Self-Supervised Pre-training on the Target Domain for Cross-Domain Person Re-identification

Junyin Zhang
junyinz@cqu.edu.cn
Chongqing University

Yongxin Ge*
yongxinge@cqu.edu.cn
Chongqing University

Xinqian Gu
xinqian.gu@vipl.ict.ac.cn
University of Chinese Academy of Sciences

Boyu Hua
byhua@cqu.edu.cn
Chongqing University

Tao Xiang
txiang@cqu.edu.cn
Chongqing University

ABSTRACT

Most existing cluster-based cross-domain person re-identification (re-id) methods only pre-train the re-id model on the source domain. Unfortunately, the pre-trained model may not perform well on the target domain due to the large domain gap between source and target domains, which is harmful to the following optimization. In this paper, we propose a novel Self-supervised Pre-training method on the Target Domain (SPTD), which pre-trains the model on both the source and target domains in a self-supervised manner. Specifically, SPTD uses different kinds of data augmentation manners to simulate different intra-class changes and constraints the consistency between the augmented data distribution and the original data distribution. As a result, the pre-trained model involves some specific discriminative knowledge on the target domain and is beneficial to the following optimization. It is easy to combine the proposed SPTD with other cluster-based cross-domain re-id methods just by replacing the original pre-trained model with our pre-trained model. Comprehensive experiments on three widely used datasets, i.e. Market1501, DukeMTMC-ReID and MSMT17, demonstrate the effectiveness of SPTD. Especially, the final results surpass previous state-of-the-art methods by a large margin.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

cross-domain; person re-identification; self-supervised learning

ACM Reference Format:
Junyin Zhang, Yongxin Ge, Xinqian Gu, Boyu Hua, and Tao Xiang. 2021. Self-Supervised Pre-training on the Target Domain for Cross-Domain Person Re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475565

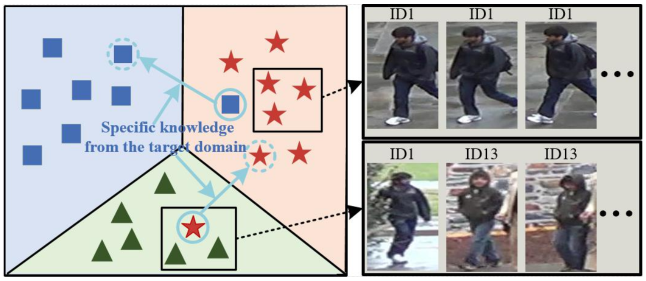


Figure 1: Pseudo label estimation according to clustering result. Different shapes denote different person identities. Exsiting cross-domain re-id methods only pre-train their models on the source domain. In this case, due to large domain bias, the pre-trained model may assign inaccurate pseudo labels for hard samples on the target domain during clustering, which is harmful to the following optimization. After the specific knowledge from the target domain is added to the pre-trained model, it is more robust for hard samples and can help to estimate better pseudo labels during clustering.

1 INTRODUCTION

Person re-identification (re-id) [26, 36, 41] aims to retrieve the images of the target person across non-overlapping cameras. It plays an important role in intelligent surveillance system. In general, the performance of CNN-based re-id models heavily relies on the number of labeled training images. However, labeling re-id data is expensive, time-consuming, and error-prone. Therefore, recent research pays more attention to cross-domain person re-id [14–16], which can obtain comparable results w.r.t. supervised methods without labeled data on the target domain.

The objective of cross-domain re-id is to transfer the discriminative knowledge from the labeled source domain to the unlabeled target domain. Most existing methods can be divided into two streams: generation-based methods [4, 20, 29] and cluster-based methods [7, 8, 33, 35]. In general, generation-based methods firstly transfer the images from the source domain to the style of the target domain by GAN [4, 20] or adversarial domain attack [29]. Then the transferred images are used to train the re-id model. Since style transferring reduces the domain gap between the source and target domains, the trained model can perform well on the target domain. In contrast, cluster-based methods firstly pre-train a re-id model on

[*] Corresponding author: Yongxin Ge.

MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475565

the source domain. Then the pre-trained model is used to extract features for the images on the target domain and clustering algorithm (e.g., k-means or DBSCAN) is adopted to assign a pseudo labels for each image. Finally, they use these pseudo labels to finetune the re-id model on the target domain. As for the cluster-based method, the discriminative ability of the fine-tuned model on the target domain depends heavily on the quality of pseudo labels. Furthermore, the quality of pseudo labels relies on the discriminative ability of the pre-trained model and the accuracy of the clustering algorithm. However, the model which is pre-trained only on the source domain may not perform well on the target domain, when the domain gap between the source and target domains is large. As shown in Fig. 1, the model only pre-trained on the source domain may assign inaccurate pseudo labels for hard samples on the target domain during clustering, which is harmful to the following optimization. After some specific discriminative knowledge from the target domain is added to the model in the pre-training stage, better pseudo labels can be estimated during clustering, which is beneficial to the following model updating.

