International Journal of Machine Learning and Cybernetics (2022) 13:255–268
https://doi.org/10.1007/s13042-021-01399-1

ORIGINAL ARTICLE

# Domain adaptive attention-based dropout for one-shot person re-identification

**Xulin Song**[1,2] · **Zhong Jin**[1,2]

Received: 23 December 2020 / Accepted: 26 July 2021 / Published online: 7 August 2021


**Abstract**
Cross-domain person re-identification (re-ID) has attracted much attention due to its wide applications in the field of computer vision and surveillance. However, the domain shift issue leads to unsatisfactory generalization performance of a model on an unseen target domain when the model is trained on the source domain. Current methods usually adopt clustering methods to assign pseudo labels for unlabeled target images, resulting in high dependence on the performance of clustering method. In this paper, we firstly focus on extracting universal domain-adaptive features by designing a domain-adaptive-attention-based-dropout (DAAD) layer. DAAD layer is achieved by a universal attention-based dropout adapter (ADA) bank to hide the most discriminative region stochastically and a domain attention module to assign weights to the two domains (source and target). Then two feature memories are introduced according to one-shot learning in which only one image is annotated for each target identity. These two memories are designed to store target features from labeled and unlabeled images, respectively. The labeled feature memory is leveraged to estimate pseudo labels for these unlabeled images while the unlabeled feature memory aims to maximize distances between all the unlabeled images and minimize distances between similar images simultaneously. Extensive experiments on three re-ID datasets (DukeMTMC-reID, Market-1501, and MSMT17) demonstrate that the proposed model is effective to improve the domain adaptation performance than existing techniques.

**Keywords** Domain adaptation · Attention · One-shot learning · Person re-identification

## 1 Introduction

Person Re-identification (re-ID) [3, 42] aims to match people across non-overlapping surveillance camera views. It embraces many applications due to its great potential for video surveillance and public security. Despite the impressive advancements have been witnessed [39] by convolutional neural networks (CNNs) [16, 17, 28, 33, 40] in recent years, person re-ID is still challenging towards its practical applications [10, 24]. The key problem is domain shift caused by illuminations, backgrounds, occlusions, and camera conditions, among many others. Concretely, these re-ID models trained on source domain suffer large performance degradation when tested on target domain [8, 12]. Here, each dataset is regarded as a distinct domain in the community of person re-identification.

To tackle the domain shift problem, there are two main solutions. The first and the most straightforward is to collect enough labeled data of target domain for supervised training. However, it is expensive and laborious to collect and annotate such a large-scale dataset, resulting in the seriously limited practicability for actual applications. The second and the most used is unsupervised domain adaptation (UDA) method which trained on labeled source domain together with unlabeled target domain then tested on target domain [26, 38]. In the conventional settings of UDA, most approaches are mainly designed for a closed-set scenario and always have an assumption that the source and target domain share the same label space, i.e. the two domains have identical classes. However, this assumption is invalid to person re-ID problem since different re-ID

✉ Zhong Jin
zhongjin@njust.edu.cn

Xulin Song
xulinsong@njust.edu.cn

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China

[2] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing, China

datasets have completely different identities. Therefore, UDA for person re-ID task is more challenging. In this work, we are dedicated to learning discriminative and robust features that domain-sensitive for effective cross-domain person re-ID.

In person re-ID, the local features of body regions without extra background noise are obviously more discriminative compared to these global features from the whole images. Attention mechanism which is designed to help the networks focus on the most discriminative regions has been widely introduced into person re-ID [4, 20, 36, 43]. The performance gain achieved by existing attention models proved that erasing these background noise can increase the performance of person re-ID. However, their cases are only suitable for one domain. In the situation of domain adaptation, only leveraging the discriminative region with completely remove background may probably lead overfitting on the source domain. Inspired by the success of these attention models, we rethink the attention to combine with domain adaptation. A model will be designed in which it not only can remove the background noise but also can address the problem of domain shift.

To learn robust and universal person features for the target domain, we propose a domain-adaptive-attention-based-dropout (DAAD) layer to extract universal domain-adaptive features and further improve re-ID performance in the target domain. DAAD layer is composed of two key components: a universal attention-based dropout adapter (ADA) bank and a domain attention (DA) module. The ADA bank generates two attention-based dropout adapters to produce domain-sensitive features. Each ADA generates a discriminative map and a drop map to help the network to focus on the most discriminative region or not. The domain attention module is designed to assign weights to the domain-sensitive features from the bank so as to enable the adapters are specialize on individual domains. In this way, DAAD layer can be represented as an atomic domain adaptation unit which utilized to build domain detectors in this work.

To predict pseudo labels for the unlabeled images in the target domain, we introduce two feature memories, namely labeled memory and unlabeled memory. The two feature memories store the features of all the labeled and unlabeled images, respectively. They are updated according to the new features learned from each iteration. Concretely, the memory that stores labeled image features is used to compute the similarity between each unlabeled image and each labeled image and then assign a pseudo label for the unlabeled image. The memory that saves the unlabeled image feature treats each image as an individual identity together with keeping clusters in the target domain. During the iteration, we adaptively maximize the distances between unlabeled images and minimize the distances between these similar images for the unlabeled images in the target as [9].

The main contributions can be summarized in the following twofolds. (1) We propose a domain-adaptive-attention-based-dropout (DAAD) layer to generate domain sensitive features. It contains a universal attention-based dropout adapter bank that can specialize on individual domains and a domain attention module to assign different network activations for different domains. DAAD layer is also plug-and-play and can predict the domain of images automatically, the domain of interest therefore is not required in advance. (2) We design two feature memories in the target domain for one-shot learning. The labeled feature memory is utilized to predict pseudo labels for the unlabeled images, the unlabeled feature memory adaptively keep clusters in the unlabeled images for the target domain.

## 2 Related works

### 2.1 Unsupervised person re-ID

Supervised re-ID methods have made great progress benefiting from the rapid advancement of deep CNNs [16, 17, 28, 40] and sufficient labeled data. In addition, graph convolutional network [25] based method is also proposed for person re-identification, in which each person image is regarded as a node of the graph. However, supervised learning is impractical in real-world applications due to intensive labeling costs. This motivates the research into an unsupervised manner. Some works [11, 21] add additional auxiliary information, eg. facial landmarks or pose estimation, for unsupervised person re-ID. Some unsupervised re-ID works focus on one dataset and need no additional information [23, 35]. In [23], a bottom-up clustering method (BUC) is proposed which jointly optimizes the CNNs and the relationship between images. Recently, transfer learning has been introduced into unsupervised person re-ID [12, 32] in which all utilize transferable information from an external source domain. This work follows the setting that addresses person re-ID by introducing a labeled source domain.

### 2.2 Unsupervised domain adaptive person re-ID

This work is related to unsupervised domain adaptation (UDA) method which transfers discriminative re-ID information from labeled source to unlabeled target domain. Some works for addressing UDA by reducing the discrepancy [14, 15] between the two domains or by learning an adversarial domain-classifier [2, 30]. However, these UDA approaches are usually based on an assumption that the class space is identical in both source domain and target domain, while this is not applicable in person re-ID community. Benefiting from the great progress of deep learning, some recent deep learning-based unsupervised domain adaptation

methods for re-ID task [8, 13, 47, 48] have emerged. In [8], it proposed a similarity preserving generative adversarial networks (SPGAN) to translate the images from source domain into target domain. Zhong et al. [47] aim to improve the generalization ability by the proposed Hetero–Homogeneous Learning (HHL) method which can achieve camera invariance and domain connectedness simultaneously. Fu et al. [13] propose a Self-similarity Grouping (SSG) method to harness the similar characteristics between images in the target domain. Zhong et al. [48] investigate into three intra-domain invariances of the target domain to achieve effective domain adaptation accuracy. In [6], an unsupervised domain adaptive person re-identification method, style transfer re-identification (STReID), is proposed to solve the potential image distinctions between different domains. In this work, the domain attention adaptive based dropout layer aims to capture the domain-sensitive feature and achieves competitive performance on domain adaptive person re-ID.

