AD-Cluster: Augmented Discriminative Clustering for Domain Adaptive Person Re-identification

Yunpeng Zhai[1,2], Shijian Lu[3], Qixiang Ye[4,6], Xuebo Shan[1,2], Jie Chen[1,6],
Rongrong Ji[5,6], Yonghong Tian[1,2,6*]
[1]School of Electronic and Computer Engineering, Peking University, China
[2]NELVT, School of EE&CS, Peking University, Beijing, China
[3]Nanyang Technological University, Singapore, [4]University of Chinese Academy of Sciences, China
[5]Xiamen University, China, [6]Peng Cheng Laboratory, China
{ypzhai, shanxb, yhtian}@pku.edu.cn, shijian.lu@ntu.edu.sg, qxye@ucas.ac.cn,
chenj@pcl.ac.cn, rrji@xmu.edu.cn

Abstract

Domain adaptive person re-identification (re-ID) is a challenging task, especially when person identities in target domains are unknown. Existing methods attempt to address this challenge by transferring image styles or aligning feature distributions across domains, whereas the rich unlabeled samples in target domains are not sufficiently exploited. This paper presents a novel augmented discriminative clustering (AD-Cluster) technique that estimates and augments person clusters in target domains and enforces the discrimination ability of re-ID models with the augmented clusters. AD-Cluster is trained by iterative density-based clustering, adaptive sample augmentation, and discriminative feature learning. It learns an image generator and a feature encoder which aim to maximize the intra-cluster diversity in the sample space and minimize the intra-cluster distance in the feature space in an adversarial min-max manner. Finally, AD-Cluster increases the diversity of sample clusters and improves the discrimination capability of re-ID models greatly. Extensive experiments over Market-1501 and DukeMTMC-reID show that AD-Cluster outperforms the state-of-the-art with large margins.

1. Introduction

Person re-identification (re-ID) aims to match persons in an image gallery collected from non-overlapping camera networks. Despite of the impressive progress of supervised methods in person re-ID [5] [50], models trained in one domain often fail to generalize well to others due to the change of camera configurations, lighting conditions, person views, etc. Domain adaptive re-ID methods that can work across domains remain a very open research challenge.

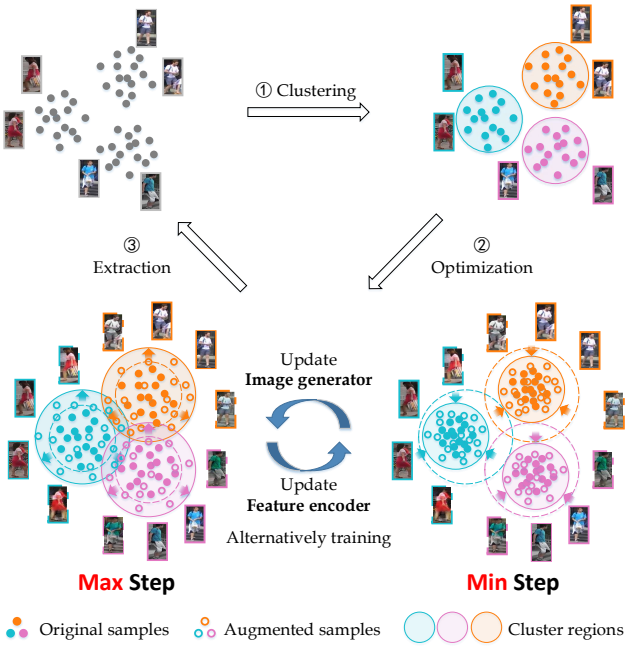


Figure 1. AD-Cluster alternatively trains an image generator and a feature encoder, which respectively Maximizes intra-cluster distance (*i.e.*, increase the diversity of sample space) and Minimizes intra-cluster distance in feature space (*i.e.*, decrease the distance in new feature space). It enforces the discrimination ability of re-ID models in an adversarial min-max manner. (Best viewed in color)

To implement domain adaptive re-ID, unsupervised domain adaptation (UDA) methods have been widely explored [44], [26], [27], [10], [45], [61], [32], [14], [52],

*Corresponding author.

[29]. One major line of UDA methods attempts to align the feature distributions of source and target domains [44], [26]. Another line of methods utilizes adversarial generative models as a style transformer to convert pedestrian images (with identity annotations) of a source domain into a target domain [27], [10], [45], [32]. The style-transferred images are then used to train a re-ID model in the target domain. Many UDA methods preserve discriminative information across domains or camera styles, but they largely ignore the unlabeled samples and so the substantial sample distributions in target domains. Recent approaches [14], [47] alleviate this problem by predicting pseudo-labels in target domains. They leverage the cluster (pseudo) labels for model fine-tuning directly but are often susceptible to noises and hard samples. This prevents them from maximizing model discrimination capacity in target domains.

In this paper, we propose an innovative augmented discriminative clustering (AD-Cluster) technique for domain adaptive person re-ID. AD-Cluster aims to maximize model discrimination capacity in the target domain by alternating discriminative clustering and sample generation as illustrated in Fig. 1. Specifically, density-based clustering first predicts sample clusters in the target domain where sample features are extracted by a re-ID model that is pre-trained in the source domain. AD-Cluster then learns through two iterative processes. First, an image generator keeps translating the clustered images to other cameras to augment the training samples while retaining the original pseudo identity labels (i.e. cluster labels). Second, a feature encoder keeps learning to maximize the inter-cluster distance while minimizing the intra-cluster distance in feature space. The image generator and the feature encoder thus compete in an adversarial min-max manner which iteratively estimate cluster labels and optimize re-ID models. Finally, AD-Cluster aggregates the discrimination ability of re-ID models through such adversarial learning and optimization.