In this paper, a novel Self-supervised Pre-training method on the Target Domain (SPTD) is proposed to introduce the discriminative knowledge from the target domain in pre-training stage for cross-domain re-id. Specifically, we use different kinds of data augmentation manners (e.g., random flip, random crop, random erase, and color distort) for the samples on target domain to simulate different intra-class changes (e.g., posture change, imperfect detection result, occlusion, and illumination variation). To make the pre-trained model adapt the intra-class changes on the target domain, SPTD constrains the consistency between the augmented data distribution and the original data distribution. As a result, the pre-trained model involves some specific discriminative knowledge on the target domain, which can decrease the error of pseudo labels estimation and is beneficial to the following optimization. Extensive ablation studies demonstrate the effectiveness of SPTD and the results on three widely used datasets surpass state-of-the-art methods.

The main contributions of this work are summarized in three folds:

- finding that the pre-trained method with specific discriminative knowledge on the target domain can reduce the error of pseudo labels estimation during fine-tuning.
- proposing a simple yet effective SPTD method, which can be easily combined with existing cluster-based cross-domain re-id approaches to improve their performance.
- achieving state-of-the-art performance on three widely used cross-domain re-id datasets.

2 RELATED WORK

2.1 Unsupervised Person Re-identification

The purpose of unsupervised re-id methods is to learn a robust model without labeled information. Previous works [17, 22] usually design robust and discriminative features by hand. Recently, deep learning methods have achieved great progress in several fields of computer vision and have been effective in handling unsupervised person re-id problems [2, 9, 31, 43]. In [37], hierarchical clustering is used to generate pseudo labels, and the hard-batch triplet loss is utilized to reduce the influence of hard examples. Wang et al. [28] consider the unsupervised person re-id task as a multi-label classification problem. Then, they design a memory-based multi-label classification loss (MMCL) to update the re-id model. Lin et al. [19] discard the clustering manner. Instead, three metric ways are proposed to compute image-level similarity. Liao et al. [18] reconsider the process of matching the feature maps. Then, they design a query-adaptive convolution (QAConv) to find local correspondences and generalize the feature maps. In [34], Wu et al. propose three different self-supervision tasks to learn a discriminative feature embedding space. Although these methods achieve great improvement, their results still are not satisfactory compared with supervised re-id methods.

2.2 Cross-Domain Person Re-identification

Cross-domain re-id aims to transfer the knowledge from a labeled source domain to an unlabeled target domain. To reduce the data bias between source and target samples, some GAN-based methods are presented to bridge the domain gap in the data level. For example, SPGAN [4] and PTGAN [32] are proposed to transfer the images from the source domain to the style of the target domain on the premise of preserving the identity labels. However, the quality of transferred images is not satisfied, which leads to poor unsupervised re-id results. In [46], Zhun et al. consider the mismatching between cameras is also one of the main reasons for the different distributions between source and target domains. Therefore, a Hetero- and Homogeneously Learning (HHL) method is introduced to enforce camera invariance. In [20], Liu et al. decompose the complicated cross-domain transfer into a set of factor-wise subtransfers in a fine-grained manner. Cluster-based methods also draw much attention in recent years and improve the state-of-the-art results for cross-domain person re-id. In [6], Fan et al. propose a PUL method to combine CNN and clustering algorithms in an overall framework, in which CNN is used to extract discriminative pedestrian features and these features are computed into different groups with assigned pseudo labels by a clustering algorithm. Fu et al. [7] divide the feature maps into three parts, in which each part is clustered into a series of groups to assign pseudo labels. In [8], an unsupervised framework, Mutual Mean-Teaching (MMT), is proposed to learn better features from the target domain via off-line refined hard pseudo labels and on-line refined soft pseudo labels in an alternative training manner. In [35], Yang et al. propose an asymmetric co-teaching framework (ACT) to dig the abundant information in hard samples by taking an adversarial pattern. Most recent methods neglect the intra-domain variations in the target domain, which has heavily degraded the model performance on the target domain. Zhong et al. [47] design three types of underlying invariance constraints in the target domain to generalize the re-id model. In [48], CBN, a BN variant, is proposed to align the distribution of images captured by different cameras, which significantly improves the performance of transfer learning for person re-id. Wang et al. [29] introduce a smoothing adversarial domain attack (SADA) method to align the source images into each camera from the target domain. Meanwhile, a p-memory reconsolidation (p-MR) is proposed to reconsolidate the source knowledge in the

