint-aware attention maps from the input view image targeting at different viewpoints. To infer multi-view features from the obtained attentive single-view features, we design a conditional generative network trained by an adversarial architecture. The network of the real multi-view data branch is only available in the training phase. Auxiliary vehicle classifiers are configured at the end of D to help match the inferred multi-view features with correct input vehicles' identities. Finally, given positive and negative vehicle pairs, a contrastive loss is designed to optimize the network for distance metric learning. (Best viewed in color.)

tance between $\mathbf{f}_i$ and $\mathbf{f}_j$ when $\mathbf{l}_{ij} = 1$ and maximize that when $\mathbf{l}_{ij} = 0$ by adopting the pairwise contrastive loss [7]. Therefore, the most significant factor to achieve effective re-ID is how to design and optimize the $F(\cdot)$, α and $T(\cdot)$.

3.2. Attentive Multi-View Inference Network

Our proposed viewpoint-aware attentive multi-view inference network mainly consists of four important components. The network architecture is illustrated in Fig. 2. Learning $F(\cdot)$ for extracting vehicles' single-view features is first addressed by training a deep CNN using vehicles' attribute labels. To obtain viewpoint-aware attention maps α for extracting core regions targeting at different viewpoints from the input view, corresponding viewpoint embeddings are adopted to attend to one intermediate layer of the F Net. Exploiting the attentive feature maps for different viewpoints as conditions, we aim to generate multi-view features by $T(\cdot)$ with an adversarial training architecture. During training, features extracted from real images in various viewpoints of the input vehicle are used for the real data branch, but this branch is no longer needed in the testing phase. The discriminator simultaneously distinguishes the generated multi-view features from the real ones and adopts auxiliary vehicle classifiers to help match the inferred features with the correct input vehicle's identities. Finally, given pairwise image inputs, a contrastive loss is configured at the end to optimize the network embedded with distance metric learning. The details of each component are clearly explained in the following four sub-sections.

3.2.1 Vehicle Feature Learning

The F Net is built with a deep CNN module for learning vehicles' intrinsic features containing vehicles' model, color and type information. Its backbone deploys five convolutional ($conv$) layers and two fully-connected (fc) layers. The first two $conv$ layers are configured with 5×5 kernels, while the following three $conv$ layers are set with 3×3 kernels. Stride is set with 4 for the first $conv$ layer and 2 for the remaining $conv$ layers. The Leaky-ReLU is set after each layer with the leak of 0.2. Detailed hyper-parameters' settings are illustrated in the bottom-left part of Fig. 2.

In addition to two 1024-dimensional fc layers connected with multi-attributes classification, one more 256-dimensional fc layer is configured for viewpoint classification. All the vehicle images are coarsely categorized into five viewpoints ($V = 5$) as front, rear, side, front-side and rear-side which are enough to describe a vehicle comprehensively. After training the F Net, we can extract viewpoint features over all the training data and easily learn five viewpoints' feature clusters by k-means clustering, thus the feature in the center of each cluster called central viewpoint feature can be obtained. These central viewpoint features are used for learning the viewpoint-aware attention model.

3.2.2 Viewpoint-aware Attention Mechanism

Visual attention models can automatically select salient regions and drop useless information from features. In the vehicle re-ID problem, our model requires to focus on the overlapped visual pattern of vehicles between the input viewpoint and the target viewpoint. For instance, to tell the difference between two similar vehicles from the front-side and rear-side viewpoints, humans usually will pay attention to their shared side appearance to discriminate whether the two vehicles are the same or not. The top-right part of Fig. 3 shows some examples. Thus, we aim to address this problem by proposing a viewpoint-aware attention model.

Fig. 3 illustrates the underlying design of our attention mechanism. In order to extract feature vectors of different regions, we select the $Conv4$ layer of the F Net since it has high-level perceptrons and keeps a large enough spatial size. Thus, the input image is represented as $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N\}$, where N is the number of image regions and $\mathbf{u}_n$ is a 256-dimensional feature vector of the n-th region. Our model performs viewpoint-aware attentions by multiple steps. Attention mechanism at each step can be considered as a building block. An attention map can be produced by learning a context vector weakly supervised by labels indicating shared appearance between the input and target viewpoints.

