CamStyle: A Novel Data Augmentation Method for Person Re-Identification

Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, *Senior Member, IEEE*, and Yi Yang

***Abstract*— Person re-identification (re-ID) is a cross-camera retrieval task that suffers from image style variations caused by different cameras. The art implicitly addresses this problem by learning a camera-invariant descriptor subspace. In this paper, we explicitly consider this challenge by introducing camera style (CamStyle). CamStyle can serve as a data augmentation approach that reduces the risk of deep network overfitting and that smooths the CamStyle disparities. Specifically, with a style transfer model, labeled training images can be style transferred to each camera, and along with the original training samples, form the augmented training set. This method, while increasing data diversity against overfitting, also incurs a considerable level of noise. In the effort to alleviate the impact of noise, the label smooth regularization (LSR) is adopted. The vanilla version of our method (without LSR) performs reasonably well on few camera systems in which overfitting often occurs. With LSR, we demonstrate consistent improvement in all systems regardless of the extent of overfitting. We also report competitive accuracy compared with the state of the art on Market-1501 and DukeMTMC-re-ID. Importantly, CamStyle can be employed to the challenging problems of one view learning and unsupervised domain adaptation (UDA) in person re-identification (re-ID), both of which have critical research and application significance. The former only has labeled data in one camera view and the latter only has labeled data in the source domain. Experimental results show that CamStyle significantly improves the performance of the baseline in the two problems. Specially, for UDA, CamStyle achieves state-of-the-art accuracy based on a baseline deep re-ID model on Market-1501 and DukeMTMC-reID. Our code is available at: https://github.com/zhunzhong07/CamStyle.**

Manuscript received April 22, 2018; revised August 26, 2018 and September 28, 2018; accepted September 28, 2018. Date of publication October 8, 2018; date of current version November 2, 2018. This work was supported in part by the National Nature Science Foundation of China under Grants 61572409, 61876159, 61806172, U1705286, and 61571188, in part by the Fujian Province 2011 Collaborative Innovation Center of TCM Health Management, in part by the Collaborative Innovation Center of Chinese Oolong Tea Industry-Collaborative Innovation Center (2011) of Fujian Province, in part by the Fund for Integration of Cloud Computing and Big Data, in part by the Innovation of Science and Education, in part by the Data to Decisions CRC (D2D CRC), and in part by the Cooperative Research Centre Programme. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Dong Xu. *(Corresponding author: Shaozi Li.)*

Z. Zhong is with the Cognitive Science Department, Xiamen University, Xiamen 361005, China, and also with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: zhunzhong@stu.xmu.edu.cn).

L. Zheng is with the Research School of Computer Science, The Australian National University, Canberra, ACT 0200, Australia (e-mail: liangzheng06@gmail.com).

Z. Zheng and Y. Yang are with the Centre for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: zdzheng12@gmail.com; yee.i.yang@gmail.com).

S. Li is with the Cognitive Science Department, Xiamen University, Xiamen 361005, China (e-mail: szlig@xmu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2018.2874313

***Index Terms*— Person re-identification, CamStyle, one-view learning, unsupervised domain adaptation.**

I. Introduction

PERSON re-identification (re-ID) [1]–[3] is a cross-camera retrieval task. Given a query person-of-interest, it aims to retrieve the same person from a database collected from multiple non-overlapping cameras. In this task, a person image often undergoes intensive changes in appearance and background. Capturing images by different cameras is a primary cause of such variations (Fig. 1). Usually, cameras differ from each other regarding resolution, environment illumination, *etc*.

In addressing the challenge of camera variations, a previous body of the literature chooses an implicit strategy. That is, to learn stable feature representations that have invariance property under different cameras. Examples in traditional approaches include KISSME [4], XQDA [5], DNS [6], *etc*. Examples in deep representation learning methods include IDE [1], SVDNet [7], TripletNet [8], *etc*.

Compared to previous methods, this paper resorts to an explicit strategy from the view of camera style data augmentation. We are mostly motivated by the need for large data volume in deep learning based person re-ID. To learn rich features which are robust to camera variations, annotating large-scale datasets is useful but prohibitively expensive. Nevertheless, if we can add more samples to the training set that are aware of the style differences between cameras, we are able to 1) address the data scarcity problem in person re-ID and 2) learn invariant features across different cameras. Preferably, this process should not cost any more human labeling, so that we can keep costs low.

Based on the above discussions, we propose a camera style (CamStyle) data augmentation method to regularize CNN training for person re-ID. In its *vanilla version*, we learn image-image translation models for each camera pair with CycleGAN [10]. With the learned CycleGAN model, for a training image captured by a certain camera, we can generate new training samples in the style of other cameras (Fig. 1 and Fig. 2). In this manner, the training set is a combination of the original training images and the style-transferred images. The style-transferred images can directly borrow the label from the original training images. During training, we use the new training set for re-ID CNN training following the baseline model in [1]. The vanilla method is beneficial in reducing over-fitting and achieving camera-invariant property, but, importantly, we find that it also introduces noise to the system (Fig. 2). This problem deteriorates its benefit under

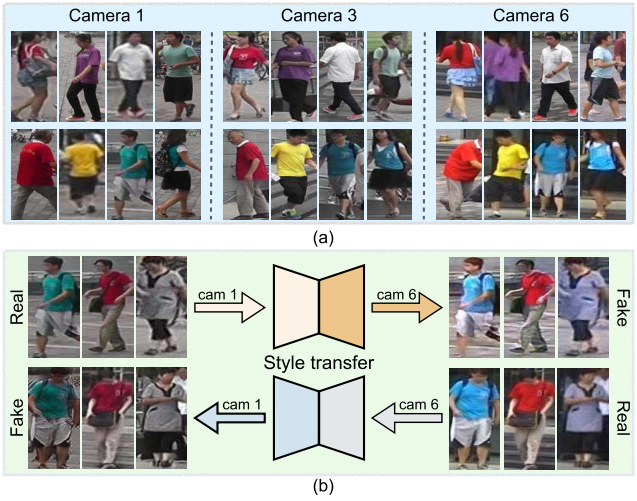


Fig. 1. (a) Example images from Market-1501 [9]. (b) Examples of camera-aware style transfer between two cameras (camera 1 and camera 6) using our method. Images in the same column represent the same person.

full-camera systems where the relatively abundant data has a lower over-fitting risk. To mitigate this problem, in the *improved version*, we further apply label smoothing regularization (LSR) [11] on the style-transferred samples, so that their labels are softly distributed during training.

The proposed approach, CamStyle, has three advantages. First, it can be regarded as a data augmentation scheme that not only smooths the camera style disparities, but also reduces the impact of CNN over-fitting. Second, by incorporating camera information, it helps learn pedestrian descriptors with the camera-invariant property. Finally, it is unsupervised, guaranteed by CycleGAN, indicating fair application potentials.

Apart from the normal person re-identification settings, this paper further identifies two important tasks which benefit from the proposed CamStyle method, *i.e.* one view (camera) learning and unsupervised domain adaptation. On the one hand, in one view learning, labeled training images are only available from one camera view, and the identities of images in other cameras are unknown. Therefore, this setting provides a very limited number of labeled images for training deep models, and the model trained on images collected from one camera may suffer from image style variations caused by other cameras. To address this problem, we employ CamStyle to generate new training images whose styles are similar to the unlabeled cameras. In this way, CamStyle produces more training data for training the re-ID model, reduces the risk of over-fitting and improves the camera-invariance property of re-ID models.

On the other hand, re-ID models trained on labeled dataset often fail to perform well on an unseen testing set. This is a domain adaptation task where we are provided with a fully labeled source dataset named as source domain. We are also provided an unlabeled target dataset which is named as target domain. In domain adaptation [12]–[15], training is conducted using both the labeled source domain data and the unlabeled target domain data. Then, testing is conducted on the target domain. Recent works mainly aim at reducing the gap between the source and target domains on intermediate feature-level [16] or on image-level [17], [18]. However, these methods only take into consideration the inter-domain variations, and overlook the intra-domain variations. The inter-domain variations refer to the variations between source domain and target domain, *e.g.*, the differences in illumination, background, clothing style between the two domains. The intra-domain variations refer to the variations within the same domain. In our paper, the target domain is composed of several subdomains, corresponding to different cameras. We find that the images captured by different cameras (subdomains) are different in their style. In fact, the image style variation between different cameras is a critical influencing factor in person re-ID. This is because in the process of testing, we attempt to find out true matched persons captured by different cameras from a given query person. Without considering the image style variations caused by target cameras, a domain adaptation model trained on the source set may only capture the overall data bias between the source and target domains and may encounter difficulties when the image style of the target cameras changes greatly. To solve this problem, CamStyle is first applied to learn camera style transfer models from source domain to each target camera. Then, each labeled source images can be style transferred to target cameras, and the generated images are utilized to train re-ID models for target domain. In this manner, our method not only reduces the gap between the source domain and target domain, but also considers the intra-domain variations in target domain.

In summary, the contributions of this work are featured in the following aspects.

- We introduce a vanilla camera-aware style transfer model for re-ID data augmentation. In few-camera systems, the improvement can be as large as 17.1%.
- We propose an improved method through applying LSR on the style-transferred samples to softly distribute their labels during re-ID training. In full-camera systems, consistent improvement is observed.
- Compared to [19], a more detailed description is presented in "Introduction" and "Related Work" sections.
- We further expand the experiments to compare and discuss the proposed method with more data augmentation techniques, and, analyze the parameters of the proposed method in more detail.
- Importantly, we extend the current system to the domain adaptation task. CamStyle can be effectively applied to this task and achieves state-of-the-art performance, which demonstrates the extendibility and effectiveness of CamStyle in domain adaptation.
- We are the first to propose the setting of one view learning in the community of person re-ID, which is very important in practice. One view learning is an open set problem in which the unlabeled views may contain classes that might be different from the labeled view. We show that the CamStyle is very useful in this setting.

