Cross-View GAN Based Vehicle Generation for Re-identification

Yi Zhou
y.zhou1@uea.ac.uk
Ling Shao
ling.shao@ieee.org

School of Computing Sciences
University of East Anglia
Norwich, UK

Abstract

Automatic vehicle re-identification (re-ID) is highly valuable and significant in public transportation systems, but has not achieved much progress since the visual appearances vary hugely across different viewpoints of a vehicle. Feature matching in this problem is extremely difficult, and traditional person re-ID algorithms cannot be suitably applied to vehicles. However, image generation by convolutional generative adversarial networks (GANs), which has obtained breakthrough progress, inspires us to generate vehicles in different viewpoints from only one visible view to tackle vehicle re-ID. In this work, we propose a new deep architecture, called *Cross-View Generative Adversarial Network* (XVGAN), to learn the features of vehicle images captured by cameras with disjoint views, and take the features as conditional variables to effectively infer cross-view images. Finally, the features of the original images are combined with the features of generated images in other views to learn distance metrics for vehicle re-ID. Our model can successfully generate realistic images in different views of the same vehicle, and contribute to re-ID on two public datasets: VeRi and VehicleID.

1 Introduction and Related Work

Intelligent vehicle surveillance techniques have been widely explored in the past decades. Most researches focus on vehicle detection [12, 23, 26, 35] and recognition [16, 28] tasks. However, a more interesting and challenging problem, called vehicle re-identification (re-ID), has not achieved much progress. Vehicle re-ID solves the problem of searching a target vehicle in many non-overlapping cameras in the public surveillance system. It can be applied to many practical scenarios such as urban surveillance and security. However, due to the special 3D structure, a vehicle usually looks highly different in varying viewpoints. On the contrary, two similar but different vehicles in the same viewpoint always look more similar than two different viewpoints of the same vehicle by machine vision. Thus, the cross-view vehicle matching task is considered extremely challenging.

A similar problem, called person re-ID [4], has been widely researched for a decade. As shown in Figure 1(a), images of the same human usually have a common appearance even though with large viewpoint variations. Conventional person re-ID algorithms focus on two aspects: designing robust and discriminative features [15, 24, 31, 32] and learning distance metrics [1, 8, 14, 30, 33] across different views. Moreover, recent widespread deep learning

© 2017. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

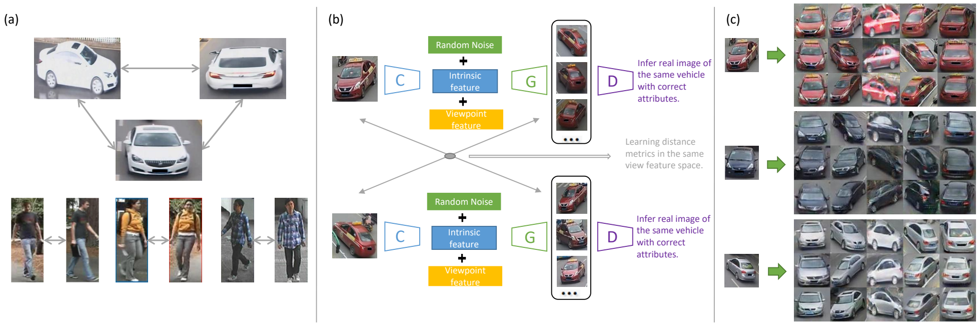

(a)
(b)
Random Noise
Intrinsic feature
Viewpoint feature
C
G
D
Infer real image of the same vehicle with correct attributes.
Learning distance metrics in the same view feature space.
Random Noise
Intrinsic feature
Viewpoint feature
C
G
D
Infer real image of the same vehicle with correct attributes.
(c)

Figure 1: **(a)** Motivation to study multi-view vehicle re-ID. The change of visual pattern across different views of a vehicle is much larger than that of a person. **(b)** Overview of the proposed XVGAN. A Classification Net is first used for learning vehicles' intrinsic features containing model, color and type information. Besides, viewpoint features are also learned. The Generative Net then takes a vehicle's intrinsic feature of the visible view, the average feature of the expected viewpoint and a random noise vector as inputs to infer images of the same vehicle in other views. The Discriminative Net distinguishes real images from synthetic samples while keeping images generated with correct vehicle attributes. Finally, the inferred vehicle images from cross-view pair data contribute to learning distance metrics for re-ID. **(c)** Generated image examples in different viewpoints for the input vehicle.

methods [13, 22, 27] successfully solve the two tasks simultaneously within one end-to-end model. However, these approaches, working well on humans, are not suitable for vehicles. Minimizing the distance between the side and front views of the same vehicle, and pushing away that of two similar vehicles both in the front view, will make a model trained without a correct and converged loss. Therefore, reasonable feature and distance metric learning should be optimized in the same manifold.