The context vector at step t can attend to certain regions of the input view by the following equation:

$$\mathbf{c}^t = \mathbf{Attention}(\mathbf{c}^{t-1}, \{\mathbf{u}_n\}_{n=1}^N, \mathbf{v}), \tag{2}$$

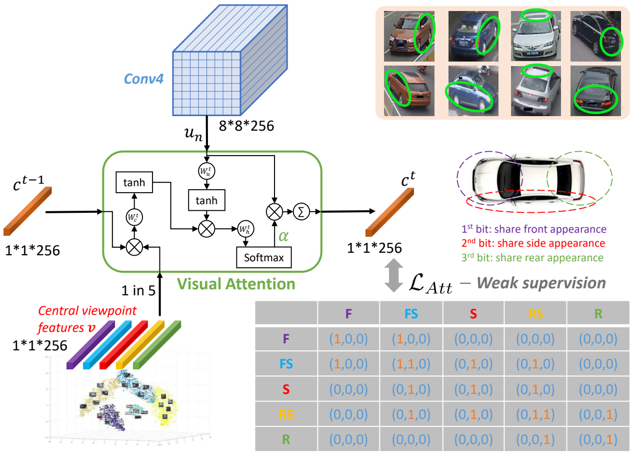
where $\mathbf{c}^{t-1}$ is the context vector at step $t-1$ and $\mathbf{v}$ denotes one of the five central viewpoint features. The soft attention mechanism is adopted that a weighted average of all the input feature vectors is used for computing the context vector. The attention weights $\{\alpha_n^t\}_{n=1}^N$ are calculated through two layer non-linear transformations and the softmax function:

$$\begin{aligned}
\mathbf{h}_n^t &= \tanh(\mathbf{W}_c^t(\mathbf{c}^{t-1} \odot \mathbf{v}) + \mathbf{b}_c^t) \odot \tanh(\mathbf{W}_u^t \mathbf{u}_n + \mathbf{b}_u^t), \\
\alpha_n^t &= \text{softmax}(\mathbf{W}_h^t \mathbf{h}_n^t + \mathbf{b}_h^t), \\
\mathbf{c}^t &= \sum_{n=1}^N \alpha_n^t \mathbf{u}_n,
\end{aligned} \tag{3}$$

where $\mathbf{W}_c^t$, $\mathbf{W}_u^t$, $\mathbf{W}_h^t$ and bias terms are learnable parameters. $\mathbf{h}_n^t$ is the hidden state and $\odot$ denotes the element-wise multiplication. The context vector $\mathbf{c}^0$ is initialized by:

$$\mathbf{c}^0 = \frac{1}{N} \sum_{n=1}^N \mathbf{u}_n. \tag{4}$$

Learning this viewpoint-aware attention model is mainly weakly supervised by the shared appearance region's labels between the input and target viewpoints. We design three-bit binary codes to encode the view-overlap information as shown in the bottom-right matrix of Fig. 3. The first bit is set as 1 when the two viewpoints share the front appearance, while the second and third bits denote whether the side and rear appearances are shared or not, respectively. The attention loss $\mathcal{L}_{Att}$ is optimized by the cross entropy. Specifically, if the input vehicle image is the front-side viewpoint and the target viewpoint is rear-side, the central viewpoint feature of rear-side will be adopted as the $\mathbf{v}$ and the supervision codes will be $(0, 1, 0)$ since the two viewpoints only share the side appearance region. Then, once the attention model is trained, it outputs an attention map only giving high response on the side appearance of the input vehicle. Moreover, for certain cases where none of the front, side or rear appearance is overlapped between viewpoint pairs (i.e. $(0, 0, 0)$), it is surprisingly observed that the top appearance would be attended, which is shown in the results of Sec. 4.2.


Conv4
u_n
8*8*256
c^{t-1}
1*1*256
tanh
W_c^t
W_u^t
tanh
W_h^t
Softmax
α
Σ
c^t
1*1*256
1 in 5
Visual Attention
Central viewpoint features v
1*1*256
1st bit: share front appearance
2nd bit: share side appearance
3rd bit: share rear appearance
$\mathcal{L}_{Att}$ – Weak supervision

	F	FS	S	RS	R
F	(1,0,0)	(1,0,0)	(0,0,0)	(0,0,0)	(0,0,0)
FS	(1,0,0)	(1,1,0)	(0,1,0)	(0,1,0)	(0,0,0)
S	(0,0,0)	(0,1,0)	(0,1,0)	(0,1,0)	(0,0,0)
RS	(0,0,0)	(0,1,0)	(0,1,0)	(0,1,1)	(0,0,1)
R	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,1)	(0,0,1)

Figure 3. The details of the viewpoint-aware attention model. The top-right part gives examples of overlapped regions of certain arbitrary viewpoint pairs.