II. Related Work

A. Deep Learning Person Re-Identification

Person re-ID can be regarded as an image retrieval [20]–[23] task. Benefit from the strong ability of deep

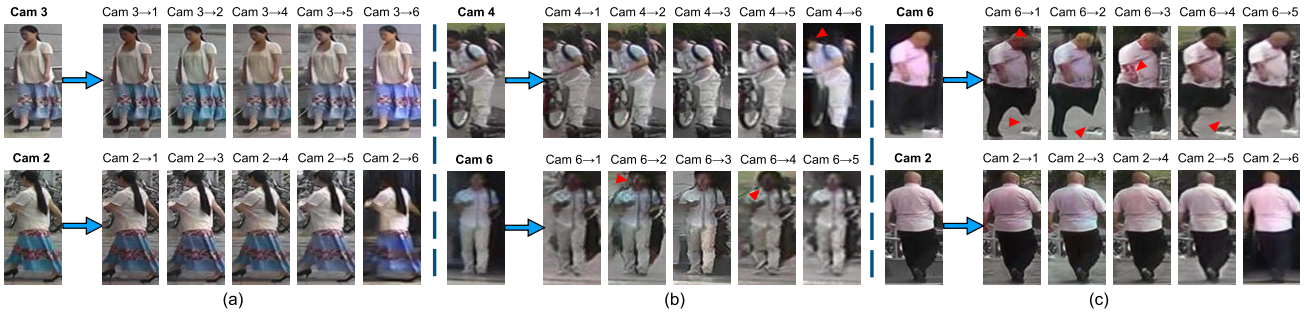


Fig. 2. Examples of style-transferred samples in Market-1501 [9]. An image captured in a certain camera is translated to styles in other 5 cameras. Despite the success cases, image-image translation noise indicated by red arrows should be considered.

learning [24]–[30], many deep learning methods [31]–[38] have been proposed in person re-ID. In [31], input image pairs are partitioned into three overlapping horizontal parts respectively, and through a siamese CNN model to learn the similarity of them using cosine distance. Later, Wu *et al.* [32] increase the depth of networks with using smaller convolution filters to obtain a robust feature. In addition, Varior *et al.* [34] merge long short-term memory (LSTM) model into a siamese network that can handle image parts sequentially so that the spatial information can be memorized to enhance the discriminative capability of the deep features.

Another effective strategy is the classification model, which makes full use of the re-ID labels [1], [7], [39]–[42]. Zheng *et al.* [1] propose the ID-discriminative embedding (IDE) to train the re-ID model as image classification which is fine-tuned from the ImageNet [43] pre-trained models. Wu *et al.* [39] propose a Feature Fusion Net (FFN) by incorporating hand-crafted features into CNN features. Recently, Sun *et al.* [7] iteratively optimize the fully connected (FC) feature with Singular Vector Decomposition and produce orthogonal weights.

B. Data Augmentation in Person Re-Identification

When a CNN model is excessively complex compared to the number of training samples, over-fitting might happen. To address this problem, many regularization methods and data augmentation methods have been proposed in the community of deep learning, such as Dropout [44] and Batch Norm [45] for regularization, and, various transformations including cropping, flipping and translation for data augmentation. Dropout is widely utilized in various recognition tasks. It randomly abandons (assigning to zero) the output of each hidden neuron with a probability in the training stage and only employs the contribution of the remaining weights in forward pass and back-propagation. Recently, several methods aim to address the over-fitting problem in person re-ID. McLaughlin *et al.* [46] improve the generalization of network by utilizing background and linear transformations to generate various samples. Recently, Zhong *et al.* [47] randomly erase a rectangle region in input image with random values which prevents the model from over-fitting and makes the model robust to occlusion. Similarity, Huang *et al.* [48] propose to augment the training data with adversarially occluded samples. The hard occluded samples are selected by a pre-trained model to further optimize the re-ID model. More related to this work, Zheng *et al.* [49] use DCGAN [50] to generate unlabeled samples, and assign them with a uniform label distribution to regularize the network. In contrast to [49], the style-transferred samples in this work are produced from real data with relatively reliable labels. In the view of pose translation, a pose-transferable framework [51] is proposed to generate novel samples with rich pose variations. The novel samples are combined with the original training samples to enhance the re-ID model.

C. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [52] have achieved impressive success in recent years, especially in image generation [50]. Recently, GANs have also been applied to image-to-image translation [10], [53]–[56], style transfer [57]–[59] and cross domain image generation [17], [60], [61]. Isola *et al.* [53] apply a conditional GANs to learn a mapping from input to output images for image-to-image translation application. The main drawback of [53] is that it requires pairs of corresponding images as training data. To overcome this problem, Liu and Tuzel [54] propose a coupled generative adversarial network (CoGAN) by employing weight-sharing networks to learn a joint distribution across domains. More recently, CycleGAN [10] introduces cycle consistency based on "pix2pix" framework in [53] to learn the image translation between two different domains without paired samples. Style transfer and cross domain image generation can also be regarded as image-to-image translation, in which the style (or domain) of input image is transferred to another while remaining the original image content. In [57], a style transfer method is introduced by separating and recombining the content and style of images. Bousmalis *et al.* [60] introduce an unsupervised GAN framework that transfers images from source domain to an analog image in target domain. Similarity, in [61], the Domain Transfer Network (DTN) is proposed by incorporating multiclass GAN loss to generate images of unseen domain, while reserving original identity. Unlike previous methods which mainly consider the quality of the generated samples, this work aims at using the style-transferred samples to improve the performance of re-ID.

D. Unsupervised Domain Adaptation

Although many researchers make efforts to normal person re-identification settings, few works [16]–[18], [62]–[64] have studied on unsupervised domain adaptation for re-ID. Peng *et al.* [62] propose to learn a discriminative representation for target domain based on asymmetric multi-task dictionary learning. Deng *et al.* [17] learn a similarity preserving generative adversarial network based on CycleGAN [10] to translate images from source domain to target domain. The translated images are utilized to train re-ID models in a supervised manner. In [16], a transferable model is proposed to jointly learn attribute-semantic and identity discriminative feature representation for target domain. These methods attempt to reduce the divergence between source domain and target domain on either the image space [17], [18] or feature space [16], [62], [63], but overlook the image style variations caused by different cameras in target domain. In this work, we explicitly consider the intra-domain image variations caused by target cameras for learning discriminative representations of target domain. HHL [64] proposes to learn camera invariance via positive pairs formed by unlabeled target samples and their camera style transferred counterparts. Differ from HHL, this work transfers source data to styles of the target cameras and directly learns the deep re-ID model with labeled transferred samples in a supervised way.

E. Camera-Specific Transfer-Based Methods

Our approach is also related to camera transfer modeling methods. To capture the appearance variations between cameras, earlier works focus on modeling the color transfer between two overlapping cameras [65]–[68]. Porikili [65] proposes to learn Brightness Transfer Function (BTF) for estimating the colors variations of a person from one camera to another. The transformation of colors is deduced by computing a correlation matrix between two histograms of persons that captured from the same scene region at the same time. Later, Javed *et al.* [66], [67] extend the BTF for non-overlapping cameras where the appearance variations are also caused by illumination and pose changes. Rather than evaluating individual histograms for each person, the CBTF approach [69] (Cumulative Brightness Transfer Function) proposes to accumulate the pixels from the whole training samples. A main shortcoming of the above BTF-based methods is that they attempt to learn the camera transfer by a single function. In fact, there may be multi-mappings for transferring a image from one camera to another. To address this shortcoming, the Implicit Camera Transfer (ICT) [70] approach is presented to learn camera transfer by a binary relation, which allows camera transfer function to be a multi-valued mapping. Later, ECT (Explicit Camera Transfer) [71] introduces to model camera appearance transfer through a single function, as well as leveraging the intra-camera samples to model appearance variations. All of the above methods required labeled training pairs for modeling camera transfer models and are applied to the small-scale datasets. This work proposes an unsupervised approach for learning camera transfer model and shows that our method is suitable for the scenario of large-scale person re-identification.

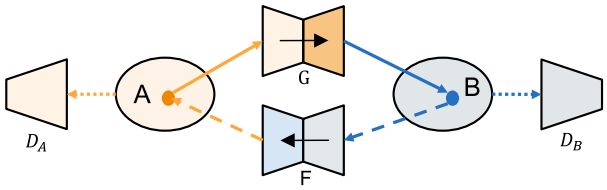


Fig. 3. The CycleGAN model includes two generators G and F, and two adversarial discriminators D_A and D_B. D_B enforces G to transfer images from A into outputs indistinguishable from domain B, and vice versa for D_A and F.

III. The Proposed Method

In this section, we first briefly look back at the CycleGAN [10] in Section III-A. We then describe the camera-aware data generation process using CycleGAN in Section III-B. The baseline and the training strategy with LSR are described in Section III-C and Section III-D, respectively.

A. CycleGAN Review

Given two datasets $\{x_i\}_{i=1}^{\mathcal{M}}$ and $\{z_j\}_{j=1}^{\mathcal{N}}$, collected from two different domains A and B, where $x_i \in A$ and $z_j \in B$, The goal of CycleGAN is to learn a mapping function $G : A \rightarrow B$ such that the distribution of images from $G(A)$ is indistinguishable from the distribution B using an adversarial loss. CycleGAN contains two mapping functions $G : A \rightarrow B$ and $F : B \rightarrow A$. Two adversarial discriminators D_A and D_B are proposed to distinguish whether images are translated from another domain. CycleGAN applies the GAN framework to jointly train the generative and discriminative models. The overall CycleGAN loss function is expressed as:

$$\begin{aligned} V(G, F, D_A, D_B) = {} & V_{GAN}(D_B, G, A, B) \\ & + V_{GAN}(D_A, F, B, A) \\ & + \lambda V_{cyc}(G, F), \end{aligned} \tag{1}$$

where $V_{GAN}(D_B, G, A, B)$ and $V_{GAN}(D_A, F, B, A)$ are the loss functions for the mapping functions G and F and for the discriminators D_B and D_A. $V_{cyc}(G, F)$ is the *cycle consistency loss* that forces $F(G(x)) \approx x$ and $G(F(z)) \approx z$, in which each image can be reconstructed after a cycle mapping. λ penalizes the importance between V_{GAN} and V_{cyc}. In the training of CycleGAN, we alternatively train the generators and discriminators and aim to optimize:

$$G^*, F^* = \arg \max_{D_A, D_B} \min_{G, F} V(G, F, D_A, D_B) \tag{2}$$

The overview of CylceGAN is shown in Fig. 3.