The main contributions of this paper can be summarized in three aspects. First, it proposes a novel discriminative clustering method that addresses domain adaptive person re-ID by density-based clustering, adaptive sample augmentation, and discriminative feature learning. Second, it designs an adversarial min-max optimization strategy that increases the intra-cluster diversity and enforces discrimination ability of re-ID models in target domains simultaneously. Third, it achieves significant performance gain over the state-of-the-art on two widely used re-ID datasets: Market-1501 and DukeMTMC-reID.

2. Related Works

While person re-ID has been extensively investigated from various perspectives, we mainly review the domain adaptive person re-ID approaches, which are largely driven by unsupervised domain adaptation (UDA) methods.

2.1. Unsupervised Domain Adaptation (UDA)

Domain alignment. UDA defines a learning problem where source domains are fully labeled while sample labels in target domains are totally unknown. To learn discriminative modes in target domains, early methods focus on learning feature/sample mapping between source and target domains [38], [42]. As an representative method, correlation alignment (CORAL) [42] pursued minimizing domain shift by aligning the mean and co-variance of source and target distributions. Recent methods [22], [2], [28] attempted reducing the domain shift by using generative adversarial networks (GANs) to learn a pixel-level transformation. The most representative CYCADA [22] transferred samples across domains at both pixel- and feature-level.

Domain-invariant features. The second line of UDA methods focuses on finding domain-invariant feature spaces [33], [31], [16], [30], [43], [17], [1]. To fulfill this purpose, Long *et al.* [30], [19] proposed the Maximum Mean Discrepancy (MMD), which maps features of both domains into the same Hilbert space. Ganin *et al.* [17] and Ajakan *et al.* [1] designed domain confusion loss to learn domain-invariant features. Saito *et al.* [39] proposed aligning distributions of source and target domains by maximizing the discrepancy of classifiers' outputs.

Pseudo-label prediction. Another line of UDA methods involves learning representations in target domains by using the predicted pseudo-label. In general, this approach uses an alternative estimation strategy: predicting pseudo-labels of samples by simultaneous modelling and optimizing the model using predicted pseudo-labels [4], [37], [40], [54]. In the deep learning era, clustering loss has been designed for CNNs and jointly learning of features, image clusters, and re-ID models in an alternative manner [8], [51], [49], [11], [24], [3], [18].

2.2. UDA for Person re-ID

To implement domain adaptive person re-ID, researchers largely referred to the above reviewed UDA methods by incorporating the characteristics of person images.

Domain alignment. In [26], Lin *et al.* proposed minimizing the distribution variation of the source's and the target's mid-level features based on Maximum Mean Discrepancy (MMD) distance. Wang *et al.* [44] utilized additional attribute annotations to align feature distributions of source and target domains in a common space. Other works enforced camera in-variance by learning consistent pairwise similarity distributions [46] or reducing the discrepancy between both domains and cameras [35].

GAN-based methods have been extensively explored for domain adaptive person re-ID [32], [61], [45], [10], [27]. HHL [61] simultaneously enforced cameras invariance and domain connectedness to improve the generalization ability of models on the target set. PTGAN [45], SPGAN [10],

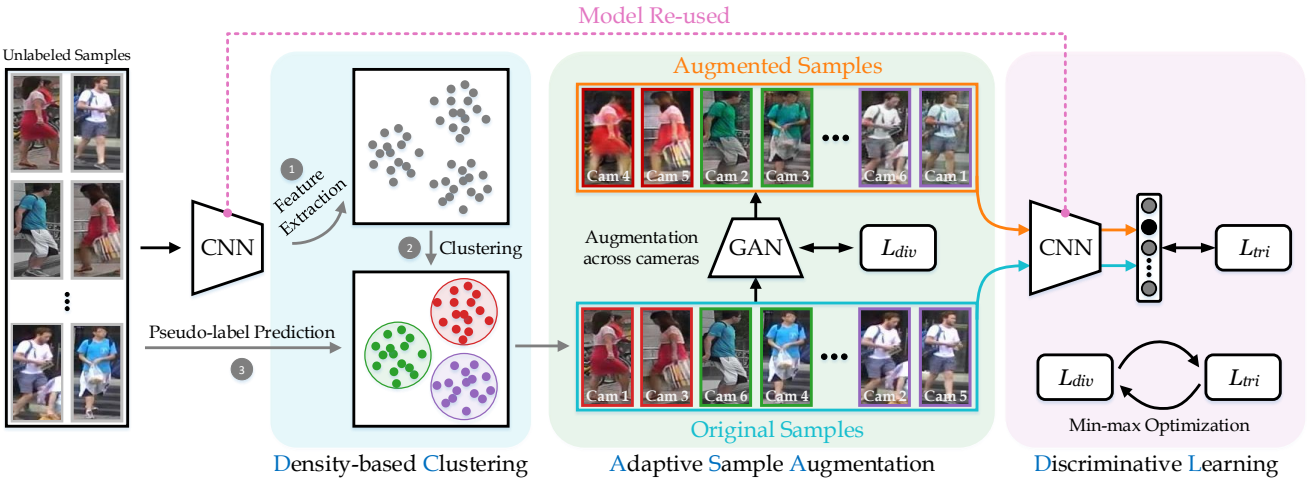


Figure 2. The flowchart of the proposed AD-Cluster: The AD-Cluster consists of three components including density-based clustering, adaptive sample augmentation, and discriminative feature learning. Density-based clustering estimates sample pseudo-labels in the target domain. Adaptive sample augmentation maximizes the sample diversity cross cameras while retaining the original pseudo-labels. Discriminative learning drives the feature extractor to minimize the intra-cluster distance. L_{div} denotes the diversity loss and L_{tri} indicates the triplet loss. (Best viewed in color)

ATNet [27], CR-GAN [6] and PDA-Net [23] transferred images with identity labels from source into target domains to learn discriminative models.

