Self-Training With Progressive Augmentation for Unsupervised Cross-Domain Person Re-Identification*

Xinyu Zhang[1] Jiewei Cao[2] Chunhua Shen[2] Mingyu You[1]
[1]Tongji University, China [2]The University of Adelaide, Australia

Abstract

Person re-identification (Re-ID) has achieved great improvement with deep learning and a large amount of labelled training data. However, it remains a challenging task for adapting a model trained in a source domain of labelled data to a target domain of only unlabelled data available. In this work, we develop a self-training method with progressive augmentation framework (PAST) to promote the model performance progressively on the target dataset. Specially, our PAST framework consists of two stages, namely, conservative stage and promoting stage. The conservative stage captures the local structure of target-domain data points with triplet-based loss functions, leading to improved feature representations. The promoting stage continuously optimizes the network by appending a changeable classification layer to the last layer of the model, enabling the use of global information about the data distribution. Importantly, we propose a new self-training strategy that progressively augments the model capability by adopting conservative and promoting stages alternately. Furthermore, to improve the reliability of selected triplet samples, we introduce a ranking-based triplet loss in the conservative stage, which is a label-free objective function based on the similarities between data pairs. Experiments demonstrate that the proposed method achieves state-of-the-art person Re-ID performance under the unsupervised cross-domain setting.

Code is available at: tinyurl.com/PASTReID

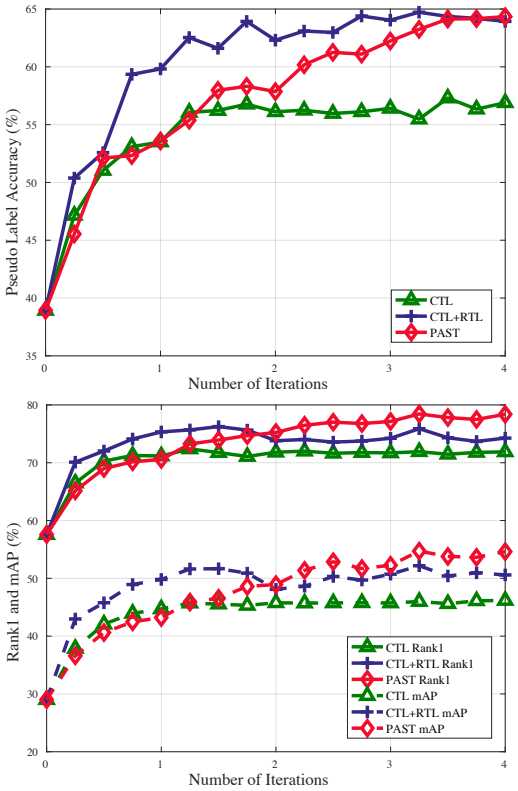

Figure 1 – Label quality *vs.* model generalization. The accuracy of pseudo labels prediction (top) and performance comparison (bottom) of different methods over training iterations. Here we use Duke [45] as the source domain and Market-1501 [44] as the target domain.

1. Introduction

Person re-identification (Re-ID) is a crucial task in surveillance and security, which aims to locate a target pedestrian across non-overlapping camera views using a probe image. With the advantages of convolutional neural networks (CNN), many person Re-ID works focus on *supervised learning* [14, 32, 39, 5, 48, 4, 6, 20, 31, 7, 26] and achieve satisfactory improvements. Despite the great success, they depend on large labelled datasets which are costly and sometime impossible to obtain.

To tackle this problem, a few *unsupervised learning* methods [36, 24, 22] propose to take advantage of abundant unlabelled data, which are easier to collect in general. Unfortunately, due to lack of supervision information, the performance of unsupervised methods is typically weak, thus being less effective for practical usages. In contrast, *unsupervised cross-domain* methods [38, 10, 36, 47, 18, 27, 12, 25, 21, 30] propose to use both labelled datasets (source domain) and unlabelled datasets (target domain). However, directly applying the models trained in the source domain to

*Work was done when X. Zhang was visiting The University of Adelaide, Australia. First two authors contributed to this work equally. C. Shen is the corresponding author: chunhua.shen@adelaide.edu.au

the target domain leads to unsatisfactory performances due to the inconsistent characteristics between the two domains, which is known as the *domain shift* problem [21]. In unsupervised cross-domain Re-ID, the problem becomes how to transfer the learned information of a pre-trained model from the source domain to the target domain effectively in an unsupervised manner.

Some domain transfer methods [47, 18, 27, 12, 25, 21, 30, 24] have taken great efforts to address this challenge, where the majority of them are based on *pseudo label* estimation [12, 30, 25]. They extract embedding features of unlabelled target datasets using the pre-trained model and apply unsupervised clustering methods (*e.g.*, k-means and DBSCAN [11]) to separate the data into different clusters. The samples in the same cluster are assumed to belong to the same person, which are adapted as pseudo labels for supervised learning. The drawback of these methods is that the performance highly depends on the clustering quality, reflecting on whether or not samples with the same identity are assigned to one cluster. In other words, the performance relies on to what extent are the pseudo labels consistent with ground truth identity labels. Since the percentage of corrupted labels largely affect the model generalization on the target dataset [42], we propose a method to improve the quality of labels in a progressive way which results in considerable improvement of model generalization on the unseen target dataset.

Here we propose a new *Self-Training with Progressive Augmentation framework* (PAST) to: 1) restrain error amplification at early training epochs when the quality of pseudo label is low; and 2) progressively incorporate more confidently labelled samples for self-training when the label quality is becoming better. PAST has two learning stages, *i.e.*, *conservative* and *promoting* stage, which consider complementary data information via different learning strategies for self-training.