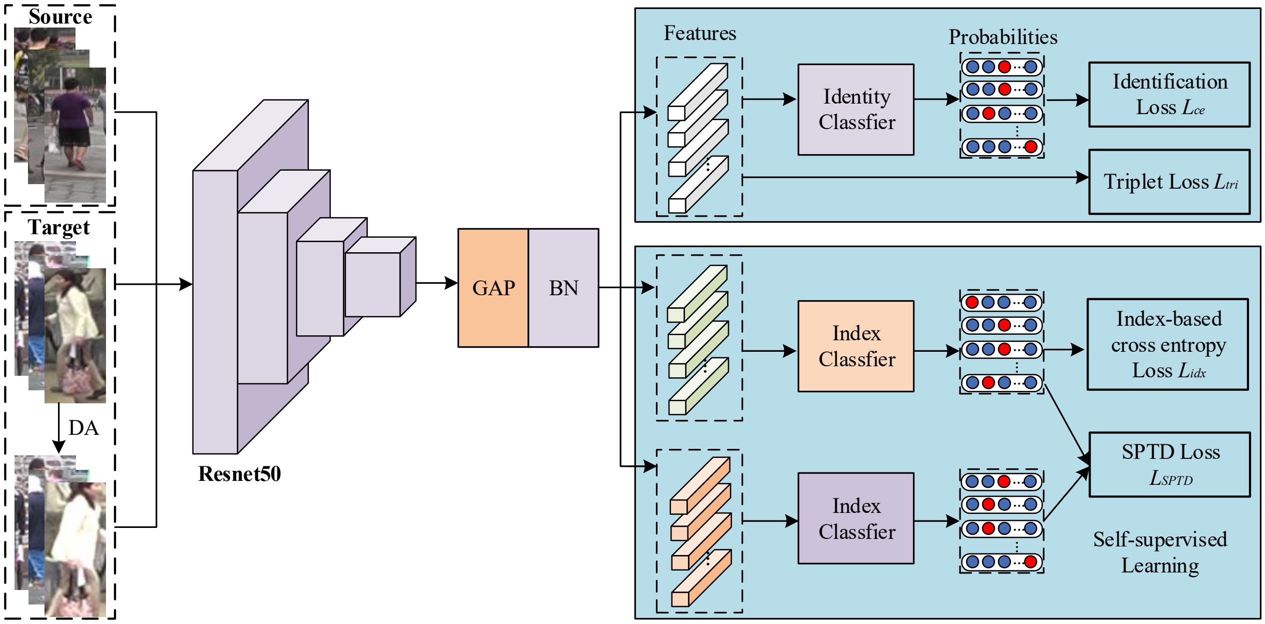


Figure 2: The Framework of SPTD. DA represents the data augmentation manners. GAP and BN denotes Global Average Pooling and Batch Normalization, respectively. SPTD takes the samples from source and target domains as inputs. As for the samples from the source domain, identification loss L_{ce} and triplet loss L_{tri} are used to extract discriminative knowledge. As for the samples from the target domain, different data augmentation manners are first applied to simulate different intra-class changes. Then, L_{idx} is used to map the extracted features of the original images to a discriminative representation space. Finally, we use L_{SPTD} to constrains the consistency between the augmented data distribution and the original data distribution in the shared representation space. Overall, the training of SPTD can be considered as a multi-task learning process.

fine-tuning stage. Compared with these methods, we focus on introducing the discriminative knowledge from the target domain to pre-train a robust re-id model by designing two self-supervised tasks on the pre-training stage, which is beneficial to estimate better pseudo labels on the fine-tuning stage.

2.3 Self-supervised Learning

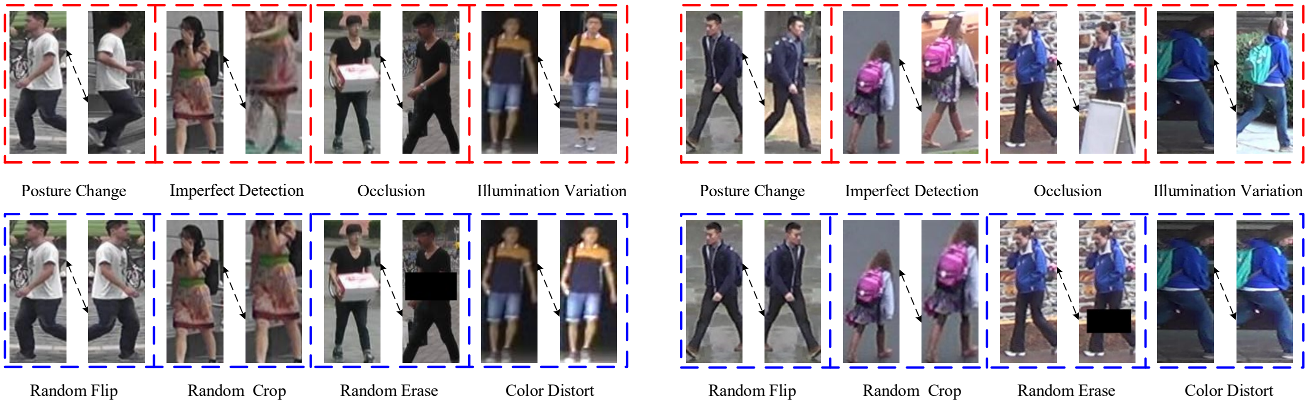
Self-supervised learning aims to learn a general feature by constructing specific auxiliary tasks to improve the performance of the target task. In [5], Doersch et al. design a pretext task to learn the spatial structure between the randomly sampled patch and the anchor patch. In [10], a self-supervised task is proposed for semantic feature learning, in which a convolutional neural network is trained to predict the rotation that is applied to the input images. Noroozi et al. [23] introduce a context-free network (CFN) to solve Jigsaw puzzles as a pretext task. It greatly improves the performance of object classification and detection. He et al. [11] present momentum contrast (MOCO) for visual representation learning, which outperforms previous self-supervised learning methods by a large margin. In [3], Hinton et al. explore various augmentation groups and design a simple framework for contrastive learning of visual representations (simCLR) to maximize agreement among different augmented views of the same data example in the latent space. In this paper, the proposed SPTD introduces the specific knowledge from the target domain on the pre-training stage, which is beneficial to the following model updating.