### 2.3 Attention model in person re-ID

Attention has been found effective in person re-ID task [4, 20, 36, 43]. Li et al. [20] propose a Harmonious Attention CNN (HA-CNN) which jointly learn soft pixel attention and hard regional attention along with the optimization of feature representations. Chen et al. [4] propose an Attentive but Diverse Network (ABD-Net) by integrating attention mechanism and diversity regularization into a network to learn robust and discriminative features. In [36], Bryan et al. directly leverage second-order feature statistics to model long-range relationship between feature maps for person re-ID. Zheng et al. [43] propose a Consistent Attentive Siamese Network (CASNet) to address cross-view matching issues, eg. spatial localization and view-invariant representation learning in person re-ID. In [19], a spatial softmax is leveraged to calculate the attention weights. The stacked multimodal attention network (SMAN) [18] utilizes the stacked multimodal attention mechanism to compute the cross-modal similarity. These methods typically utilize attention to enlarge the representation power in one domain, we propose the DAAD layer which provides a new viewpoint on learning domain-sensitive features in source and target domains for person re-ID.

### 2.4 One-shot re-ID

There are some recent works for one-shot person re-identification [1, 13, 35] which only leverage one labeled image for each identity and the rest images are unlabeled during training. The transfer local relative distance comparison (t-LRDC) [44] method introduces the one-shot group-based verification to address the open-world person re-identification. Wu et al. [35] propose a progressive model that gradually predicts the pseudo labels for the unlabeled data for person re-ID with one labeled image. In [13], Fu et al. exploit the similarity between unlabeled images to generate clusters automatically and propose a clustering-guided self-similarity Grouping (SSG) approach to conduct one-shot domain adaptation for person re-ID. Bak *et al.* [1] split the person images into texture and color by leveraging a single pair of ColorChecker images which capture the differences between camera color distributions. This work differs from these works in that we introduce two feature memories to store features from the labeled images and the unlabeled images during training.

## 3 The proposed method

This paper aims to address domain adaptation problem for one-shot person re-ID. In the context of one-shot learning in re-ID, there is a fully labeled source domain $S = \{(x_i^s, y_i^s)\}_1^{N_S}$ which includes $N_S$ person images and each image $x_i^s$ associated with a corresponding identity $y_i^s \in \{1, 2, \ldots, P_S\}$, where $P_S$ is the number of identities in the source domain. There also is a target domain $T = \{x_i^t\}_1^{N_T}$ with $N_T$ person images. Specially, $T$ can be split into two sub-datasets $T = T_L \cup T_U$ where $T_L$ is the dataset with one labeled image for each identity and $T_U$ is the dataset with the remaining unlabeled images. In general, identities from the source domain and the target domain are completely different and they also have different distributions. The goal of this work is to leverage the labeled source domain data, the labeled one-shot target data and the unlabeled target data to learn discriminative embeddings to generalizes well on the target test data.

### 3.1 Overview of the proposed framework

The framework of the proposed method is illustrated in Fig. 1. In our method, the inputs are sampled from the following three aspects, namely labeled source images, labeled one-shot target images and unlabeled target images. First, the inputs are fed into a backbone network ResNet-50 [16] which has pre-trained on ImageNet [7]. Concretely, we keep the layers of ResNet-50 till the pooling-5 layer. Then, a domain-adaptive-attention-based-dropout (DAAD) layer is plugged to further extract domain-related discriminative features. DAAD layer will be elaborated in Fig. 2. Furthermore, the features are all fed into an FC-Block which consists of global average pooling (GAP), fully connected layer (FC), batch normalization (BN), and ReLU activation. Finally, features from the labeled source are fed into a classification module followed with an $N_S$-dimensional FC layer and a softmax activation for supervised learning. The features from the target then fed into the following two components for one-shot

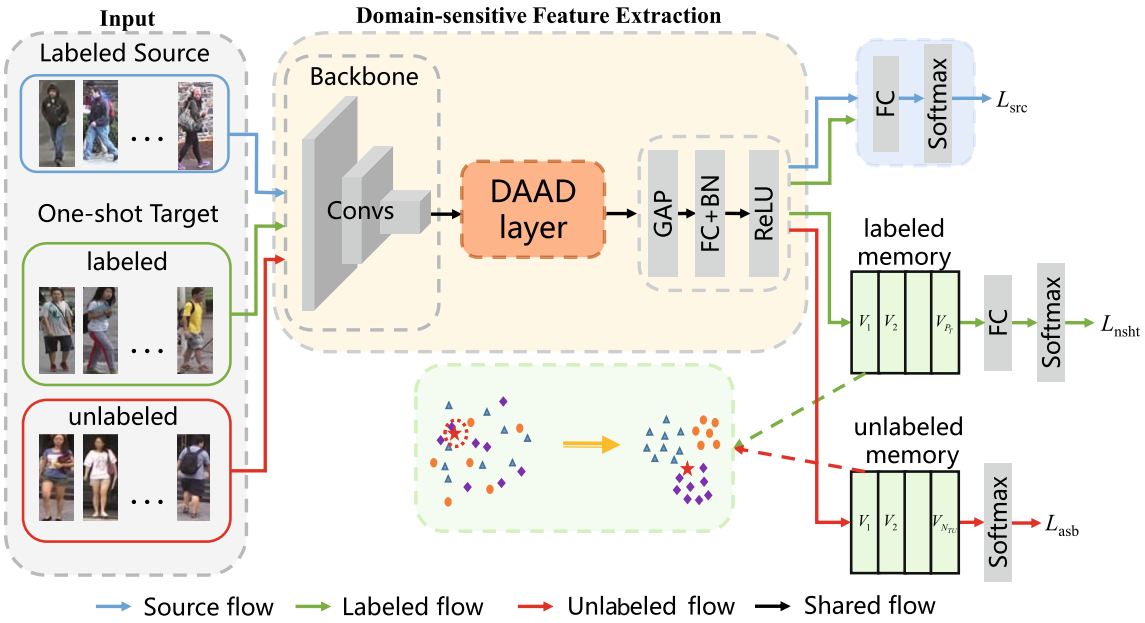


**Fig. 1** The architecture of our proposed approach. DAAD layer denotes the proposed domain adaptive attention-based dropout layer. GAP means global average pooling and FC is fully connection.

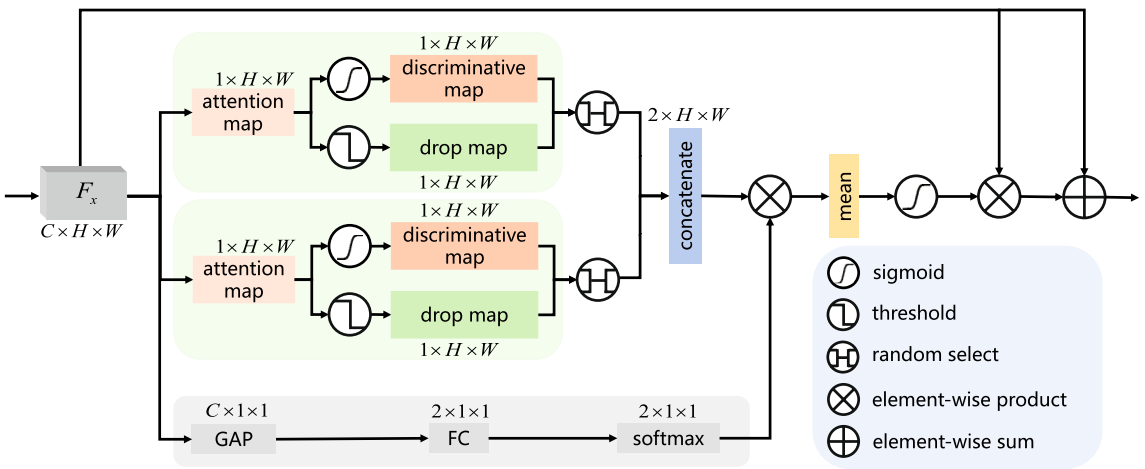


**Fig. 2** The architecture of our proposed domain-adaptive-attention-based-dropout (DAAD) layer.

learning. Each component contains a memory module to save the up-to-date features for the labeled and unlabeled target images, respectively. The labeled memory is used to estimate pseudo label for the unlabeled image. While the unlabeled memory is designed to estimate the similarities between images in each mini-batch and the whole unlabeled images saved in the memory. Both the two memory modules have the same mechanism and will be elaborated in the following sections.