Conservative Stage. As shown in Figure 1, the percentage of correctly labelled data is low at first due to the domain shift. In this scenario, we need to select *confidently labelled* examples to reduce the label noise. We consider the similarity score between images as a good indicator of confidence measure. Besides the widely used clustering-based triplet loss (CTL) [17], which is sensitive to the quality of pseudo labels generated from the clustering method, we propose a novel label-free loss function, *ranking-based triplet loss (RTL)*, to better capture the characteristic of data distribution in the target domain.

Specifically, we calculate the ranking score matrix for the whole target dataset and generate triplets by selecting the positive and negative examples from top ranked images for each anchor. The triplets are then fed into the model and trained with the proposed RTL. In the conservative stage, we mainly consider the local structure of data distribution which is crucial for avoiding model collapse when the label quality is mediocre at early learning epochs.

Promoting Stage. However, as the number of training triplets dramatically grows in large datasets and triplets only focus on local information, the learning process with triplet loss inevitably becomes instability and suffers from the suboptimal result, as shown by the "CTL" and "CTL+RTL" in Figure 1. To remedy this issue, we propose to use the global distribution of data points for network training at the promoting stage. Specifically, we treat each cluster as a class and convert the learning process into a classification problem. Softmax cross-entropy loss is used to force different categories staying apart for encouraging inter-class separability. After the promoting stage, the model is prone to be more stable which facilitates learning the discriminative features. Since the error is most likely amplified when training on images with extremely corrupted labels using the softmax cross-entropy loss, we employ this stage following the conservative learning stage and carry out these two stages interchangeably. With this alternate process, our proposed PAST framework can stabilize the training process and progressively improve the capability of model generalization on the target domain.

To summarize, our main contributions are as follows:

1) We present a novel self-training with progressive augmentation framework (PAST) to solve the unsupervised cross-domain person Re-ID problem. By executing the two-stage self-training process, namely, conducting conservative and promoting stage alternately, our method considerably improves the model generalization on unlabelled target-domain datasets.

2) We propose a ranking-based triplet loss (RTL), solely relying on similarity scores of data points, to avoid selecting triplet samples with unreliable pseudo labels.

3) We take advantage of global data distribution for model training with softmax cross-entropy loss, which is beneficial for training stability and promoting the capability of model generalization.

4) Experimental results on three large-scale datasets indicate the effectiveness of our proposed method on the task of unsupervised cross-domain person Re-ID.

2. Related Work

Supervised Person Re-ID. Most existing deep person Re-ID methods follow a supervised setting. They mainly focus on well-designed model architectures [32, 8, 37, 33, 43, 39, 7, 4], additional attributions [5, 29, 48, 6] and metric learning [14, 17, 23, 46]. Although significant progress has been obtained by these methods, they all require a large amount of labelled training data, which is costly to obtain due to the huge data volume and drastic appearance changes among different people.

Unsupervised Person Re-ID. To alleviate the above limi-

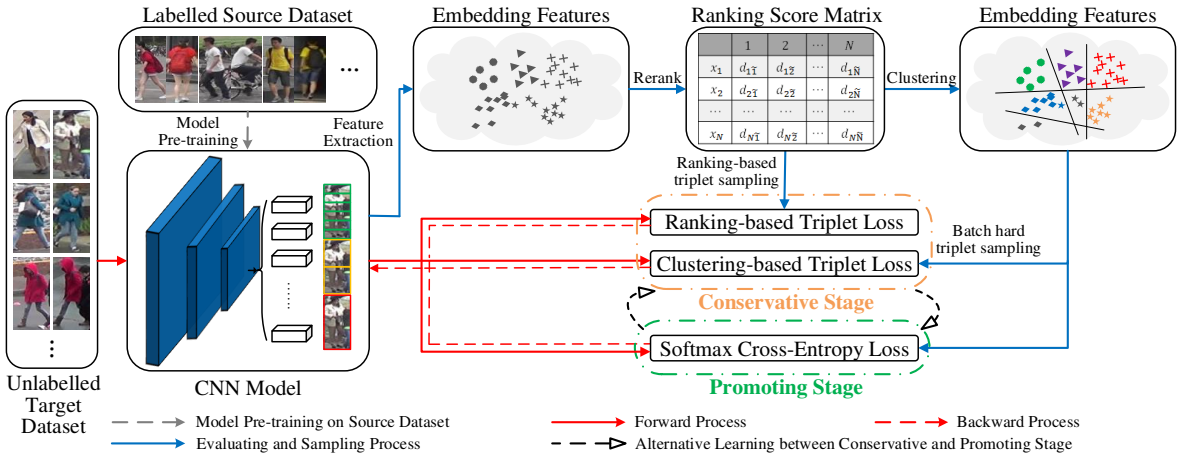

Figure 2 – The overview of our self-training framework with progressive augmentation (PAST). The model is pre-trained on the labelled source dataset. During training, we first carry out a sampling process, which consists of extracting embedding features of unlabelled target dataset with the current model and calculating the ranking score matrix with Eq. (2). We then assign pseudo labels to training samples via HDBSCAN [3] clustering method. After that, we conduct conservative stage by using clustering-based triplet loss (CTL) and the proposed ranking-based triplet loss (RTL) to update the model. In promoting stage, the softmax cross-entropy loss is employed to further improve the capability of the model. Note that the conservative stage and promoting stage alternate iteratively during the whole learning process. For Re-ID evaluation, we extract the embedding features for both query and gallery images and use the cosine distance for ranking.

tation, unsupervised person Re-ID methods [40, 24, 22, 34, 35] are proposed to make full use of large-scale unlabelled data. Most of them exploit cross-view identity-specific information to capture discriminate features [40, 35] or adopt clustering methods to separate unlabelled images into different classes [22, 24]. However, there is till a large performance gap between supervised Re-ID methods and unsupervised ones.