3 THE PROPOSED METHOD

3.1 Approach Overview

Most cluster-based cross-domain re-id methods adopt a pre-trained model as the initialized model to estimate pseudo labels for the unlabeled samples from the target domain with the help of a clustering algorithm. However, due to large domain bias, the re-id model only pre-trained on the source domain does not involve specific knowledge from the target domain, thus inaccurate pseudo labels may be assigned for the hard samples on the target domain. In this paper, we propose a novel Self-supervised Pre-training method on the Target Domain (SPTD) to optimize the model on both source and target domains in the pre-training stage. The proposed SPTD aims to add some specific discriminative knowledge from the target domain to the pre-trained model to achieve better pseudo labels estimation. The framework of SPTD is illustrated in Fig. 2.

Given a labeled source dataset $\mathcal{S}$ and an unlabeled target dataset $\mathcal{T}$, our goal is to take advantage of both $\mathcal{S}$ and $\mathcal{T}$ to pre-train a discriminative backbone model $f(\bullet|\theta)$ (θ denotes the weights). To



(a) Market1501

(b) DukeMTMC-ReID

Figure 3: Examples of hard sample pairs with same identity in four cases on both (a) Market1501 and (b) DukeMTMC-ReID. Appearance changes of the same person caused by different disturbed conditions (circled by red boxes) and the corresponding data augmentation manner to simulate these disturbed conditions (circled by blue boxes).

achieve this, a self-supervised learning method, i.e. SPTD, is designed to introduce the specific knowledge from the target domain to the backbone model. Specifically, we simulate different intra-class changes by different data augmentation manners for the samples from $\mathcal{T}$ and constrain the consistency between the augmented data distribution and the original data distribution in a shared representation space. As a result, some prior knowledge from the target domain can be added to the pre-trained model, which is beneficial to the following optimization. Details of the proposed methods are introduced in the following subsection.

Algorithm 1 Self-supervised Pre-training

```
Input: labeled source domain $\mathcal{S}$, Unlabeled Target domain $\mathcal{T}$
Output: A learned backbone model $f(\bullet|\theta)$
for $e = 1 \longrightarrow$ Max epoch do
    for $r = 1 \longrightarrow$ Max iteration do
        Sampling mini-batch $b^s$ and $b^t$ from $\mathcal{S}$ and $\mathcal{T}$
        Constructing augmented images $b_a^t$ from $b^t$
        Feeding $b^s$ to model and computing $L_{tri}$ and $L_{ce}$
        Feeding $b^t$ and $b_a^t$ to model and computing $L_{idx}$
        and $L_{SPTD}$
        Updating parameters $\theta$ with overall loss $\mathcal{L}$
    end for
end for
```

3.2 Self-supervised Pre-training on the Target Domain

In re-id task, as shown in Fig. 3, some disturbed conditions, e.g., posture change, imperfect detection, occlusion, and illumination variation, may lead to large appearance changes across the samples of the same person, i.e. large intra-class variation. Due to large domain bias, the model only pre-trained on the source domain may be not robust enough against large intra-class variation on the target domain. In this case, inaccurate pseudo labels may be assigned, which is harmful to the following optimization. To alleviate this, we use different data augmentation manners, i.e. random flip, random crop, random erase, and color distort to simulate posture change, imperfect detection, occlusion, and illumination variation, respectively. Then, SPTD constrains the consistency of the augmented data distribution and the original one. Specifically, given an image $x_i^t (1 \leqslant i \leqslant N)$ (N is the number of images on this dataset) from the target domain, we denote the augmented image as $\bar{x}_i^t$. Since there is no identity label on the target domain, to map these two types of data to a shared representation space and align them, we consider the image index i of each original image x_i^t on the target domain as its label and denote it by $\{y_i^t | 1 \leqslant i \leqslant N\}$. To introduce x_i^t on the pre-training stage, we fomulate index-based cross-entropy loss as:

$$L_{idx} = \frac{1}{B}\sum_{i=1}^{B}(L_{ce}(O_{\phi}(f(x_i^t|\theta)), y_i^t)) \tag{1}$$

where O_{ϕ} denotes the index classifier and ϕ represents the weights of O_{ϕ}, B is the batch size and L_{ce} represents cross entropy loss. Then, SPTD loss L_{SPTD} is defined as the KL divergence between the probability distribution of the augmented data and the original data:

$$L_{SPTD} = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N}\mathcal{R}_i(j) log \frac{\mathcal{R}_i(j)}{Q_i(j)} \tag{2}$$

where $\mathcal{R}_i(j)$ and $Q_i(j)$ are the predicted probability on the j-th class of the original image x_i^t and the augmented image $\bar{x}_i^t$, respectively.

3.3 Loss Function

The re-id model is pre-trained on both source and target domain. As for the samples from the source domain, triplet loss with hard example mining L_{tri} [13] and identification loss, i.e. softmax-based cross entropy loss L_{ce} are used to extract discriminative knowledge from the source domain. As for the samples from the target domain, L_{idx} and L_{SPTD} are used to introduce specific knowledge form the target domain to the pre-trained model. Therefore, the final loss function in the pre-training stage can be formulated as:

$$\mathcal{L} = L_{tri} + L_{ce} + L_{idx} + L_{SPTD} \tag{3}$$

The process of self-supervised pre-training is shown in Algorithm 1.

Most previous works only use L_{tri} and L_{ce} as their loss functions to pre-train their models. In contrast, we introduce two self-supervised constraints, i.e. $L_{idx} + L_{SPTD}$, to introduce the data from the target domain for re-id model pre-training. With the specific discriminative knowledge from the target domain, our SPTD can estimate better pseudo labels for hard samples in fine-tuning stage. Note that the proposed method only focuses on the pre-training stage and can be combined with any other cluster-based cross-domain re-id methods.

4 EXPERIMENTS

4.1 Experimental settings

We evaluate the proposed method on three widely used datasets, namely, Market1501 [45], DukeMTMC-ReID [25] and MSMT17 [32]. Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP) are used as the evaluation metrics.

Datasets. Market1501 [45] contains 32,668 images of 1501 identities, which are captured by 6 cameras on campus. Among them, 12,936 images from 751 identities are used for training, and the remaining 750 identities are used for testing. As for DukeMTMC-ReID [25], it contains 36,411 images of 1404 persons, which are captured by 8 cameras. Among them, 16,522 images are used for training, and the other 2,228 and 17,661 images are divided into query and gallery sets, respectively. MSMT17 [32] includes 126,411 person images from 4,101 identities collected by 15 cameras. Specifically, 32621 images of 1041 identities and 93820 images of 3060 identities are used as the training and testing sets, respectively.

Implementation Details. We adopt ResNet-50 [12] as the backbone, which is pre-trained for 80 epochs and fine-tuned for 40 epochs. Adam optimizer is adopted to optimize the networks with a weight decay of 0.0005 for both pre-training and fine-tuning. In pre-traing stage, 64 person images of 16 identities (4 for each identity) are contained on each mini-batch. The learning rate is initialized as 0.00035 and divides 10 of its previous value on the $40th$ and $70th$ epoch. During fine-tuning, MMT [8] is used as the baseline method. We fix the learning rate to 0.00035 for overall 40 training epochs. K-means and DBSCAN are used as clustering algorithms to generate the pseudo labels for model updating, respectively. More details, the number of pseudo classes is set as 500, 700, 1500 for Market-1501, DukeMTMC-ReID and MSMT17 when using k-means as the clustering algorithm on the proposed SPTD. The mini-batch of target domain contains 64 person images of 16 identities (4 for each identity), which needs to be re-organized with updated hard pseudo labels after each epoch. All the images are resized into 256 × 128. We only adpot randomly erasing in target domain fine-tuning and implement our method on PyTorch framework. Details are consistent with MMT [8].

Table 1: Comparison with several state-of-the-arts on DukeMTMC-ReID. Specifically, 'M→D' denotes Market1501 and DukeMTMC-ReID as the source and target domains, respectively.