### 3.2 Domain adaptive attention based dropout layer

In this section, we will introduce our proposed domain-adaptive-attention based-dropout (DAAD) layer which has domain sensitivity and is able to adapt to different domains automatically as shown in Fig. 2. DAAD layer focuses on extracting universal domain-adaptive features from the source and target domains. It is achieved by a universal attention-based dropout adapter (ADA) bank to hide the

most discriminative region stochastically and a domain attention module to assign weights to the source and target domain. In the proposed universal DAAD layer, all parameters and computations are shared across domains, and a single network processes all domains all the time. The ADA bank and domain attention module will be elaborated in the following subsections.

**Universal AD Adapter Bank** is used to construct a universal representation space. It is implemented by concatenating outputs of the two individual attention-based dropout adapters

$$F_{UB} = \left[F_{AD}^{1}, F_{AD}^{2}\right] \in \mathbb{R}^{2\times W\times H} \tag{1}$$

where $F_{AD}^{i}$ is the output of each attention-based dropout adapter that will be elaborated next. Regarding to the two domains, i.e. source and target, there will have two adapters. Each adapter has the same structure and projects the inputs to the statistics of a particular domain.

**Attention-based dropout adapter** induces the module to learn the most discriminative region and the entire region of person simultaneously. It first produces an attention map from the input, then generates a discriminative map and a drop map. As discussed in [5], the discriminative map rewards the most discriminative region to improve the classification accuracy. Meanwhile, the drop map penalizes the most discriminative region to induce the module to cover the entire region of people. In the training process, one of the two maps is stochastically selected.

Given the input feature map $F_x \in \mathbb{R}^{C\times H\times W}$, where $C$, $H$, $W$ are the number of channels, height, and width, respectively. First, we get the attention map $A \in \mathbb{R}^{1\times H\times W}$ by leveraging average pooling on $F_x$ along with the channel-wise. Then the discriminative map $M_{dcm} \in \mathbb{R}^{1\times H\times W}$ is generated by utilizing the sigmoid function on the attention map. In this way, the intensity in the most discriminative region is near to 1 and the discriminative features can be preserved. To get the discriminative map and the drop map, we introduce two hyper-parameters, $\eta$ and *drop_rate*, where $\eta$ commands the region size that to be dropped while *drop_rate* controls the frequency that the drop map is used. Based on the attention map, we use $\eta$ to get a drop threshold $T_{drop} = I_{max} \times \eta$ where $I_{max}$ is the maximum intensity in the attention map $A$. Then the drop map $M_{drop} \in \mathbb{R}^{1\times H\times W}$ is generated by setting the intensity to 0 if it is larger than $T_{drop}$. Otherwise, it is set to 1. After this operation, the most discriminative region is hided when the intensity is 0 in the drop map. However, if the drop map is applied in each iteration, the module cannot observe the most discriminative region during training and would result in poor classification performance. As a remedy, the drop map is selected stochastically according to the *drop_rate*. The drop map and the discriminative map are applied alternatively. When the drop map is applied, the discriminative map is hidden and vice versa.

**The domain attention module** is proposed to achieve domain sensitivity. It produces a domain-sensitive set of weights to combine the proposed two attention-based dropout adapters. Specifically, the domain attention module learns to assign network activations to different domains and soft-rout their responses by the domain attention module. This enables the adapters to specialize on individual domains.

To get the weights in proper dimensions, we first apply a global average pooling on $F_x$ to remove the influence of spatial position. Then follows with a liner layer and softmax layer as in Fig. 2

$$W_{DA} = \text{softmax}\left(W_l F_{gap}\left(F_x\right)\right) \tag{2}$$

where $F_{gap}$ denotes the global average pooling operation and $W_l$ is the weights in the FC layer. After the softmax layer, $W_{DA}$ is utilized to weight the two adapters $F_{UB}$, then the domain adaptive responses are generated

$$F_{DA} = \text{Mean}(F_{UB} W_{DA}) \in \mathbb{R}^{1\times H\times W} \tag{3}$$

where Mean$(\cdot)$ is average operation on the feature map.

Inspired by the residual learning and identity mapping in the ResNet [4] which is known as identity skip connection (shortcut), we then utilize the identity skip connection for avoiding the gradient degradation. Therefore, the output of DAAD layer can be denoted as

$$F_{out} = F_x + F_x \otimes \sigma\left(F_{DA}\right) \tag{4}$$

where $\sigma$ is the sigmoid activation and $\otimes$ denotes the channel-wise multipication.

### 3.3 Feature memory module

Inspired by [9, 48], we introduce the memory module to predict the pseudo label along with calculating the similarities between the unlabeled target images during training.

Considering we have a target domain contains $N_T$ images and $P_T$ person. After the FC-Block, each image is equipped with a $d$-dimensional vector. The labeled memory $M_L \in \mathbb{R}^{P_T\times d}$ and the unlabeled memory $M_U \in \mathbb{R}^{N_{TU}\times d}$ will save the corresponding features and be updated after each training epoch. $P_T$ is the number of the identities in the target domain and also is the number of images in one-shot labeled target domain dataset. $N_{TU} = N_T - P_T$ denotes the number of the rest unlabeled images in target domain. Both the memories initialized with zeros. Since the labeled memory and the unlabeled memory have the same setting except for their size. For simplicity, we record both memories as $M$. The index in $M$ corresponds to the image index in the labeled

or unlabeled dataset, namely, $M[i]$ is the feature of the *ith* image. In the forward-propagation, we obtain stored features that can be used to compute the similarity between images within mini-batch and labeled or unlabeled images from the memory. During the back-propagation, the memory $M$ is updated as [9, 48] by $M[i] = \lambda \times M[i] + (1-\lambda) \times \phi(x_i)$, $\lambda \in [0,1]$ is a hype-parameter which controls the update rate of memory, and $\phi(x_i)$ is features extracted from the current iteration. According to our experiment results, we find that an increasing $\lambda$ is superior to a fixing $\lambda$. In the initialization, $M$ is set to zeros and a smaller $\lambda$ is set to accelerate the update rate at the beginning of training.