By aligning feature and/or appearance, the above methods can preserve well the discriminative information from source domains; however, they largely ignore leveraging the unlabeled samples in target domains, which hinder them from maximizing the model discrimination capacity.

Pseudo-label prediction. Recently, the problem about how to leverage the large number of unlabeled samples in target domains has attracted increasing attention [14], [52], [29], [47], [48], [62]. Clustering [14], [57], [55], [15] and graph matching [52] methods have been explored to predict pseudo-labels in target domains for discriminative model learning. Reciprocal search [29] and exemplar-invariance approaches [48] were proposed to refine pseudo labels, taking camera-invariance into account concurrently.

Existing approaches have explored cluster distributions in the target domain. On the other hand, they still face the challenge on how to precisely predict the label of hard samples. The hard/difficult samples are crucial to a discriminative re-ID model but they often confuse clustering algorithms. We address this issues by iteratively generating and including diverse and representative samples in the target domain, which enforces the discrimination capability of re-ID models effectively.

3. The Proposed Approach

Under the context of unsupervised domain adaptation (UDA) for person re-ID, we have a fully labeled source domain $\{X_s, Y_s\}$ that contains N_s person images of M identities in total in the source domain. X_s and Y_s denote the sample images and identities in the source domain, respectively, where each image $x_{s,i}$ is associated with an identity $y_{s,i}$. In addition, we have an unlabeled target domain $\{X_t\}$ that contains N_t person images. The identities of images in the target domain are unavailable. The goal of AD-Cluster is to learn a re-ID model that generalizes well in the target domain by leveraging labeled samples in the source domain and unlabeled samples in the target domain.

3.1. Overview

AD-Cluster consists of two networks including a CNN as the feature encoder f and a Generative Adversarial Network (GAN) as the image generator g as shown in Fig. 2. The encoder f is first trained using labeled samples in the source domain with cross-entropy loss and triplet loss [21]. In the target domain, unlabelled sample are represented by features that are extracted by f, where density-based clustering groups them to clusters and uses the cluster IDs as the pseudo-labels of the clustered samples. With each camera being a new domain with different styles, g translates each sample of the target domain to other cameras and this generates identity-preserving samples with increased diversity. After that, all samples in the target domain together with those generated are fed to re-train the feature encoder f. The generator g and encoder f thus learn in an adversarial min-max manner iteratively, where g keeps generating identity-preservative samples to maximize the intra-cluster variations in the sample space whereas f learns discriminative representation to minimize the intra-cluster variations

in the feature space as illustrated in Fig. 1.

3.2. UDA Procedure

Supervised learning in source domain: In the source domain, the CNN-based person re-ID model is trained by optimizing classification and ranking loss [21]:

$$\mathcal{L}_{src} = \mathcal{L}_{cls} + \mathcal{L}_{tri}. \tag{1}$$

For a batch of samples, the classification loss is defined by

$$\mathcal{L}_{cls} = -\frac{1}{n_s}\sum_{i=1}^{n_s} \log p(y_{s,i}|x_{s,i}), \tag{2}$$

where n_s, i and s denote the number of images in a batch, image index and source domain, respectively. $p(y_{s,i}|x_{s,i})$ is the predicted probability of image $x_{s,i}$ belonging to $y_{s,i}$.

The ranking triplet loss is defined as

$$\begin{aligned}\mathcal{L}_{tri} = \sum_{i=1}^{n_s}[m + \|f(x_{s,i}) - f(x_{s,i^+})\|_2 \\ - \|f(x_{s,i}) - f(x_{s,i^-})\|_2],\end{aligned} \tag{3}$$

where x_{s,i^+} denotes the samples belonging to the same person with $x_{s,i}$. x_{s,i^-} denotes the samples belonging to different persons with $x_{s,i}$. m is a margin parameter [21].

Density-based clustering in target domain: In each learning iteration, density-based clustering [12] is employed in the target domain for pseudo-label prediction. The clustering procedure includes three steps: (1) Extracting convolutional features for all person images. (2) Computing a distance matrix with k-reciprocal encoding [60] for all training samples and then performing density-based clustering to assign samples into different groups. (3) Assigning pseudo-labels Y_t' to the training samples X_t according to the groups they belong to.

Adaptive sample augmentation across cameras: Due to the domain gap, the pseudo-labels predicted by density-based clustering suffer from noises. In addition, the limited number of training samples in the target domain often leads to the low diversity of samples in each cluster. These two factors make it difficult to learn discriminative representation in the target domain.

To address these issues, we propose to augment samples in the target domain with a GAN to aggregate sample diversity. The used GAN should possess the following two properties: (1) Generating new person images from existing ones while preserving the original identities; (2) Providing additional invariance such as camera configurations, lighting conditions, and person views.