Method	M → D			
	mAP	rank-1	rank-5	rank-10
UMDL [24]	7.3	18.5	31.4	37.6
TJ-AIDL [30]	23.0	44.3	59.6	65.0
SPGAN [4]	22.3	41.1	56.6	63.0
PTGAN [32]	-	27.4	-	50.7
ATNet [20]	24.9	45.1	59.5	64.2
CFSM [1]	27.3	49.8	-	-
PDA-Net [16]	45.1	63.2	77.0	82.5
PCB-PAST [40]	54.3	72.4	-	-
ECN [47]	40.4	63.3	-	-
SSG [7]	53.4	73.0	80.6	83.2
SADA [29]	55.8	74.5	85.3	88.7
AD-Cluster [38]	54.1	72.6	82.5	85.5
CBN [48]	44.9	68.0	80.0	83.9
ACT [35]	54.5	72.4	-	-
MPLP+MMCL [28]	51.4	72.4	82.9	85.0
MMT [8] (k-means)	65.1	78.0	88.8	92.5
MMT [8] (DBSCAN)	62.3	76.3	87.7	91.2
NRMT [42]	62.2	77.8	86.9	89.5
B-SNR+GDS-H [15]	59.7	76.7	-	-
MEB-Net [39]	66.1	79.6	88.3	92.2
UNRN [43]	69.1	82.0	90.7	93.5
GLT [44]	69.2	82.0	90.2	92.8
SPTD (k-means)	**69.4**	**82.1**	**90.8**	**93.6**
SPTD (DBSCAN)	**72.1**	**83.3**	**92.7**	**94.4**
Supervised (BOT[21])	**75.9**	**86.2**	-	-

4.2 Comparisons with state-of-the-art methods

We compare the proposed SPTD with some existing state-of-the-art approaches in Tab. 1 and Tab. 2. It can be seen that the proposed method outperforms all of the state-of-the-art methods by a large margin. Specifically, we first compare three generation-based methods, SPGAN [4], PTGAN [32] and ATNet [20] on 'D → M'. Compared with the best generation-based representation ATNet, SPTD (k-means) improves 53.7% in mAP and 35.4% in rank-1. The mAP and rank-1 of SPTD (DBSCAN) can be lead to 56.5% and 36.3% improvement over ATNet. Especially, we further compare recent state-of-the-art cluster-based methods including SSG [7], SADA [29], AD-Cluster [38], ACT [35], MPLP+MMCL [28], MMT [8], NRMT [42], B-SNR+GDS-H [15], MEB-Net [39], UNRN [43], GLT [44] with the proposed SPTD. Our method achieves the best performance on both 'M → D' and 'D → M' settings. Especially, SPTD (k-means) leads to 4.1% rank-1 and 4.3% mAP improvement over MMT (k-means) and SPTD (DBSCAN) leads to 7.0% rank-1 and 9.8% mAP improvement over MMT (DBSCAN) on 'M → D'. Similarly, on 'D → M' setting, SPTD (k-means) improve 3.4% rank-1 and 8.1% mAP compared with MMT (k-means) and SPTD (DBSCAN) significantly improve 2.5% rank-1 and 8.3% mAP compared with MMT (DBSCAN).

Table 2: Comparison with several state-of-the-arts on Market1501. Specifically, 'D→M' denotes DukeMTMC-ReID and Market1501 as the source and target domains, respectively.

Method	D → M			
	mAP	rank-1	rank-5	rank-10
UMDL [24]	12.4	34.5	52.6	59.6
TJ-AIDL [30]	26.5	58.2	74.8	81.1
SPGAN [4]	22.8	51.5	70.1	76.8
PTGAN [32]	-	38.6	-	66.1
ATNet [20]	25.6	55.7	73.2	79.4
CFSM [1]	28.3	61.2	-	-
PDA-Net [16]	47.6	75.2	86.3	90.2
PCB-PAST [40]	54.6	78.4	-	-
ECN [47]	43.0	75.1	-	-
SSG [7]	58.3	80.0	90.0	92.4
SADA [29]	59.8	83.0	91.8	94.1
AD-Cluster [38]	68.3	86.7	94.4	96.5
CBN [48]	52.0	81.7	91.9	94.7
ACT [35]	60.6	80.5	-	-
MPLP+MMCL [28]	60.4	84.4	92.8	95.0
MMT [8] (k-means)	71.2	87.7	94.9	96.9
MMT [8] (DBSCAN)	73.8	89.5	96.0	97.6
NRMT [42]	71.7	87.8	94.6	96.5
B-SNR+GDS-H [15]	72.5	89.3	-	-
MEB-Net [39]	76.0	89.9	96.0	97.5
UNRN [43]	78.1	91.9	96.1	97.8
GLT [44]	79.5	**92.2**	96.5	97.8
SPTD (k-means)	**79.3**	91.1	**96.9**	**97.9**
SPTD (DBSCAN)	**82.1**	92.0	**97.1**	**98.0**
Supervised (BOT[21])	**85.7**	**94.1**	**-**	**-**

Table 3: Comparison with several state-of-the-arts on MSMT17. Specifically, 'M → MS' denotes Market1501 and MSMT17 as the source and target domains, respectively.

