A Novel Unsupervised Camera-aware Domain Adaptation Framework for Person Re-identification

Lei Qi[1], Lei Wang[2*], Jing Huo[1], Luping Zhou[3], Yinghuan Shi[1], and Yang Gao[1*]

[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]School of Computing and Information Technology, University of Wollongong
[3]School of Electrical and Information Engineering, The University of Sydney

Abstract

Unsupervised cross-domain person re-identification (Re-ID) faces two key issues. One is the data distribution discrepancy between source and target domains, and the other is the lack of label information in target domain. They are addressed in this paper from the perspective of representation learning. For the first issue, we highlight the presence of camera-level sub-domains as a unique characteristic of person Re-ID, and develop "camera-aware" domain adaptation to reduce the discrepancy not only between source and target domains but also across these sub-domains. For the second issue, we exploit the temporal continuity in each camera of target domain to create discriminative information. This is implemented by dynamically generating online triplets within each batch, in order to maximally take advantage of the steadily improved feature representation in training process. Together, the above two methods give rise to a novel unsupervised deep domain adaptation framework for person Re-ID. Experiments and ablation studies on benchmark datasets demonstrate its superiority and interesting properties.

1. Introduction

Person re-identification (Re-ID) matches images of the same identity captured by different cameras of non-overlapping views [4]. In unsupervised cross-domain person Re-ID, labeled data is only available in source domain, while all data in target domain is unlabeled [20]. It aims to learn an effective model to conduct Re-ID in target domain.

Unsupervised cross-domain person Re-ID faces two key issues. One is the data distribution discrepancy of source and target domains, caused by the variations such as body pose, view angle, illumination, image resolution, occlusion and background. The other is the lack of label information in target domain, due to time-consuming or even infeasible manual annotation in real applications. Tangible progress has been made to address them [24, 20, 25, 6, 31, 2], as reviewed in next section. However, unsupervised cross-domain person Re-ID is still far from satisfactory, especially when compared with the supervised counterpart [1, 29, 11].

This paper aims to better address the above two issues from the perspective of representation learning. We try to learn a deep feature representation (or equally, a shared subspace) in which the discrepancy of source and target domains is more effectively reduced and the discriminative information in target domain is more effectively reflected.

First, in most practical applications, the camera ID of a frame can be readily obtained[1]. The frames from the same camera exhibit a common visual style (e.g., in terms of background, image resolution, viewing angle, and so on). This camera-specific style has recently started attracting attention in person Re-ID (e.g., image-image translation)[31]. The presence of camera-level *sub-domains* is a unique characteristic of cross-domain person Re-ID. Nevertheless, it has not been well exploited from the perspective of representation learning. To utilize this characteristic, we propose a novel *camera-aware* domain adaptation. It emphasizes that with the learned representation, the distribution discrepancy across camera-level sub-domains shall also be sufficiently reduced, dealing with the discrepancy of source and target domains at a finer level. This idea is realized by adversarial learning with a novel criterion called *cross-domain camera equiprobability*.

Second, temporal information is usually available for the

---
[1]For example, camera ID is provided for each frame in benchmark datasets of person Re-ID, as will be shown in the experiment.

images from a camera (e.g., via frame ID) in person Re-ID. Temporally close frames more likely correspond to the same subject, which has long been used as an important cue for video analysis. Nevertheless, it has not been well exploited for unsupervised cross-domain person Re-ID to handle the lack of label information in target domain. In this paper, temporal information is jointly used with image distances to generate, in an unsupervised manner, triplets from each camera in target domain. A smart scheme is proposed to better seek true positive and true negative images. More importantly, instead of generating triplets offline, we dynamically generate triplets online in each batch during training. This allows us to fully take advantage of the steadily improved feature representation to produce better triplets. In turn, these triplets help networks to learn better feature representations, forming a positive loop to exploit discriminative information from target domain.

Together, the above two improvements give rise to a novel domain adaptation framework for unsupervised cross-domain person Re-ID. As will be shown, both improvements are essential and it is their joint effort that makes the proposed framework really function. For clarity, the contributions in this work are summarized as follows.

First, a camera-aware domain adaptation method is developed by considering the unique presence of camera-level sub-domains in person Re-ID. To the best of our survey, our adversarial learning based method is the first one, among those aiming to learn better feature representation, to integrate source and target domains at this fine level.

Second, an unsupervised online in-batch triplet generation method is proposed to explore the underlying discriminative information in unlabeled target domain. Through high-quality triplets, it provides important information to boost the performance of the whole framework.

Last, both theoretical analysis and experimental studies are conducted to illustrate the proposed camera-aware domain adaptation. The results and ablation studies demonstrate the superiority of the proposed framework and its interesting properties.

2. Related Work

Unsupervised domain adaptation. Unsupervised cross-domain person Re-ID is related to unsupervised domain adaptation, a more general technique handling unlabeled target domain with the help of labeled source domain. In the literature, most unsupervised domain adaptation methods learn a common mapping between source and target distributions. Several methods based on the maximum mean discrepancy (MMD) have been proposed [14, 15, 27, 22]. Long *et al.* [14] use a new deep adaptation network, where hidden representations of all task-specific layers are embedded in a Reproducing Kernel Hilbert space. To transfer a classifier from source domain to target do-

main, the work in [15] jointly learns adaptive classifiers between the two domains by a residual function. In [8, 3], autoencode-based methods are investigated to explore the discriminative information in target domain. Recently, adversarial learning [7, 26, 21] has been applied to domain adaptation. Ganin *et al.* [7] propose the gradient reversal layer (GRL) to pursue the same distribution between source and target domains. Inspired by generative adversarial networks (GANs), Tzeng *et al.* [21] leverage a GAN loss to match the data distributions of source and target domains.