In the testing stage, we are provided with two translation models G and F. Given an input image from domain A, we apply G to transfer it to the style of domain B, and vice versa for input image from domain B and F. In this way, we could generate images in the style of domain B (or A) that are transferred from domain A (or B).

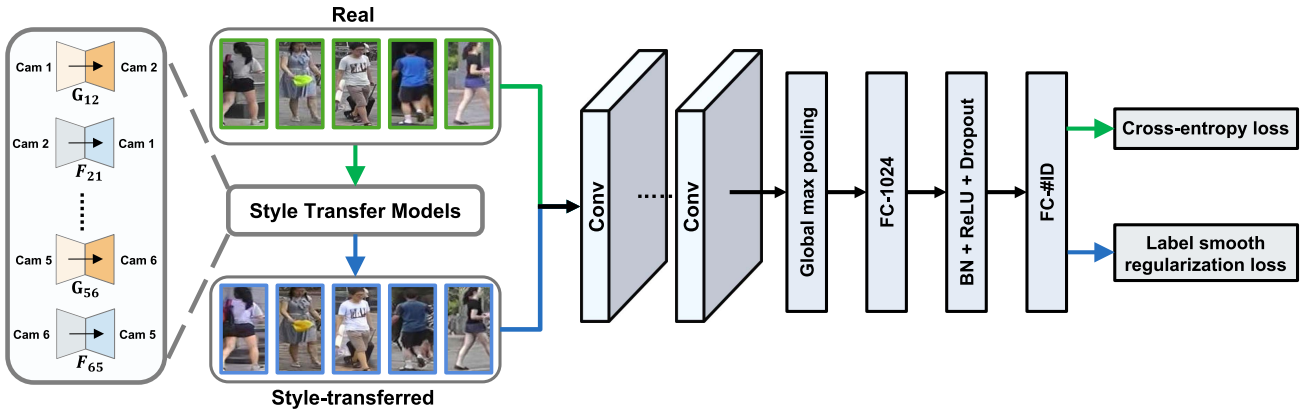


Fig. 4. The framework of the proposed method. The camera-aware style transfer models are learned from the real training data between different cameras. For each real image, we can utilize the trained transfer models to generate images which fit the styles of target cameras. Subsequently, real images (green boxes) and style-transferred images (blue boxes) are combined to train the re-ID CNN. The cross-entropy loss and the label smooth regularization (LSR) loss are applied to real images and style-transferred images, respectively.

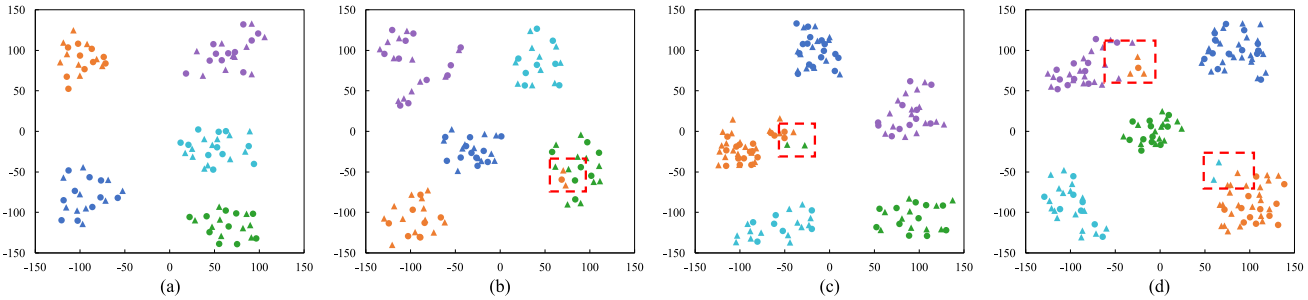


Fig. 5. Barnes-Hut t-SNE [72] visualization on Market-1501. We randomly select real training images of 700 identities to train the re-ID model and visualize the real samples (R, dots) and their fake (style-transferred) samples (F, triangles) of a rest 20 identities. In each figure, different colors represent different identities. We observe 1) fake samples generally overlay with the real samples (see (a), (b), (c) and (d)), laying the foundation of their data augmentation mechanism; 2) noisy fake data exist now and then (red boxes in (a), (b) and (c)), which needs regularization techniques such as LSR. Best viewed in color.

B. Camera-Aware Image-Image Translation

In this work, we employ CycleGAN to generate new training samples: the styles between different cameras are considered as different domains. Given a re-ID dataset containing images collected from L different camera views, our method is to learn image-to-image translation models for each camera pair with CycleGAN. To encourage the style-transfer to preserve the color consistency between the input and output, we add the *identity mapping loss* [10] in the CycleGAN loss function (Eq. 1) to enforce the generator to approximate an identity mapping when using real images of the target domain as input. The *identity mapping loss* can be expressed as:

$$V_{identity}(G, F) = E_{x\backsim p_x}[\|F(x) - x\|_1] + E_{z\backsim p_z}[\|G(z) - z\|_1], \quad (3)$$

Specifically, for training images, we use CycleGAN to train camera-aware style transfer models for each pair of cameras. Following the training strategy in [10], all images are resized to 256×256. We use the same architecture for our camera-aware style transfer networks as CycleGAN. The generator contains 9 residual blocks and four convolutions, while the discriminator is 70×70 PatchGANs [53].

With the learned CycleGAN models, for a training image collected from a certain camera, we generate $L-1$ new training samples whose styles are similar to the corresponding cameras (examples are shown in Fig. 2). In this work, we call the generated image as **style-transferred** image or **fake** image. In this manner, the training set is augmented to a combination of the original images and the style-transferred images. Since each style-transferred image preserves the content of its original image, the new sample is considered to be of the same identity as the original image. This allows us to leverage the style-transferred images as well as their associated labels to train re-ID CNN in together with the original training samples.

Discussions: As shown in Fig. 5, the working mechanism of the proposed data augmentation method mainly consists in: 1) the similar data distribution between the real and fake (style-transferred) images, and 2) the ID labels of the fake images are preserved. In the first aspect, the fake images fill up the gaps between real data points and marginally expand the class borders in the feature space. This guarantees that the augmented dataset generally supports a better characterization of the class distributions during embedding learning. The second aspect, on the other hand, supports the usage of supervised learning [1], a different mechanism from [49] which leverages unlabeled GAN images for regularization.

C. Baseline Deep re-ID Model

Given that both the real and fake (style-transferred) images have ID labels, we use the ID-discriminative embedding (IDE) [1] to train the re-ID CNN model. Using the Softmax loss, IDE regards re-ID training as an image classification task. We use ResNet-50 [26] as backbone and follow the training strategy in [1] for fine-tuning on the ImageNet [43] pre-trained model. Different from the IDE proposed in [1], we discard the last 1000-dimensional classification layer and add two fully connected (FC) layers. The output of the first FC layer has 1024 dimensions named as "FC-1024", followed by batch normalization [45], ReLU and Dropout [44]. The addition "FC-1024" follows the practice in [7] which yields improved accuracy. The output of the second FC layer, is C-dimensional, where C is the number of classes in the training set. In our implementation, all input images are resized to 256×128. The network is illustrated in Fig. 4.

D. Training With CamStyle

Given a new training set composed of real and fake (style-transferred) images (with their ID labels), this section discusses the training strategies using the CamStyle. When we view the real and fake images equally, *i.e.,* assigning a "one-hot" label distribution to them, we obtain a *vanilla version* of our method. On the other hand, when considering the noise introduced by the fake samples, we introduce the *full version* which includes the label smooth regularization (LSR) [11].

1) Vanilla Version: In the vanilla version, each sample in the new training set belongs to a single identity. During training, in each mini-batch, we randomly select M real images and N fake images. The loss function can be written as,

$$\mathcal{L} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_R^i + \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_F^j, \tag{4}$$

where $\mathcal{L}_R$ and $\mathcal{L}_F$ are the cross-entropy loss for real images and fake images, respectively. The cross-entropy loss function can be formulated as,

$$\mathcal{L}_{Cross} = -\sum_{c=1}^{C}\log(p(c))q(c), \tag{5}$$

where C is the number of classes, and $p(c)$ is the predicted probability of the input belonging to label c. $p(c)$ is normalized by the softmax layer, so $\sum_{c=1}^{C} p(c) = 1$. $q(c)$ is the ground-truth distribution. Since each person in the training set belongs to a single identity y, $q(c)$ can be defined as,

$$q(c) = \begin{cases} 1 & c = y \\ 0 & c \neq y. \end{cases} \tag{6}$$

So minimizing the cross entropy is equivalent to maximizing the probability of the ground-truth label. For a given person with identity y, the cross-entropy loss in Eq. 5 can be rewritten as,

$$\mathcal{L}_{Cross} = -\log p(y). \tag{7}$$

Because the similarity in overall data distribution between the real and fake data, the vanilla version is able to improve the baseline IDE accuracy under a system with a few cameras, as to be shown in Section V.

2) Full Version: The style-transferred images have a positive data augmentation effect, but also introduce noise to the system. Therefore, while the vanilla version has merit in reducing over-fitting under a few-camera system in which, due to the lack of data, over-fitting tends to occur, its effectiveness is compromised under more cameras. The reason is that when data from more cameras is available, the over-fitting problem is less critical, and the problem of transfer noise begins to appear.