Method	M → MS			
	mAP	rank-1	rank-5	rank-10
PTGAN [32]	2.9	10.2	-	24.4
ECN [47]	8.5	25.3	36.3	42.1
SSG [7]	13.2	31.6	-	49.6
MMT [8] (k-means)	22.9	49.2	63.1	68.8
MMT [8] (DBSCAN)	24.0	50.1	63.5	69.3
NRMT [42]	19.8	43.7	56.5	62.2
UNRN [43]	25.3	52.4	64.7	69.7
GLT [44]	26.5	56.6	67.5	72.0
SPTD (k-means)	**31.5**	**61.3**	**74.0**	**78.6**
SPTD (DBSCAN)	**30.6**	**57.5**	**70.1**	**75.0**

Table 4: Comparison with several state-of-the-arts on MSMT17. Specifically, 'D → MS' denotes DukMTMC-ReID and MSMT17 as the source and target domains, respectively.

Method	D → MS			
	mAP	rank-1	rank-5	rank-10
PTGAN [32]	3.3	11.8	-	27.4
ECN [47]	10.2	30.2	41.5	46.8
SSG [7]	13.3	32.2	-	51.2
MMT [8] (k-means)	23.3	50.1	63.9	69.8
MMT [8] (DBSCAN)	25.1	52.9	66.3	71.3
NRMT [42]	20.6	45.2	57.8	63.3
UNRN [43]	26.2	54.9	67.3	70.6
GLT [44]	27.7	59.5	70.1	74.2
SPTD (k-means)	**32.9**	**62.8**	**75.2**	**79.8**
SPTD (DBSCAN)	**32.1**	**60.2**	**72.7**	**77.5**

We also conduct experiments on MSMT17, which is a larger and more challenging dataset. Several works report their performance on MSMT17, i.e., PTGAN [32], ECN [47], SSG [7], MMT [8], NRMT [42], UNRN [43] and GLT [44]. As shown in Tab. 3 and Tab. 4, our SPTD (k-means) achieves 61.3% rank-1 and 31.5% mAP, which significantly outperforms MMT (k-means) by 12.1% rank-1 and 8.6% mAP on 'M → MS'. Similarly, on 'D → MS' setting, SPTD (k-means) leads to 12.7% rank-1 and 9.6% mAP improvement over MMT (k-means). In short, these comparisons demonstrate that pre-trained model with specific discriminative knowledge from target domain is beneficial to estimate better pseudo labels.

Table 5: The results conducted by pre-trained models with different self-supervised losses on the setting of 'M → D'.

Method	L_{idx}	L_{SPTD}	mAP	rank-1
baseline			28.8	45.8
SPTD	✓		33.9	49.9
SPTD	✓	✓	34.6	52.5

4.3 Effectiveness of L_{idx} and L_{SPTD}

To verify the effectiveness of two designed self-supervised losses, we reproduce the proposed SPTD with different loss functions and the results are shown in Tab. 5. It can be seen that the addition of index-based cross-entropy loss L_{idx} significantly improves the performance of baseline by 3.9% rank-1 and 4.3% mAP. Meanwhile, self-supervised pre-training on the target domain loss L_{SPTD} further improves the performance to 52.5% rank-1 and 34.6% mAP. This comparison demonstrates that L_{idx} and L_{SPTD} can introduce the specific knowledge from the target domain in the pre-training stage and the pre-trained model is more robust to the hard samples of the target domain. Therefore, our SPTD can estimate far better pseudo labels, which is beneficial to the following updating.

4.4 Effectiveness of different augmentation manners.

We adopt four different augmentation manners, i.e, random flip (RF), random erase (RE), color distort (CD) and random crop (RC) to simulate the posture change, imperfect detection, occlusion, and illumination variation on the target domain. To verify their effectiveness, we reproduce the proposed SPTD with different augmentation manners and the results are shown in Tab. 6. Our best model is compared with a baseline and other configurations with each of

Table 6: The performance of the proposed SPTD with different augumentation manners on both 'M → D' and 'D → M' settings. 'RF, RE, CD, and RC' denotes random flip, random erase, color distort and random crop, respectively.

Stage	Method	RF	RE	CD	RC	M → D				D → M			
						mAP	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10
Pre-training	baseline					28.8	45.8	61.1	67.0	30.0	59.4	74.4	80.2
	SPTD	✓				31.7	48.6	67.3	73.5	29.8	59.7	78.0	83.9
		✓	✓			34.6	51.5	69.1	74.6	31.2	60.6	78.8	83.9
		✓	✓	✓		34.5	51.8	**69.4**	**75.0**	32.2	62.5	79.9	85.1
		✓	✓	✓	✓	**34.6**	**52.5**	69.1	74.6	**33.1**	**64.0**	**81.2**	**86.2**
Fine-tuning	baseline(k-means)					65.1	78.0	88.8	92.5	71.2	87.7	94.9	**96.9**
	SPTD(k-means)	✓	✓	✓	✓	**69.4**	**82.1**	**90.8**	**93.6**	**79.3**	**91.1**	**95.1**	96.8
	baseline(DBSCAN)					62.3	76.3	87.7	91.2	73.8	89.5	96.9	97.9
	SPTD(DBSCAN)	✓	✓	✓	✓	**72.1**	**83.3**	**92.7**	**94.4**	**82.1**	**92.0**	**97.1**	**98.0**