Unsupervised Cross-Domain Person Re-ID. Recently, researchers pay intensive attention to unsupervised cross-domain person Re-ID algorithms [38, 10, 36, 47, 18, 27, 12, 25, 21, 30] which leverages the labelled data in the source domain. They all focus on overcoming domain shift so as to learn domain-invariant feature representation.

Among these existing works, PTGAN [38] and SPGAN [10] transfer source images into target-domain style by CycleGAN and then use translated images to train a model. Another line of unsupervised cross-domain person Re-ID works [36, 47, 25, 18] combine other auxiliary information as an assistant task to improve the model generalization. For instance, TFusion [25] integrates spatio-temporal patterns to improve the Re-ID precision, while EANet [18] uses pose segmentation. TJ-AIDL [36] learns an attribute-semantic and identity discriminative feature representation space simultaneously, which can be transferred to any new target domain for re-id tasks. Similar to supervised learning, these domain adaptation approaches suffer from the need of collecting attribute annotations.

Beyond the above methods, in general, some approaches [12, 30, 25, 28] focus on estimating pseudo identity labels on the target domain so as to learn deep models in a supervised manner. Image matching [1, 2] and clustering methods are used to generate a series of training data which are used to update networks with an embedding loss (*e.g.*, triplet loss [17] or contrastive loss) [30, 25] or classification loss (*e.g.*, softmax cross-entropy loss) [12]. However, embedding loss functions suffer from the limitation of sub-optimal results and slow convergence, while classification loss highly depends on the quality of pseudo labels. While the work in [41] introduces a simple domain adaptation framework which also use both triplet loss and softmax cross-entropy loss jointly, it aims at solving the one-shot leaning problem.

3. Our Method

For unsupervised cross-domain person Re-ID, the problem that we concentrate on is how to learn robust feature representations for unlabelled target datasets using the prior knowledge from the labelled source datasets. In this section, we present the proposed *self-training with progressive augmentation framework* (PAST) in detail.

3.1. Framework Overview

The overview of our proposed PAST is described in Figure 2, which has two main components: *conservative stage* and *promoting stage*.

We first train a CNN model M using labelled source training dataset S in a supervised manner. Then, this pre-trained model is utilized to extract features $\mathbf{F}$ of all training images on the target domain T. In the conservative stage, based on the ranking score matrix D_R computed on the above image features $\mathbf{F}$, we can generate a more reliable training set T_U via the HDBSCAN [3] clustering method

(other clustering methods can be employed here too). This updated training set T_U is a subset of the whole training data T. Combining with two triplet-based loss functions, *i.e.*, *clustering-based triplet loss (CTL)* and the proposed *ranking-based triplet loss (RTL)*, local relationship in the target domain can be captured from triplets formed by the current training set T_U for model optimization. After that, we extract features $\mathbf{F}_U$ on the updated training set T_U via using the updated model M. In the promoting stage, with the new features $\mathbf{F}_U$ from the conservative stage, we propose to employ softmax cross-entropy loss for further optimizing the network. At this stage, the global distribution of the training set is considered to improve the discrimination of feature representation. Finally, the capability of model generalization is improved gradually by training the model with the conservative stage and promoting stage alternately. The details of PAST are described in Algorithm 1.

3.2. Conservative Stage

The task of unsupervised cross-domain Re-ID is to develop a method that is able to learn robust features on unlabelled target domain, where the objective is to get same samples together and push different samples far from each other. Triplet loss [47, 30, 25] has been proved to be able to discover meaningful underlying local structure of data distribution by generating reliable triplets of the target data. Different from the supervised setting, pseudo labels are assigned to unlabelled samples, which is more difficult to construct high-quality triplets. Therefore, our goal is to design a learning strategy to not only generate reliable samples but also improve the model performance.

In practice, we conduct the following procedure in the conservative stage. At the beginning, on the whole training dataset T: $\{x_1, x_2, ..., x_N\}$, we extract features $\mathbf{F}$: $\{\mathbf{f}(x_1), \mathbf{f}(x_2), ..., \mathbf{f}(x_N)\}$ from the current model M, and adopt the k-reciprocal encoding [46], which is a variation of the Jaccard distance between nearest neighbors sets, to generate the distance matrix D as:

$$\begin{gathered} D = [\mathbf{D}_J(x_1)\ \mathbf{D}_J(x_2)\ \ldots\ \mathbf{D}_J(x_N)]^T, \\ \mathbf{D}_J(x_i) = [d_J(x_i, x_1)\ d_J(x_i, x_2)\ \ldots\ d_J(x_i, x_N)], \\ \forall i \in \{1,\ 2,\ \ldots,\ N\}, \end{gathered} \tag{1}$$

where $\mathbf{D}_J(x_i)$ represents the distance vector of one specific person x_i with all training images. $d_J(x_i, x_j)$ is the Jaccard distance between sample x_i and x_j.

Since smaller distance means greater similarity between images, we sort every distance vector $\mathbf{D}_J(x_i)$ in ascending order, yielding a ranking score matrix D_R:

$$\begin{gathered} D_R = [\mathbf{D}_R(x_1)\ \mathbf{D}_R(x_2)\ \ldots\ \mathbf{D}_R(x_N)]^T, \\ \mathbf{D}_R(x_i) = [d_J(x_i, \widetilde{x}_1)\ d_J(x_i, \widetilde{x}_2)\ \ldots\ d_J(x_i, \widetilde{x}_N)], \\ \forall i \in \{1,\ 2,\ \ldots,\ N\}, \end{gathered} \tag{2}$$

where $\mathbf{D}_R(x_i)$ is a sorted copy of $\mathbf{D}_J(x_i)$. Given a specific sample x_i, $\widetilde{x}_j$ in $d_J(x_i, \widetilde{x}_j)$ represents the j-th most similar sample.