### 3.4 Supervised learning for the labeled source domain

Due to the available identities in the source domain, we can train the source domain in the supervised learning way. Given the $N_S$ source domain images $\left\{(x_1^s, y_1^s), (x_2^s, y_2^s), \ldots, \left(x_{N_S}^s, y_{N_S}^s\right)\right\}$, $S$ denotes the source domain. As a classification task, the probability that image $x_i^s$ belongs to the identity $y_i^s$ can be formulated as the common used softmax function

$$p\left(y_i^s | x_i^s\right) = \frac{\exp\left(\phi\left(x_i^s; W\right)\right)}{\sum_j^{P_S} \exp\left(\phi\left(x_j^s; W\right)\right)} \tag{5}$$

where $\phi(\cdot; W)$ denotes the $i$th logit of the output from the network given image $x_i^s$, $P_S$ denotes the number of identities in the source domain and $W$ is the network-related weights. Then the supervised loss by cross-entropy loss is defined as

$$L_{src} = -\frac{1}{N_S} \sum_{i=1}^{N_S} \log p\left(y_i^s | x_i^s\right) \tag{6}$$

### 3.5 One-shot learning for the target domain

In this subsection, we introduce one-shot (OS) learning for the target domain. After feature extraction for the target images, there are two branches and each has a memory. For the unlabeled images within each mini-batch, the first branch aims to predict pseudo label through leveraging the labeled features stored in the labeled memory, the second branch minimizes the similarity among all the unlabeled images together with maximizing the similarity between similar images.

There is the target domain $T = T_L \cup T_U$, where $T_L = \left\{x_1^{tl}, \ldots, x_{P_T}^{tl}\right\}$, $T_U = \left\{x_1^{tu}, \ldots, x_{N_{TU}}^{tu}\right\}$. As described in Fig. 1, we obtain the labeled features $F_{TL} = \left\{f_1^{tl}, \ldots, f_{P_T}^{tl}\right\}$ and unlabeled features $F_{TU} = \left\{f_1^{tu}, \ldots, f_{N_{TU}}^{tu}\right\}$ from FC-block. Based on the labeled memory $M_L \in \mathbb{R}^{P_T \times d}$, we can calculate the similarity between each unlabeled image within mini-batch and all the labeled images. Here, we utilize the cosine distance as similarity, where the bigger distance means the more similarity according to cosine similarity. Then the similarity matrix can be denoted as

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1P_T} \\ s_{21} & s_{22} & \cdots & s_{2P_T} \\ \cdots & \cdots & \ddots & \cdots \\ s_{b1} & s_{b2} & \cdots & s_{bP_T} \end{bmatrix} \tag{7}$$

where $b$ is the size of each mini-batch. For each image within mini-batch, we assign the label of the most similar labeled image as its pseudo label

$$y_i = \arg\max_{j=1,\ldots,P_T} s_{ij} \quad (i = 1, \ldots, b) \tag{8}$$

However, it is not reliable only use the label of its nearest neighbor as pseudo label. Suppose the pseudo label of image $x_i^{tu}$ from Eq. (7) is $y_i$, so the unlabeled image is the most unlikely to share the same identity with another unlabeled image who has the smallest distance to the labeled image labeled $y_i$. For a specific labeled image, the unlabeled image with the largest distance and the unlabeled image with the smallest distance have the minimum probability to share the same identity. To enlarge the confidence of pseudo label, we introduce a confidence term as

$$s_{\min(i)} = \min_{k=1,\ldots,b} s_{ky_i} \tag{9}$$

$$w_i = 1 - \frac{s_{\min(i)}}{s_{iy_i}} \tag{10}$$

when $s_{min(i)}$ and $s_{iy_i}$ are closer, $w_i$ will be closer to 0. That means, the pseudo label for $x_i^{tu}$ is not reliable. Otherwise, if $s_{min(i)}$ is far away from $s_{iy_i}$, then $w_i$ will be close to 1 and the pseudo label is reliable. So the loss function for one-shot prediction will be

$$L_{nsht} = -\frac{1}{N_{TU}} \sum_{i=1}^{N_{TU}} w_i \log p\left(k | x_i^t\right) \tag{11}$$

where $p(k|x_i^t) = \frac{\exp(\phi(x_i^t; W) \cdot (M_L[i])^T / \beta)}{\sum_j^{N_{TU}} \exp(\phi(x_j^t; W) \cdot (M_L[j])^T / \beta)}$, $\phi(\cdot; W)$ represents the embedding feature extracted from the network, $W$ denotes the network weights and $M_L[i]$ is the $i$th labeled image feature stored in the labeled memory. The $\beta \in (0, 1]$ denotes temperature fact to balance the distribution.

The goal of the second branch is to minimize the similarity among all the unlabeled images while maximizing the similarities between similar images. Inspired by [9], we introduce the adaptive selection mechanism. To minimize the similarity among all the unlabeled images, each unlabeled image is assumed as an individual class

$$L_{\max} = -\frac{1}{N_{TU}} \sum_{i=1}^{N_{TU}} \log p(i|x_i^t) \tag{12}$$

Here, the index of each unlabeled image is regarded as its pseudo label.

For each unlabeled image, there may exist some positive images that share the same identity. Here, we assume that the unlabeled image and its neighborhoods belong to the same identity. Neighborhoods are adaptively selected by a similarity threshold $\gamma$, i.e. the image will be selected as neighborhood only if its distance to the given unlabeled image is larger than the similarity threshold $\gamma$. The selection is formulated as [9] by minimizing the loss

$$L_{as} = -\frac{1}{N_{TU}} \sum_{i=1}^{N_{TU}} \sum_{j=1}^{N_{TU}} v_i^j \log p\left(j|x_i^t\right) \tag{13}$$

where $v_i^j \in \{0,1\}$ is the selection indication vector. If $s\left(\phi\left(x_i^t;W\right),M[j]\right) > \gamma$, $v_i^j = 1$ and indicates that $j$-th image is selected into neighborhood. Otherwise, it is not selected into neighborhood. Here, the similarity threshold $\gamma$ is set to 0.55. By minimizing $L_{as}$, similar images in the target are forced to stay closer.

To ensure the number of neighborhoods for each unlabeled image is almost equal, the balance loss is formulated by integrating an balance term in Eq. (13)

$$L_{asb} = -\frac{1}{N_{TU}} \sum_{i=1}^{N_{TU}} \frac{\mathbb{I}\left(\left\|v_i\right\|_1 > 1\right)}{\left\|v_i\right\|_1 \log \left\|v_i\right\|_1} \sum_{j=1}^{N_{TU}} v_i^j \log p\left(j|x_i^t\right) \tag{14}$$

where $\left\|v_i\right\|_1$ represents the number of neighborhoods for image $x_i^t$, and $\mathbb{I}(\cdot)$ means the binary selector function. Especially, $\left\|v_i\right\|_1$=1 means the image has no selected neighborhoods and $\mathbb{I}(\cdot)$=0, while $\left\|v_i\right\|_1 > 1$ means the image has selected neighborhoods and then $\mathbb{I}(\cdot) = 1$. The sum of the losses between images and its neighborhoods depends on the number of neighborhoods. When the number of selected neighborhoods is too large, the sum of their losses is large. The balance term $\frac{1}{\left\|v_i\right\|_1 \log \left\|v_i\right\|_1}$ is therefore punished heavily and the sum of the losses is decreased. Otherwise, it punishes slightly. As a consequence, each image will attract relatively similar number of neighborhoods.

### 3.6 Final loss

The final loss function combines the supervised learning for source domain, the one-shot loss for estimating pseudo label and the adaptive selection with balance strategy loss

$$L = (1-\alpha)L_{src} + \alpha\left(L_{nsht} + L_{asb}\right) \tag{15}$$

where $\alpha \in [0,1]$ controls the importance of the source loss and the target loss. $L_{src}$ is the supervised loss of the source domain, $L_{nsht}$ denotes the one-shot loss for the target domain, and $L_{asb}$ represents the neighbor selection loss. Algorithm 1 shows the detailed training procedure of the proposed method.