The transfer noise arises from two causes. 1) CycleGAN does not perfectly model the transfer process, so errors occur during image generation. 2) Due to occlusion and detection errors, there exists noisy samples in the real data, transferring these noisy samples to fake data may produce even more noisy samples. In Fig. 5, we visualize some examples of the deep feature of real and fake data on a 2D space. Most of the generated samples are distributed around the original images. When transfer errors happen (see Fig. 5(c) and Fig. 5(d)), the fake sample will be a noisy sample and be far away from the true distribution. When a real image is a noise sample (see Fig. 5(b) and Fig. 5(d)), it is far away from the images with the same labels, so its generated samples will also be noisy. This problem reduces the benefit of generated samples under full-camera systems where the relatively abundant data has a lower over-fitting risk.

To alleviate this problem, we apply the label smoothing regularization (LSR) [11] on the style-transferred images to softly distribute their labels. That is, we assign less confidence on the ground-truth label and assign small weights to the other classes. The re-assignment of the label distribution of each style-transferred image is written as,

$$q_{LSR}(c) = \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{C} & c = y \\ \dfrac{\epsilon}{C} & c \neq y, \end{cases} \tag{8}$$

where $\epsilon \in [0, 1]$. When $\epsilon = 0$, Eq. 8 can be reduced to Eq. 6. Then, the cross-entropy loss in Eq. 5 is re-defined as,

$$\mathcal{L}_{LSR} = -(1-\epsilon)\log p(y) - \frac{\epsilon}{C}\sum_{c=1}^{C}\log p(c) \tag{9}$$

For real images, we do not use LSR because their labels correctly match the image content. Moreover, we experimentally show that adding LSR to the real images does not improve the re-ID performance under full-camera systems (see Section V-D). So for real images, we use the one-hot label distribution. For style-transferred images, we set $\epsilon = 0.1$, the loss function $\mathcal{L}_F = \mathcal{L}_{LSR}(\epsilon = 0.1)$.

Discussions: Recently, Zheng *et al.* [49] propose the label smoothing regularization for outliers (LSRO) to use the unlabeled samples generated by DCGAN [50]. In [49], since the generated images do not have labels, a uniform label distribution is assigned to the generated samples, *i.e.*, $\mathcal{L}_{LSR}(\epsilon = 1)$. Comparing with LSRO [49], our system has two differences. 1) Fake images are generated according to camera styles. The usage of CycleGAN ensures that the generated images

Fig. 6. Examples generated by our method and DCGAN in [49].

remain the main characteristics of the person (Fig. 6 provides some visual comparisons). 2) Labels in our systems are more reliable. We use LSR to address a small portion of unreliable data, while LSRO [49] is used under the scenario where no labels are available.

IV. CamStyle in Unsupervised Domain Adaptation and One View Learning

In this section, we first describe the problem of unsupervised domain adaptation (UDA) in Section IV-A. Then the implementation of CamStyle in UDA is introduced in Section IV-B. We present the implementation of CamStyle in one view learning in Section IV-C.

A. Problem Definition of Domain Adaptation

In the problem of unsupervised domain adaptation (UDA) in person re-identification (re-ID), we are provided with a full-labeled source set $\{X_s, Y_s\}$ including N_s person images. Each image x_s corresponds to an identity y_s, where $y_s \in \{1, 2, ..., C_s\}$, and C_s is the number of identities. Meanwhile, there are also provided N_t unlabeled target images x_t from unlabeled target set $\{X_t\}$. Images in the target set are collected from L_t camera views and the identities of them are unknown. The object of domain adaptation approaches are to train a re-ID model that generalizes well on target domain testing set with both labeled source training images and unlabeled target training images. Next, we will introduce a new unsupervised domain adaptation approach by using the proposed CamStyle.

B. CamStyle Domain Adaptation

Similar to the implementation of CamStyle in supervision person re-ID introduced in Section III, the CamStyle based domain adaptation approach consists of two steps: 1) source-target camera style transfer for training images generation, and 2) training re-ID model with CamStyle.

1) Source-Target Camera Style Transfer: Assume we are provided full labeled source domain images and unlabeled target domain images. The target domain images are collected from L_t camera views. We consider the styles under different target cameras as different domains. The goal of our method is to learn image-image translation models that translate labeled images from source domain to each target camera. We use CycleGAN to train L_t camera style transfer models for each pair of source domain and target camera. Thus, with the learned CycleGAN models, for each source domain image, we can generate L_t new training images whose styles are similar to each target camera, respectively. Each generated image remains the identity of the original source image and can be used in supervised learning for the target domain.

2) Training re-ID Model With CamStyle: With the camera style transferred images that associated labels, we employ the training strategy proposed in Section III to learn the re-ID model for target domain.

Discussions: Recently, Deng *et al.* [17] and Wei *et al.* [18] propose to learn a preserving generative adversarial network (GAN) based on CycleGAN to translate images from source domain to target domain. The generated images are applied to train re-ID model for target domain. Both of these two approaches only consider the inter-domain image variations, while ignoring the intra-domain style variations caused by target cameras. Compared with these two methods, our method jointly considering the image style variations of inter-domain and intra-domain in the source-target image translation process. Visual comparison of image-image translation between CamStyle and SPGAN proposed by Deng *et al.* [17] are shown in Fig. 7.

TABLE I

The Number of Labeled Identities (#ID) and Number of Labeled Images per Identity (#Img/ID) Under Each Camera on Market-1501 and DukeMTMC-reID

Dataset	**Market-1501**		**DukeMTMC-reID**	
Labeled data	#ID	#Img/ID	#ID	#Img/ID
Camera-**1**	652	3.09	404	6.95
Camera-**2**	541	3.16	378	7.96
Camera-**3**	694	3.9	201	5.41
Camera-**4**	241	3.82	165	8.45
Camera-**5**	576	4.06	218	7.73
Camera-**6**	558	5.82	348	10.63
Camera-**7**	-	-	217	6.13
Camera-**8**	-	-	265	5.68

C. CamStyle in One View Learning

In fact, it is easier to label person identities from one camera view than across disjoint cameras. However, there may have a few labeled samples for each identity under one camera. Table I shows the details of the number of identities and number of images per identities under each camera on Market-1501 [9] and DukeMTMC-reID [49]. Specially, there are less than 6 images per identities under each camera on Market-1501. When training the re-ID model with samples captured from one camera, the re-ID model may suffer from lack of training samples and be sensitive to style variations caused by different cameras. Note that, one view (camera) learning can also be considered as a special domain adaptation problem, where labeled data collected from one camera is the source domain and unlabeled images collected other cameras are target domains. It is an open-set problem in which the unlabeled views may include identities that might be different from the labeled view.

To overcome these challenges, we extend CamStyle to the problem of one view learning. Specifically, our approach can

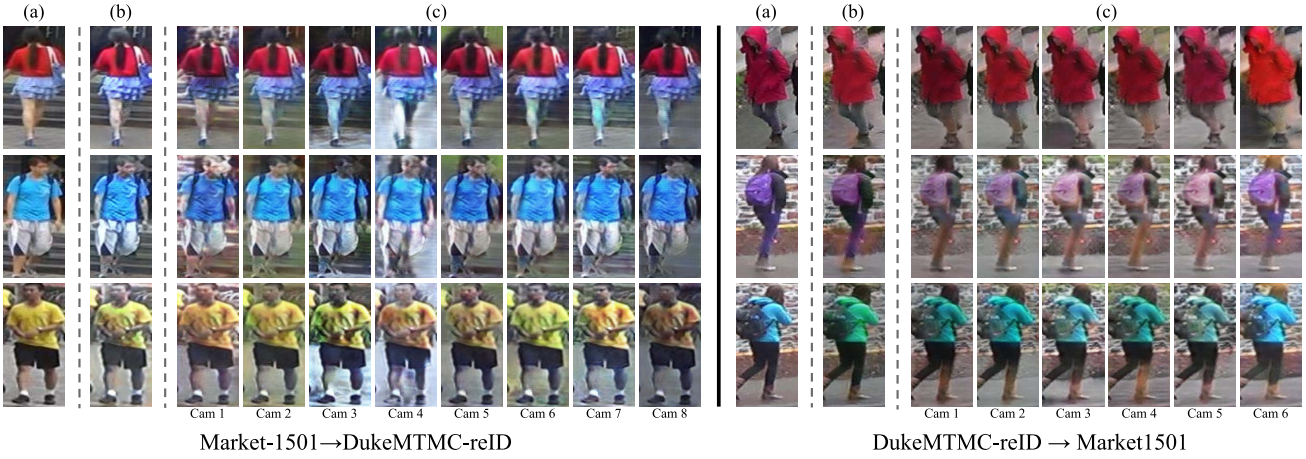


Fig. 7. Visual examples of image-image translation in domain adaptation. (a) original source image, (b) style-transferred images generated by SPGAN [17], (c) style-transferred images generated by CamStyle. Our method (CamStyle) generates various images that are similar to the styles of target cameras.

use all samples under different cameras to learn camera style transfer models without any identity annotation and generates new samples for each labeled image collected from one camera. The generated images and original labeled training images are formed the new training data. We use the augmented labeled data for training re-ID model. In this way, CamStyle not only reduces the risk of over-fitting caused by lack of training data, but also improves the camera-invariance property of re-ID model.

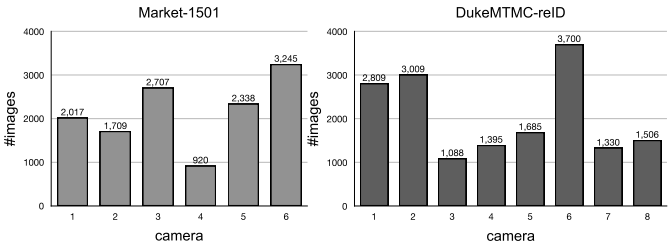


Fig. 8. The number of training samples under different cameras on Market-1501 and DukeMTMC-reID.

V. Experiment

A. Datasets

We evaluate our method on Market-1501 [9] and DukeMTMC-reID [49], [73], because both datasets 1) are large-scale and 2) provide camera labels for each image.