Then, we apply a hierarchical density-based clustering algorithm (HDBSCAN) [3] on D_R to split the whole training images into different clusters. Each cluster is considered as a specific class, in which samples of the same cluster can be assigned to the same pseudo label. Note that some images are discarded since there is no corresponding cluster for them. Thus, images with assigned labels are used as the updated training set T_U to further optimize the model. We combine two types of triplet loss functions together to update the model, *i.e.*, clustering-based triplet loss and ranking-based triplet loss, as described below.

Clustering-based Triplet Loss (CTL). Batch hard triplets mining [17] is proposed to mine the relations among samples within a mini-batch. Following the setting in [17], we randomly sample P clusters and K instances of each cluster to compose a mini-batch with size of PK. For each anchor image x_a, the corresponding hardest positive sample x_p and the hardest negative sample x_n within the batch are selected to form a triplet. Since the pseudo labels are from a clustering method, we name this loss function as *clustering-based triplet loss* (CTL), which is formulated as follows:

$$\begin{aligned} L_{CTL} &= \sum_{a=1}^{PK} [m + ||\mathbf{f}(x_a) - \mathbf{f}(x_p)||_2 - ||\mathbf{f}(x_a) - \mathbf{f}(x_n)||_2]_+ \\ &= \sum_{i=1}^{P} \sum_{a=1}^{K} [m + \overbrace{\max_{p=1\ldots K} ||\mathbf{f}(x_{i,a}) - \mathbf{f}(x_{i,p})||_2}^{\text{hardest positive}} \\ &\quad \underbrace{- \min_{\substack{n=1\ldots K \\ j=1\ldots P \\ j \neq i}} ||\mathbf{f}(x_{i,a}) - \mathbf{f}(x_{j,n})||_2}_{\text{hardest negative}}]_+, \end{aligned} \tag{3}$$

where $x_{i,j}$ is a data point representing the j-th image of the i-th cluster in the batch. $\mathbf{f}(x_{i,j})$ is the feature vector of $x_{i,j}$. m is the margin between positive and negative pairs.

Ranking-based Triplet Loss (RTL). However, it is clear that the effectiveness of CTL highly depends on the quality of label estimation, which is subjected to whether the clustering result is correct or not. To avoid this dependence, we propose a Ranking-based Triplet Loss (RTL), which only makes full use of the ranking score matrix D_R. Since there is no need to estimate the labels of images, it is a label-free method for reflecting the relationships between data pairs.

Specifically, given a training anchor x_a, positive sample x_p is randomly selected from the top η nearest neighbors according to the ranking score vector $\mathbf{D}_R(x_a)$, and negative sample x_n is from the location $(\eta, 2\eta]$. In addition, instead of hard margin in CTL, we introduce a *soft margin* based on the relative ranking position of x_p and x_n, which is benefi-

cial to different scales of intra-class variation. The formula of RTL is shown as:

$$L_{RTL} = \sum_{a=1}^{PK} [\frac{|P_p - P_n|}{\eta} m + ||\mathbf{f}(x_a) - \mathbf{f}(x_p)||_2 - ||\mathbf{f}(x_a) - \mathbf{f}(x_n)||_2]_+, \quad (4)$$

where the selected anchors in each batch are the same as CTL. m is the margin same as Eq. (3). η is the maximum of ranking position for positive sample selection. P_p and P_n are the ranking positions of x_p and x_n with respect to x_a.

To summarize, we optimize the network using the combination of CTL and RTL to better capture the local structure of data distribution. Our final triplet-based loss function in conservative stage is shown in Eq. (5):

$$L_C = L_{RTL} + \lambda L_{CTL}, \quad (5)$$

where λ is the loss weight to trade off the influence of two loss functions. Experiments show that this combined loss function improves the capability of model representation.

3.3. Promoting Stage

Since triplet-based loss functions only focus on the data relation within each triplet, the model will be prone to instability and stuck into a sub-optimal local minimum. To alleviate this problem, we propose to apply classification loss to further improve model generalization by taking advantage of global information of data distribution of training samples. In the promoting stage, a fully-connected layer is added at the end of the model as a classification layer, which is initialized according to the features of current training set. Softmax cross-entropy loss is used as the objective function, which is formulated as:

$$L_P = -\sum_{i=1}^{PK} \log \frac{e^{W_{\hat{y}_i}^T x_i}}{\sum_{c=1}^{C} e^{W_c^T x_i}}, \quad (6)$$

where $\hat{y}_i$ is the pseudo label of the sample x_i. C is the number of clusters from the HDBSCAN clustering method with updated training set T_U.

Feature-based Weight Initialization for Classifier. Due to the variation of cluster results in each iteration, the newly added classifier needs be re-trained after the HDBSCAN clustering. Instead of random initialization, we exploit the mean features of each cluster as the initial parameters. Specifically, for each cluster c, we calculate the mean feature $\overline{F}_c$ by averaging all the embedding features of its elements. The parameters $\mathbf{W}$ of the classifier are initialized as the following formula:

$$W_c = \overline{F}_c, \; c \in \{1, 2, \ldots, C\}, \quad (7)$$

where $\mathbf{W} \in \mathbb{R}^{d \times C}$. $W_c \in \mathbb{R}^d$ is the c-th column of $\mathbf{W}$ and d is the feature dimensionality. An advantage of this initialization is that we can use the previous information to avoid the fluctuation of accuracy caused by random initialization, which is useful for the convergence of model training. Please refer to the appendix for the comparison.