Image Generation is another hot topic which has achieved great success on many vision tasks such as text-to-image generation [21, 29], image style transformation [3, 17], and super resolution [10]. Inspired by this idea, we propose a cross-view generative adversarial network (XVGAN) in this paper, to infer vehicle images across different viewpoints and further contribute to the re-ID problem. Figure 1(b) sketches an overview of the XVGAN which comprises three sub-networks: Classification (C), Generative (G) and Discriminative (D) Nets. C Net can learn intrinsic features of the input vehicle image. Conditioned on these features as well as a pre-computed viewpoint feature of the expected view, G and D Nets aim to generate real images of the same vehicle in other viewpoints. Examples of the inferred samples by XVGAN are illustrated in Figure 1(c). After inference, the features extracted from the input and output views are concatenated for further learning distance metrics between vehicles across different views. The following two parts summarize some related works on image generation and vehicle re-ID.

Image Generation. The original GAN [5] is proposed with a deconvolutional network for generating images from noise and a convolutional network for discriminating real or fake samples. InfoGAN [2], DCGAN [20], and text-to-image GAN [21] are all excellent follow-up works to investigate better models for different tasks. Alternatively, variational autoencoders (VAE) [7] optimize the encoder and decoder networks by minimizing ℓ_2 distance between the generated images and the posterior distribution. Moreover, [25] reconstructs the unseen views only with a CNN, and [34] implements view synthesis by appearance flow.

Vehicle Re-ID. Most vehicle re-ID algorithms usually adopt license plate-based approaches [9, 11] which are only effective in the front and rear views. Besides, deep relative distance learning [16], adopting coupled clusters' loss and a mixed difference network structure, is designed for learning the difference between similar vehicles based on a new VehicleID dataset. Another work [19] also introduces a large-scale VeRi dataset, and employs visual feature, license plate and spatial-temporal information to explore the re-ID task. However, these methods consider limited viewpoints and only exploit original views.

The **main contributions** of our work are highlighted as follows: **(1)** A novel deep XVGAN architecture is proposed for generating cross-view vehicle images from an input view. Moreover, the model is extended to solve the multi-view vehicle re-ID problem. **(2)** Extensive experiments are carried out to show the superiority of the XVGAN both on the vehicle image generation quality and re-ID performance.

2 Cross-View GAN Based Vehicle Re-ID

In this section, we present the advantages of generative adversarial networks (GANs), and propose a novel deep convolutional GAN architecture for cross-view vehicle images to contribute to the vehicle re-ID task.

2.1 Generative Adversarial Nets

GAN is an unsupervised machine learning method, which achieves great success in image generation tasks. It consists of a generative model G and a discriminative model D competing against each other in a two-player min-max game. The generative network takes a latent random vector z from a uniform distribution as input to generate samples. The $p_z(z)$ is expected to converge to a target true data distribution $p_{data}(x)$, where x is a real image. Meanwhile, the discriminative network aims to distinguish the real data from synthesized samples. These two networks are simultaneously optimized by the following problem:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{x\sim p_z(z)}[log(1-D(G(z)))]. \tag{1}$$

It has been proved [6] that a global optimum can be obtained when p_G well converges to p_{data}, if G and D have enough capacity. Compared to other generative models, GAN has few restrictions (e.g. no Markov chain and variational bound is needed relative to Boltzmann machines and VAEs). Moreover, since G is very poor in the early training stage and D can easily reject synthesized samples with high confidence, $logD(G(z))$ is maximized for better training G rather than minimizing $log(1-D(G(z)))$.