To fulfill these purposes, we employ StarGAN [7] to augment person images which can preserve the person identities while generating new images in multiple camera styles. The image generation procedure roots in the results of density-based clustering. Suppose there are K cameras in total in the target domain. A StarGAN model is first trained which enables image-image translation between each camera pair. Using the learned StarGAN model, for an image $x_{t,i}$ with pseudo-label $y_{t,i}$, we generate K augmented images $\{x_{t,i}^{(1)}, y_{t,i}\}, \{x_{t,i}^{(2)}, y_{t,i}\}, ..., \{x_{t,i}^{(K)}, y_{t,i}\}$, which have the pseudo-label $y_{t,i}$ with $x_{t,i}$ and similar styles as the images in camera $1, 2, ..., K$, respectively. In this way, the sample number in each cluster increases by a factor of $K-1$. The augmented images together with original images in target domain are used for discriminative feature learning, according to Eq. 3.

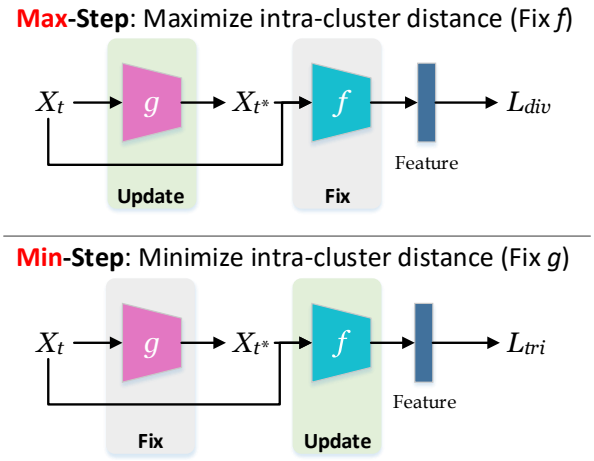


Figure 3. The proposed adversarial min-max learning: With a fixed feature encoder f, the generator g learns to generate samples that maximizes intra-cluster distance. With a fixed generator g, the feature encoder f learns to minimize the intra-cluster distance and maximize the inter-cluster distance under the guide of triplet loss.

3.3. Min-Max Optimization

Although the adaptive sample augmentation enforces the discrimination ability of re-ID models, the sample generation procedure is completely independent from the clustering and feature learning which could lead to insufficient sample diversity across cameras.

To fuse the adaptive data augmentation with discriminative feature learning, we propose an adversarial min-max optimization strategy as illustrated in Fig. 3. Specifically, we alternatively train an image generator and a feature encoder that maximize sample diversity and minimize intra-cluster distance for each mini-batch, respectively.

Max-Step: Star-GAN [7] is employed as an image generator (g) for a given feature encoder (f). In the procedure, the summation of Euclidean distances between samples and their cluster centers is defined as cluster diversity $\mathcal{D}_{div}$. For

each sample, the diversity is defined as

$$\mathcal{D}_{div}(x_{t,i}) = \left\| f(g(x_{t,i})) - \frac{1}{\sum_{j=1}^{n_t} a(i,j)} \sum_{j=1}^{n_t} a(i,j) f(x_{t,i}) \right\|_2, \tag{4}$$

where $a(i,j)$ indicates whether sample $x_{t,i}$ and $x_{t,j}$ belong to the same person or not. $a(i,j) = 1$ when $y_{t,i} = y_{t,j}$, otherwise $a(i,j) = 0$.

For a batch of sample, a diversity loss is defined as

$$\mathcal{L}_{div} = \frac{1}{n_t} \sum_{i=1}^{n_t} e^{-\lambda \mathcal{D}_{div}(x_{t,i})}, \tag{5}$$

where λ is hyper-parameter. We use a negative exponent function to prevent $\mathcal{D}_{div}$ from growing too large so as to preserve the identity of the augmented person images. According to Eq. 4 and Eq. 5, maximizing the sample diversity $\mathcal{D}_{div}$ in a cluster is equal to minimizing the loss, as

$$\arg\max_{g} \mathcal{D}_{div} \Leftrightarrow \arg\min_{g} \mathcal{L}_{div}. \tag{6}$$

$\mathcal{L}_{div}$ is combined with loss of StarGAN to optimize the generator g while augmenting samples.

Min-Step: Given a fixed generator g, the feature encoder f learns to minimize the intra-cluster distance while maximizing inter-cluster distance in feature space under the constraint of triplet loss, which is defined as

$$\begin{aligned} \mathcal{L}_{tri} = \sum_{i=1}^{n_t} [m + \|f(x_{t,i}) - f(x_{t,i^+})\|_2 \\ - \|f(x_{t,i}) - f(x_{t,i^-})\|_2], \end{aligned} \tag{7}$$

where x_{t,i^+} denotes the samples belonging to the same cluster with $x_{t,i}$. x_{t,i^-} denotes the samples belonging to different clusters with $x_{t,i}$. m is a margin parameter. Specifically, we choose all the positive samples and the hardest negative sample to construct the triplets for each anchor sample, with a mini-batch of both original and generated sample images. The objective function is defined by

$$\arg\min_{f} \mathcal{D}_{div} \Leftrightarrow \arg\min_{f} \mathcal{L}_{tri}. \tag{8}$$

When g keeps producing more diverse samples with features far away from the cluster centers, f will be equipped with stronger discrimination ability in the target domain, as illustrated in Fig. 4. **Algorithm 1** shows the detailed training procedure of the proposed AD-Cluster.

