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

Yunpeng Zhai\textsuperscript{1,2}, Shijian Lu\textsuperscript{3}, Qixiang Ye\textsuperscript{4,6}, Xuebo Shan\textsuperscript{1,2}, Jie Chen\textsuperscript{1,6}, Rongrong Ji\textsuperscript{5,6}, Yonghong Tian\textsuperscript{1,2,6}\textsuperscript{*}

\textsuperscript{1}School of Electronic and Computer Engineering, Peking University, China
\textsuperscript{2}NELVT, School of EE&CS, Peking University, Beijing, China
\textsuperscript{3}Nanyang Technological University, Singapore, \textsuperscript{4}University of Chinese Academy of Sciences, China
\textsuperscript{5}Xiamen University, China, \textsuperscript{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.

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

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

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 representations 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 $\{x_{t,i}^{(1)}, y_{t,i}\}, \{x_{t,i}^{(2)}, y_{t,i}\}, \ldots, \{x_{t,i}^{(K)}, y_{t,i}\}$, which have the pseudo-label $y_{t,i}$ with $x_{t,i}$ and similar styles as the images in camera $1, 2, \ldots, K$, 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.

### 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 $D_{div}$. For
each sample, the diversity is defined as
\[
D_{\text{div}}(x_{t,i}) = \left\| f(g(x_{t,i})) - \frac{1}{\sum_{j=1}^{n_{t}} a(i,j)} \sum_{j=1}^{n_{t}} a(i,j) f(x_{t,j}) \right\|_2^2
\]
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
\[
L_{\text{div}} = \frac{1}{n_t} \sum_{i=1}^{n_t} e^{-\lambda D_{\text{div}}(x_{t,i})},
\]
where \(\lambda\) is hyper-parameter. We use a negative exponent function to prevent \(D_{\text{div}}\) 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 \(D_{\text{div}}\) in a cluster is equal to minimizing the loss, as
\[
\arg \max_g D_{\text{div}} \Leftrightarrow \arg \min_g L_{\text{div}}.
\]

\(L_{\text{div}}\) 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
\[
L_{\text{tri}} = \sum_{i=1}^{n_t} [m + \|f(x_{t,i}) - f(x_{t,i}^+)\|_2 - \|f(x_{t,i}) - f(x_{t,i}^-)\|_2],
\]
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
\[
\arg \min_f D_{\text{div}} \Leftrightarrow \arg \min_f L_{\text{tri}}.
\]
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 \(S\), target domain dataset \(T\)

**Output**: Feature encoder \(f\)

1. Pre-train feature encoder \(f\) on \(S\) by optimizing Eq. 1.
2. for each clustering iteration do
   3. Extract features \(F = f(T)\).
   4. Cluster training samples in target domain using \(F\).
   5. for each mini-batch \(B \subset T\) do
      6. Max-step: train image generator \(g\) by \(B\).
      7. Min-step: train feature encoder \(f\) by \(\{B, g(B)\}\).
   8. end for
3. end for

4. return Feature encoder \(f\)

![False Positive (at the beginning)](a) Iteration 0
![False Negative (at the beginning)](b) Iteration 15
![True Positive](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) appear in only one camera. The rest 408 (considered as distractors) appear in at least two cameras and the rest 408 (considered as distractors) appear in only one camera.

**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
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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>DukeMTMC-reID → Market-1501</th>
<th>Market-1501 → DukeMTMC-reID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-5</td>
</tr>
<tr>
<td>LOMO [25]</td>
<td>27.2</td>
<td>41.6</td>
</tr>
<tr>
<td>Bow [58]</td>
<td>35.8</td>
<td>52.4</td>
</tr>
<tr>
<td>UMDL [34]</td>
<td>34.5</td>
<td>52.6</td>
</tr>
<tr>
<td>PTGAN [45]</td>
<td>38.6</td>
<td>-</td>
</tr>
<tr>
<td>PUL [13]</td>
<td>45.5</td>
<td>60.7</td>
</tr>
<tr>
<td>SPGAN [10]</td>
<td>51.5</td>
<td>70.1</td>
</tr>
<tr>
<td>CAMEL [53]</td>
<td>54.5</td>
<td>-</td>
</tr>
<tr>
<td>ATNet [27]</td>
<td>55.7</td>
<td>73.2</td>
</tr>
<tr>
<td>MMFA [26]</td>
<td>56.7</td>
<td>75.0</td>
</tr>
<tr>
<td>SPGAN+LMP [10]</td>
<td>57.7</td>
<td>75.8</td>
</tr>
<tr>
<td>TJ-AIDL [44]</td>
<td>58.2</td>
<td>74.8</td>
</tr>
<tr>
<td>CamStyle [63]</td>
<td>58.8</td>
<td>78.2</td>
</tr>
<tr>
<td>HHL [61]</td>
<td>62.2</td>
<td>78.8</td>
</tr>
<tr>
<td>ECN [62]</td>
<td>75.1</td>
<td>87.6</td>
</tr>
<tr>
<td>UDAP [41]</td>
<td>75.8</td>
<td>89.5</td>
</tr>
<tr>
<td>AD-Cluster (Ours)</td>
<td><strong>86.7</strong></td>
<td><strong>94.4</strong></td>
</tr>
</tbody>
</table>

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 $6 \times 10^{-5}$ 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 camera 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
### Table 2. Ablation studies of AD-Cluster: 

<table>
<thead>
<tr>
<th>Methods</th>
<th>DukeMTMC-reID → Market-1501</th>
<th>Market-1501 → DukeMTMC-reID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-5</td>
</tr>
<tr>
<td>Supervised Model (upper bound)</td>
<td>91.9</td>
<td>97.4</td>
</tr>
<tr>
<td>Direct Transfer</td>
<td>46.3</td>
<td>63.8</td>
</tr>
<tr>
<td>Baseline</td>
<td>73.8</td>
<td>85.7</td>
</tr>
<tr>
<td>Baseline+ASA</td>
<td>83.3</td>
<td>93.6</td>
</tr>
<tr>
<td>Baseline+ASA+DL</td>
<td>86.7</td>
<td>94.4</td>
</tr>
</tbody>
</table>

*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.*

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 performance. For example, the rank-1 accuracy is increased by 9% and 6% compared to Baseline+ASA and Baseline respectively.
person Re-ID performance beyond the Baseline+ASA. Take the transfer DukeMTMC-reID \(\rightarrow\) 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 \(\rightarrow\) Market1501 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 \(\lambda\) 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 \(\lambda\) usually leads to a higher cluster diversity, which further leads to better Re-ID performance. On the other hand, \(\lambda\) 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 demonstrate 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