**Algorithm 1** Training procedure of the proposed method

**Input:** Labeled Source domain $S = \{(x_i^s, y_i^s)\}_1^{N_S}$,
One-shot target domain $T = \{x_i^t\}_1^{N_T} = T_L \cup T_U$,
where $T_L = \{(x_i^{tl}, y_i^{tl})\}_1^{P_T}$, $T_U = \{x_i^{tu}\}_1^{N_T - P_T}$.
**Output:** Adapted Model $\phi(\cdot;W)$.

```
randomly initialize W
initialize feature memories M_L and M_U with zeros.
for each epoch do
    pre-train the backbone network with DAAD layer on S by optimizing Eq. (6).
    for each mini-batch do
        extract target labeled feature F^tl = {f_i^tl}_1^b, f_i^tl = φ(x_i^tl; W)
        extract target unlabeled feature F^tu = {f_i^tu}_1^b, f_i^tu = φ(x_i^tu; W)
        calculate distance d(M_L; F^tu), d(M_U; F^tu)
        assign pseudo labels for unlabeled images according to d(M_L; F^tu) and OS learning.
        select neighbors according to d(M_U; F^tu) for each unlabeled image.
    end for
    train the model φ(·; W) with S, T_L and T_U.
    update the memory M_L and M_U.
end for
```

## 4 Experiments

### 4.1 Datasets and evaluation protocol

**Datasets** We evaluate our method on three re-ID datasets, DukeMTMC-ReID [45], Market-1501 [41] and MSMT17 [34]. DukeMTMC-ReID [45] is a sub-dataset of DukeMTMC [27] and collected from eight cameras for 1404 identities. It has 16522 training images with 702 identities, while 2228 query images and 17661 gallery images with the rest 702 identities. Market-1501 is composed of 1501 identities from six camera views. There are 12936 images of 751 identities for training, 3368 query images and 19732 gallery images of 750 identities for the test. MSMT17 is the largest public-available benchmark currently. It includes 4101 identities from a 15-camera system. The training set contains 32621 images from 1041 identities and the test set includes 11659 images in query and 82161 images in gallery.

**Evaluation protocol** During training, we leverage a dataset as labeled source domain while another dataset as the unlabeled target domain. In testing, we evaluated our method on the target test set by Mean Average Precision (MAP) and Cumulative Matching Characteristic (CMC) which used rank1, rank5, rank10 as metrics. To be fair, we did not utilize the re-Ranking [46] algorithm.

**Table 1** The impact on performance of the number of adapters in DAAD layer.

| Adapter number | Duke → Market | | Market→ Duke | |
|---|---|---|---|---|
| | MAP(%) | R1(%) | MAP(%) | R1(%) |
| 1 | 59.3 | 83.4 | 46.0 | 67.6 |
| 2 | **59.4** | **84.4** | **47.6** | **69.0** |
| 3 | 57.3 | 82.6 | 46.6 | 68.0 |
| 4 | 57.7 | 82.8 | 47.3 | 68.6 |

The best performance is shown in bold

**Table 2** The impacts on performance with applying DAAD layer in different network levels

| DAAD layer pluged | Duke → Market | | Market→ Duke | |
|---|---|---|---|---|
| | MAP (%) | R1 (%) | MAP (%) | R1 (%) |
| N/A | 57.0 | 82.2 | 45.9 | 68.0 |
| layer-1+ | 58.1 | 82.5 | 45.4 | 67.6 |
| layer-2+ | 57.9 | 83.0 | 46.1 | 68.1 |
| layer-3+ | 58.7 | 83.4 | 45.3 | 66.3 |
| layer-4+ | **59.4** | **84.4** | **47.6** | **69.0** |

"N/A" denotes that the proposed DAAD layer is not plugged in the network

The best performance is shown in bold

### 4.2 Implementation details

The proposed method is implemented on Pytorch. With the backbone ResNet-50 [16] pre-trained on ImageNet [7], we plug our DAAD layer after the layer-4 layer of ResNet-50 for demonstrating its re-ID performance. After the DAAD layer, we obtain features by adding an FC-Block which consists of a fully-connected layer, batch normalization, ReLU activation, and dropout. After feature extraction, the images from the source are fed into a classifier which composed of FC and softmax for supervised learning. For these target images, the labeled images are stored in a labeled feature memory to estimate pseudo labels for the unlabeled images. The unlabeled images will be stored in an unlabeled feature memory. Initially, images are all resized to $256 \times 128$ . During training, we perform random cropping, random flipping and random erasing for both source and target domains. In order to enhance the robustness and reduce the impact of camera invariance, we adopt CamStyle [49] as data augmentation in the target domain.

As is well known, there is at least one image can be acquired from each camera of each identity. So we randomly select one image from camera_1 of each identity to constitute the labeled data set. During training, We fixed the first two residual layers of ResNet-50 to save the GPU memory. The last stride size of the last residual layer is set to 1 to enlarge the spatial resolution. In our experiments, the number of training iteration and the batch size are set to 50 and 128 respectively. The weight decay and momentum for Stochastic Gradient Descent (SGD) optimizer are set to $5 \times 10^{-4}$ and 0.9, respectively. The learning rate was initially set to 0.1 and will multiply 0.1 after 40 epochs. We also have the following hyper-parameters. Without specification, we set the initial update rate $\lambda$ to 0.01 for both memories, and then increases it linearly as $\lambda = 0.01 \times$ epoch. We set $\eta$ to 0.85 for the region size to drop. For the importance of these losses, it is controlled by $\alpha$ which is set to 0.3. The temperature fact $\beta$ is set to 0.05.

In the community of person re-identification, the dataset usually include training set and test set, and the test set consists of query set and gallery set. Based on the above model design and parameter setting, the training process is performed by feeding the labeled source training set images and the one-shot target training images into the model. According to the experimental results, the model tends to converge after 50 epochs training iteratively. Then the target testing set (query and gallery) is fed into the optimized model to test the performance by compute the similarity between query images and the gallery images.

### 4.3 Parameter analysis

In this subsection, we leverage the pre-trained ResNet-50 as our backbone network. We analyze the impact on performance of the position that DAAD layer is plugged and these important hyper-parameters in our method. During our experiments, we change one parameter while keeping the others fixed.

**Table 3** Performance according to *drop_rate*

| Drop Map (%) | Discriminative Map (%) | Duke → Market | | Market → Duke | |
|---|---|---|---|---|---|
| | | MAP (%) | R1 (%) | MAP (%) | R1 (%) |
| 0 | 100 | 57.5 | 82.2 | 45.8 | 67.5 |
| 10 | 90 | 57.8 | 83.3 | 45.1 | 67.1 |
| 15 | 85 | **59.4** | **84.4** | **47.6** | **69.0** |
| 20 | 80 | 56.6 | 82.7 | 45.9 | 67.8 |
| 30 | 70 | 57.7 | 82.7 | 46.3 | 67.2 |
| 40 | 60 | 56.9 | 81.8 | 45.2 | 67.1 |
| 50 | 50 | 56.6 | 81.1 | 45.1 | 67.2 |
| 60 | 40 | 56.3 | 80.8 | 44.9 | 67.0 |
| 70 | 30 | 55.7 | 81.2 | 43.8 | 66.5 |
| 80 | 20 | 55.4 | 80.5 | 44.0 | 66.1 |
| 90 | 10 | 55.8 | 80.3 | 43.2 | 65.8 |
| 100 | 0 | 52.1 | 81.3 | 42.0 | 65.4 |
| N/A | N/A | 57.0 | 82.2 | 45.9 | 68.0 |
| 15 | N/A | 57.5 | 82.7 | 44.6 | 66.8 |
| N/A | 85 | 58.2 | 83.3 | 45.7 | 67.6 |

"N/A" represents either the "Drop Map" or the "Discriminative Map" is not applied or both of them are not applied

The best performance is shown in bold

**Impact of the number of adapters in DAAD layer** We claim that two adapters are used in DAAD layers according to the number of domains. To verify the suppose, the impact on performance of the number of adapters in DAAD layer is summarized in Table 1. We can conclude that in both cases, Duke transfer to Market and Market transfer to Duke, the best performance can be obtained when the number of adapters is equal to two. This suggest that the best performance emerges when the number of the adapters is identical with the number of domains.