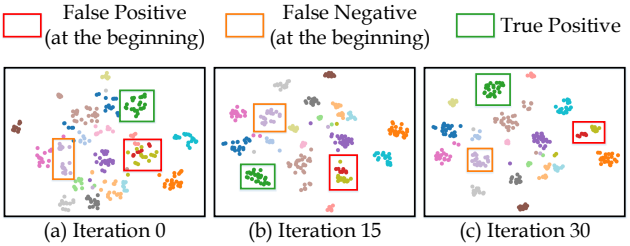
4. Experiments

We detail the implementation and evaluation of AD-Cluster. During the evaluation, ablation studies, parameter analysis, and comparisons with other methods are provided.

Algorithm 1 Training procedure of AD-Cluster

Input: Source domain dataset $\mathbf{S}$, target domain dataset $\mathbf{T}$
Output: Feature encoder f

```
Pre-train feature encoder f on S by optimizing Eq. 1.
for each clustering iteration do
    Extract features F = f(T).
    Cluster training samples in target domain using F.
    for each mini-batch B ⊂ T do
        Max-step: train image generator g by B.
        Min-step: train feature encoder f by {B, g(B)}.
    end for
end for
return Feature encoder f
```



(a) Iteration 0 (b) Iteration 15 (c) Iteration 30

Figure 4. The sparsely and incorrectly distributed person image features of different identities are grouped to more compact and correct clusters through the iterative clustering process. (Best viewed in color with zoom in.)

4.1. Datasets and Evaluation Metrics

The experiments were conducted over two public datasets Market1501 [58] and DukeMTMC-ReID [36] [59] by using the evaluation metrics Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP).

Market1501 [58]: This dataset contains 32,668 images of 1,501 identities from 6 disjoint surveillance cameras. Of the 32,668 person images, 12,936 images from 751 identities form a training set, 19,732 images from 750 identities (plus a number of distractors) form a gallery set, and 3,368 images from 750 identities form a query set.

DukeMTMC-ReID [36] [59]: This dataset is a subset of the DukeMTMC. It consists of 16,522 training images, 2,228 query images, and 17,661 gallery images of 1,812 identities captured using 8 cameras. Of the 1812 identities, 1,404 appear in at least two cameras and the rest 408 (considered as distractors) appear in only one camera.

4.2. Implementation Details

We adopt the ResNet-50 [20] as the backbone network and initialize it by using parameters pre-trained on the ImageNet [9]. During training, the input image is uniformly resized to 256×128 and traditional image augmentation is performed via random flipping and random erasing. For each identity from the training set, a mini-batch of size 256

Methods	DukeMTMC-reID → Market-1501				Market-1501 → DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [25]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
Bow [58]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
UMDL [34]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PTGAN [45]	38.6	-	66.1	-	27.4	-	50.7	-
PUL [13]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
SPGAN [10]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
CAMEL [53]	54.5	-	-	26.3	-	-	-	-
ATNet [27]	55.7	73.2	79.4	25.6	45.1	59.5	64.2	24.9
MMFA [26]	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
SPGAN+LMP [10]	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
TJ-AIDL [44]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle [63]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL [61]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
ECN [62]	<u>75.1</u>	<u>87.6</u>	<u>91.6</u>	<u>43.0</u>	<u>63.3</u>	<u>75.8</u>	<u>80.4</u>	<u>40.4</u>
UDAP [41]	*75.8*	*89.5*	*93.2*	*53.7*	*68.4*	*80.1*	*83.5*	*49.0*
AD-Cluster (Ours)	**86.7**	**94.4**	**96.5**	**68.3**	**72.6**	**82.5**	**85.5**	**54.1**

Table 1. Comparison of the proposed AD-Cluster with state-of-the-art methods: For the transfers DukeMTMC-reID → Market-1501 and Market-1501 → DukeMTMC-reID, the proposed AD-Cluster significantly outperforms all state-of-the-art methods over all evaluation metrics. The top-three results are highlighted with bold, italic, and underline fonts, respectively.

is sampled with P = 32 randomly selected identities and K = 8 (original to augmented samples ratio = 3:1) randomly sampled images for computing the hard batch triplet loss.

In addition, we set the margin parameter at 0.5 and use the SGD optimizer to train the model. The learning rate is set at 6×10^{-5} and momentum at 0.9. The whole training process consists of 30 iterative min-max clustering process, each of which consists of 70 training epochs.

Our network was implemented on a PyTorch platform and trained using 4 NVIDIA Tesla K80 GPUs (each with 12GB VRAM).

4.3. Comparisons with State-of-the-Arts

We compare AD-Cluster with state-of-the-art unsupervised person ReID methods including: 1) LOMO [25] and BOW [58] that used hand-crafted features; 2) UMDL [34], PUL [13] and CAMEL [53] that employed unsupervised learning; and 3) nine UDA-based methods including PTGAN [45], SPGAN [10], ATNet [27], CamStyle [63], HHL [61], and ECN [62] that used GANs; MMFA [26] and TJ-AIDL [44] that used images attributes; and UDAP [41] that employed clustering. Table 1 shows the person Re-ID performance while adapting from Market1501 to DukeMTMC-reID and vice versa.