2.2 XVGAN

We propose a new cross-view GAN (XVGAN), which consists of three main networks, to generate vehicles across different viewpoints. A Classification Net aims to learn vehicle features including type, color, some unique patterns and the viewpoint information. Conditioned on these features, the Generative Net takes the intrinsic vehicle features of input views, random vectors and central viewpoint features of expected views as inputs to synthesize vehicle

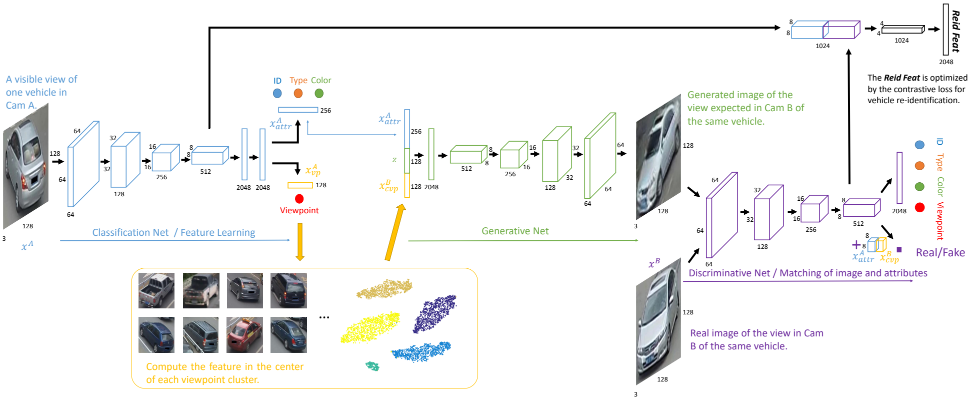

A visible view of one vehicle in Cam A.
x^A
Classification Net / Feature Learning
ID Type Color
256
x^A_{attr}
x^A_{vp}
128
Viewpoint
2048 2048
512
Compute the feature in the center of each viewpoint cluster.
...
z
x^B_{cvp}
Generative Net
Generated image of the view expected in Cam B of the same vehicle.
x^B
Real image of the view in Cam B of the same vehicle.
Discriminative Net / Matching of image and attributes
ID Type Color Viewpoint
Real/Fake
1024
Reid Feat
The ***Reid Feat*** is optimized by the contrastive loss for vehicle re-identification.

Figure 2: Network Architecture of the XVGAN. The yellow part is for learning the central features of five main viewpoints by k-means clustering. The Discriminative Net is not only for generative adversarial training, but also supervised by vehicles' multi-attributes to help the Generative Net infer real images with correct vehicle attributes in certain views. The final convolutional layers in the Classification Net and the Discriminative Net are concatenated in depth for further learning to measure distances by contrastive loss.

images in certain views of the same input vehicles. The Discriminative Net distinguishes the generated samples from the real images, and simultaneously tries to match the inferred vehicle images with correct attributes and viewpoints. Finally, the features of original views are concatenated with the features of inferred views to further learn distance metrics for re-ID. The network architecture is illustrated in Figure 2. The final learned $ReidFeat$ can be directly adopted for measuring distances between vehicles across different views.

Feature learning and viewpoint clustering. Formally, x^A denotes the input vehicle images in camera A. The trunk architecture of the Classification Net consists of 4 convolutional layers (kernel size = 5, padding = 2 and stride = 2) and 2 fully-connected layers. The Leaky-ReLU is set after each layer. Then, we configure two layers for learning the 256-dimensional x^A_{attr} by multi-attributes classification and the 128-dimensional x^A_{vp} by viewpoint classification separately, since we expect viewpoint information of view A weakened in x^A_{attr}, but strengthened in x^A_{vp}. The viewpoints of all vehicles are coarsely categorized into five groups: front, rear, side, front-side and rear-side. During training the Classification Net, the loss of viewpoint classification can be fast and well converged. Thus, we can easily learn five viewpoints' feature clusters from all the training data by k-means clustering, and compute the feature in the center of each cluster, x^B_{cvp}, as a condition to generate views in camera B.

Matching-aware conditional vehicle GAN. The conditional generator is defined as G: $\mathbb{R}^F \times \mathbb{R}^Z \times \mathbb{R}^V \rightarrow \mathbb{R}^I$, where F is the dimension of x^A_{attr}, Z is for random noise, V is for x^B_{cvp} and I is for images. Besides, the discriminator is denoted as D: $\mathbb{R}^I \rightarrow \{0,1\} \times \prod L_i$, where $i = \{1: ID,\ 2: Type,\ 3: Color,\ 4: Viewpoint\}$. L_i denotes the range of each label. The optimization of the G and D in Eq. 1 can be reformulated as:

$$\mathcal{L}_D = E_{x\sim p_{data}(x)}[logD(x)] - \sum_{i=1}^{4} log(D(x), l_i), \tag{2}$$
$$\mathcal{L}_G = E_{x\sim p_z(z);\; x^A_{attr},\; x^B_{cvp}\sim p_{data}(x^A_{attr},\; x^B_{cvp})}[log(1 - D(G(x^A_{attr}, z, x^B_{cvp})))].$$