Market-1501 [9] contains 32,668 labeled images of 1,501 identities collected from 6 camera views. Images are detected using deformable part model [74]. The dataset is split into two fixed parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing. There are on average 17.2 images per identity in the training set. In testing, 3,368 hand-drawn images from 750 identities are used as queries to retrieve the matching persons in the database. Single-query evaluation is used.

DukeMTMC-reID [49] is a newly released large-scale person re-ID dataset. It is collected from 8 cameras and comprised of 36,411 labeled images belonging to 1,404 identities. Similar to Market-1501, it consists of 16,522 training images from 702 identities, 2,228 query images from the other 702 identities and 17,661 database images. We use rank-1 accuracy and mean average precision (mAP) for evaluation on both datasets. The details of the number of training samples under each camera are shown in Fig. 8.

B. Experiment Settings

1) Camera-Aware Style Transfer Model: Following Section III-B, given a training set captured from L camera views, we train a camera-aware style transfer (CycleGAN) model for each pair of cameras. Specifically, we train (6 choose 2) $=$ 15 and (8 choose 2) $=$ 28 CycleGAN models for Market-1501 and DukeMTMC-reID, respectively. During training, we resize all input images to 256×256 and use the Adam optimizer [75] to train the models from scratch with $\lambda = 10$ for all the experiments. We set the batch size $= 1$. The learning rate is 0.0002 for the Generator and 0.0001 for the Discriminator at the first 30 epochs and is linearly reduced to zero in the remaining 20 epochs. In camera-aware style transfer step, for each training image, we generated $L - 1$ (5 for Market-1501 and 7 for DukeMTMC-reID) extra fake training images with their original identity preserved as augmented training data.

2) Baseline CNN Model for re-ID: In the training of baseline method, we follow the training strategy in [1]. Specifically, we keep the aspect ratio of all images and resize them to 256×128. Unless otherwise specified, two data augmentation methods, random cropping and random horizontal flipping, are employed during training. We set the probability of performing random flipping to 0.5 and random cropping to 1. The dropout probability p is set to 0.5. We use ResNet-50 [26] as backbone, in which the second fully connected layer has 751 and 702 units for Market-1501 and DukeMTMC-reID, respectively. The learning rate starts with 0.01 for ResNet-50 base layers and 0.1 for the two new added full connected layers. We use the SGD solver to train re-ID model and set the batch size to 128. The learning rate is divided by 10 after 40 epochs, we train 50 epochs in total. In testing,

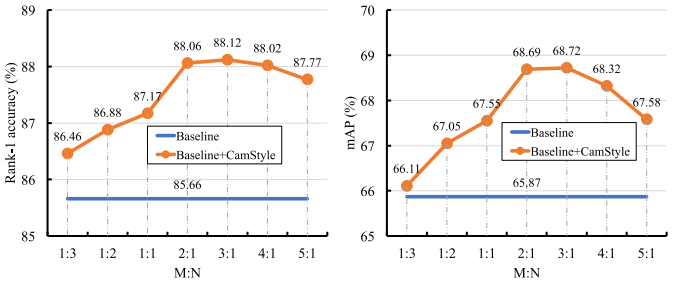


Fig. 9. Evaluation with different ratio of real data and fake data ($M : N$) in a training mini-batch on Market-1501.

we extract the output of the Pool-5 layer as image descriptor (2,048-dim) and use the Euclidean distance to compute the similarity between images.

3) Training CNN With CamStyle: For training of CamStyle, we adopt the same setting as training of the baseline model, except that we randomly select M real images and N fake (style-transferred) images in a training mini-batch. If not specified, we set $M : N = 3 : 1$. Note that, since the number of fake images is larger than that of real images, in each epoch, we use all the real images and randomly selected a $\frac{N}{M} \times \frac{1}{L-1}$ proportion of all fake images. Unless otherwise specified, we apply random cropping and random flipping on both real and fake data during training.

C. Parameter Analysis

An important parameter is involved with CamStyle, *i.e.*, the ratio of $\frac{M}{N}$, where M and N indicate the number of real and fake (style-transferred) training samples in the mini-batch. This parameter encodes the fraction of fake samples used in training. By varying this ratio, we show the experimental results in Fig. 9. It can be seen that, CamStyle with different $\frac{M}{N}$ consistently improves over the baseline. When using more fake data than real data ($M : N < 1$) in each mini-batch, our approach slightly gains about 1% improvement in rank-1 accuracy. On the contrary, when $M : N > 1$, our approach yields more than 2% improvement in rank-1 accuracy. The best performance is achieved when $M : N = 3 : 1$.

In Fig. 10 we evaluate the impact of ϵ defined in Eq. 9. When $\epsilon = 0$, the fake data is trained with cross-entropy loss, and the improvement of "baseline+CamStyle" is limited compared to "baseline". When applying LSR on fake data, the rank-1 accuracy and mAP improve with the increase of ϵ and achieve the best results when ϵ is between 0.07 to 0.15. For example, "baseline + CamStyle" yields rank-1 accuracy of 88.6% when $\epsilon = 0.7$. This is +2.94% higher than "baseline" (85.66%). A larger value of ϵ is likely to be harmful to the performance, because fake images still preserve the original image content to some extent. As such, we should set a high confidence to class of the original image. Taking the above considerations into account, we select $\epsilon = 0.1$ from the range of 0.07 to 0.15.

D. Variant Evaluation

1) Baseline Evaluation Under Different Camera Systems: To fully present the effectiveness of CamStyle, our baseline

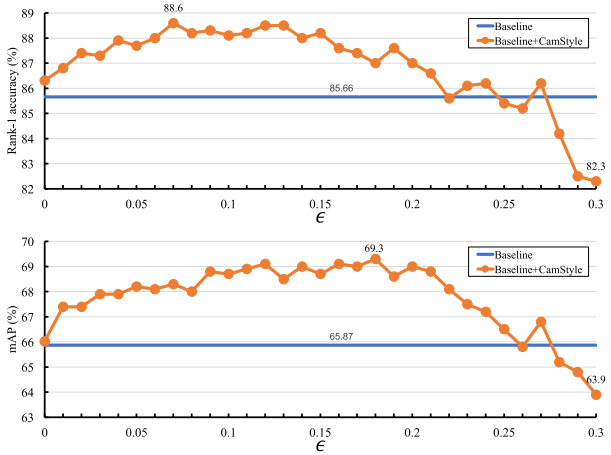


Fig. 10. Results on Market-1501 using different values of ϵ.

TABLE II
PERFORMANCE EVALUATION ON MARKET-1501 USING DIFFERENT LOSS FUNCTIONS. CROSSE: CROSS-ENTROPY, LSR: LABEL SMOOTH REGULARIZATION [11]

Training data	$\mathcal{L}_R$	$\mathcal{L}_F$	Rank-1	mAP
Real	CrossE	None	85.66	65.87
Real	LSR	None	85.21	65.60
Real+Fake	CrossE	CrossE	86.31	66.02
Real+Fake	CrossE	LSR	**88.12**	**68.72**
Real+Fake	LSR	LSR	87.11	67.90

systems consist of 2, 3, 4, 5, 6 cameras for Market-1501 and 2, 3, 4, 5, 8 cameras for DukeMTMC-reID, respectively. In a system with 3 cameras, for example, the training and testing sets both have 3 cameras. In Fig. 11, as the number of cameras increases, the rank-1 accuracy increases. This is because 1) more training data are available and 2) it is easier to find a rank-1 true match when more ground truths are present in the database. In the full-camera (6 for Market-1501 and 8 for DukeMTMC-reID) baseline system, the rank-1 accuracy is 85.6% on Market-1501 and is 72.3% on DukeMTMC-reID.

2) Investigation Into the Effect of Vanilla CamStyle Under Different Camera Systems: We evaluate the effectiveness of the vanilla method (without LSR) in Fig. 11 and Table II. We have two observations. First, in systems with 2 cameras, *vanilla CamStyle* yields significant improvement over the baseline CNN. On Market-1501 with 2 cameras, the improvement reaches +17.1% (from 43.2% to 60.3%). On DukeMTMC-reID with 2 cameras, the rank-1 accuracy is improved from 45.3% to 54.8%. This indicates that the few-camera systems, due to the lack of training data, are prone to over-fitting, so that our method exhibits an impressive system enhancement. Second, as the number of camera increases in the system, the improvement of *vanilla CamStyle* becomes smaller. For example, in the 6-camera system on Market-1501, the improvement in rank-1 accuracy is only +0.7%. This indicates that 1) the over-fitting problems becomes less severe in this full system and that 2) the noise brought by CycleGAN begins to negatively affect the system accuracy.