Nevertheless, for a person Re-ID task, the distribution discrepancy also exists at the camera level. It will not be effectively reduced when only the overall domain-level discrepancy is concerned. In this sense, directly applying existing unsupervised domain adaptation methods to a person Re-ID task may not be the best option.

Unsupervised cross-domain person Re-ID. As previously mentioned, most existing methods on this topic address two issues, *i.e.,* reducing data distribution discrepancy between two domains and generating discriminative information for target domain. In the literature, methods have been developed to learn a shared subspace or dictionary across domains [19, 24, 17, 20]. However, these methods are not based on deep learning and thus do not fully explore the high-level semantics in images. Recently, several deep-learning-based methods [31, 2, 23, 16] have been seen. In [16, 23], generating pseudo labels for target images is investigated. Lv *et al.* [16] propose an unsupervised incremental learning algorithm, aided by the transfer learning of spatio-temporal patterns of pedestrians in target domain. In [23], the proposed approach simultaneously learns an attribute-semantic and identity-discriminative feature representation in target domain. However, when generating discriminative information for target domain, the above methods do not utilize temporal continuity of images in each camera. Moreover, the generation of information is usually conducted offline and separately, instead of on the fly during training. All of these will be improved in our framework.

Recently, generating extra training images for target domain has become popular [25, 6, 31, 2]. Wei *et al.* [25] impose constraints to maintain the identity in image generation. The approach in [6] enforces the self-similarity of an image before and after translation and the domain-dissimilarity of a translated source image and a target image. Zhong *et al.* [31] propose to seek camera invariance by using unlabeled target images and their camera-style transferred counterparts as positive matching pairs. Besides, it views source and target images as negative pairs for the domain connectedness. Note that these methods have attempted to deal with the distribution discrepancy at the camera level. Differently, they reduce the discrepancy through the approach of image generation, rather than learning better representation as in this paper. As will be demonstrated

in the experiment, our approach can produce better person Re-ID performance in target domain than these methods.

3. The Proposed Framework

Our framework consists of three objectives, including i) classification of the labeled images in source domain; ii) camera-aware domain adaptation via adversarial learning; and iii) enforcing discrimination information in target domain. The first objective is not our focus and is implemented by following the literature [21, 14]. The second and third objectives are detailed in Sections 3.1 and 3.2.

3.1. Camera-aware domain adaptation

In person Re-ID, images of different cameras form a set of unique units. This is also reflected in cross-domain discrepancy. Merely reducing the overall discrepancy of source and target domains will not effectively handle the camera-level discrepancy, and this could adversely affect the quality of learned feature representation. We put forward a camera-aware domain adaptation to map the images of different cameras into a shared subspace. To achieve this, a camera-aware adversarial learning (CAL in short) method is developed. Unlike conventional adversarial learning dealing with two domains [7], CAL deals with multiple sub-domains (i.e., cameras in source and target domains).

Let X_s and X_t be the training images in source and target domains and $X = [X_s, X_t]$. The camera IDs (i.e., the label of each camera class) of the images in X are denoted by Y_c. Let C_s and C_t denote the number of the cameras in source and target domains, respectively, and $C = C_s + C_t$. Adversarial learning involves the optimization of discriminator and generator. As commonly seen, the discriminator in this work is optimized by a cross-entropy loss defined on the C camera classes in source and target domains as

$$\min_D \mathcal{L}_{\mathrm{CAL-D}}(X, Y_c, B) =$$
$$\min_D \left[\mathbb{E}_{(x,y_c)\sim(X,Y_c)} \left(-\sum_{k=1}^{C} \delta(y_c - k) \log D(B(x), k) \right) \right],$$
(1)

where x denotes an image, y_c the camera class label of x, and $\delta(\cdot)$ the Dirac delta function. B denotes the backbone network, and $B(x)$ is the feature representation of x. D denotes the discriminator and $D(B(x), k)$ denotes the prediction score for x with respect to the k-th camera class.

The generator in this work is the backbone network B. Due to the special need to deal with multiple camera classes, the optimization of B becomes tricky. We first investigate the gradient reversal layer (GRL) technique [7]. By showing its limitations for this task, we propose a new criterion "cross-domain camera equiprobability" (CCE) for our task.