**The impact of applying DAAD layer at different layers.** Since the DAAD layer is plug-and-play, we plug DAAD layer from low-level layers to high-level layers. The impacts on performance are recorded in Table 2. From Table 2, applying DAAD layer can increase both the MAP and rank-1 accuracy when Duke is transferred to Market. Especially, MAP and rank-1 accuracy can be increased except for plugging it after layer-1 when Market is transferred to Duke. The best performance can be achieved by applying DAAD layer after layer-4. This indicates that high-level layers are more related to class-specific features while lower-level feature maps are general.

**The effect of drop_rate** Evaluation with different values of drop_rate on MAP and rank-1 accuracy displayed as in Table 3. We observed that the best performance with MAP and rank-1 accuracy can be achieved when drop_rate is 0.15. While the MAP and rank-1 accuracy have a large degradation when the drop map is leveraged at every iteration, *i.e.* the probability of drop map is 100%. This is because the model never observed the most discriminative regions. Meanwhile, We also observed that the MAP and rank-1 accuracy decreases as the drop_rate increases. The performance is poor when the discriminative map is adopted with probability 100% due to the model overfits on the source domain. To evaluate the effect of the discriminative map

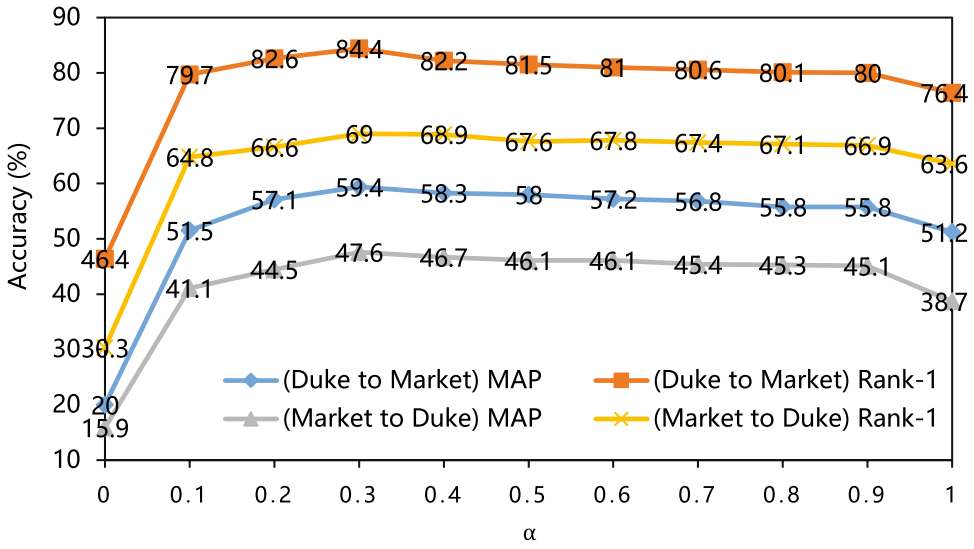


**Fig. 3** Effect on performance of the losses importance: $\alpha$

**Table 4** Comparison of various methods under different settings on DukeMTMC-ReID (Duke) and Market-1501 (Market)

| Methods | Market (M, %) | | | | | Duke (D, %) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Src. | MAP | R1 | R5 | R10 | Src. | MAP | R1 | R5 | R10 |
| Sup | N/A | 69.4 | 87.6 | 95.5 | 97.2 | N/A | 57.8 | 75.6 | 87.3 | 90.6 |
| DT | D | 20.0 | 46.4 | 63.7 | 70.6 | M | 15.9 | 30.3 | 45.4 | 52.5 |
| AE [9] | D | 58.0 | 81.6 | 91.9 | 94.6 | M | 46.7 | 67.9 | 79.2 | 83.6 |
| Ours (DAAD+) | D | 58.4 | 83.8 | 92.6 | 94.8 | M | 47.3 | 68.6 | 79.8 | 83.3 |
| Ours (OS+) | D | 57.0 | 82.2 | 92.3 | 94.7 | M | 45.9 | 68.0 | 78.8 | 82.2 |
| Ours (DAAD+OS, DAADOS) | D | **59.4** | **84.4** | **92.9** | **95.3** | M | **47.6** | **69.0** | **79.7** | **83.6** |

*Sup* supervised learning, *DT* directly apply the source-trained model on the target domain, *DAAD* domain Adaptive Attention based Dropout, *OS* one-shot learning, *Src* source domain

The best performance is shown in bold

and the drop map, we deactivating them respectively. From the lower part of Table 3, we conclude that only applying the drop map or the discriminative map with the best probability achieves relatively lower MAP and rank-1 accuracy than applying them at the same time.

**The importance of losses** $\alpha$ To investigate the effect by $\alpha$ in Eq. (15), different values are employed as shown in Fig. 3. When $\alpha = 0$, the model is regarded as the direct transfer (DT) that the model is only trained on the labeled source. It is obvious that our method ($\alpha > 0$) significantly improves the performance at all these values. Besides, our method outperforms the DT method by a large margin when the model is only trained on the unlabeled target ($\alpha = 1$). Besides, the best performance can be achieved when $\alpha = 0.3$ both for Duke is transferred to Market and Market is transferred to Duke. This demonstrates the effectiveness of our method.

### 4.4 Performance evaluation

To evaluate the effectiveness of our proposed DAAD layer and the one-shot adaptive learning, we conduct ablation studies in Table 4. "Sup" stands for the experiment in the supervised learning manner that ground truth labels are available in the target domain, "DT" denotes the experiment that directly apply the source trained model on the target domain. All the other experimental parameter settings are same as the proposed DAADOS. The performance of Sup is regarded as the upper bound of our method because it is trained with all the labels. The upper bound specifies the best performance which our method may achieve. From the upper part of Table 4, the upper bound in our method is 69.4% in MAP and 87.6%

**Table 5** Comparison of proposed method with these state-of-the-art domain adaptive methods on DukeMTMC-reID (Duke), Market-1501 (Market)

| Methods | Duke → Market (%) | | | | Market → Duke (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | MAP | R1 | R5 | R10 | MAP | R1 | R5 | R10 |
| UMDL[29] | 12.4 | 34.4 | 52.6 | 59.6 | 7.3 | 18.5 | 31.4 | 37.6 |
| PTGAN[34] | – | 38.6 | – | 66.1 | – | 27.4 | – | 50.7 |
| PUL[12] | 20.5 | 45.5 | 60.7 | 66.7 | 16.4 | 30.0 | 43.4 | 48.5 |
| SPGAN[8] | 22.8 | 51.5 | 70.1 | 76.8 | 22.3 | 41.1 | 56.6 | 63.0 |
| SPGAN+LMP[8] | 26.7 | 57.7 | 75.8 | 82.4 | 26.2 | 46.4 | 62.3 | 68.0 |
| MMFA[22] | 27.4 | 56.7 | 75.0 | 81.8 | 24.7 | 45.3 | 59.8 | 66.3 |
| TJ-AIDL[32] | 26.5 | 58.2 | 74.8 | 81.1 | 23.0 | 44.3 | 59.6 | 65.0 |
| HHL[47] | 31.4 | 62.2 | 78.8 | 84.0 | 27.2 | 46.9 | 61.0 | 66.7 |
| CamStyle[49] | 27.4 | 58.8 | 78.2 | 84.3 | 25.1 | 48.4 | 62.5 | 68.9 |
| ECN[48] | 43.0 | 75.1 | 87.6 | 91.6 | 40.4 | 63.3 | 75.8 | 80.4 |
| SSG[13] | 58.3 | 80.0 | 90.0 | 92.4 | 53.4 | 73.0 | 80.6 | 83.2 |
| AE[9] | 58.0 | 81.6 | 91.9 | 94.6 | 46.7 | 67.9 | 79.2 | 83.6 |
| ACT[37] | 60.6 | 80.5 | – | – | 54.5 | 72.4 | – | – |
| MMCL[31] | 60.4 | 84.4 | 92.8 | 95.0 | 51.4 | 72.4 | 82.9 | 85.0 |
| Ours (DAADOS) | **59.4** | **84.4** | **92.9** | **95.3** | **47.6** | **69.0** | **79.7** | **83.6** |