Since the target is to infer multi-view features containing all the five viewpoints' information from the input view, as illustrated in the green curly brackets of Fig. 2, we extract the input view's $Conv4$ feature maps and output corresponding attention maps $\{\alpha_v\}_{v=1}^V$ for other four viewpoints. The feature maps of the input view are masked by different viewpoints' attention maps. Then, these intermediate attentive feature maps $\{\mathbf{x}_v\}_{v=1}^V$ are concatenated as conditional embeddings to further infer multi-view features.

3.2.3 Adversarial Multi-view Feature Learning

Traditional adversarial learning models employ a generative net and a discriminative net which are two competing neural networks. The generative net usually takes a latent random vector z from a uniform or Gaussian distribution as input to generate samples, while the discriminative net aims to distinguish the real data x from generated samples. The $p_z(z)$ is expected to converge to a target real data distribution $p_{data}(x)$. In this paper, we propose a conditional feature-level generative network to infer real multi-view features from the attentive features of single-view inputs.

Instead of generating real images by normal GANs, our model aims to transform single-view features into multi-view features by a generative model. Two networks for both the fake path and the real path are designed as G_f and G_r, respectively. The input of G_f is the concatenated attentive feature $\{\mathbf{x}_v\}_{v=1}^V$ of the input single-view image in which the noise is embedded in the form of dropout. The input of G_r is the real features $\{\bar{\mathbf{x}}_v\}_{v=1}^V$ of images from different viewpoints of the same vehicle identity with G_f. The G_r is designed mainly for better fusing and learning a real high-level multi-view feature of the input vehicle.

Since we do not need to generate images by gradually enlarging the spatial size of feature maps but infer high-level multi-view features, G_f and G_r are proposed with residual transformation modules rather than adopting deconvolutional layers. The residual transformation module consists of four residual blocks whose hyper-parameters are shown in Fig. 2. The advantage of using residual blocks is that the networks can better learn the transformation functions and fuse features of different viewpoints by a deeper perceptron. Moreover, G_f and G_r have the same architecture but do not share the parameters since they are set with different purposes. We tried to set G_f and G_r sharing parameters, but the model failed to converge since the inputs of the two paths have a huge difference.

The discriminative net D employs a general fully convolutional network to distinguish the real multi-view features from the generated ones. Rather than maximizing the output of the discriminator for generated data, the objective of feature matching [22] is employed to optimize G_f to match the statistics of features in an intermediate layer of D. The adversarial loss is defined in the following equation:

$$\begin{aligned}\mathcal{L}_{Advers} &= \max_D(\mathbb{E}(\log(D(G_r(\{\bar{\mathbf{x}}_v\}_{v=1}^V)))) \\ &+ \mathbb{E}(\log(1 - D(G_f(\{\mathbf{x}_v\}_{v=1}^V))))) \\ &+ \min_{G_f} ||\mathbb{E}(D_m(G_r(\{\bar{\mathbf{x}}_v\}_{v=1}^V))) - \mathbb{E}(D_m(G_f(\{\mathbf{x}_v\}_{v=1}^V)))||_2^2,\end{aligned} \tag{5}$$

where m means the m^{th} layer in D ($m = 4$ in our setting). Moreover, D is trained with auxiliary vehicles' multi-attributes classification to better match inferred multi-view features with input vehicles' identities. The architecture of D is shown in Fig. 2. The second *conv* layer is concatenated with the input single-view feature maps to better optimize the conditioned G_f and D. Then, we apply two more *conv* layers to output the final multi-view feature $\mathbf{f}_{MV\_Reid}$ which is a 2048-dimensional feature vector. The final *conv* layer deploys the 4×4 kernels while others use 3×3 kernels. For all the *conv* layers in G_f, G_r and D, we adopt Leaky-ReLU activation and batch normalization. The pre-activation proposed in [8] is implemented for residual blocks.

In the training phase, in addition to optimizing the $\mathcal{L}_{Advers}$, the $\mathcal{L}_{Reid}$ defined in Sec. 3.1 is configured to make the model learning with distance metrics given positive and negative vehicle image pairs. Learning $\mathcal{L}_{Reid}$ is based on the $\mathbf{f}_{MV\_Reid}$ inferred from the single-view input rather than corresponding real multi-view inputs. Our distance metric learning is more reasonable since the generated multi-view feature space is viewpoint-invariant. In the testing phase, only single-view inputs are available. Given any image pair in arbitrary viewpoints, each image can pass forward the F, G_f and D to infer the $\mathbf{f}_{MV\_Reid}$ containing all viewpoints' information of the input vehicle, then the Euclidean distance between the pair can be computed for the final re-ID ranking.