GRL-based adaptation scheme. The gradient reversal layer (GRL) [7] is commonly used to reduce distribution
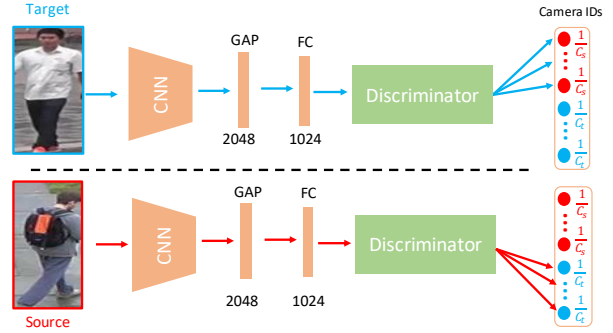

Figure 1. Illustration of training the backbone network B with the proposed cross-domain camera equiprobability (CCE) loss in our CAL method at each iteration. FC and GAP stand for fully connected layer and global average pooling. The top of the figure shows that each image in target domain (in blue) is required to be equiprobably misclassified into all camera classes in source domain (in red). The bottom similarly shows the case for each image in source domain. The discriminator is set as a 128-d FC layer.

discrepancy of two domains by *maximizing* the domain discrimination loss (i.e., Eq.(1)). A direct extension of GRL to our task leads to optimizing the generator B as

$$\min_B \mathcal{L}_{\mathrm{CAL-B}}(X, Y_c, D) \triangleq \max_B \mathcal{L}_{\mathrm{CAL-D}}(X, Y_c, D)$$
$$= \min_B \left[\mathbb{E}_{(x,y_c)\sim(X,Y_c)} \sum_{k=1}^{C} \delta(y_c - k) \log D(B(x), k) \right],$$
(2)

where for consistency it is written as minimizing the negative discriminator loss.

To train the backbone network B with Eq.(2), we insert GRL between B and D as in the literature [7]. During forward propagation, GRL is simply an identity transform. During backpropagation, GRL reverses (i.e., multiplying by a negative constant) the gradients of the domain discriminator loss with respect to the network parameters in feature extraction layers and pass them backwards. This GRL-based adaptation scheme can somehow work to reduce distribution discrepancy across different cameras (i.e., sub-domains), as will be experimentally demonstrated shortly.

However, this scheme has a drawback. Maximizing the discriminator loss only enforces an image not to be classified into its true camera class. It will appear to be "equivalently good" for this optimization as long as an image is classified into any wrong camera classes, including those from the *same* domain. In this case, this scheme will not be able to effectively pull source and target domains together. The larger the discrepancy between source and target domains is, the more pronounced this issue could be.

CCE-based adaptation scheme. In this scheme, we enforce "preferred misclassification patterns" to maximize the discriminator loss. Noting that the primary goal of cross-domain person Re-ID is to reduce the distribution discrep-

ancy of source and target domains, we require that with the learned feature representation, an image from a camera class in source domain shall be wrongly classified into a camera class in target domain and vice versa. As we do not have any bias on the camera classes in the opposite domain, it is required that the image shall be misclassified into them with equal probability, as illustrated in Fig. 1. This is where the name "cross-domain camera equiprobability" (CCE) comes from. This scheme effectively avoids considering the misclassification among the camera classes of the same domain, which is unfavourably done in the GRL scheme, and makes specific efforts to pull source and target domains together. Also, the CCE-based scheme can practically lead to alleviating the discrepancy across all cameras in target domain, as demonstrated in the experiment.

Let $\mathcal{S}$ and $\mathcal{T}$ denote source and target domains. Formally, the CCE loss on an image x can be expressed as

$$\mathcal{L}_{\text{CCE}}(x) = \begin{cases} -\frac{1}{C_s} \sum_{i=1}^{C_s} \log(D(B(x), i)), & x \in \mathcal{T} \\ -\frac{1}{C_t} \sum_{j=1}^{C_t} \log(D(B(x), j)), & x \in \mathcal{S} \end{cases} \quad (3)$$

For x in target domain, $D(B(x), i)$ denotes the predicted probability that x belongs to the ith camera class in source domain. Similar definition applies to $D(B(x), j)$ for x in source domain. In this way, the optimization for training the backbone network B (as a generator) is defined as

$$\min_B \mathcal{L}_{\text{CAL-B}}(X, D) = \min_B \mathbb{E}_{x \sim X} \mathcal{L}_{\text{CCE}}(x, D). \quad (4)$$

During adversarial training, D and B are trained in an alternate way. Each iteration consists of two steps: i) Temporarily fixing the weights of B, D is trained by Eq.(1) to predict the camera ID of each image; ii) Temporarily fixing the weights of D, B is trained by Eq.(4) to learn feature representation. This is repeated till convergence. Note that traditional two-domain adversarial learning [7, 21] is just a special case of this CCE-based scheme when there is exactly one camera class in each of source and target domains.

Remarks. A question may arise: why do we prefer the CCE criterion rather than simply requiring an image from each camera class to be equiprobably misclassified into all the other (i.e., $C_s + C_t - 1$) camera classes? This is because in person Re-ID the cross-domain discrepancy is usually more significant than the within-domain counterpart and affects the performance more. By enforcing the cross-domain camera equiprobability, the CCE criterion puts higher priority on reducing the former discrepancy and therefore is a better option. Its effectiveness and advantage will be validated in the experimental study.