The input of the generator G is the concatenation of the x^A_{attr}, x^B_{cvp} and a random noise prior $z \sim \mathcal{N}(0,1)$. x^A_{attr} can be regarded as an intrinsic feature learned from the original view A without much viewpoint information, while x^B_{cvp} is a central viewpoint feature in the expected view B. A fully-connected layer is set for better fusing the three vectors and then four deconvolutional layers are adopted for generating synthesized vehicle samples. The hyper-parameter settings of the Generative Net are reverse to that of the Classification Net. Moreover, batch normalization is operated on all the layers.

The discriminator D takes the generated samples and the real images in view B as inputs. The main trunk of the Discriminator has the similar structure in Classification Net. Meanwhile, to match the inferred images with the same attributes of original vehicles in view A and correct viewpoint in view B, we add a fully-connected layer and simultaneously optimize the whole Discriminative Net by multi-label classification. The viewpoint label is for view B, which is different from the viewpoint in Classification Net for view A. Batch normalization and Leaky-ReLU are adopted for all the layers in the discriminator as well. Moreover, for better optimizing the conditioned G and D, we also replicate the x^A_{attr} and x^B_{cvp} embeddings spatially and do concatenation in depth when the spatial size is 8×8, and perform a 1×1 convolution afterwards.

Distance learning for re-ID. In addition to training the XVGAN for image generation, we extend the architecture to simultaneously optimize for the vehicle re-ID problem. Define a pair of images (x^A_q, x^B_p) as positive if they are two views from the same vehicle and (x^A_q, x^B_n) as negative if they are from different vehicles. We feed forward the image pairs to our XVGAN in a Siamese-like way. Take a positive pair as an example, the mapped feature pair (f^A_q, f^B_p) from the last convolutional layer in Classification Net and the inferred $(\hat{f}^B_q, \hat{f}^A_p)$ from the last convolutional layer in Discriminative Net are concatenated as $(f_q = f^A_q + \hat{f}^B_q, f_p = f^B_p + \hat{f}^A_p)$. Conversely, (f_q, f_n) is for negative pairs. Then, f_q, f_p and f_n, including both the features from the original images and the inferred samples, can be further adopted for learning distance metrics by minimizing a contrastive loss $\mathcal{L}_{reid}$ [6] to shorten the distance between the same vehicle and maximize that between different vehicles. Moreover, a convolutional layer (kernel size = 3, padding = 1 and stride = 2) and a fully-connected layer are configured at the end. Our distance metric learning is more reasonable since it is optimized in the same feature manifold by combining the original and generated views.

In the inference phase, the Classification Net first predicts viewpoints of the query vehicle and each candidate in the gallery set. Then, the corresponding central viewpoint features for each other as well as the intrinsic features are adopted to generate cross-view images. Finally, the 2048-dimensional *ReidFeat* is used to measure distance for ranking.

Implementation Settings. The random noise z is set as 128-dimensional, sampled from uniform distribution. The spatial size of generated images is 128×128. Similar to [20], we adopt the ADAM Optimizer with the learning rate of 0.0002 and the momentum of 0.5. We set the mini-batch size as 64, and trained our model for 500 epochs on a GPU server configured with four GTX TITAN X cards.

3 Experiments

In this section, we first qualitatively evaluate the generation ability of the XVGAN compared to three baselines. Moreover, we explore the performance on vehicle re-ID compared to some state-of-the-arts on two public vehicle re-ID datasets. All the experiments are implemented based on the deep learning framework TensorFlow [20].

3.1 Datasets and Evaluation Protocol

To well explore our proposed cross-view GAN on vehicle generation for re-ID, we evaluate the XVGAN on a large-scale real vehicle surveillance dataset: VeRi [19]. The VeRi dataset, containing 776 different vehicles, is collected by 20 cameras along a circular road within a city area. The training set consists of 576 vehicles with 37,781 images and the test set has 200 vehicles with 11,579 images. We adopt the ID, color, type and viewpoint information of training data to optimize our model. The viewpoint labels followed in [28] are labeled by ourselves. Moreover, since the license plate and spatial-temporal relation annotations have not been released, we do not include them in our experiments.