3.2.4 Optimization

The training scheme for VAMI consists of four steps. In the first step, the F Net for vehicle feature learning is trained using Softmax classifiers. Then, the computed five central viewpoint features are used for training the viewpoint-aware attention model by $\mathcal{L}_{Att}$. In the second step, the G_r for learning the real multi-view features from five viewpoints' inputs needs to be pre-trained by auxiliary vehicles' multi-attributes classification together with D. Otherwise, optimizing the G_f, G_r and D together at the early stage will make the $\mathcal{L}_{Advers}$ unstable since the fused real data distribution in the adversarial architecture has not been shaped. Once the G_r is trained, we fix it. In the following step, the conditioned G_f and D nets can be optimized by $\mathcal{L}_{Advers}$ to infer multi-view features from single-view inputs. Finally, the pairwise loss $\mathcal{L}_{Reid}$ is added to fine-tune the whole network except for F and G_r to learn distance metrics, since at the early training stage the inferred multi-view features are poor so that the $\mathcal{L}_{Reid}$ cannot contribute to the optimization.

4. Experiments

We first qualitatively demonstrate the viewpoint-aware attention model. Then, ablation studies and comparisons with state-of-the-art vehicle re-ID methods are evaluated on the VeRi [15] and VehicleID [13] datasets.

4.1. VeRi-776 and Training Details

Experiments are mainly conducted on the VeRi-776 dataset since each vehicle has multiple viewpoints' images so that we can fully evaluate the effectiveness of our VAMI. The VeRi dataset contains 776 different vehicles captured in 20 cameras. The whole dataset is split into 576 vehicles with 37,778 images for training and 200 vehicles with 11,579 images for testing. An additional set of 1,678 images selected from the test vehicles are used as query images. We strictly follow the evaluation protocol proposed in [15]. During training, for those vehicles without certain viewpoints, neighboring viewpoints are substituted. Since an image-to-track search is proposed, in addition to the Cumulative Matching Characteristic (CMC) curve, a mean average precision (mAP) is also adopted for evaluation [15].

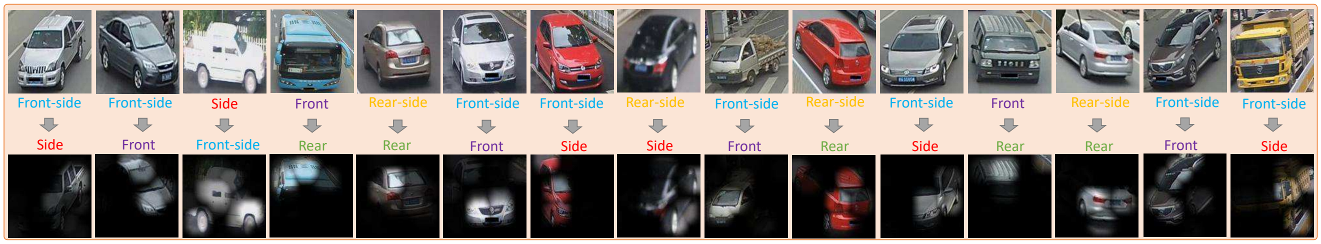


Figure 4. Viewpoint-aware attention maps. The upper row shows the input images and the bottom row shows the output attention maps. The highly-responded region is obtained by the input view attended with the central viewpoint feature of the target viewpoint.

To train the model, the ADAM Optimizer is adopted with the empirical learning rate of 0.0002 and the momentum of 0.5. The mini-batch size is set as 128. Training of the F Net and viewpoint-aware attention model are stopped after 30 and 35 epochs, respectively, when the losses converge to stable values. Moreover, we first pre-train the G_r and D by 50 epochs and then start the adversarial learning with the G_f for 200 epochs. Finally, we randomly combine 10k positive pairs and 30k negative pairs and add the $\mathcal{L}_{Reid}$ loss for a joint training by additional 50 epochs.

4.2. Qualitative Attention Map Results

Before evaluating the re-ID results, we first qualitatively demonstrate the effectiveness of the viewpoint-aware attention model. Fig. 4 shows some examples of attention maps achieved by our model. For instance, if the viewpoint of the input image is front-side and the target viewpoint is side, the central viewpoint feature of the side view will be used to attend to the side appearance region of the input view image. Then, only the feature in this region is selected for further multi-view feature inference. The effectiveness of this attention model for multi-view vehicle re-ID has been evaluated by the ablation study in Sec. 4.3.2.