Algorithm 1: The Self-training with Progressive Augmentation Framework (PAST)

Input : labelled source domain dataset S; whole unlabelled target domain training dataset T; CNN model M; maximum iteration $I_{\max}$; HDBSCAN clustering method; minimal samples in each cluster for HDBSCAN $S_{\min}$.
Output: Model M.
Initialization: Initialize model M on S; Initial selected training set $T_U = T$.

```
for i = 1 to I_max do
    Conservative Stage:
    Extract embedding features F on training data T from M;
    Compute ranking score matrix D_R on whole training data T
      with F according to Eq. (2);
    Update training set T_U using HDBSCAN(D_R; S_min);
    Update model M using T_U according to Eq. (5);
    Extract embedding features F_U on T_U from M;
    Promoting Stage:
    Initialize classifier based on F_U according to Eq. (7);
    Update model M using T_U according to Eq. (6);
end
```

3.4. Alternate Training

In this paper, we develop a simple yet effective self-training strategy which can capture both the local and global structures of unlabelled training images. That is, the conservative stage and the promoting stage are conducted alternately. At the beginning, the model is trained using the local relations between data points alone, so that the difficulty of error amplification brought by classification loss can be prevented. After several training steps in the conservative stage, the ability of model representation and the quality of clusters are more trusty. Next, we use the softmax cross-entropy loss in the promoting stage to further augment the capability of the model, which is useful to avoid model falling into local optimum caused by triplet-based loss functions in the conservative stage. The updated model is then used as the initial state for conservative stage and the model is trained using these two stages alternately. As the training goes on, the model generalization is progressively improved, allowing to learn more discriminate feature representation of training images. The details of this two-stage alternate self-training method are presented in Algorithm 1. We also provide a visualization of this alternate self-training process on improving the quality of clusters in Figure 3.

4. Experiments

We evaluate the proposed PAST on the unsupervised cross-domain person Re-ID task. Three large-scale person Re-ID datasets are tested, namely Market-1501 [44], DukeMTMC-Re-ID [45], and CUHK03 [19].

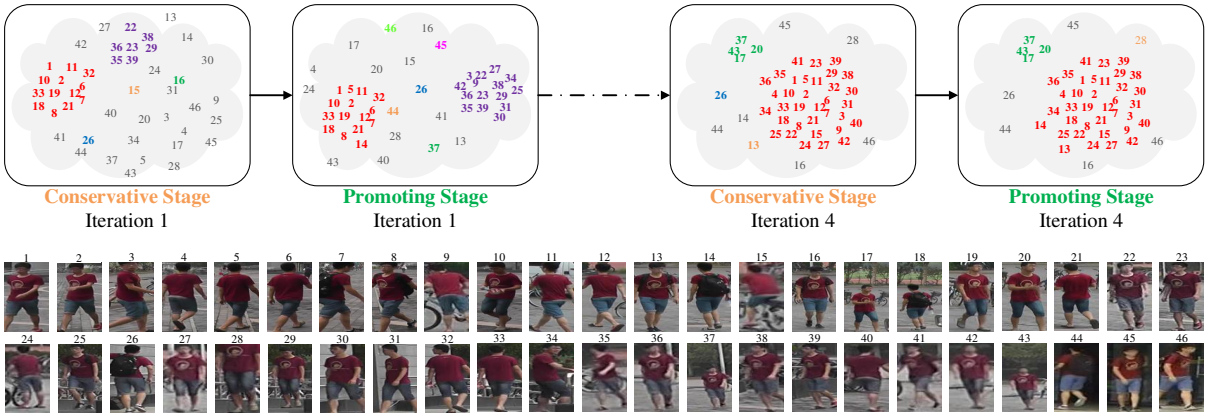

Figure 3 – Illustration of PAST progressively improves the clustering quality of one identity. All the 46 images belong to the same person but their labels are unknown in our scenario. At each iteration, image IDs with the same color denote that they are assigned to the same clusters (pseudo labels) by the clustering method. Gray image IDs mean the samples do not belong to any cluster and thus are not used for training. From iteration 1 to iteration 4, more samples are selected for training and the pseudo labels become more reliable.

Market-1501 [44] contains 32,668 labelled images of 1,501 identities taken by 6 cameras, where pedestrians are detected and cropped by the Deformable Part Model (DPM) [13]. The dataset is split into a training set with 12,936 images of 751 identities and a test set with 19,732 images of 750 identities.

DukeMTMC-Re-ID [45] consists of 36,411 labelled images belonging to 1,404 identities observed by 8 camera views. It has 16,522 images of 702 identities for the training set and the remaining 19,889 images of 702 identities for the test set. Hereafter, the term "*Duke*" also refers to this dataset for simplicity.

CUHK03 [19] is composed of 14,096 images from 1,467 identities captured by 2 cameras. There are two types of pedestrian bounding boxes available: manually cropped and DPM detected [13]. Here we only use the DPM labels in our experiment for a fair comparison. Following Market-1501 and Duke, the new train/test evaluation protocol [46] of CUHK03 is used: 7,365 images of 767 identities for training and 6,732 images of 700 identities for testing.