The best performance is shown in bold

**Table 6** Comparison of proposed method with these state-of-the-art domain adaptive methods on MSMT17

| Methods | Duke → MSMT17 | | | | Market-1501 → MSMT17 | | | |
|---|---|---|---|---|---|---|---|---|
| | MAP | R1 | R5 | R10 | MAP | R1 | R5 | R10 |
| PTGAN[34] | 3.3 | 11.8 | – | 27.4 | 2.9 | 10.2 | – | 24.4 |
| ECN[48] | 10.2 | 30.2 | 41.5 | 46.8 | 8.5 | 25.3 | 36.3 | 42.1 |
| SSG[13] | 13.3 | 32.2 | – | 51.2 | 13.2 | 31.6 | – | 49.6 |
| AE[9] | 11.7 | 32.3 | 44.4 | 50.1 | 9.2 | 25.5 | 37.3 | 42.6 |
| Ours (DAADOS) | **13.6** | **37.5** | **49.1** | **54.4** | **11.7** | **33.3** | **44.8** | **50.0** |

The best performance is shown in bold

in rank-1 accuracy for Market, while it is 57.8% and 75.6% in MAP and rank-1 accuracy for Duke. We also record the performance about direct transfer which trains on source domain and tests on target domain. We observed from Table 4 that direct transfer has a serious performance drop. Concretely, the model trained on Duke direct transfer to Market achieves 20.0% in MAP and 46.4% in rank-1 accuracy, the performance has 49.4% and 43.2% drop in MAP and rank-1 accuracy respectively. The same phenomenon occurs when trains the model on Market while tests on Duke.

In the lower part of Table 4, we utilize AE [9] as our baseline. Based on AE, we plugged our proposed DAAD layer with one-shot learning. By applying the DAAD layer or the one-shot learning, the performance gain can be achieved both on Market and Duke. While the best performance is achieved by applying them at the same time. Specially, our method with DAAD layer and one-shot learning improves the performance by 1.4% and 2.8% in MAP and rank-1 accuracy when the model is trained on Duke while tested on Market. Similarly, we have the performance gain 0.9% in MAP and 1.1% in rank-1 accuracy on Duke. This demonstrates that DAAD layer with one-shot learning is an effective way to improve the performance for person re-ID.

### 4.5 Comparison with state-of-the-art

We compare our method with these state-of-the-art domain adaptive methods on three datasets DukeMTMC-reID, Market-1501 and MSMT17 as displayed in Table 5, Table 6. Table 5 summaries the state-of-the-art domain adaptive methods on DukeMTMC-reID and Market-1501. As for MSMT17 results are shown in Table 6.

Table 5 reports the comparisons tested on DukeMTMC-reID, Market-1501. To be fair, we only compare the methods that training and test are all on the same domains. Specially, UMDL [29] and PUL [34] are initialized with the model trained on the labeled source domain, while trained on the unlabeled target domain. The following ten methods, PTGAN [34], SPGAN [8], SPGAN+LMP [8], MMFA [22], TJ-AIDL [32], HHL [47], CamStyle [49], ECN [48], SSG [13] and AE [9], are the standard domain adaptive methods that trained on both the labeled source and the unlabeled target domains. As can be seen, our method is competitive and outperforms almost all the existing domain adaptive methods except SSG [13]. Specially, ECN [48], SSG [13] and AE [9] are all clustered-based methods while their performance is lagged behind the proposed method. Concretely, our method achieves MAP = 59.0% and Rank-1 accuracy is 84.0% when trained on labeled DukeMTMC-reID and Market-1501 that with one-shot setting, tested on Market-1501. We also obtain MAP = 46.8% and rank-1=68.8% vice-versa. Compared to AE [9] which we adopt as baseline, our method exceeds AE by 1.0% in MAP and 2.8% in Rank-1 accuracy tested on Market-1501. When tested on Duke, the MAP and rank-1 accuracy are 0.9% and 1.1% higher than AE, respectively. From Table 5, the performance of the proposed method is superior to most methods. It is also comparable to the new published state-of-the-art methods, eg. ACT [37] and MMCL [31].

The proposed method is also evaluated on MSMT17 which is a more challenging and larger dataset. As MSMT17 is newly public-available, there are only a few domain adaptive methods that experimented on MSMT17. The comparison results are displayed in Table 6. Obviously, our method surpasses the current best method SSG [13] by 0.3% and 5.3% in MAP and rank-1 accuracy when Duke adaptives to MSMT17, 1.7% in rank-1 accuracy when Market adaptives to MSMT17.

### 4.6 Visualization of DAAD features

The feature maps from DAAD are visualized as [50]. In Fig. 4, the left part is visualization of Duke images while Market images are on the right part. Each image is followed with a feature map from DAAD layer. The query column denotes the original query image. Positive match has the same identity with the query while negative match has the different identity with query but with the similar appearance. For images and its positive match, their feature maps have the similar importance to corresponding regions. For

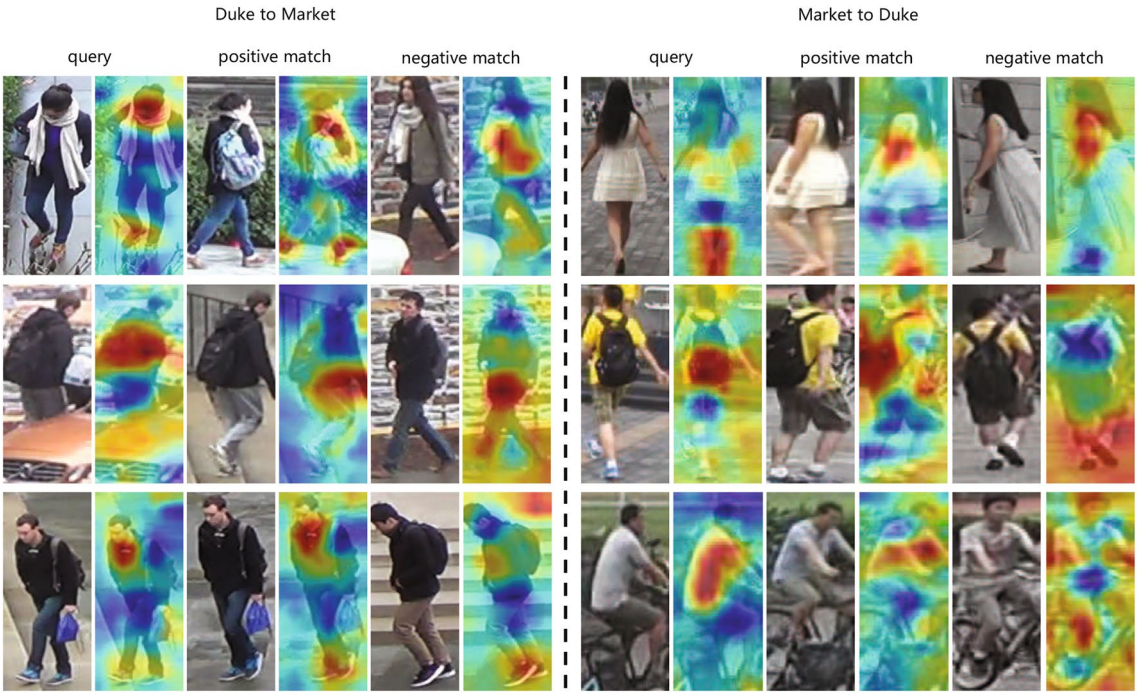


**Fig. 4** Visualization of DAAD features. The warmer color present the greater weight

instance, query and its positive match in the first row of Fig. 4 in Duke focus on the foot region. However, its negative match mostly focus on the upper body. The same phenomenon can be find in other images. This demonstrate that our DAAD layer is effective to find the image which has the same identity with query.