Theoretical analysis of CCE. At last, we provide theoretical analysis to gain more insight on this new criterion. The full analysis is provided in the supplement material. **Proposition.** *Let $\mathcal{S}$ and $\mathcal{T}$ denote source and target domains. x^s and x^t are the images from the two domains;*

$p_s(x)$ and $p_t(x)$ are their probability density functions; and C_s and C_t are the number of camera classes in these two domains. Let $p(x|\mathcal{C}_i^s)$ and $p(x|\mathcal{C}_i^t)$ be the class-conditional density functions of the ith camera class in source and target domains, respectively. It can be proved (see supplement material) that ideally, minimizing the CCE loss will lead to

$$\begin{aligned} p(x^s|\mathcal{C}_i^t) &= p_s(x^s), \quad \forall x^s \in \mathcal{S}; \; i = 1, \cdots, C_t. \quad (5) \\ p(x^t|\mathcal{C}_i^s) &= p_t(x^t), \quad \forall x^t \in \mathcal{T}; \; i = 1, \cdots, C_s. \\ p_s(x) &= p_t(x), \quad \forall x \in \mathcal{S} \cup \mathcal{T}. \end{aligned}$$

This result indicates that in the learned shared space:

i) For any image in source domain, it will not feel the discrepancy among the C_t camera classes in target domain. Its class-conditional density function value for these camera classes (e.g., $p(x^s|\mathcal{C}_i^t)$) just equals its density function value in its own domain (e.g., $p_s(x^s)$). The above conclusion also applies to any image in target domain in a similar but reverse way.

ii) The data distributions of source and target domains, $p_s(x)$ and $p_t(x)$, will become identical and the overall cross-domain distribution discrepancy can be removed.

Meanwhile, it is worth mentioning that minimizing the CCE loss does not theoretically guarantee that an image (in either source or target domain) will not feel the distribution discrepancy among the camera classes in *its own* domain. Nevertheless, jointly considering the above three proved equalities, it can be reasonably expected (see the supplement) that the above situation could be observed in practice. This will be experimentally demonstrated shortly.

3.2. Unsupervised online triplet generation

Only reducing cross-domain distribution discrepancy is insufficient, even though the above camera-aware domain adaptation is deployed. Rather, maintaining the intrinsic properties of target domain is equally essential. Otherwise, the distribution of target domain could be arbitrarily altered just for reducing the distribution discrepancy, significantly degrading the Re-ID performance in target domain. To avoid this, this framework explores the underlying discriminative information in target domain.

This information is explored in the form of image triplet, consisting of an anchor image, a positive image (i.e., sharing the same identity as the anchor) and a negative image (i.e., having a different identity). When selecting positive and negative images, we not only consider the distance between images but also jointly utilize the temporal information among images, which can often be obtained via the frame ID in person Re-ID. More importantly, we generate triplets online in each batch during training. This allows triplet generation to effectively take advantage of the steadily improved feature representation to produce better triplets. Note that the triplet generation is carried out in an unsupervised manner and only needed in training process.

Given a camera in target domain, all of its images are sorted temporally into a list. From this list, p non-overlapping fragments are randomly selected to construct a batch. Each fragment consists of q images, and the batch therefore contains $n (= p \times q)$ images in total. Each of the n images is used as an anchor image to create triplets in turn. To begin with, a pairwise distance matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ is computed for the n images, with the feature representation learned by the network so far. To generate triplets for an anchor image I_a, we develop the following rules.

Above all, according to $\mathbf{M}$, sort all the $(n-1)$ images (excluding I_a) in the batch in ascending order of the distance from I_a. The obtained list is denoted by $S(I_a)$.

Positive image selection. To be selected as positive, an image must meet both the requirements: i) it is within the top-k positions of $S(I_a)$, and ii) it is from the same fragment as I_a. The first requirement ensures that this image is indeed similar to I_a in terms of feature representation, while the second one further increases its likelihood of positiveness with temporal information. Jointly using these two requirements helps us to select highly likely (not guaranteed though) true positive images. In implementation, k is empirically set. The total number of selected positive images is denoted by k_p. Note that k_p could be zero, meaning that this anchor cannot find any positive images by the above rule. In this case, this anchor will not be taken into account.

Negative image selection. Starting from the head of the list $S(I_a)$, each image I is checked in turn against the following conditions: 1) I is *not* from the same fragment of I_a, and 2) no image in the fragment of I has previously been selected as negative. That is, the negative images are selected as the nearest neighbours of the anchor from the fragments *other than* the anchor's, with the *condition* that each of these negative images shall be, respectively, collected from a *different* fragment. Such a rule is designed to deal with the case that the same person may reappear in two or more fragments. Requiring each negative sample to reside in different fragments well reduces (although cannot entirely avoid) the chance to mis-select a truly positive sample as negative. The total number of selected negative images is denoted by k_n.

Once triplets are generated in a training batch, we can train the backbone network via a triplet loss defined as

$$\mathcal{L}_{\text{Triplet}} = \sum_{a=1}^{n} w_a [\bar{d}_p(I_a) - \bar{d}_n(I_a) + m]_+, \qquad (6)$$

where w_a is zero if I_a has no positive images and one otherwise. $[t]_+$ equals t if $t > 0$ and zero otherwise. m is the margin. $\bar{d}_p(I_a) = \frac{1}{k_p} \sum_{i=1}^{k_p} d(I_a, I_p^i)$ and $\bar{d}_n(I_a) = \frac{1}{k_n} \sum_{i=1}^{k_n} d(I_a, I_n^i)$ are the average distances of the positive and negative samples from the anchor, respectively. Using average distances here helps to mitigate the adverse effect
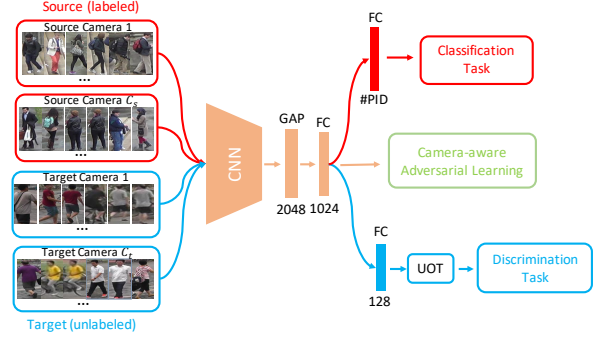

Figure 2. Illustration of the proposed unsupervised camera-aware domain adaptation framework, where FC and GAP denote fully connected layer and global average pooling. #PID and UOT denote the total number of person classes in source domain and unsupervised online triplet generation, respectively.

in case any positive or negative sample is wrongly selected. $d(\cdot, \cdot)$ is just the distance used for the distance matrix $\mathbf{M}$.