To optimize the contrastive loss for re-ID, we randomly generate 30,000 positive pairs and 70,000 negative pairs for training. In the test phase, we strictly follow the evaluation protocol proposed in [19]. An image-to-track search is adopted instead of conventional image-to-image fashion. The experiments are evaluated on 1,678 query images and 2,021 gallery tracks. In addition to the Cumulative Matching Characteristic (CMC) curve widely employed in person Re-Id, a mean average precision (mAP) is also used.

3.2 Cross View Generation with Correct Attributes and Viewpoints

Before evaluating the final re-ID performance, we first present the vehicle image generation results by our XVGAN compared to some baselines. To validate the effectiveness of each designed component in the XVGAN, we carefully evaluate three corresponding baselines explained in detail as follows.

VAE Architecture. To explore that GAN based framework can better recover the training distribution and be asymptotically consistent, we first set the Variational Auto-Encoder (VAE) architecture as a baseline. In this method, the Discriminative Net is discarded, and the Classification Net and Generative Net work as the encoder and decoder, respectively. All the convolutional, fully-connected and deconvolutional layers have exactly same settings with the XVGAN for a fair comparison. In addition to the KL divergence, we adopt ℓ_2 norm for optimizing the decoder loss.

Purely Attributes-conditioned GAN. Conventional GANs usually condition on class labels or semantic attributes. However, more descriptive intrinsic visual features are not taken advantage of if the Generative Net only takes noise and some discrete values of attributes as inputs. To validate the superiority of our cross-view image-conditioned GAN, we evaluate a baseline of GAN purely conditioned on discrete attributes. In this method, after predicting the attribute labels of a vehicle in an original view image by the Classification Net, the ID, type, color and the expected viewpoint labels are directly encoded to a 384-dimensional (256+128) vector and then concatenated with the random noise vector. The structures of the Generative and Discriminative Nets are same as that of the XVGAN.

XVGAN without Matching-awareness. The goal of XVGAN is not only to generate real vehicle images, but also to infer images with correct attributes and viewpoints. A gen-

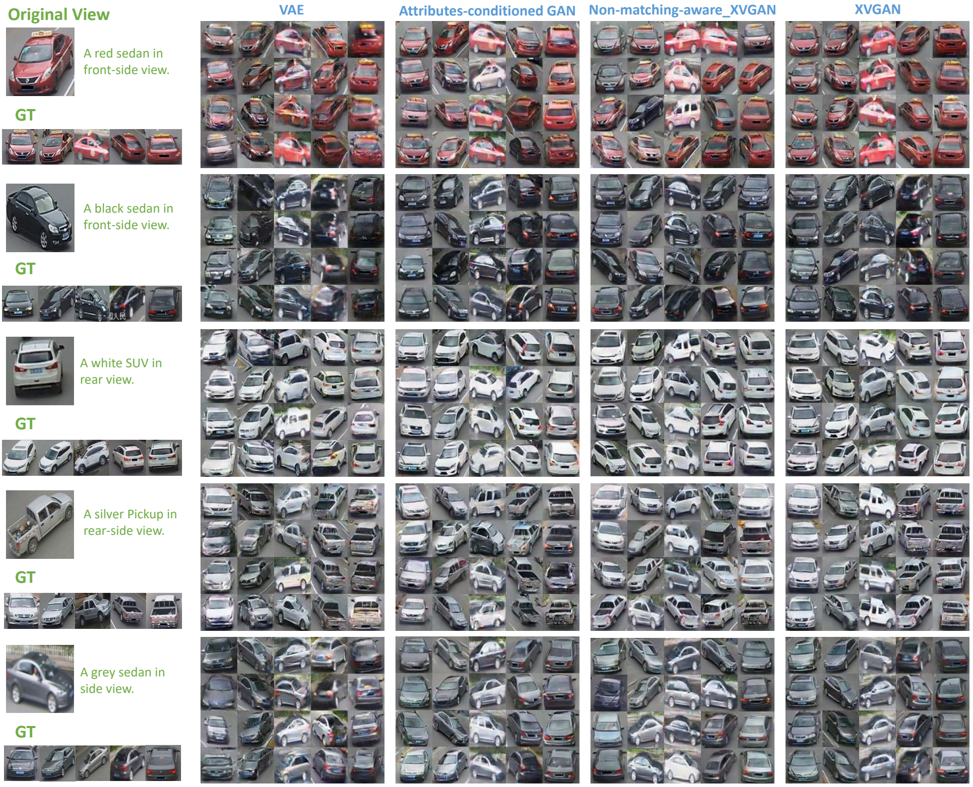

Original View
A red sedan in front-side view.
GT
A black sedan in front-side view.
GT
A white SUV in rear view.
GT
A silver Pickup in rear-side view.
GT
A grey sedan in side view.
GT
VAE
Attributes-conditioned GAN
Non-matching-aware_XVGAN
XVGAN