4.1. Implementation Details

Model and Preprocessing. We adopt PCB [32] as our feature extractor, in which ResNet-50 [16] pretrained on ImageNet [9] without the last classification layer is used as the backbone. Similar to the EANet [18], we use nine regions for feature representation. Instead of using the part-aligned pooling [18], we directly use even-part pooling (like PCB) for simplification. The dimensionality of each embedding layer is set to 256. In addition, we append a specific classification layer composed by one fully connected layer after each embedding layer in the promoting stage. The number of classes changes according to the number of clusters generated by the HDBSCAN clustering. All input images are resized to $384 \times 128 \times 3$. Note that we only use random flipping for data augmentation.

Training Settings. We use the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. Unless otherwise noted, we set the batch size to 64 and the number of iterations to 4 for all experiments. Instead of using the same learning rates for both conservative and promoting stage, we found that setting different learning rates for these stages can work better. The reason is that the parameters from the conservative stage should be updated slowly in order to alleviate the negative effects bought by the incorrect pseudo labels. Specifically, the learning rates are initialized to 10^{-4} for the backbone layers and 2×10^{-4} for the embedding layers in conservative stage, In promoting stage, the newly added classification layer uses an initial learning rate of 10^{-3} and all other layers use 5×10^{-5} instead. After the 3-rd iteration, all learning rates are multiplied by 0.1. The margins m of Eq. (3) and Eq. (4) are set to 0.3.

Evaluation Settings. For performance evaluation, feature vectors from the embedding layers of nine parts are normalized separately and are concatenated as the final representation. Given a query image, we calculate its cosine distances with all gallery images for ranking. We use the cumulative match characteristic (CMC) curve [15] and mean average precision (mAP) [44] as the evaluation metrics. The CMC curve shows the probabilities that a query appears in the candidate lists with various sizes. For a single query, the average precision (AP) is computed from the area under its precision-recall curve. The mAP is then calculated as the mean value of APs of all queries. Note that the single-shot setting [32] is adopted in all experiments.

4.2. Ablation Study

In this subsection, we aim to thoroughly analyse the effectiveness of each component in our PAST framework.

Effectiveness of the Conservative Stage. As shown in

Method	Stage	M→D		D→M	
		Rank-1	mAP	Rank-1	mAP
PCB* [32] (DT)	-	42.73	25.70	57.57	29.01
PCB-R* [46]	-	49.69	39.38	59.74	41.93
PCB-R-CTL	C	68.18	49.06	71.88	46.17
PCB-R-RTL	C	70.69	52.02	72.65	47.62
PCB-R-CTL+RTL	C	71.63	52.05	74.26	50.59
PCB-R-PAST	C+P	**72.35**	**54.26**	**78.38**	**54.62**

Table 1 – The effectiveness of conservative stage and promoting stage in our proposed Self-training with Progressive Augmentation Framework (PAST). D→M represents that we use Duke [45] as source domain and Market-1501 [44] as target domain. $*$ denotes that the results are produced by us. **DT** means Direct Transfer from PCB with 9 regions. **R** means applying k-reciprocal encoding method [46]. **CTL** represents clustering-based triplet loss [17], while **RTL** is our proposed ranking-based triplet loss. Our **PAST** framework consists of conservative stage and promoting stage that are denoted by **C** and **P** respectively.

Table 1, we conduct several experiments to verify the effectiveness of CTL, RTL and the combination of these two triplet loss functions on the task of M→D and D→M. First, only with CTL, we improve the performance by 18.49% and 12.14% at Rank-1 accuracy compared with the results from k-reciprocal encoding method [46] on M→D and D→M respectively. Second, we observe that containing only our proposed RTL, the Rank-1 accuracy and mAP increase by 21% and 12.64% for M→D, while 12.91% and 5.69% on D→M. This obvious improvement shows that both CTL and RTL are useful for increasing model generalization. And CTL obtains slightly lower performance than RTL. Then, as described in Eq. (5), we combine CTL and RTL together to jointly optimize model in our conservative stage. It is clear that we achieve better results on both M→D and D→M. Especially for D→M, we gain 2.38% and 4.42% on Rank-1 and mAP comparing to using CTL alone, which shows the significant benefit of our RTL. After the conservative stage, the model adapt to the target domain more appropriately.

Effectiveness of the Promoting Stage. However, as illustrated in Figure 1, there is no further gains even with more training iterations when only using triplet-based loss functions. We believe that it is because during conservative stage, the model only sees local structure of data distribution brought by triplet samples. Thus, in our PAST framework, we employ softmax cross-entropy loss as the objective function in the promoting stage to train the model with the conservative stage alternately. Refer to Table 1 again, compared with only using conservative stage, our PAST can further improve mAP and Rank-1 by 2.21% and 0.72% on M→D task, and 4.03% and 4.12% for D→M. Meanwhile, from Figure 3, the quality of clusters is also improved with our PAST framework. This shows that the promoting stage does play an important role in model generalization.

Through the above experiments, different components in our PAST have been evaluated and verified. We show that our PAST framework is not only beneficial for improving model generation but also for refining clustering quality.

Method	Cluster	M→D		D→M	
		Rank-1	mAP	Rank-1	mAP
PCB-R-CTL	K	44.84	26.93	54.39	29.94
	D	53.73	36.27	67.41	42.42
	H	68.18	49.06	71.88	46.17
PCB-R-CTL+RTL	K	53.99	34.46	56.26	32.73
	D	67.91	49.08	72.54	48.06
	H	71.63	52.05	74.26	50.59
PCB-R-PAST	K	68.94	49.97	75.48	51.39
	D	71.90	53.07	75.62	51.70
	H	**72.35**	**54.26**	**78.38**	**54.62**

Table 2 – The comparison of different clustering methods. **K**, **D** and **H** represents K-means, DBSCAN [11] and HDBSCANRank1HDBSCAN clustering method respectively.