Finally, although $\mathbf{M}$ can be simply calculated by Euclidean distance, more advanced measures can be readily used. This work uses a re-ranking algorithm [30] to improve $\mathbf{M}$, in order to generate even better triplets and further improve the person Re-ID performance in target domain.

3.3. The overall proposed framework

Recall that this framework consists of an adversarial task across the cameras (sub-domains) in source and target domains, a discrimination task for target domain, and a classification task for source domain, as shown in Fig. 2. The overall loss function in a training batch is expressed as

$$\min_D \mathcal{L}_{\text{CAL}-D}(X, Y_c, B) =$$
$$\min_D \left[\mathbb{E}_{(x,y_c) \sim (X,Y_C)} \left(-\sum_{k=1}^{C} \delta(y_c - k) \log D(B(x), k) \right) \right],$$
$$\min_B \mathcal{L}(X, Z_s, D) =$$
$$\min_B \left(\mathcal{L}_{\text{Cross}}(X_s, Z_s) + \lambda_1 \mathcal{L}_{\text{Triplet}}(X_t) + \lambda_2 \mathcal{L}_{\text{CAL}-B}(X, D) \right),$$
$$(7)$$

where $\mathcal{L}_{\text{Cross}}$, $\mathcal{L}_{\text{Triplet}}$ and $\mathcal{L}_{\text{CAL}-B}$ are the cross-entropy loss for source domain, the triplet loss for target domain, and the adversarial loss for B. λ_1 and λ_2 are the trade-off parameters. Z_s is the person IDs of X_s in source domain. To caculate $\mathcal{L}_{\text{Triplet}}$, one camera in target domain is randomly chosen to construct the training batch and generate triplets at each iteration.

In this framework, ResNet-50 [9] is used as backbone network. Global average pooling (GAP) is used to obtain the 2048-d feature representation. To do person Re-ID in target domain, the 2048-d feature representation is extracted for each query and gallery images, and an L_2 normalization is applied. Euclidean distance is calculated to rank the gallery images for a query.

4. Experiments

4.1. Datasets and settings

We evaluate our approach on two large-scale benchmark datasets: Market1501 [28], DukeMTMC-reID (Duke in short) [29]. **Market1501** contains 1,501 persons with 32,668 images from six cameras. Among them, 12,936 images of 751 identities are used for training. For evaluation, there are 3,368 and 19,732 images in the query and gallery sets. **DukeMTMC-reID** has 1,404 persons from eight cameras, with 16,522 training images of 702 identities, 2,228 queries, and 17,661 gallery images. Both camera ID and Frame ID information are available on Market1501 and DukeMTMC-reID. To evaluate person Re-ID performance, we use Rank-1 accuracy and mAP [28]. On Market1501, there are single- and multi-query evaluation protocols. We use the more challenging single-query protocol.

For training CAL, we randomly select 64 images from each of source and target domains in a batch. The 64 images from source domain are also used to train the classification component. To generate triplets, we set p (i.e., number of temporal fragments) and q (i.e., number of images per fragment) to 12 and 10. k and k_n used in the selection of positive and negative samples are set as 5 and 2. The margin of triplet loss, m, is 0.3. λ_1 and λ_2 in Eq.(7) are set as 1.

The proposed model is trained with the SGD optimizer in a total of 100 epochs. The initial learning rates of the fine-tuned parameters (those in the pre-trained ResNet-50 on ImageNet [5]) and the new parameters (those in the newly added layers) are 0.1 and 0.01, respectively. When the number of epochs exceeds 80, we decrease the learning rates by a factor of 0.1. The size of the input images is 256×128.

Note that the **baseline** model (denoted by BL) in this experiment represents the ResNet-50 [9] with additional 1024-d fully connected (FC) layer and cross-entropy loss. UOT and UOT(eud) denote two unsupervised online triplet generation variants in which the aforementioned re-ranking algorithm [30] (our default setting) and Euclidean distance are used to compute the distance matrix $\mathbf{M}$, respectively. CAL-GRL and CAL-CCE denote the proposed camera-aware adversarial learning (CAL) implemented by the GRL and CCE schemes, respectively. For clarity, we use UCDA-GRL[2] and UCDA-CCE to represent BL+UOT+CAL-GRL and BL+UOT+CAL-CCE, respectively. Note that for the proposed framework, camera IDs and frame IDs are only needed in training, but not in the test stage.