4.3. Ablation Studies

4.3.1 Effect of Multi-View Inference

The primary contribution needed to be investigated is the effectiveness of the multi-view feature inference for vehicle re-ID. We compare the VAMI with three baselines. The first one simply adopts the feature of the original input view image extracted from the second fully-connected layer of the F Net. The second one adds a $\mathcal{L}_{Reid}$ to learn distance metrics based on the single-view features. Moreover, we also drop the $\mathcal{L}_{Reid}$ of the VAMI as a baseline to explore the improvement by metric learning on the multi-view features. Fig. 6(a) illustrates CMC curves of different approaches.

As shown in the upper half of Table 1, the mAP increases 13.3% by inferred multi-view features compared with original single-view features. Such a huge improvement shows the proposed multi-view inference indeed benefits the vehicle re-ID from arbitrary viewpoints. Optimizing $\mathcal{L}_{Reid}$ on the multi-view feature space has a further gain of 8.19% which shows the distance metric learning performed on the viewpoint-invariant feature space is more suitable. The ranking results demonstrate the similar tendency. Moreover, Fig. 5 compares qualitative results that most gallery candidates ranked in top positions by directly adopting single-view feature based learning usually have similar viewpoints with query ones, but more candidates with different viewpoints can be proposed by our VAMI and correct hits in the top-10 ranks become much more as well. The last two rows in the column of results by VAMI give the failure cases where some gallery candidates have highly similar appearances from the same viewpoint with the query vehicle images. In Fig. 5, the features of images of 20 test vehicles are also visualized. It shows samples of the same vehicle in different viewpoints are scattered based on the single-view features, but clustered after multi-view inference by VAMI.

Table 1. Evaluation (%) of effectiveness of the multi-view inference (MV Infer.) and adversarial network (Advers. Net.).

	Baselines	mAP	r=1	r=5	r=20
MV Infer.	Single-view feat	28.64	63.52	78.69	87.13
	Single-view feat + $\mathcal{L}_{Reid}$	32.59	66.21	80.63	89.86
	Multi-view feat	41.94	71.51	85.69	93.66
	Multi-view feat + $\mathcal{L}_{Reid}$	50.13	77.03	90.82	97.16
Advers. Net.	Regular objective for G_f	41.59	74.26	86.51	92.55
	No auxiliary classifiers for D	33.43	69.54	79.34	88.46
	Regular CNN for G_f and G_r	42.89	72.95	84.66	92.82
	ℓ_2 loss	34.96	67.92	82.60	91.48
	VAMI	50.13	77.03	90.82	97.16

4.3.2 Effect of Attention Model

To transform features across different viewpoints, we only need to attend to regions of the input view containing the appearance overlapped with target viewpoints while ignoring other useless regions. The viewpoint-aware attention model is dropped to explore its significance in this baseline. The input view's $Conv4$ feature maps are simply concatenated with a same size noise volume as the input for the G_f. Table 2 shows the mAP largely decreases 10.01% if we drop the attention model. However, the non-attentive multi-view feature inference module can still outperform the single-view based baselines.

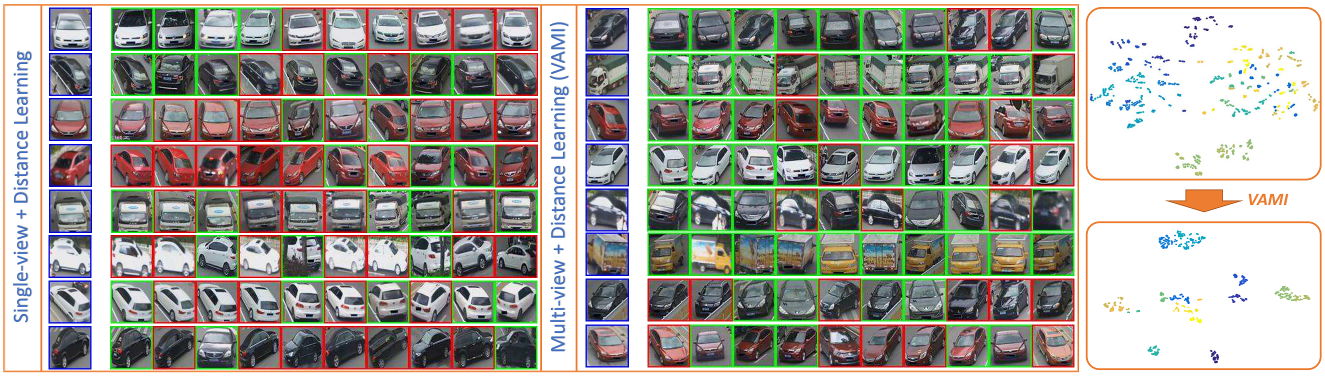
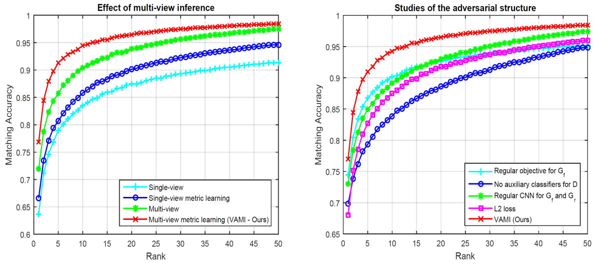