Figure 3: Qualitative examples of generated vehicle images in different viewpoints from only one original visible view. The left column shows the original vehicle images and their corresponding groundtruths in five viewpoints. The right four columns show generated samples by VAE, Attributes-conditioned GAN, XVGAN without matching-awareness and XVGAN.

erated image with mismatched vehicle attributes or viewpoints cannot contribute to the final re-ID performance. The multi-attributes learning of generated views is configured for the constraint in the Discriminative Net. To prove the effectiveness of this design, an ablation experiment of dropping the multi-attributes learning is conducted.

Figure 3 demonstrates some qualitative examples of generated vehicle images by our XVGAN compared to three baselines. According to the results, we have the following observations and analysis. First, the VAE architecture generated much more blurred images compared to GAN-based frameworks, even though most correct attributes and viewpoints can be successfully generated. For the attributes-conditioned GAN, we find that the intra-variations within the synthesized same color and type vehicles are large, since the detailed intrinsic features of the visible original view images are not exploited. In other words, from an input view image, the attributes-conditioned GAN can generate many similar vehicles rather than exactly the same target, thus, its improvement for the re-ID task is also limited. Moreover, the XVGAN without the matching-awareness constraint can generate real vehicle images, however, some attributes and viewpoints of the inferred views frequently mismatch the original vehicle. Finally, our XVGAN can synthesize highly real vehicle images with correct attributes and viewpoints in most cases. Its effectiveness for vehicle re-ID is further investigated in the next two sections.

3.3 Re-identification on the VeRi Dataset

To evaluate the re-ID performance, in addition to the image generation based models, we also compare to five traditional one-view based methods which only exploit the features of original views to measure distances across different views. One is simply adopting the second fully-connected layer in the Classification Net of XVGAN. The LOMO [15] feature is a highly successful handcrafted feature adopted for person re-ID. Moreover, the deep person re-ID model of domain guided dropout (DGD) [27] is also transferred to vehicles by re-training on [16, 19]. The GoogLeNet feature extracted from the model fine-tuned on vehicles in [28], is a solid deep representation containing rich semantic vehicle attributes information. A weighted combination of SIFT, Color Name and GoogLeNet features, proposed as FACT in [18, 19], can well discriminate vehicles in joint domains. Besides, since the VAE baseline does not have the Discriminative Net, we concatenate the last convolution layer of the encoder and the first convolution layer of the decoder instead, and then learn the *ReidFeat* by the contrastive loss. Furthermore, we also compare to a view synthesis method by appearance flow (AppFlow) [34]. Since only one visible input view is available in the vehicle re-ID test phase, we adopt the single-input view network of AppFlow.

One-view based Re-ID						Multi-view based Re-ID					
Methods	mAP	Top-1	Top-5	Top-20	Top-50	Methods	mAP	Top-1	Top-5	Top-20	Top-50
LOMO	9.03	23.89	40.32	58.61	73.96	AppFlow	18.91	53.09	69.78	80.56	88.93
GoogLeNet	17.58	51.98	66.79	78.77	86.37	VAE	16.68	46.52	61.41	74.94	84.25
FACT	18.54	52.35	67.16	79.97	87.09	Attr-GAN	17.53	52.75	66.13	78.16	86.36
XVGAN-C	18.42	51.14	64.50	79.42	91.51	XVGAN-w/o-M	20.07	55.21	69.85	82.87	91.76
DGD	17.96	50.51	68.86	80.05	87.62	**XVGAN**	**24.65**	**60.20**	**77.03**	**88.14**	**93.95**

Table 1: mAP and matching rate (%) at rank-1, 5, 20 and 50 on VeRi Dataset. Attr-GAN denotes Attribute-conditioned GAN, and w/o-M is abbreviation of without Matching-awareness. XVGAN-C only adopts the feature in the Classification Net.