4.2. Comparison with the state-of-the-art methods

We compare our approach with seven state-of-the-art unsupervised cross-domain person Re-ID approaches. Among them, there are two non-deep-learning-based methods (UMDL [19] and UJSDL [20]) and five deep-learning-

based methods. The latter includes two recent pseudo-label-generation-based methods (TFusion [16] and TJ-AIDL [23]) and three recent image-generation-based approaches (PTGAN [25], SPGAN [6] and HHL [31]). Market1501 and Duke are in turn used as source and target domains to compare these methods. Table 1 reports the result. As seen, our approach (i.e., UCDA-CCE) consistently achieves the best results in both settings. For the setting of "Duke→Market1501", it obtains 34.5% and 64.3% in mAP and Rank-1 accuracy, outperforming all the methods in comparison. In the setting of "Market1501→Duke", UCDA-CCE still excels. Particularly, compared with HHL [31], the state-of-the-art by using image generation to reduce the camera-level discrepancy in target domain, UCDA-CCE gains 9.5% (36.7 vs. 27.2) in mAP and 8.5% (55.4 vs. 46.9) in Rank-1 accuracy.

This result shows the advantage of our approach over the pseudo-label-generation-based and image-generation-based methods. It effectively alleviates the data distribution discrepancy at the camera level through representation learning. We will demonstrate this property in Section 4.4.

Table 1. Comparison with the state-of-the-art methods of unsupervised Re-ID on Market1501 and DukeMTMC-reID (Duke).

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
UMDL [19]	12.4	34.5	7.3	18.5
UJSDL [20]	-	50.9	-	32.2
TFusion [16]	-	60.8	-	-
TJ-AIDL [23]	26.5	58.2	23	44.3
PTGAN [25]	-	38.6	-	27.2
SPGAN+LMP [6]	26.7	57.7	26.2	46.4
HHL [31]	31.4	62.2	27.2	46.9
UCDA-GRL	30.9	60.4	31.0	47.7
UCDA-CCE (Ours)	**34.5**	**64.3**	**36.7**	**55.4**

Our approach can be easily extended to semi-supervised person Re-ID. That is, when within each camera in target domain, the labels of person identity become available for each frame (e.g., obtained by tracking algorithms or manual annotation), our approach can utilize this information by simply changing our unsupervised online triplet generation (UOT) to traditional triplet generation [10]. We name this setting SOT where "S" stands for semi-supervised.[3] Our approach is compared with TAUDL [12], a recent state-of-the-art method for video-based person Re-ID. Note that TAUDL utilizes the person identity labels within each camera in target domain but does not employ cross-domain adaptation from a source domain. This comparison aims to show that via CAL, our approach can effectively utilize source domain to produce better Re-ID features in target domain than TAUDL. This is validated in Table 2. Compared with TAUDL, our approach (BL+SOT+CAL-GRL/CAL-CCE) achieves considerable improvement in both tasks of

---

[2]UCDA is short for Unsupervised Camera-aware Domain Adaptation.

[3]This setting is called semi-supervised because person identity labels are only available *within* each individual camera but not across all cameras.

"Market1501→Duke" and "Duke→Market1501".

Table 2. Comparison with the state-of-the-art approach of semi-supervised person Re-ID on Market-1501 and Duke.

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
TAUDL [12]	41.2	63.7	43.5	61.7
BL+SOT+CAL-GRL	46.6	72.2	44.3	62.0
BL+SOT+CAL-CCE (Ours)	**49.6**	**73.7**	**45.6**	**64.0**

Table 3. Comparison of camera-aware adversarial learning (CAL) and domain-aware adversarial learning (DAL) in unsupervised and semi-supervised settings on Market-1501 and Duke.

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
BL+UOT+DAL	25.5	54.1	26.2	42.4
BL+UOT+CAL-GRL	30.9	60.4	31.0	47.7
BL+UOT+CAL-CCE (Ours)	**34.5**	**64.3**	**36.7**	**55.4**
BL+SOT+DAL	40.2	67.3	34.8	52.2
BL+SOT+CAL-GRL	46.6	72.2	44.3	62.0
BL+SOT+CAL-CCE (Ours)	**49.6**	**73.7**	**45.6**	**64.0**

To validate the effectiveness of the proposed camera-aware adversarial learning (CAL), we compare our approach with the state-of-the-art domain adaptation methods. They use domain-aware adversarial learning (DAL) [7, 21] to conduct adversarial learning between source and target domains only. We implement DAL by ourselves to ensure a fair comparison. As in Table 3, DAL (BL+UOT+DAL and BL+SOT+DAL) is inferior to the proposed CAL in both unsupervised and semi-supervised settings. There is a large gap between them in both experiments. This result further demonstrates the advantage of CAL by considering the discrepancy across all camera-level sub-domains.

4.3. On the effectiveness of CAL and UOT

In the following ablation studies, we provide more details on the effectiveness of the two proposed components: camera-aware adversarial learning (CAL) and unsupervised online triplet generation (UOT) in Tables 4 and 5.

Effectiveness of CAL. First, BL+UOT+CAL-CCE consistently outperforms BL+UOT+CAL-GRL in Table 4. The former improves 3.9% (64.3 vs. 60.4) and 7.7% (55.4 vs. 47.7) on Rank-1 accuracy in "Duke→Market1501" and "Market1501→Duke". This shows that the proposed CCE-based scheme can overcome the drawback of the GRL-based one, as analyzed in Section 3.1. Second, incorporating the proposed CAL (via either BL+UOT+CAL-CCE or BL+UOT+CAL-GRL in Table 4) greatly improves over BL+UOT. This validates the effectiveness of CAL in helping reduce camera-level distribution discrepancy to learn better feature representation. Also, we validate the effectiveness of CAL in the semi-supervised setting. As seen in Table 5, BL+SOT+CAL-CCE improves BL+SOT by 10.4% (49.6 vs. 39.2) and 6.4% (45.6 vs. 39.2) on mAP in "Duke→Market1501" and "Market1501→Duke".