As Table 1 shows, LOMO and BOW using hand-crafted features do not perform well. UMDL [34], PUL [13] and CAMEL [53] derive image features through unsupervised learning, and they perform clearly better than LOMO and BOW under most evaluation metrics. The UDA-based methods further improve the person Re-ID performance in most cases. Specifically, UDAP performs much better than other methods as it employed the distribution of clusters in the target domains.The performance of the UDA methods using GAN is diverse. In particular, ECN performs better than most methods using GANs because it enforces cameras invariance and domain connectedness.

In addition, AD-Cluster performs significantly better than all compared methods. As Table 1 shows, AD-Cluster achieves a rank-1 accuracy of 86.7% and an mAP of 68.3% for the unsupervised adaptation DukeMTMC-reID → Market1501, which outperforms the state-of-the-art (by UDAP) by 10.9% and 14.6%, respectively. For Market1501 → DukeMTMC-reID, AD-Cluster obtains a rank-1 accuracy of 72.6% and an mAP of 54.1% which outperforms the state-of-the-art (by UDAP) by 4.2% and 5.1%, respectively.

Note that AD-Cluster improves differently for the two adaptations in reverse directions between the two datasets. This can also be observed for most existing methods as shown in Table 1. We conjecture that this is because the large variance of samples in DukeMTMC-reID caused more clustering noise, which reduces the effectiveness of pseudo-label prediction and hinders the model adaptation.

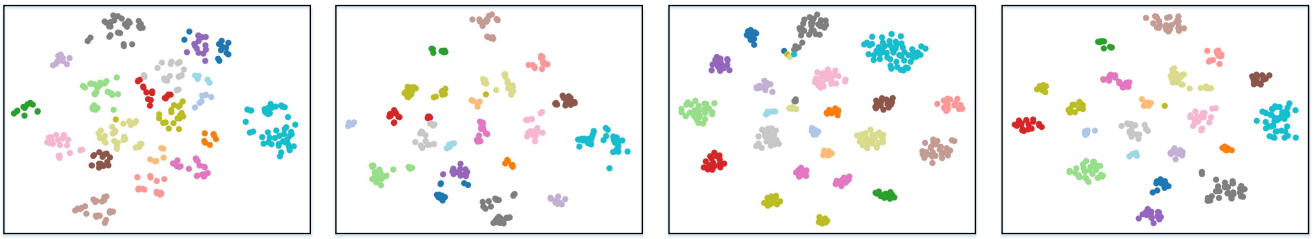
4.4. Ablation Studies

Extensive ablation studies are performed to evaluate each component of AD-Cluster as shown in Table 2.

Baseline, the Upper and Lower Bounds: We first derive the upper and lower performance bounds for the ablation studies as shown in Table 2. Specifically, the upper bounds of Re-ID performance are derived by the *Supervised*

Methods	DukeMTMC-reID → Market-1501				Market-1501 → DukeMTMC-reID			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
Supervised Model (upper bound)	91.9	97.4	98.4	81.4	82.8	92.2	94.9	69.8
Direct Transfer	46.3	63.8	71.2	21.3	28.0	42.9	49.4	14.2
Baseline	73.8	85.7	89.0	51.0	68.6	79.3	82.2	49.0
Baseline+ASA	83.3	93.6	95.7	62.8	71.5	81.1	84.2	52.7
Baseline+ASA+DL	86.7	94.4	96.5	68.3	72.6	82.5	85.5	54.1

Table 2. Ablation studies of AD-Cluster: *Supervised Models:* Re-ID models trained by using the labelled training images of the target domain; *Direct Transfer:* Re-ID models trained by using the labelled training images of the source domain; *Baseline:* Baseline Re-ID models trained via Density-based Clustering [12]; *Baseline+ASA:* Baseline model plus the proposed Adaptive Sample Augmentation; *Baseline+ASA+DL:* Baseline model plus the proposed Sample Augmentation and Discriminative Feature Learning.



(a) Direct Transfer (J = 0.0556) (b) Density-based Clustering (J = 0.1674) (c) Sample Augmentation (J = 0.2205) (d) Discriminative Learning (J = 0.2555)

Figure 5. Comparison of sample distributions on Market-1501 dataset with different transfer techniques: J denotes the ratio between inter-class scatter and intra-class scatter and a larger J means better transfer. (Best viewed in color)

Models which are trained by using labelled target-domain training images and evaluated over the target-domain test images. The lower performance bounds are derived by the *Direct Transfer* models which are trained by using the labelled source-domain training images and evaluated over the target-domain test images. We can observe huge performance gaps between the *Direct Transfer* models and the *Supervised Models* due to the domain shift. Take the Market-1501 as an example. The rank-1 accuracy of the supervised model reaches up to 91.9% but it drops significantly to 46.3% for the directly transferred model which is trained by using the DukeMTMC-reID training images.

In addition, Table 2 gives the performance of *Baseline* models which are transfer models as trained by iterative density-based clustering as described in [41]. As Table 2 shows, the *Baseline* model outperforms the *Direct Transfer* model by a large margin. For example, the rank-1 accuracy improves from 46.3% to 73.8% and from 28.0% to 68.6%, respectively, while evaluated over the datasets Market1501 and DukeMTMC-reID. This shows that the density-based clustering in the *Baseline* can group samples of same identities to any irregular distributions by utilizing the density correlation. At the same time, we can observe that there are still large performance gaps between the *Baseline* models and the *Supervised Models*, $e.g.$, a drop of 30% in mAP while transferring from DukeMTMC-reID to Market1501.