Single-view + Distance Learning
Multi-view + Distance Learning (VAMI)
VAMI

Figure 5. The left part compares qualitative results (Top-10 ranks) of single-view feature+$\mathcal{L}_{Reid}$ and our VAMI. Blue boxes are query vehicles, while the green and red boxes denote correct hits and incorrect ones, respectively. The right part visualizes the spaces of the original single-view features and the multi-view features inferred by our VAMI.



(a) (b)

Figure 6. (a) CMC results of evaluation of the multi-view inference. (b) CMC results of studies of the adversarial structure.

Table 2. Evaluation (%) of attention model. k is the number of the attention step. n is the noise rate in the form of dropout.

Baselines	mAP	r = 1	r = 5	r = 20
VAMI w/o attention	40.12	69.31	82.81	91.34
VAMI	50.13	77.03	90.82	97.16
k = 1	43.78	73.86	88.27	95.42
k = 2	50.13	77.03	90.82	97.16
k = 3	45.60	74.54	88.31	95.72
k = 4	41.05	70.23	83.44	91.90
n = 0.1	42.33	71.76	86.13	93.94
n = 0.2	43.25	73.32	87.98	95.10
n = 0.3	50.13	77.03	90.82	97.16
n = 0.4	47.68	75.19	88.97	96.04
n = 0.5	45.54	74.62	88.42	95.68
n = 0.6	41.59	71.49	85.45	93.30

Our attention model can be built by k steps in depth. Thus, the variable k is evaluated to explore the best performance. Table 2 shows the highest mAP is achieved when k equals 2. Moreover, since the noise is provided in the form of dropout for G_f, we also study its effect on the attentive multi-view feature inference by varying the dropout rate n. Through experiments, dropout rate as 0.3 gets the best re-ID results. Too small noise embedding makes the generator become deterministic, while too much noise weakens the attentive features degrading the model to the non-attentive version. Neither of such two cases gets satisfactory performance.

4.3.3 Adversarial Structure Studies

To study each designed component in the adversarial multi-view feature generation module, we explore them individually to compare the results.

Regular objective for G_f. Traditional objective for updating the generator in an adversarial architecture is to directly maximize the output of the discriminator, which usually overtrains the discriminator. The training of the G_f and D is unstable. We compare it as a baseline with our objective of feature matching.

No auxiliary classifiers for D. Configuring auxiliary vehicles' classifiers can make the adversarial networks learned with vehicles' intrinsic features. It helps to match the inferred multi-view features with correct identities of the input vehicles. Otherwise, the generated features can be close to real features but do not match input identities, which is a destructive weakness for the re-ID task.

Regular CNN for G_f and G_r. To evaluate the advantage of the designed residual transformation module, we compare it with a regular CNN without identity mapping. In this baseline, we configure convolutional layers with the same settings of hyper-parameters both for the G_f and G_r.

ℓ_2 loss. Our overall aim is to transform single-view features into multi-view features. To explore the ability of generating real features by our adversarial network, an ℓ_2 loss is adopted instead of $\mathcal{L}_{Advers}$ at f_{MV_Reid} to minimize the distance between features of the single-view and real multi-view paths. Noise embedding in G_f is dropped in this case.

Fig. 6(b) compares the CMC curves of different baselines. As shown in the bottom half of Table 1, the final proposed VAMI outperforms all the baselines by large margins, which validates each carefully designed component is effective to benefit the multi-view vehicle re-ID task.

Table 3. Comparisons (%) with state-of-the-art re-ID methods. Methods in the last three rows include spatial-temporal (ST) information.