Effectiveness of UOT. Adding the proposed unsupervised online triplets into the baseline (i.e., BL+UOT) clearly

improves the baseline (BL), as seen in Table 4. This confirms the benefit of exploiting discriminative information from target domain via the proposed UOT. In addition, we test two schemes to compute the distance matrix $\mathbf{M}$ in Section 3.2, via Euclidean distance (BL+UOT(eud)) and the default re-ranking algorithm (BL+UOT), respectively. As seen, BL+UOT does achieve better performance.

In addition, we are interested in what if CAL is used alone without the UOT component. Both BL+CAL-GRL and BL+CAL-CCE are investigated at the bottom of Table 4. As seen, merely using CAL is insufficient. They do not show sufficient improvement over BL+UOT or even BL+UOT(eud). BL+CAL-CCE even fails in the case of "Duke→Market1501", although showing some improvement on "Market1501→Duke". This result is explained as follows. Duke has more cameras than Market1501. Also, in the literature, when the same Re-ID model is applied to them, the performance on Duke is usually inferior to that on Market1501 [13, 1, 29, 11]. This shows that Market1501 is a relatively easier dataset than Duke. When Market1501 is target domain and the more challenging Duke is source domain, its distribution could be altered significantly in order to fit the Duke's, if no discriminative information from target domain is used as a regularizer. Meanwhile, the situation will become less critical in the other way round. This is because Duke's data distribution is more complicated and therefore has sufficient "capacity" to fit Market1501's.

The above result clearly shows the necessity of UOT in our framework. Working together, the two proposed components produce the best performance in Table 4.

Table 4. Performance of the proposed framework when using different components on Market1501 and DukeMTMC-reID (Duke).

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
BL	19.4	47.1	21.3	38.4
BL+UOT(eud)	23.6	51.0	24.1	40.2
BL+UOT	27.4	55.5	27.5	44.3
BL+UOT+CAL-GRL	30.9	60.4	31.0	47.7
BL+UOT+CAL-CCE (Ours)	**34.5**	**64.3**	**36.7**	**55.4**
BL+CAL-GRL	20.5	47.6	22.7	41.4
BL+CAL-CCE	8.4	27.6	23.8	45.4

Table 5. Effectiveness of CAL in the semi-supervised setting on Market-1501 and DukeMTMC-reID (Duke).

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
BL+SOT	39.2	65.9	39.2	56.6
BL+SOT+CAL-GRL	46.6	72.2	44.3	62.0
BL+SOT+CAL-CCE (Ours)	**49.6**	**73.7**	**45.6**	**64.0**

4.4. Further evaluation of the proposed framework

We examine the inter-domain (between source and target domains) and inter-camera (across all cameras in target domain) discrepancy to validate the effectiveness of CAL, as reported in Table 6. Since our goal is to obtain better feature representation for target domain, we focus on target domain

Table 6. The data distribution discrepancy of inter-domain (between source and target domains) and inter-camera (across all cameras in target domain) on the task of Duke→Market1501. Note that a smaller value indicates better performance in this table.

	BL	DAL	CAL-GRL	CAL-CCE
Inter-domain ($\times 10^3$)	1.48	1.22	1.28	1.25
Inter-camera ($\times 10^2$)	6.76	5.97	4.36	2.90

when examining the discrepancy across all cameras. In this experiment, we measure the inter-domain discrepancy by the distance $d_{\mathrm{inter-domain}} = \left\| \overline{X}_s - \overline{X}_t \right\|_2$, where $\overline{X}_s$ and $\overline{X}_t$ denote the sample mean of source and target domains, respectively. To measure the inter-camera discrepancy, we use $d_{\mathrm{inter-camera}} = \frac{1}{C_t} \sum_{c=1}^{C_t} \left\| \overline{X}_{t,c} - \overline{X}_t \right\|_2$, where $\overline{X}_{t,c}$ is the sample mean of the cth camera class in target domain and C_t is the total number of cameras in target domain. These distances are calculated in Table 6. First, for the inter-domain discrepancy, DAL, CAL-GRL and CAL-CCE are all smaller than BL. This validates that they are all able to reduce the discrepancy between source and target domains. Particularly, since DAL specifically focuses on the overall domain-level discrepancy, its distance is smaller than those of CAL-GRL and CAL-CCE. In addition, CAL-GRL has a slightly larger value than CAL-CCE. This is consistent with the analysis on the drawback of CAL-GRL in Section 3.1. Second, for the inter-camera discrepancy, both CAL-GRL and CAL-CCE achieve smaller distances than DAL because DAL does not consider camera-level discrepancy. Besides, this experiment shows that CAL-CCE achieves the smallest distance, showing its best capability in reducing the discrepancy across the cameras in target domain. Additionally, we visualize the data distributions obtained by the feature representation from BL and CAL-CCE in Fig. 3. The result further illustrates the effectiveness of CAL-CCE.

Table 7. Comparison of the CCE-based scheme and the AOE-based scheme on Market1501 and DukeMTMC-reID (Duke).