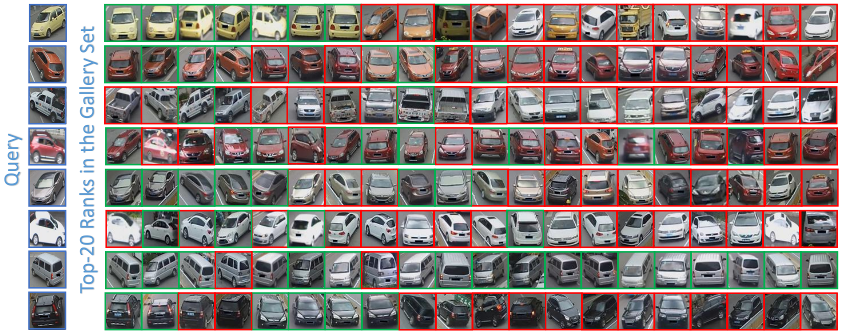
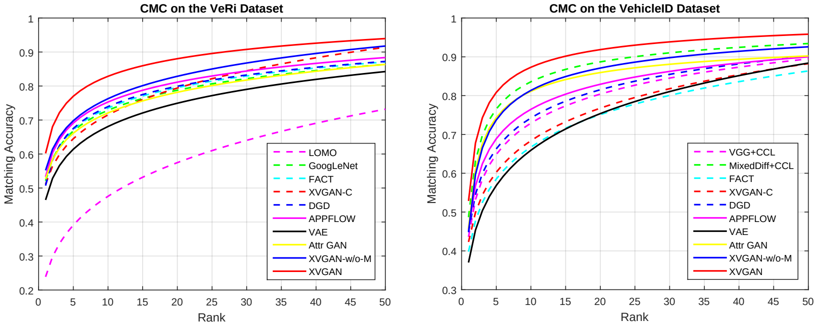


Figure 4: Qualitative success and failure examples of top-20 rank on the VeRi dataset. Green boxes mean correct hits, while red boxes denote wrong ones.

Table 1 illustrates the mAP and matching rate results by different methods. The FACT feature achieves the highest mAP of 18.54% among the five one-view based methods. How-



(a) (b)

Figure 5: CMC results evaluated on two datasets. The solid lines denote image generation models, while the dashed lines are methods only exploiting the original one view.

ever, our XVGAN beats FACT by 6.11%. The improvement shows the effectiveness of the cross-view generation can indeed contribute to the vehicle re-ID problem in disjoint views. Moreover, without the design of matching-aware multi-label supervision in the Discriminative Net, the mAP of XVGAN decreases by 4.58%. Also, neither of VAE-based generation model and purely attribute-conditioned GAN achieves satisfactory performance. Thus, each component of design in the XVGAN is proved to be significant for re-ID. The matching rates of XVGAN at top-1, 5, 20, 50 are consistently higher than those of other baselines. The detailed comparison of CMC curves is shown in Figure 5(a). Besides, Figure 4 demonstrates qualitative examples of top-20 ranks for some query vehicles. We can observe that images of the same vehicle with large viewpoint variations compared to the query one can be successfully distinguished from many similar candidates in most cases. However, some false hits still exist usually caused by homogeneous visual patterns from very similar candidates in the same viewpoint.

3.4 Re-identification on the VehicleID Dataset

To make our method more convincing and show its generalizability, we evaluate the XVGAN on another large-scale vehicle dataset: VehicleID [16]. All the vehicles in the VehicleID dataset are captured in up to only two viewpoints: front and rear. The dataset is divided into the training set with 110,178 images of 13,134 vehicles and the test set with 111,585 images of 13,133 vehicles. A coupled clusters loss (CCL) and a mixed difference network structure (Mixed Diff) for vehicle re-ID are also introduced in [16]. Following its evaluation protocol, we conduct the image-to-image search. One image is randomly selected for each vehicle in the test set to construct the gallery set with the size of 800. Other images are adopted as query ones. The experiment is carried out 10 times to obtain the final results.

One-view based Re-ID					Multi-view based Re-ID				
Methods	Top-1	Top-5	Top-20	Top-50	Methods	Top-1	Top-5	Top-20	Top-50
VGG+CCL	43.62	64.84	80.12	89.29	AppFlow	45.52	69.45	82.02	90.15
Mixed Diff+CCL	48.93	75.65	88.47	93.37	VAE	37.62	56.84	75.36	88.38
FACT	39.85	58.47	74.98	86.36	Attr-GAN	45.83	74.33	85.94	90.23
XVGAN-C	42.31	60.26	76.77	88.24	XVGAN-w/o-M	44.91	73.48	87.04	92.58
DGD	44.72	66.68	81.35	90.31	**XVGAN**	**52.89**	**80.84**	**91.86**	**95.83**

Table 2: Matching accuracies (%) at rank-1, 5, 20 and 50 on the two-view VehicleID Dataset.