Adaptive Sample Augmentation: We first evaluated the adaptive sample augmentation as described in Section 3.2. For this experiment, we designed a network *Baseline+ASA* that just incorporates the adaptive sample augmentation into the *Baseline* that performs transfer via iterative density-based clustering. As shown in Table 2, adaptive sample augmentation improves the re-ID performance significantly. For DukeMTMC-reID → Market1501, the *Baseline+ASA* achieves a rank-1 accuracy of 83.3% and an mAP of 62.8% which are higher than the *Baseline* by 9.5% and 11.8%, respectively. The contribution of the proposed sample augmentation can also be observed in the perspective of sample distributions in the feature space as illustrated in Fig. 5(c), where the including of the proposed sample augmentation improves the sample distribution greatly as compared with density-based clustering as shown in Fig. 5(b).

The large performance improvements can be explained by the effectiveness of the augmented samples. Specifically, the iterative injection of ID-preserving cross-camera images helps to reduce the feature distances of person images within the same cluster ($i.e.$, the intra-cluster distances) and increase that of different clusters ($i.e.$, the inter-cluster distances) simultaneously.

Discriminative Learning: We evaluated the the discriminative learning component as described in Section 3.3. For this experiment, we designed a new network *Baseline+ASA+DL* that further incorporates discriminative learning into the *Baseline+ASA* network as described in the previous subsection. As shown in Table 2, the incorporation of discriminative learning consistently improves the

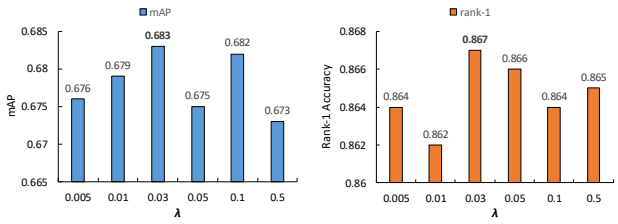


Figure 6. The min-max attenuation coefficient λ in Eq. 5 affects both mAP and rank-1 accuracy (evaluated on Market-1501).

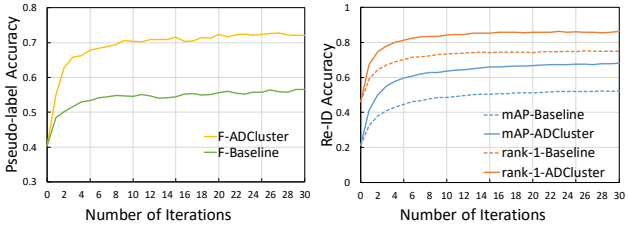


Figure 7. Iterative min-max clustering outperforms density-based clustering consistently for both accuracy of pseudo-label prediction on the left and mAP & rank-1 accuracy of person Re-ID on the right (for DukeMTMC-reID → Market1501).

person Re-ID performance beyond the *Baseline+ASA*. Take the transfer DukeMTMC-reID → Market1501 as an example. The *Baseline+ASA+DL* achieves a rank-1 accuracy of 86.7% and an mAP of 68.3% which outperforms the corresponding *Baseline+ASA* by 3.4% and 5.5%, respectively. The superior performance of the proposed discriminative learning can also be observed intuitively in the perspective of sample distributions in feature space as shown in Fig. 5(d). The effectiveness of the discriminative learning can be largely attributed to the min-max clustering optimization that alternately trains the image generator to generate more diverse samples for maximizing the sample diversity and the feature encoder for minimizing the intra-class distance.

From another perspective, it can be seen that *Baseline+ASA+DL* (*i.e.*, the complete AD-Cluster model) outperforms the *Baseline* by up to 13% in rank-1 accuracy and 17% in mAP, respectively. This demonstrates the effectiveness of the proposed ID-preserving cross-camera sample augmentation and discriminative learning in UDA-based person Re-ID. In addition, we can observe that the performance of *Baseline+ASA+DL* becomes even close to the *Supervised Models*. For example, the *Baseline+ASA+DL* achieves a rank-1 accuracy of 86.7% for the transfer DukeMTMC-reID → Market-1501 which is only 5.2% lower than the corresponding Supervised Model.

Specificity of AD-Cluster. The performance of the AD-Cluster is related to the sample generation method. In this work, we generate cross-camera images by using StarGAN which theoretically can be replaced by any other ID-preserving generators. The key is how well the re-ID model can learn camera style in-variance via generating new samples. The AD-Cluster could thus be influenced by two factors: the quality of generated samples and the strength of camera style in-variance of the sample distribution in the target domain. These variances explain the different improvements by AD-Cluster over different adaptation tasks.

4.5. Discussion

The min-max attenuation coefficient λ in Eq. 5 will affect the ID-preserving min-max clustering and so the person Re-ID performance. We studied this parameter by setting it to different values and checking the person Re-ID performance. Fig. 6 shows experimental results on Market-1501. Using a smaller λ usually leads to a higher cluster diversity, which further leads to better Re-ID performance. On the other hand, λ should not be very small for the target of identity preservation. Experiments show that AD-Cluster performs best when $\lambda = 0.03$. We also evaluate the accuracy of the pseudo-labels that are predicted during the iterative min-max clustering, as well as how the person Re-ID performance evolves during this process. Fig. 7 (left) shows that the f-score of the predicted pseudo-labels keeps improving during the iterative clustering process. Additionally, the proposed min-max clustering outperforms the density-based clustering [12] significantly in both mAP and rank-1 accuracy as shown in the right graph in Fig. 7.