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
AOE	29.6	59.9	31.5	51.3
CCE (Ours)	**34.5**	**64.3**	**36.7**	**55.4**

As analyzed in the *Remarks* of Section 3.1, why do we prefer the CCE-based scheme rather than equiprobably mixing a camera class with all the other ($C_s + C_t - 1$) ones? In this experiment, we name this setting AOE, standing for "all others equiprobability." We compare AOE and CCE in Table 7. As shown, the CCE-based scheme outperforms the AOE-based scheme in both tasks. This is consistent with our previous analysis. It indicates that during reducing camera-level discrepancy, giving the reduction of cross-domain discrepancy higher priority is beneficial because it is usually more significant than the within-domain counterpart and affects the performance more.

Finally, we validate the advantage of the unsupervised *online* triplet (UOT) generation by comparing it with the

(a) Baseline (BL)   (b) CAL-CCE (Ours)

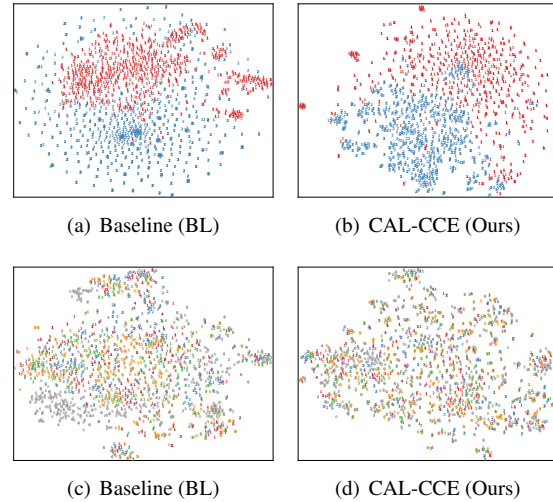(c) Baseline (BL)   (d) CAL-CCE (Ours)

Figure 3. Visualization of data distributions at the domain-level and camera-level via t-SNE [18]. The features of each image are extracted by the baseline (BL) and CAL-CCE in the task of "DukeMTMC-reID→Market1501", respectively. The top shows the distributions of source and target domains (i.e., inter-domain).The bottom illustrates the distribution of each camera class in target domain (i.e., inter-camera on Market1501), where different colors denote different camera classes. As seen, the proposed CAL-CCE effectively "mixes" the two domains and the camera classes as expected. Illustration of the methods of DAL and CAL-GRL is in the supplement material.

offline way. To generate offline triplets, we use the same method in Section 3.2 with the mere difference that they are generated by the feature from the baseline model before training starts. In total, 88,719 and 100,742 triplets are generated on Market1501 and Duke. We randomly select 40 triplets in each batch (i.e., 120 samples, giving the same batch size as the online method) to train our model. As seen in Table 8, offline method is clearly inferior. It confirms the advantage of online method by utilizing the steadily improved features in training, as discussed in Section 3.2.

Table 8. Comparison of offline and online triplet generation on Market-1501 and DukeMTMC-reID (Duke).

Method	Duke→Market1501		Market1501→Duke	
	mAP	Rank-1	mAP	Rank-1
Offline Triplets	13.5	37.1	10.2	21.1
Online Triplets (Ours)	**27.4**	**55.5**	**27.5**	**44.3**

5. Conclusion

This paper proposes a novel deep domain adaptation framework to address two key issues in unsupervised cross-domain person Re-ID. It clearly shows that when pursuing better feature representation for person Re-ID, considering camera-level domain discrepancy is beneficial. Also, exploring discriminative information from unlabeled target domain is equally, if not more, important. Only when these two components are adequately resolved, unsupervised cross-domain person Re-ID can become promising.

References

[1] Ejaz Ahmed, Michael J. Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 7

[2] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[3] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2

[4] Kuan-Wen Chen, Chih-Chuan Lai, Pei-Jyun Lee, Chu-Song Chen, and Yi-Ping Hung. Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras. *IEEE Transactions Multimedia (TMM)*, 2011. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 6

[6] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6

[7] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, 2015. 2, 3, 4, 7

[8] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 6

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 6

[11] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 7

[12] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *European Conference on Computer Vision (ECCV)*, 2018. 6, 7

[13] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 7

[14] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015. 2, 3

[15] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 2

[16] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[17] Andy Jinhua Ma, Jiawei Li, Pong C. Yuen, and Ping Li. Cross-domain person reidentification using domain adaptation ranking svms. *IEEE Transactions on Image Processing (TIP)*, 2015. 2

[18] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 2008. 8

[19] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and YongHong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6

[20] Lei Qi, Jing Huo, Xiaocong Fan, Yinghuan Shi, and Yang Gao. Unsupervised joint subspace and dictionary learning for enhanced cross-domain person re-identification. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 2018. 1, 2, 6

[21] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4, 7

[22] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv*, 2014. 2

[23] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. *arXiv*, 2018. 2, 6

[24] Xiaojuan Wang, Wei-Shi Zheng, Xiang Li, and Jianguo Zhang. Cross-scenario transfer person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2016. 1, 2

[25] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6

[26] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. *arXiv*, 2018. 2

[27] Xu Zhang, Felix X. Yu, Shih-Fu Chang, and Shengjin Wang. Deep transfer network: Unsupervised domain adaptation. *arXiv*, 2015. 2

[28] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 6

[29] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 6, 7

[30] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Reranking person re-identification with k-reciprocal encoding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 6

[31] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6