Comparison with Different Clustering Methods. We evaluate three different clustering methods, *i.e.*, k-means, DBSCAN [11] and HDBSCAN [3] in the conservative stage. The performance of utilizing these clustering methods under different settings are specified in Table 2. For k-means, the number of cluster centroids k is set to 702 and 751 on target data of Market-1501 and Duke respectively, which is the same as the number of identities of source training data. It is clear that HDBSCAN performs better than k-means and DBSCAN under either only using conservative stage or whole PAST framework. For instance, using HDBSCAN can achieve mAP 54.26% and Rank-1 72.35% for M→D task in PAST framework, which are 4.29% and 3.41% higher than using k-means, and 1.19% and 0.45% than using DBSCAN. In addition, we also observe that whatever clustering method we use, our PAST framework always outperforms only using conservative stage. This means that on the one hand, HDBSCAN clustering method has more powerful effect in our framework; on the other hand, our PAST framework indeed provides improvement of feature representation on target domain.

4.3. Comparison with State-of-the-art Methods

Following evaluation setting in [18, 47], we compare our proposed PAST framework with state-of-the-art unsupervised cross-domain methods, shown in Table 3. It can be seen that only using conservative stage with CTL and RTL for training, the performance is already competitive with other cross-domain adaptive methods. For example, although EANet [18] proposes complex part-aligned pooling and combines pose segmentation to provide more information for adaptation, our conservative stage still outperforms it by 3.93% in Rank-1 and 4.05% in mAP when testing on M→D. Moreover, our PAST framework surpasses all previous methods by a large margin, which achieves 54.26%, 54.62%, 57.34%, 51.79% in mAP and 72.35%, 78.38%, 79.48%, 69.88% in Rank-1 accuracy for M→D, M→D, C→M, C→D. We can also prove that it is useful to alternately use conservative and promoting stage by comparing with the last two rows in Table 3. Especially, our PAST can

Method	M→D		D→M		C→M		C→D	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
UMDL [27]'16	18.5	7.3	34.5	12.4	-	-	-	-
PUL [12]'18	30.0	16.4	45.5	20.5	41.9	18.0	23.0	12.0
PTGAN [38]'18	27.4	-	38.6	-	31.5	-	17.6	-
SPGAN [10]'18	46.4	26.2	57.7	26.7	-	-	-	-
TJ-AIDL [36]'18	44.3	23.0	58.2	26.5	-	-	-	-
HHL [47]'18	46.9	27.2	62.2	31.4	56.8	29.8	42.7	23.4
ARN [21]'18	60.2	33.4	70.3	39.4	-	-	-	-
EANet [18]'19	67.7	48.0	78.0	51.6	66.4	40.6	45.0	26.4
Theory [30]'18	68.4	49.0	75.8	53.7	-	-	-	-
PCB* [32] (DT)'18	42.73	25.70	57.57	29.01	51.43	27.28	29.40	16.72
PCB-R* [46]	49.69	39.38	59.74	41.93	55.91	38.95	35.19	26.89
PCB-R-CTL+RTL (Ours)	71.63	52.05	74.26	50.59	77.70	54.36	65.71	46.58
PCB-R-PAST (Ours)	**72.35**	**54.26**	**78.38**	**54.62**	**79.48**	**57.34**	**69.88**	**51.79**

Table 3 – Comparison with state-of-the-art methods under unsupervised cross-domain setting. In each column, the **1st** and 2nd highest scores are marked by **red** and blue respectively. D, M, C represent Duke [45], Market-1501 [44] and CUHK03 [19] respectively.

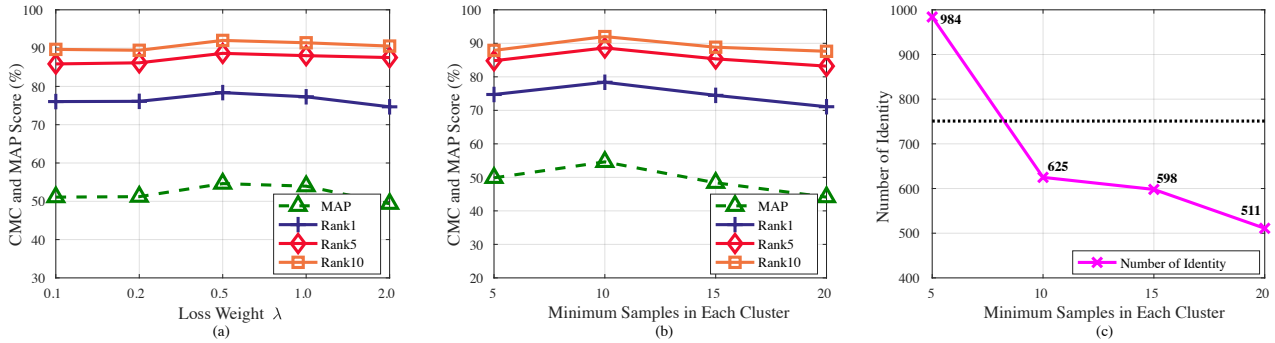

Figure 4 – Analysis of hyper parameters on D→M setting. (a): The impact of the loss weight λ; (b): The impact of the minimum samples $S_{\min}$ at each cluster in HDBSCAN clustering method; (c): The number of clusters from HDBSCAN with different minimum sample $S_{\min}$.

improve 4.71% and 5.21% in Rank-1 and mAP for C→D compared with only using conservative stage.

4.4. Parameter Analysis

We conduct additional experiments to evaluate the parameter sensitivity.