VeRi					VehicleID									
Settings	Query = 1678, Test = 11579				Settings	Test Size = 800			Test Size = 1600			Test Size = 2400		
Methods	mAP	r = 1	r = 5	r = 20	Methods	r = 1	r = 5	r = 20	r = 1	r = 5	r = 20	r = 1	r = 5	r = 20
LOMO [11]	9.78	23.87	39.14	57.47	LOMO [11]	19.76	32.01	45.04	18.85	29.18	39.87	15.32	25.29	35.99
DGD [28]	17.92	50.70	67.52	79.93	DGD [28]	44.80	66.28	81.52	40.25	65.31	76.76	37.33	57.82	70.25
GoogLeNet [29]	17.81	52.12	66.79	78.77	GoogLeNet [29]	47.88	67.18	78.46	43.40	63.86	74.99	38.27	59.39	72.08
FACT [15]	18.73	51.85	67.16	79.56	FACT [15]	49.53	68.07	78.54	44.59	64.57	75.30	39.92	60.32	72.92
XVGAN [41]	24.65	60.20	77.03	88.14	XVGAN [41]	52.87	80.83	91.86	49.55	71.39	81.73	44.89	66.65	78.04
SiameseVisual [23]	29.48	41.12	60.31	79.87	VGG+CCL [13]	43.62	64.84	80.12	39.94	62.98	76.07	35.68	56.24	68.41
OIFE [26]	48.00	65.92	87.66	96.63	MixedDiff+CCL [13]	48.93	75.65	88.47	45.05	68.85	79.88	41.05	63.38	76.62
VAMI (Ours)	**50.13**	**77.03**	**90.82**	**97.16**	VAMI (Ours)	**63.12**	**83.25**	**92.40**	**52.87**	**75.12**	**83.49**	**47.34**	**70.29**	**79.95**
SiameseCNN+PathLSTM [23]	58.27	83.49	90.04	96.03	No ST information	-	-	-	-	-	-	-	-	-
SiameseVisual([23])+STR([15])	40.26	54.23	74.97	91.68		-	-	-	-	-	-	-	-	-
VAMI (Ours) + STR([15])	**61.32**	**85.92**	**91.84**	**97.70**		-	-	-	-	-	-	-	-	-

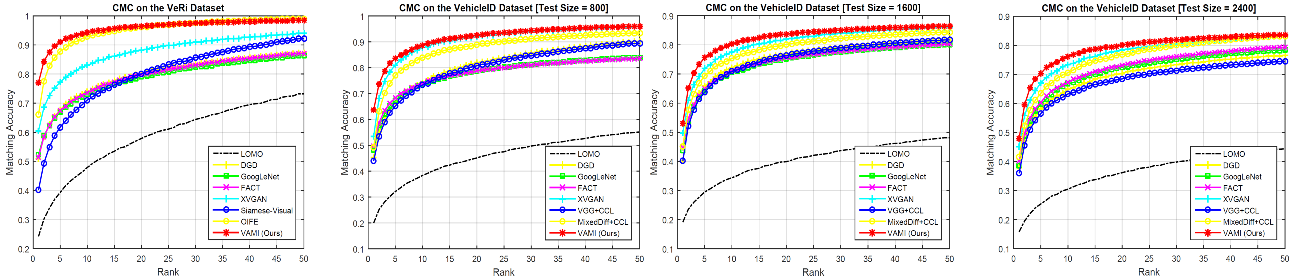


Figure 7. CMC curve comparisons of different vision-based vehicle re-ID methods.

4.4. Comparisons with State-of-the-arts

4.4.1 Evaluation on the VeRi-776 dataset

We compare the VAMI with state-of-the-art vehicle re-ID methods. LOMO [11] is a hand-crafted local feature first proposed for person re-ID. It aims to address the problem against viewpoint and illumination variations. DGD [28] is a method which can learn generic and robust deep features with data from multiple domains. We transfer the model from humans to vehicles by re-training on the VeRi and VehicleID datasets. The GoogLeNet fine-tuned on the CompCars dataset [29] is able to extract great visual descriptors containing rich semantic features for vehicles. FACT [15], consisting of SIFT, Color Name and GoogLeNet features, is proposed to discriminate vehicles in joint domains. XVGAN [41] proposes to generate cross-view images from the input view of a vehicle, then combines the original and generated views to compute distances. Siamese-Visual [23] is proposed to learn vehicles' features computing a pairwise visual similarity by classification cross-entropy loss. Moreover, OIFE [26] aims to align local region features of different viewpoints based on key points.

The CMC curves are shown in Fig. 7 and detailed mAP results are listed in Table 3. In all the vision-based methods, our VAMI achieves the best performance over the second place which also has an orientation-based region module, with 2.13% mAP increase. The key point alignment of OIFE does not work well for large viewpoint variations. The Siamese-Visual simply adopts a pairwise deep CNN for distance metric learning but does not include vehicles' semantic attributes learning. XVGAN focuses on image generation so that the re-ID results are limited by the blurred image quality and small resolution. All the other methods are hugely beaten by the VAMI. Moreover, we also combine our model with the spatial-temporal relations (STR [15]), which still outperforms other ST-based methods.