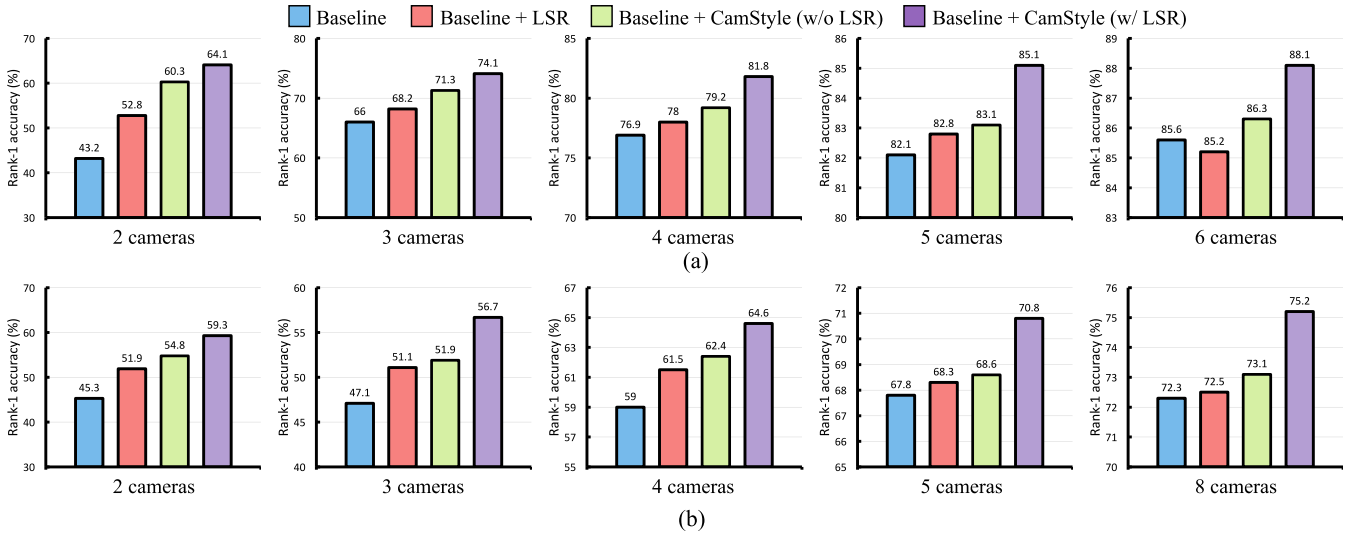


Fig. 11. Comparison of different methods on Market-1501 and DukeMTMC-reID, *i.e.*, baseline, baseline + LSR, baseline + CamStyle *vanilla* (w/o LSR), baseline + CamStyle (w/ LSR). Rank-1 accuracy is shown. Five systems are shown, which have 2, 3, 4, 5, 6 cameras for Market-1501 and 2, 3, 4, 5, 8 cameras for DukeMTMC-reID, respectively. We show that CamStyle (with LSR) yields consistent improvement over the baseline.

TABLE III
IMPACT ANALYSIS OF USING DIFFERENT CAMERAS FOR TRAINING CYCLEGANS ON MARKET-1501. WE ADOPT THE 6-CAMERA SYSTEM. WE START FROM USING THE 1ST AND 2ND CAMERAS, AND THEN GRADUALLY ADD OTHER CAMERAS FOR TRAINING CYCLEGANS

Method	Rank-1	mAP
Baseline	85.66	65.87
Baseline+CamStyle (1+2)	87.20	67.64
Baseline+CamStyle (1+2+3)	87.32	68.53
Baseline+CamStyle (1+2+3+4)	87.42	68.23
Baseline+CamStyle (1+2+3+4+5)	87.85	68.51
Baseline+CamStyle (1+2+3+4+5+6)	**88.12**	**68.72**

3) Evaluation on the Effectiveness of LSR for CamStyle: As previously described, when tested in a system with more than 3 cameras, *vanilla CamStyle* achieves less improvement than the 2-camera system. We show in Fig. 11 and Table II that using the LSR loss on the fake images achieves higher performance than cross-entropy. As shown in Table II, using cross-entropy on style-transferred data improves the rank-1 accuracy to 86.31% under full-camera system on Market-1501. Replacing cross-entropy with LSR on the fake data increases the rank-1 accuracy to 88.12%.

In particular, Fig. 11 and Table II show that using LSR alone on the real data does not help much, or even decrease the performance on full-camera systems. For example, the rank-1 accuracy drops from 88.12% (using cross-entropy loss for real data and using LSR for fake data) to 87.11% when using LSR on both real and fake data. *Therefore, the fact that CamStyle with LSR improves over the baseline is not attributed to LSR alone, but to the interaction between LSR and the fake images.* By this experiment, we justify the necessity of using LSR on the fake images.

4) Investigation Into the Impact of Using Different Cameras for Training Camera-Aware Style Transfer Models: In Table III, we show that as using more cameras to train camera-aware style transfer models, the rank-1 accuracy is improved from 85.66% to 88.12%. Particularly, our method obtains +1.54% improvement in rank-1 accuracy even only using the 1th and 2th camera to train camera-aware style transfer model. In addition, when training cameras style transfer models with using 5 cameras, it has the rank-1 accuracy of 87.85%, which is 0.27% lower than of using 6 cameras. This shows that even using a part of the cameras to learn camera-aware style transfer models, our method can yield approximately equivalent results to using all the cameras.

5) Analysis of Different Data Augmentation Methods: To further validate the CamStyle, we further compare it with other data augmentation methods, random flipping + random cropping (RF + RC), Random Erasing (RE) [47], random rotation, random scaling, random shearing, Gaussian noise and color changing [46]. Because different data augmentation methods are distinct in their working mechanism, it is expected that the best performance of different methods is achieved under different values of M:N. Therefore, for fair comparison, we set the parameter M:N to the value which data augmentation method yields the best performance. M and N represent the number of real data and fake data in a training batch, respectively. Here, the probability of performing data augmentation is N:(M + N). We set the probability of performing random flipping and random erasing to 0.5, and of random cropping to 1. Random rotation is performed by rotating sampled within 5 degrees with a random probability of 0.5. In random scaling, we resize the image to 95%-100% of the original size and then pad zeros to the image border. The probability of random scaling is 0.3. Random shearing is applied by tilting sampled within 5 degrees with a random probability of 0.5. For Gaussian noise method, we add different levels of Gaussian noise to an image. Specifically, we randomly add Gaussian noise to 5% of the image pixels. The probability of adding Gaussian noise is set to 0.3. For color changing, we first transform the image color space from RGB to HSV and then add a random value (sampled from $[-0.2, 0.2]$) to the hue channel of all pixels. The image is then transformed back to

Fig. 12. Example images generated by different data augmentation methods.

TABLE IV
COMPARISON COMBINATIONS BETWEEN DIFFERENT DATA AUGMENTATION METHODS ON MARKET-1501. THE FIRST COLUMN INDICATES THE NUMBER OF AUGMENTATIONS METHODS EMPLOYED ON THE BASELINE MODEL. **RF + RC**: RANDOM FLIP + RANDOM CROP, **RE**: RANDOM ERASING [47]

# Augm.	RF+RC	RE	CamStyle	Rotation	Shear	Scale	Colour	Noise	Rank-1	mAP
0	-	-	-	-	-	-	-	-	84.15	64.10
1	✓	-	-	-	-	-	-	-	85.66	65.87
	-	✓	-	-	-	-	-	-	86.83	68.50
	-	-	✓	-	-	-	-	-	85.01	64.86
	-	-	-	✓	-	-	-	-	85.70	66.31
	-	-	-	-	✓	-	-	-	84.02	64.20
	-	-	-	-	-	✓	-	-	84.52	64.33
	-	-	-	-	-	-	✓	-	84.45	63.95
	-	-	-	-	-	-	-	✓	84.25	64.41
2	✓	✓	-	-	-	-	-	-	87.65	69.91
	✓	-	✓	-	-	-	-	-	88.12	68.72
	✓	-	-	✓	-	-	-	-	84.62	63.99
	-	✓	✓	-	-	-	-	-	87.89	69.10
	-	✓	-	✓	-	-	-	-	87.41	68.77
	-	-	✓	✓	-	-	-	-	87.62	67.38
3	✓	✓	✓	-	-	-	-	-	**89.49**	**71.55**
4	✓	✓	✓	✓	-	-	-	-	88.32	69.15

RGB space. This operation is able to change the brightness of images. We perform color changing with a probability of 0.4. Example images generated by different data augmentation methods are shown in Fig. 12.

The comparative results of the above data augmentation methods are shown in Table IV. First, to understand the performance of each individual data augmentation technique, we train CNN baseline using each data augmentation method individually. As show in Table IV, the rank-1 accuracy of baseline is 84.15% when no data augmentation is used. When only applying RF + RC, RE, CamStyle, or random rotation, rank-1 accuracy is increased to 85.66%, 86.83%, 85.01% and 85.70%, respectively. We observe that RE achieves the largest individual improvement in performance. However, random

TABLE V
COMPARISON WITH STATE OF THE ART ON THE MARKET-1501 DATASET. IDE* REFERS TO IMPROVED IDE WITH THE TRAINING SCHEDULE IN THIS PAPER. **RE**: RANDOM ERASING [47]

Method	Rank-1	mAP
BOW [9]	34.40	14.09
LOMO+XQDA [5]	43.79	22.22
DNS [6]	61.02	35.68
IDE [1]	72.54	46.00
Re-rank [72]	77.11	63.63
DLCE [29]	79.5	59.9
DF [73]	81.0	63.4
SSM [74]	82.21	68.80
SVDNet [7]	82.3	62.1
GAN [45]	83.97	66.07
PDF [75]	84.14	63.41
TriNet [8]	84.92	69.14
DJL [76]	85.1	65.5
PSE [77]	87.7	69.0
HA-CNN [78]	91.2	75.7
IDE*	85.66	65.87
IDE*+CamStyle	88.12	68.72
IDE*+CamStyle+RE [43]	**89.49**	**71.55**

TABLE VI
COMPARISON WITH STATE OF THE ART ON DUKEMTMC-REID. IDE* REFERS TO IMPROVED IDE WITH THE TRAINING SCHEDULE DESCRIBED IN THIS PAPER. **RE**: RANDOM ERASING [47]

Method	Rank-1	mAP
BOW+kissme [9]	25.13	12.17
LOMO+XQDA [5]	30.75	17.04
IDE [1]	65.22	44.99
GAN [45]	67.68	47.13
OIM [79]	68.1	47.4
APR [80]	70.69	51.88
PAN [81]	71.59	51.51
TriNet [8]	72.44	53.50
SVDNet [7]	76.7	56.8
IDE*	72.31	51.83
IDE*+CamStyle	75.27	53.48
IDE*+CamStyle+RE [43]	**78.32**	**57.61**

shearing, random scaling, color changing and Gaussian noise fail to improve the performance over the baseline. Then, we investigate the combinations of RF + RC, RE, CamStyle and random rotation. If we combine CamStyle with RF + RC, RE or random rotation, we observe consistent improvement over their individual usage. Nevertheless, the combination of RF + RC and random rotation fails to further improve the performance compared to their individual employment. The best performance is achieved when RF + RC, RE, and CamStyle are used together. In fact, CamStyle and some other augmentation techniques focus on different aspects of CNN invariance. In this respect, our results show that CamStyle is well complementary to these data augmentation methods. Particularly, combining CamStyle with RF + RC and RE, we are able to achieve 89.49% rank-1 accuracy on Market-1501.