## 5 Conclusion

In this paper, a domain-adaptive-attention-based-dropout (DAAD) layer according to one-shot learning is proposed for cross-domain person re-identification. By erasing the most discriminative regions stochastically, the DAAD layer can infer the original domain for input imanges automatically and then domain-sensitive features can be extracted. In addition, one-shot learning makes the model estimate pseudo labels for unlabeled images in target domain with high confidence. Both the above two designs can improve the performance in cross-domain person re-ID. Experiment results on DukeMTMC-ReID, Market-1501, and MSMT17 datasets demonstrate that the proposed method improves the performance for cross-domain person re-ID.

**Acknowledgements** This work is partially supported by National Natural Science Foundation of China under Grant nos. 61872188, U1713208, 61972204, 61672287, 61861136011, 61773215.

## References

1. Bak S, Carr P (2017) One-shot metric learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2990–2999
2. Bousmalis K, Silberman N, Dohan D, Erhan D, Krishnan D(2017) Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3722–3731
3. Chen KW, Lai CC, Lee PJ, Chen CS, Hung YP (2011) Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras. IEEE Trans Multimed 13(4):625–638
4. Chen T, Ding S, Xie J, Yuan Y, Chen W, Yang Y, Ren Z, Wang Z (2019) Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 8351–8361
5. Choe J, Shim H (2019) Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2219–2228
6. Chong Y, Peng C, Zhang J, Pan S (2021) Style transfer for unsupervised domain-adaptive person re-identification. Neurocomputing (NC) 422:314–321
7. Deng J, Dong W, Socher R, Li L.J, Li K, Fei-Fei L(2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 248–255
8. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J (2018) Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 994–1003

9. Ding Y, Fan H, Xu M, Yang Y (2020) Adaptive exploration for unsupervised person re-identification. ACM Trans Multimed Comput Commun Appl 16(1):1551–6857
10. Dong H, Lu P, Zhong S, Liu C, Ji Y, Gong S (2018) Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning. Neurocomputing 307:25–37
11. Dong X, Yu S.I, Weng X, Wei S.E, Yang Y, Sheikh Y (2018) Supervision-by-registration: an unsupervised approach to improve the precision of facial landmark detectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 360–368
12. Fan H, Zheng L, Yan C, Yang Y (2018) Unsupervised person re-identification: clustering and fine-tuning. ACM Trans Multimed Comput Commun Appl (TOMM) 14(4):1–18
13. Fu Y, Wei Y, Wang G, Zhou Y, Shi H, Huang TS (2019) Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 6112–6121
14. Ganin Y, Lempitsky V (2015) Unsupervised domain adaptation by backpropagation. In: International conference on machine learning (ICML), pp 325–333
15. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A.J (2007) A kernel method for the two-sample-problem. In: Advances in neural information processing systems, pp 513–520
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
17. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
18. Ji Z, Wang H, Han J, Pang Y (2020) SMAN: stacked multimodal attention network for cross-modal image-text retrieval. IEEE Trans Cybern (TCYB). https://doi.org/10.1109/TCYB.2020.2985716
19. Ji Z, Xiong K, Pang Y, Li X (2019) Video summarization with attention-based encoder-decoder networks. IEEE Trans Circuits Syst Video Technol (TCSVT) 30(6):1709–1717
20. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
21. Li Z, Tang J (2015) Unsupervised feature selection via nonnegative spectral analysis and redundancy control. IEEE Trans Image Process 24(12):5343–5355
22. Lin S, Li H, Li C.T, Kot A.C (2018) Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. arXiv preprint arXiv:1807.01440
23. Lin Y, Dong X, Zheng L, Yan Y, Yang Y (2019) A bottom-up clustering approach to unsupervised person re-identification. In: Proceedings of the AAAI conference on artificial intelligence vol 33, pp 8738–8745
24. Lin Y, Guo F, Cao L, Wang J (2016) Person re-identification based on multi-instance multi-label learning. Neurocomputing 217:19–26
25. Liu H, Xiao Z, Fan B, Zeng H, Zhang Y, Jiang G (2021) PrGCN: probability prediction with graph convolutional network for person re-identification. Neurocomputing (NC) 423:57–70
26. Rezaei S, Tahmoresnezhad J, Solouk V (2020) A transductive transfer learning approach for image classification. Int J Mach Learn Cybern 12(3):747–762
27. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. Springer, New York, pp 17–35
28. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
29. Song L, Cheng W, Zhang L, Bo D, Wang X (2018) Unsupervised domain adaptive re-identification: theory and practice. ArxivPrint, arXiv:1807.11334
30. Tzeng E, Hoffman J, Saenko K, Darrell T (2017) Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7167–7176
31. Wang D, Zhang S (2020) Unsupervised person re-identification via multi-label classification. In: IEEE computer society conference on computer vision and pattern recognition (CVPR), pp 10981–10990
32. Wang J, Zhu X, Gong S, Li W (2018) Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2275–2284
33. Wang X, Bao A, Cheng Y, Yu Q (2019) Weight-sharing multi-stage multi-scale ensemble convolutional neural network. Int J Mach Learn Cybern 10(7):1631–1642
34. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
35. Wu Y, Lin Y, Dong X, Yan Y, Bian W, Yang Y (2019) Progressive learning for person re-identification with one example. IEEE Trans Image Process 28(6):2872–2881
36. Xia B.N, Gong Y, Zhang Y, Poellabauer C (2019) Second-order non-local attention networks for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3760–3769
37. Yang F, Li K, Zhong Z, Luo Z, Sun X, Cheng H, Guo X, Huang F, Ji R, Li S (2020) Asymmetric co-teaching for unsupervised cross domain person re-identification. In: the AAAI conference on artificial intelligence (AAAI), pp 12597–12604
38. Yu C, Wang J, Chen Y, Qin X (2019) Transfer channel pruning for compressing deep domain adaptation models. Int J Mach Learn Cybern 10(11):3129–3144
39. Zhang X, Luo H, Fan X, Xiang W, Sun Y, Xiao Q, Jiang W, Zhang C, Sun J (2017) Alignedreid: surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184
40. Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6848–6856
41. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
42. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. ArxivPrint, arXiv:1610.02984
43. Zheng M, Karanam S, Wu Z, Radke RJ (2019) Re-identification with consistent attentive siamese networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5735–5744
44. Zheng W, Gong S, Xiang T (2016) Towards open-world person re-identification by one-shot group-based verification. IEEE Trans Pattern Anal Mach Intell (TPAMI) 38(3):591–606. https://doi.org/10.1109/TPAMI.2015.2453984
45. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE international conference on computer vision, pp 3754–3762

46. Zhong Z, Zheng L, Cao D, Li S (2017) Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1318–1327
47. Zhong Z, Zheng L, Li S, Yang Y (2018) Generalizing a person retrieval model hetero-and homogeneously. In: Proceedings of the European conference on computer vision (ECCV), pp 172–188
48. Zhong Z, Zheng L, Luo Z, Li S, Yang Y (2019) Invariance matters: exemplar memory for domain adaptive person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 598–607
49. Zhong Z, Zheng L, Zhenga Z, Li S, Yang Y (2019) Camstyle: a novel data augmentation method for person re-identification. IEEE Trans Image Process (TIP) 28(3):1176–1190
50. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929