4.4.2 Evaluation on the VehicleID dataset

The VehicleID dataset consists of the training set with 110,178 images of 13,134 vehicles and the test set with 111,585 images of 13,133 vehicles. However, the dataset only contains two viewpoints: front and rear. Thus, we drop the attention model and transfer the multi-view feature inference module into a two-view version. The $Conv4$ layer of the input view is concatenated with a same size noise volume for the input of G_f. Corresponding real images of each vehicle's two viewpoints are set for the G_r in the training phase. Table 3 shows our proposed multi-view feature inference obtains dominant performance over other approaches consistently in three settings of the gallery size.

5. Conclusion

In this paper, we proposed the VAMI model to address the challenging multi-view vehicle re-ID task. The VAMI adopts a viewpoint-aware attention model and the adversarial training architecture to implement effective multi-view feature inference from single-view input. Extensive experiments show that the VAMI can achieve promising results and outperform state-of-the-art vehicle re-ID methods.

References

[1] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao. Fast person re-identification via cross-camera semantic binary transformation. In *CVPR*, 2017. 1

[2] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *The IEEE Conference on CVPR*, 2017. 1, 2

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016. 2

[4] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *The IEEE Conference on CVPR*, 2017. 2

[5] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person reidentification*, volume 1. 2014. 1

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in NIPS*, pages 2672–2680, 2014. 2

[7] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006. 3

[8] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016. 5

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on CVPR*, 2017. 2

[10] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *The IEEE Conference on CVPR*, 2017. 1

[11] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on CVPR*, pages 2197–2206, 2015. 8

[12] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *The IEEE Conference on CVPR*, 2017. 1

[13] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *IEEE Conference on CVPR*, pages 2167–2175, 2016. 2, 5, 8

[14] X. Liu, W. Liu, H. Ma, and H. Fu. Large-scale vehicle reidentification in urban surveillance videos. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pages 1–6, 2016. 2

[15] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*, pages 869–884. Springer, 2016. 1, 2, 5, 8

[16] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in NIPS*, pages 2204–2212, 2014. 2

[17] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of The International Conference on Machine Learning*, pages 2642–2651, 2017. 2

[18] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In *The IEEE ICCV*, 2017. 2

[19] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multiscale deep learning architectures for person re-identification. In *The IEEE ICCV*, 2017. 1

[20] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *In International Conference on Learning Representations*, 2016. 2

[21] Y. Rao, J. Lu, and J. Zhou. Attention-aware deep reinforcement learning for video face recognition. In *The IEEE ICCV*, 2017. 2

[22] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in NIPS*, pages 2234–2242, 2016. 5

[23] 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 *The IEEE ICCV)*, 2017. 1, 2, 8

[24] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. In *The IEEE ICCV*, 2017. 1

[25] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *The IEEE Conference on CVPR*, 2017. 2

[26] 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 reidentification. In *The IEEE ICCV*, 2017. 1, 2, 8

[27] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai. Rgb-infrared cross-modality person re-identification. In *The IEEE ICCV*, 2017. 1

[28] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on CVPR*, 2016. 1, 8

[29] L. Yang, P. Luo, C. Change Loy, and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on CVPR*, pages 3973–3981, 2015. 8

[30] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *The IEEE ICCV*, 2017. 1

[31] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *The IEEE Conference on CVPR*, 2016. 1

[32] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *The IEEE Conference on CVPR*, 2017. 1

[33] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *The IEEE ICCV*, 2017. 1

[34] F. Zheng and L. Shao. Learning cross-view binary identities for fast person re-identification. In *IJCAI*, pages 2399–2406, 2016. 1

[35] F. Zheng, Y. Tang, and L. Shao. Hetero-manifold regularisation for cross-modal hashing. *IEEE Transactions on PAMI*, 2016. 1

[36] H. Zheng, J. Fu, T. Mei, and J. Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *The IEEE ICCV*, 2017. 2

[37] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *The IEEE Conference on CVPR*, 2017. 1

[38] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *The IEEE Conference on CVPR*, 2017. 1

[39] J. Zhou, P. Yu, W. Tang, and Y. Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *The IEEE ICCV*, 2017. 1

[40] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *The IEEE Conference on CVPR*, 2017. 1

[41] Y. Zhou and L. Shao. Cross-view gan based vehicle generation for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2017. 2, 8

[42] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma. Structured attentions for visual question answering. In *The IEEE ICCV*, 2017. 2

[43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE ICCV*, 2017. 2