Since random cropping, random flipping, random erasing, random rotation, random shearing, random scaling, Gaussian noise and color changing are based on basic image processing operations, the computational cost is not heavy in CNN training. For CamStyle, the style transfer models and fake images

TABLE VII

EVALUATION ON THE PERFORMANCE OF LEARNING FROM ONE CAMERA FOR PERSON RE-IDENTIFICATION. IN TRAINING, LABELED SAMPLES ARE AVAILABLE FROM ONLY ONE CAMERA. IN TESTING, WE USE QUERY AND GALLERY SAMPLES UNDER ALL CAMERAS

Dataset	**Market-1501**				**DukeMTMC-reID**			
Method	Baseline		Baseline+CamStyle		Baseline		Baseline+CamStyle	
Labeled data	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
All	85.66	65.87	88.12	68.72	71.31	51.83	75.27	53.48
Camera-**1**	37.02	14.70	**54.54**	**26.24**	29.80	14.30	**48.47**	**24.91**
Camera-**2**	35.75	13.99	**55.34**	**27.83**	29.08	14.25	**54.94**	**30.82**
Camera-**3**	43.20	19.38	**67.04**	**38.63**	29.17	13.15	**48.25**	**23.91**
Camera-**4**	31.86	11.96	**45.78**	**20.31**	23.65	10.86	**38.69**	**18.74**
Camera-**5**	39.16	16.23	**60.78**	**31.67**	24.51	11.52	**48.38**	**25.14**
Camera-**6**	36.49	15.13	**61.91**	**32.63**	33.75	17.92	**48.38**	**26.63**
Camera-**7**	-	-	-	-	33.93	16.48	**50.58**	**27.05**
Camera-**8**	-	-	-	-	29.89	14.57	**51.62**	**28.47**
Average	37.24	15.23	**57.56**	**29.55**	29.22	14.13	**48.66**	**25.70**

TABLE VIII

METHODS COMPARISON USING DUKE / MARKET AS SOURCE, AND USING MARKET / DUKE AS TARGET. NOTE THAT, WE EMPLOY RANDOM CROPPING AND RANDOM FLIPPING ON BOTH REAL DATA AND FAKE DATA DURING TRAINING

Methods	Training Set	DukeMTMC-reID → Market-1501					Market-1501 → DukeMTMC-reID				
		Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Baseline	Target	85.6	93.7	95.8	97.5	65.8	72.3	84.1	88.1	90.9	51.8
Baseline	Source	44.6	62.5	69.6	76.5	20.6	32.9	49.5	54.8	61.7	16.9
Basel.+CycleGAN	Sour. (Domain-Domain)	49.9	67.1	74.2	80.5	22.6	39.2	54.8	60.5	66.2	20.1
Basel.+CamStyle	Sour. (Domain-Camera)	**58.8**	**78.2**	**84.3**	**88.8**	**27.4**	**48.4**	**62.5**	**68.9**	**74.4**	**25.1**

can be trained and generated off-the-shelf. Thus, we only need to directly load the corresponding fake images in the training of re-ID model.

E. Comparison With State-of-the-Art Methods

We compare our method with the state-of-the-art methods on Market-1501 and DukeMTMC-reID in Table V and Table VI, respectively. First, using our baseline training strategy, we obtain a strong baseline (IDE*) on both datasets. Specifically, IDE* achieves 85.66% for Market-1501 and 72.31% for DukeMTMC-reID in rank-1 accuracy. Compared with published IDE implementations [1], [7], [49], IDE* has the best rank-1 accuracy on Market-1501.

Then, when applying CamStyle on IDE*, we obtain competitive results compared with the state of the art. Specifically, we achieve **rank-1 accuracy = 88.12% for Market-1501, and rank-1 accuracy = 75.27% for DukeMTMC-reID.** On Market-1501, our method has higher rank-1 accuracy than PDF [79], TriNet [8], PSE [81] and DJL [80]. The mAP of our method is slightly lower than TriNet [8] by 0.42% and lower than HA-CNN [82] by 4.15%. HA-CNN utilizes a more sophisticated network than ours. On DukeMTMC-reID, the mAP of our method is lower than SVDNet [7] by 3.32%.

Further combining CamStyle with Random Erasing data augmentation [47] (RF + RC is already implemented in the baseline), our final rank-1 performance arrives at **89.49%** for Market-1501 and **78.32%** for DukeMTMC-reID.

F. Learning From One View

Table VII shows the results of our method when using only labeled samples collected from one camera to train CNN re-ID model. Note that, we employ random cropping and random flipping on both real data and fake data during training. With training images only collected from one view, the performance of baseline significantly drops. For example, the rank-1 accuracy of baseline is 85.66% when trained with all labeled data on Market-1501, but drops to 37.02% when trained with labeled data from camera-1. With CamStyle, the performance is significantly improved in all cases. For example, our method has 57.56% rank-1 accuracy and 48.66% rank-1 accuracy averaged on all settings on Market-1501 and DukeMTMC-reID, respectively. This is 20.32% higher on Market-1501 and 19.44% higher on DukeMTMC-reID than baseline.

G. Implemented CamStyle in Domain Adaptation

1) Evaluation of Baseline on Performance: The results of baseline are reported in Table VIII. When trained on and tested on the target set, the baseline (IDE) produced high performance on both datasets. However, the performance drops significantly when the re-ID model is trained on the source set and directly tested on the target set. For example, the baseline re-ID model trained and tested on DukeMTMC-reID gives rank-1 accuracy of 72.3%, but drops to 32.9% when trained on Market-1501. The main reason is the difference in data distribution between different datasets.

2) Investigation Into the Effectiveness of CamStyle for Domain Adaptation: An effective strategy for reducing the divergence of different datasets is to translate the labeled images from source domain to target domain by training a CycleGAN between domains [17], [18]. In this way, each source image is mapped to one fake image in the style of target domain. The translated fake images remain the

TABLE IX
UNSUPERVISED PERSON RE-ID PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET-1501. CAMEL [86] USES MULTI-QUERY SETTING, THE OTHER METHODS USE SINGLE-QUERY SETTING

Methods	Duke → Market-1501			
	R-1	R-5	R-10	mAP
LOMO [5]	27.2	41.6	49.1	8.0
Bow [9]	35.8	52.4	60.3	14.8
UMDL [58]	34.5	52.6	59.6	12.4
PTGAN [18]	38.6	-	66.1	-
PUL [83]	45.5	60.7	66.7	20.5
SPGAN [17]	51.5	70.1	76.8	22.8
CAMEL [82]	54.5	-	-	26.3
SPGAN+LMP [17]	57.7	75.8	82.4	26.7
TJ-AIDL [16]	58.2	74.8	81.1	26.5
IDE*+CamStyle	**58.8**	**78.2**	**84.3**	**27.4**
IDE*+CamStyle+LMP [17]	**64.7**	**80.2**	**85.3**	**30.4**

TABLE X
UNSUPERVISED PERSON RE-ID PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON DUKEMTMC-REID

Methods	Market-1501 → Duke			
	R-1	R-5	R-10	mAP
LOMO [5]	12.3	21.3	26.6	4.8
Bow [9]	17.1	28.8	34.9	8.3
UMDL [58]	18.5	31.4	37.6	7.3
PTGAN [18]	27.4	-	50.7	-
PUL [83]	30.0	43.4	48.5	16.4
SPGAN [17]	41.1	56.6	63.0	22.3
TJ-AIDL [16]	44.3	59.6	65.0	23.0
SPGAN+LMP [17]	46.4	62.3	68.0	**26.2**
IDE*+CamStyle	**48.4**	**62.5**	**68.9**	25.1
IDE*+CamStyle+LMP [17]	**51.7**	**67.0**	**72.8**	**27.7**

original identities and are used to train re-ID model in a supervised way. As shown in Table VIII, the baseline + CycleGAN consistently improves the performance when using the domain-domain translated training images. For example, the baseline + CycleGAN improves the rank-1 accuracy from 44.6% to 49.9% when trained on DukeMTMC-reID and tested on Market-1501. When considering the image variations of target cameras, our method (CamStyle) translates labeled images from source domain to each target camera. Compared to baseline + CycleGAN, the baseline + CamStyle considers the image variations of target cameras and generates more training images capturing the styles of different target cameras. As reported in Table VIII, our method clearly outperforms the baseline and baseline + CycleGAN by a large margin. For example, when using Market-1501 as source set and tested on DukeMTMC-reID, our method is 9.2% and 5% higher than baseline + CycleGAN in rank-1 accuracy and mAP, respectively.

3) Comparison With the State of the Art: We compare the proposed CamStyle with the state-of-the-art unsupervised learning methods on Market-1501 and DukeMTMC-reID in Table IX and Table X, respectively. First, we compare our method with two hand-crafted methods, *i.e.* BOW [5], [9]. These two hand-crafted features are directly employed on target testing set without training. Both of them fail to obtain competitive results.

Then, we compare our method with three unsupervised methods including [62], CAMEL [86], and PUL [87]. These unsupervised methods exploit the unlabeled data on target domain for training re-ID model and achieve higher results than hand-crafted methods. For example, when using Market-1501 as source set and tested on DukeMTMC-reID, PUL [87] achieves 30.0% in rank-1 accuracy, outperforming the BOW [9] by 12.9%.