CMC curves are illustrated in Figure 5(b). As shown in Table 2, XVGAN increases the top-1 and 5 matching rates by 3.96% and 5.19%, respectively, compared to the second place Mixed Diff+CCL. The FACT feature performs poorly on this dataset, since neither of its components can be discriminative for small inter-variations between vehicles in the single viewpoint. The large margin between XVGAN and XVGAN-C strongly proves the significance of the contribution by the cross-view inference. Moreover, VAE also gets low accuracies because the generated cross-view images are blurred, losing details. The top-1 rate by Attr-GAN or XVGAN-w/o-M is 7.06% or 7.98% lower than that of XVGAN, respectively. Therefore, GAN models, conditioned on only discrete attribute labels or without matching-awareness design, are ineffective for vehicle re-ID.

4 Conclusion

In this paper, we proposed a novel deep XVGAN to implement cross-view vehicle generation and contribute to the multi-view vehicle re-ID task neglected by the computer vision community. Extensive experimental results showed that our model could both achieve satisfactory performance on matching-aware vehicle image generation and re-ID compared to some baselines. In future work, more solid models are being studied to transfer this framework to practical applications.

Acknowledgment

This work is collaboratively funded by Createc, Cockermouth, UK and University of East Anglia, Norwich, UK.

References

[1] Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, and Ling Shao. Fast person re-identification via cross-camera semantic binary transformation. In *CVPR*, 2017.

[2] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in NIPS*, pages 2172–2180, 2016.

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on CVPR*, pages 2414–2423, 2016.

[4] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person re-identification*, volume 1. 2014.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in NIPS*, pages 2672–2680, 2014.

[6] Raia Hadsell, Sumit Chopra, and Yann 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.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *In ICLR*, 2014.

[8] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[9] Mahmood Ashoori Lalimi, Sedigheh Ghofrani, and Des McLernon. A vehicle license plate detection method using region and edge based methods. *Computers & Electrical Engineering*, 39:834–845, 2013.

[10] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016.

[11] RT Lee, KC Hung, and HS Wang. Real time vehicle license plate recognition based on 2d haar discrete wavelet transform. *International Journal of Scientific & Engineering Research*, 3:1–6, 2012.

[12] Bo Li, Tianfu Wu, and Song-Chun Zhu. Integrating context and occlusion for car detection by hierarchical and-or model. In *ECCV*, pages 652–667, 2014.

[13] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[14] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[15] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on CVPR*, pages 2197–2206, 2015.

[16] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *Proceedings of the IEEE Conference on CVPR*, pages 2167–2175, 2016.

[17] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in NIPS*, pages 469–477, 2016.

[18] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu. Large-scale vehicle re-identification in urban surveillance videos. In *Multimedia and Expo (ICME), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

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

[20] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *In ICLR*, 2016.

[21] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd ICML*, volume 3, 2016.

[22] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *ECCV*, pages 475–491, 2016.

[23] Zehang Sun, George Bebis, and Ronald Miller. On-road vehicle detection: A review. *IEEE TPAMI*, 28(5):694–711, 2006.

[24] Shoubiao Tan, Feng Zheng, Li Liu, Jungong Han, and Ling Shao. Dense invariant feature based support vector ranking for cross-camera person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[25] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Single-view to multi-view: Reconstructing unseen views with a convolutional network. In *ECCV*, pages 322–337, 2016.

[26] Xuezhi Wen, Ling Shao, Wei Fang, and Yu Xue. Efficient feature selection and classification for vehicle detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[27] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249–1258, 2016.

[28] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on CVPR*, pages 3973–3981, 2015.

[29] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stack gan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016.

[30] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. *CVPR*, 2016.

[31] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151, 2014.

[32] Feng Zheng and Ling Shao. Learning cross-view binary identities for fast person re-identification. In *IJCAI*, pages 2399–2406, 2016.

[33] Feng Zheng, Yi Tang, and Ling Shao. Hetero-manifold regularisation for cross-modal hashing. *IEEE Transactions on PAMI*, 2016.

[34] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, pages 286–301, 2016.

[35] Yi Zhou, Li Liu, Ling Shao, and Matt Mellor. Dave: a unified framework for fast vehicle detection and annotation. In *ECCV*, pages 278–293, 2016.