5. Conclusion

This paper presents an augmented discriminative clustering (AD-Cluster) method for domain adaptive person re-ID. With density-based clustering, we introduce adaptive sample augmentation to generate more diverse samples and a min-max optimization scheme to learn more discriminative re-ID model. Experiments demonstrates the effectiveness of adaptive sample augmentation and min-max optimization for improving the discrimination ability of deep re-ID model. Our approach not only produces a new state-of-the-art in UDA accuracy on two large-scale benchmarks but also provides a fresh insight for general UDA problems. We expect that the proposed AD-Cluster will inspire new insights and attract more interests for better UDA-based recognition [15] and detection [56] in the near future.

Acknowledgement

This work is partially supported by grants from the National Key R&D Program of China under grant 2017YFB1002400, the National Natural Science Foundation of China under contract No. 61825101, No. U1611461, No. 61836012 and No. 61972217.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *CoRR*, abs/1412.4446, 2014.

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE CVPR*, pages 95–104, 2017.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

[4] Minmin Chen, Kilian Q. Weinberger, and John Blitzer. Co-training for domain adaptation. In *NeurIPS*, pages 2456–2464, 2011.

[5] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *IEEE ICCV*, pages 8351–8361, 2019.

[6] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. In *IEEE ICCV*, 2019.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE CVPR*, 2018.

[8] Adam Coates and Andrew Y. Ng. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700, pages 561–580. Springer, 2012.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009.

[10] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *IEEE CVPR*, 2018.

[11] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *NeurIPS*, pages 766–774, 2014.

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[13] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *TOMCCAP*, 14(4):83:1–83:18, 2018.

[14] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *CoRR*, abs/1705.10444, 2017.

[15] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S. Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *IEEE ICCV*, 2019.

[16] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37, pages 1180–1189, 2015.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

[18] Kamran Ghasedi, Xiaoqian Wang, Cheng Deng, and Heng Huang. Balanced self-paced learning for generative adversarial clustering network. In *IEEE CVPR*, 2019.

[19] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, June 2016.

[21] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1994–2003, 2018.

[23] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *IEEE ICCV*, 2019.

[24] Renjie Liao, Alexander G. Schwing, Richard S. Zemel, and Raquel Urtasun. Learning deep parsimonious representations. In *NeurIPS*, pages 5076–5084, 2016.

[25] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE CVPR*, June 2015.

[26] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex C. Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *BMVC*, 2018.

[27] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *IEEE CVPR*, 2019.

[28] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NeurIPS*, pages 469–477, 2016.

[29] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *IEEE ICCV*, pages 2448–2457, 2017.

[30] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In Francis R. Bach and David M. Blei, editors, *ICML*, volume 37, pages 97–105, 2015.

[31] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiaguang Sun, Yuchen Guo, and Philip S. Yu. Transfer sparse coding for robust image representation. In *IEEE CVPR*, pages 407–414, 2013.

[32] Jianming Lv and Xintong Wang. Cross-dataset person re-identification using similarity preserved generative adversarial networks. In Weiru Liu, Fausto Giunchiglia, and Bo Yang, editors, *KSEM*, pages 171–183, 2018.

[33] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE ICCV*, pages 5716–5726, 2017.

[34] 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 CVPR*, June 2016.

[35] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *IEEE ICCV*, 2019.

[36] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *IEEE ECCV Workshops*, 2016.

[37] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *NeurIPS*, pages 46–54, 2013.

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[39] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE CVPR*, 2018.

[40] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *NeurIPS*, pages 2110–2118, 2016.

[41] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *CoRR*, abs/1807.11334, 2018.

[42] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, pages 2058–2065, 2016.

[43] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

[44] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE CVPR*, 2018.

[45] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE CVPR*, 2018.

[46] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Unsupervised person re-identification by camera-aware similarity consistency learning. In *IEEE ICCV*, 2019.

[47] Jinlin Wu, Shengcai Liao, Zhen Lei, Xiaobo Wang, Yang Yang, and Stan Z. Li. Clustering and dynamic sampling based unsupervised domain adaptation for person re-identification. In *IEEE ICME*, pages 886–891, 2019.

[48] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *IEEE CVPR*, 2018.

[49] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *ICML*, volume 48, pages 478–487, 2016.

[50] Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Xiaodong Xie, and Wen Gao. Attention driven person re-identification. *Pattern Recognition*, 86:143–155, 2019.

[51] Jianwei Yang, Devi Parikh, and Dhruv Batra. Joint unsupervised learning of deep representations and image clusters. In *IEEE CVPR*, pages 5147–5156, 2016.

[52] Mang Ye, Andy Jinhua Ma, Liang Zheng, Jiawei Li, and Pong C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *IEEE ICCV*, pages 5152–5160, 2017.

[53] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE ICCV*, 2017.

[54] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *IEEE CVPR*, pages 3801–3809, 2018.

[55] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *IEEE ICCV*, 2019.

[56] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *Advances in Neural Information Processing Systems*, pages 147–155, 2019.

[57] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE ICCV*, 2015.

[59] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE ICCV*, 2017.

[60] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE CVPR*, 2017.

[61] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. In *ECCV*, pages 176–192, 2018.

[62] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE CVPR*, 2019.

[63] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camstyle: A novel data augmentation method for person re-identification. *IEEE TIP*, 28(3):1176–1190, 2019.