Finally, we compare our method with recently proposed state-of-the-art domain adaptation methods, including PTGAN [18], SPGAN [17] and TJ-AIDL [16]. Our method obtains competitive results compared with the state-of-the-art approaches. Specifically, our method achieves **rank-1 accuracy = 58.8% and mAP = 27.4%** when using DukeMTMC-reID as source set and tested on Market-1501, and achieves **rank-1 accuracy = 48.4% and mAP = 25.1%** when trained on Market-1501 and tested on DukeMTMC-reID. On the one hand, our methods give higher rank-1 accuracy than PTGAN [18], SPGAN [17] and TJ-AIDL [16] in all settings. On the other hand, when tested on DukeMTMC-reID, our method has slightly lower mAP than SPGAN + LMP [17] which employed local max pooling (LMP) in the testing phase. Moreover, we further apply LMP in our approach during testing phase. With LMP, our approach gains further improvement. Specifically, the rank-1 accuracy of our approach is higher than SPGAN + LMP [14] by 7% and 5.3% when tested on Market-1501 and DukeMTMC-reID, respectively.

VI. CONCLUSION

In this paper, we propose CamStyle, a new data augmentation method for deep person re-identification. The camera-aware style transfer models are learned for each pair of cameras with CycleGAN, which are used to generate new training images from the original ones. The real images and the style-transferred images form the new training set. Moreover, to alleviate the increased level of noise induced by CycleGAN, label smooth regularization (LSR) is applied on the generated samples. Experiments on the Market-1501 and DukeMTMC-reID datasets show that our method can effectively reduce the impact of over-fitting, and, when combined with LSR, yields consistent improvement over the baselines. We show that our method is complementary to other data augmentation techniques. In addition, CamStyle can be implemented in other important person re-identification tasks, such as one view learning and domain adaptation. Specially, our method obtains state-of-the-art results in unsupervised domain adaptation for person re-identification. In the future work, we will improve the scalability of our approach to deal with the increasing scale of cameras by multi-domain image-to-image translation methods, such as StarGAN [55] and MUNIT [56].

ACKNOWLEDGEMENTS

The authors thank Wenjing Li for encouragement.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann. (2016). "Person re-identification: Past, present and future." [Online]. Available: https://arxiv.org/abs/1610.02984

[2] 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.

[3] Y.-C. Chen, W.-S. Zheng, P. C. Yuen, and J. Lai, "An asymmetric distance model for cross-view feature mapping in person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1661–1675, Aug. 2015.

[4] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. CVPR*, 2012, pp. 2288–2295.

[5] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015, pp. 2197–2206.

[6] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. CVPR*, 2016, pp. 1239–1248.

[7] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. ICCV*, 2017, pp. 3800–3808.

[8] 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

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015, pp. 1116–1124.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2242–2251.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, 2016, pp. 2818–2826.

[12] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 7167–7176.

[13] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, 2015, pp. 1180–1189.

[14] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[15] F. Liu, J. Lu, and G. Zhang, "Unsupervised heterogeneous domain adaptation via shared fuzzy equivalence relations," *IEEE Trans. Fuzzy Syst.*, to be published, doi: 10.1109/TFUZZ.2018.2836364.

[16] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. CVPR*, 2018, pp. 2275–2284.

[17] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. CVPR*, 2018, pp. 994–1003.

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

[19] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. CVPR*, 2018, pp. 5157–5166.

[20] L. Zheng, S. Wang, and Q. Tian, "Coupled binary embedding for large-scale image retrieval," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3368–3380, Aug. 2014.

[21] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

[22] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 1–13, 2016.

[23] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1056–1069, Mar. 2016.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[27] X. Dong, Y. Yan, M. Tan, Y. Yang, and I. W. Tsang, "Late fusion via subspace search with consistency preservation," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 518–528, Jan. 2019, doi: 10.1109/TIP.2018.2867747.

[28] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2018.2844853.

[29] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2494–2502, Dec. 2016.

[30] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Macro-micro adversarial network for human parsing," in *Proc. ECCV*, 2018, pp. 418–434.

[31] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. ICPR*, 2014, pp. 34–39.

[32] L. Wu, C. Shen, and A. van den Hengel. (2016). "PersonNet: Person re-identification with deep convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1601.07255

[33] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, 2017, Art. no. 13.

[34] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. ECCV*, 2016, pp. 135–153.

[35] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. ICCV*, 2017, pp. 5409–5418.

[36] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[37] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[38] Y.-C. Chen, W.-S. Zheng, and J. Lai, "Mirror representation for modeling view-specific transform in person re-identification," in *Proc. IJCAI*, 2015, pp. 3402–3408.

[39] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proc. WACV*, 2016, pp. 1–8.

[40] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *Proc. CVPR*, 2018, pp. 5177–5186.

[41] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. CVPR*, 2017, pp. 1367–1376.

[42] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, 2016, pp. 868–884.

[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[46] N. McLaughlin, J. M. Del Rincon, and P. Miller, "Data-augmentation for reducing dataset bias in person re-identification," in *Proc. AVSS*, 2015, pp. 1–6.

[47] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. (2017). "Random erasing data augmentation." [Online]. Available: https://arxiv.org/abs/1708.04896

[48] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. CVPR*, 2018, pp. 5098–5107.

[49] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. ICCV*, 2017, pp. 3754–3762.

[50] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016, pp. 1–16.

[51] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. CVPR*, 2018, pp. 4099–4108.

[52] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 1125–1134.

[54] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NIPS*, 2016, pp. 469–477.

[55] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR*, 2018, pp. 8789–8797.

[56] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. ECCV*, 2018, pp. 172–189.

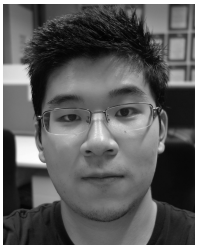
[57] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. CVPR*, 2016, pp. 2414–2423.

[58] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[59] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *Proc. CVPR*, 2018, pp. 379–388.
[60] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel–level domain adaptation with generative adversarial networks," in *Proc. CVPR*, 2017, pp. 3722–3731.
[61] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proc. ICLR*, 2017, pp. 1–14.
[62] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. CVPR*, 2016, pp. 1306–1315.
[63] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. CVPR*, 2015, pp. 325–333.
[64] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proc. ECCV*, 2018, pp. 172–188.
[65] F. Porikli, "Inter-camera color calibration by correlation model function," in *Proc. ICIP*, 2003, pp. 133–136.
[66] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," in *Proc. CVPR*, 2005, pp. 26–33.
[67] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Comput. Vis. Image Understand.*, vol. 109, no. 2, pp. 146–162, 2008.
[68] G. Lian, J.-H. Lai, C. Y. Suen, and P. Chen, "Matching of tracked pedestrians across disjoint camera views using CI-DLBP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 7, pp. 1087–1099, Jul. 2012.
[69] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions," in *Proc. BMVC*, 2008, pp. 64.1–64.10.
[70] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch, "Learning implicit transfer for person re-identification," in *Proc. ECCV*, 2012, pp. 381–390.
[71] T. Avraham and M. Lindenbaum, "Learning appearance transfer for person re-identification," in *Person Re-Identification*. London, U.K.: Springer, 2014, pp. 231–246.
[72] L. van der Maaten, "Accelerating t-SNE using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, Oct. 2014.
[73] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV Workshop*, 2016, pp. 17–35.
[74] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[75] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[76] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with *k*-reciprocal encoding," in *Proc. CVPR*, 2017, pp. 3652–3661.
[77] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, 2017, pp. 3239–3248.
[78] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. CVPR*, 2017, pp. 2530–2539.
[79] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. ICCV*, 2017, pp. 3980–3989.
[80] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. IJCAI*, 2017, pp. 2194–2200.
[81] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. CVPR*, 2018, pp. 420–429.
[82] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, 2018, pp. 2285–2294.
[83] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. CVPR*, 2017, pp. 3376–3385.
[84] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. (2017). "Improving person re-identification by attribute and identity learning." [Online]. Available: https://arxiv.org/abs/1703.07220
[85] Z. Zheng, L. Zheng, and Y. Yang. (2017). "Pedestrian alignment network for large-scale person re-identification." [Online]. Available: https://arxiv.org/abs/1707.00408
[86] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. ICCV*, 2017, pp. 994–1002.
[87] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 14, no. 4, 2018, Art. no. 83.

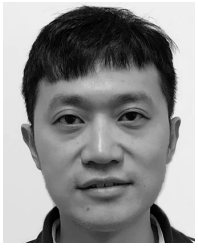
Zhun Zhong received the M.S. degree in computer science and technology from the China University of Petroleum, Qingdao, China, in 2015. He is currently pursuing the Ph.D. degree with Xiamen University. He is also a Joint Ph.D. Student with the University of Technology Sydney. His research interests include person re-identification and domain adaptation.

Liang Zheng received the B.E. degree in life science and the Ph.D. degree in electronic engineering from Tsinghua University, China, in 2010 and 2015, respectively. He was a Post-Doctoral Researcher with the Center for Artificial Intelligence, University of Technology Sydney, Australia. He is currently a Lecturer and a Computer Science Futures Fellow with the Research School of Computer Science, The Australian National University. His research interests include image retrieval, classification, and person re-identification.



Zhedong Zheng received the B.S. degree from Fudan University, China, in 2016. He is currently pursuing the Ph.D. degree with the University of Technology Sydney, Australia. His research interests include image retrieval and person re-identification.

Shaozi Li (SM'18) received the B.S. degree from Hunan University, the M.S. degree from Xi'an Jiaotong University, and the Ph.D. degree from the National University of Defense Technology. He currently serves as the Chair and a Professor with the Cognitive Science Department, Xiamen University. His research interests cover artificial intelligence and its applications, moving objects detection and recognition, machine learning, computer vision, and multimedia information retrieval. He has directed and completed more than 20 research projects, including several national 863 programs, National Nature Science Foundation of China, and Ph.D. Programs Foundation of Ministry of Education of China. He is a Senior Member of ACM and the China Computer Federation (CCF). He is a Vice Director of the Technical Committee on Collaborative Computing of CCF and the Fujian Association of Artificial Intelligence.



Yi Yang received the Ph.D. degree in computer science from Zhejiang University. He was a Post-Doctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, from 2011 to 2013. He is currently a Professor with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, School of Software, University of Technology Sydney. His research interests include multimedia, computer vision, and data science.