Analysis of the Loss Weight λ. λ is a hyper parameter which is used to trade off the effect between RTL and CTL. We evaluate the impact of λ, which is sampled from $\{0.1, 0.2, 0.5, 1.0, 2.0\}$, on the task of D→M. The results are shown in Figure 4 (a). We observe that the best result is obtained when λ is set to 0.5. Note that the value of λ has limited impacts on the model performance.

Analysis of the Minimum Samples $S_{\min}$. In addition, we analyse how the number of minimum samples ($S_{\min}$) for every cluster in HDBSCAN clustering affects the Re-ID results. We test the impact of $\{5, 10, 15, 20\}$ minimum samples on the performance of our PAST framework on D→M setting. As shown in Figure 4 (b), we can see that setting S_{min} to 10 yields superior accuracy. Meanwhile, different $S_{\min}$ has large variance on the final number of pseudo identities from HDBSCAN. We believe that it is because samples from the same class will be separated to several clusters when $S_{\min}$ is too small, while low-density classes will be abandoned if $S_{\min}$ is too large. This can be verified from Figure 4 (c), the number of identity from HDBSCAN with minimum sample 10 is 625, which is the closest one to the true value 751 on the Market-1501 training set.

5. Conclusion

In this paper, we have presented a self-training framework with progressive augmentation process (PAST) for unsupervised cross-domain person Re-ID with two learning stages. In conservative stage, we mainly focus on mining local information via triplet-based loss functions. Specially, the proposed ranking-based triplet loss makes full use of the similarity score between instances to select confident triplets, which is beneficial for avoiding model degeneration affected by poor pseudo label quality on unseen data. Then we propose to take advantage of the global data distribution in promoting stage via classification loss to further alleviate the instability caused by the former stage. These two stages alternate iteratively to improve the quality of pseudo labels and model generalization on unlabelled data. Extensive experiments show that our PAST achieves state-of-the-art unsupervised cross-domain Re-ID performance. In the future, we plan to extend the proposed method to other unsupervised cross-domain applications, such as face recognition and image retrieval.

Acknowledgements This work was in part supported by the Natural Science Foundation of Shanghai, China under Grand #17ZR1431500.

References

[1] JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.

[2] Jia-Wang Bian, Yu-Huan Wu, Ji Zhao, Yun Liu, Le Zhang, Ming-Ming Cheng, and Ian Reid. An evaluation of feature matchers forfundamental matrix estimation. In *Proc. British Machine Vis. Conf.*, 2019.

[3] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Proc. Pacific-Asia. Conf. Knowledge discovery & data mining*, pages 160–172, 2013.

[4] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2109–2118, 2018.

[5] Di Chen, Shanshan Zhang, Wanli Ouyang, Jian Yang, and Ying Tai. Person search via a mask-guided two-stream cnn model. In *Proc. Eur. Conf. Comp. Vis.*, pages 734–750, 2018.

[6] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proc. AAAI Conf. Artificial Intell.*, 2017.

[7] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2590–2600, 2017.

[8] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1335–1344, 2016.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 994–1003, 2018.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. ACM SIGKDD Int. Conf. Knowledge discovery & data mining*, volume 96, pages 226–231, 1996.

[12] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimedia Computing, Communications, and Applications*, 14(4):83, 2018.

[13] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[14] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[15] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE Int. Workshop on Performance Evaluation for Tracking and Surveillance*, volume 3, pages 1–7, 2007.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, 2016.

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[18] Houjing Huang, Wenjie Yang, Xiaotang Chen, Xin Zhao, Kaiqi Huang, Jinbin Lin, Guan Huang, and Dalong Du. Eanet: Enhancing alignment for cross-domain person re-identification. *arXiv preprint arXiv:1812.11369*, 2018.

[19] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 152–159, 2014.

[20] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2285–2294, 2018.

[21] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. pages 172–178, 2018.

[22] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *Proc. AAAI Conf. Artificial Intell.*, volume 2, pages 1–8, 2019.

[23] Giuseppe Lisanti, Iacopo Masi, Andrew D Bagdanov, and Alberto Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1629–1642, 2014.

[24] Zimo Liu, Dong Wang, and Huchuan Lu. Stepwise metric promotion for unsupervised video person re-identification. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2429–2438, 2017.

[25] Jianming Lv, Weihang Chen, Qing Li, and Can Yang. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7948–7956, 2018.

[26] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton Van Den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1846–1855, 2015.

[27] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1306–1315, 2016.

[28] Tong Shen, Dong Gong, Wei Zhang, Chunhua Shen, and Tao Mei. Regularizing proxies with multi-adversarial training for unsupervised domain-adaptive semantic segmentation. *arXiv preprint arXiv:1907.12282*, 2019.

[29] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1179–1188, 2018.

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

[31] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3800–3808, 2017.

[32] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proc. Eur. Conf. Comp. Vis.*, pages 480–496, 2018.

[33] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *Proc. IEEE Int. Conf. Advanced Video Signal-based Surveillance*, pages 1–6, 2017.

[34] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised learning of generative topic saliency for person re-identification. 2014.

[35] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong. Towards unsupervised open-set person re-identification. In *Proc. IEEE Int. Conf. Image Process.*, pages 769–773, 2016.

[36] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2275–2284, 2018.

[37] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1470–1478, 2018.

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 79–88, 2018.

[39] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2119–2128, 2018.

[40] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 994–1002, 2017.

[41] Andrew Zhai and Hao-Yu Wu. Making classification competitive for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[43] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[44] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1116–1124, 2015.

[45] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3754–3762, 2017.

[46] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1318–1327, 2017.

[47] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proc. Eur. Conf. Comp. Vis.*, pages 172–188, 2018.

[48] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5157–5166, 2018.