the following components added: RF, RE, CD and RC. Results show that all these components are required to achieve the best performance, and RE and CD are especially important. Specifically, RC adopted on SPTD leads to 2.9% mAP and 2.9% rank-1 improvement over the previous setting on 'M → D'. Similarly, on the 'D → M', the proposed method with the RC augmentation manner improves 1.4% mAP and 0.9% rank-1 compared with the previous setting.

In the fine-tuning stage, compared with baseline (kmeans), the rank-1 and mAP are improved by 4.1% and 4.3% on 'M → D'. On the 'D → M' setting, the rank-1 and mAP can be lead to 3.4% and 8.1% improvement over baseline (kmeans). Compared with baseline (DBSCAN), the rank-1 and mAP are improved by 7.0% and 9.8% on 'M → D'. On the 'D → M' setting, the rank-1 and mAP can lead to 2.5% and 8.3% improvement over baseline (DBSCAN). These comparisons demonstrate that four augmentation manners can introduce the specific discriminative knowledge from the target domain to improve the performance of the pre-trained model. Then, a robust pre-trained model can estimate better pseudo labels in the fine-tuning stage.

Table 7: The results conducted by pre-trained models with different self-supervised methods on the setting of 'M → D'.

Method	mAP	rank-1	rank-5	rank-10
MMT [11]	28.8	45.8	61.1	67.0
SimCLR [3]	30.2	46.4	63.2	69.3
SPTD	34.6	52.5	69.1	74.6

4.5 Effectiveness of different self-supervised methods.

In this section, we attempt to replace L_{idx} and L_{SPTD} with the contrastive learning loss, i.e. SimCLR, and results are shown in Tab. 7. It can be seen that SimCLR achieves 46.4% rank-1 and 30.2% mAP on 'M → D' after pretraining. Although SimCLR improves the performance of MMT, the results of SimCLR are much lower than that of our SPTD. This comparison demonstrates that the specific knowledge from the target domain is beneficial for pre-train a robust re-id model and our SPTD is an effective way to introduce the specific knowledge from the target domain.

4.6 Visualization

The visualization of feature map. We visualise feature maps of baseline/SPTD in Fig. 4. It can be seen that baseline highlights background, which may lead to that images with the same identity are assigned to different clusters. With SPTD introducing the discriminative knowledge from the target domain in pre-training stage, the feature maps focus on more foreground regions. Specifically, in the first row, two pedestrians are occluded by cars. In those cases, baseline pays more attention to the occluded region, which is harmful to pseudo labels estimation. In contrast, image features extracted by the proposed SPTD unconcern the occluded regions. This comparison demonstrates that with SPTD introducing discriminative knowledge, the pre-trained model is more robust for the samples from the target domain.

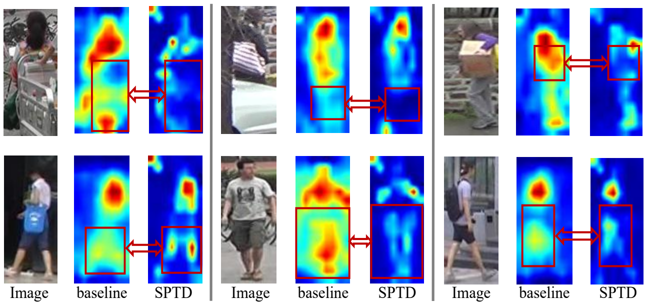


Figure 4: Visualisation of the feature maps on both Market1501 and DukeMTMC-ReID datasets. Baseline highlights more background regions. The proposed SPTD reduce the influence of background regions. The comparisons of the regions in the red boxes demonstrate the superiority of the proposed approach. (Best viewd in color)

The visualization of feature distribution. We also visualise feature distributions of baseline and SPTD on DukeMTMC-ReID dataset, using t-SNE [27]. As shown in Fig. 5, due to lacking the specific discriminative knowledge from the target domain in the pre-training stage, the image features with the same identity extracted by baseline are incompact, which may lead to assigning inaccurate pseudo labels during clustering. In contrast, the image features with the same identity extracted by the proposed method are more compact and the re-id model can estimate better pseudo labels in the fine-tuning stage. This comparison demonstrates that specific discriminative knowledge from the target domain is important for pseudo labels estimation. Therefore, we can conclude that introducing the specific discriminative knowledge from the target domain in the pre-training stage can also improve the final performance of cluster-based methods.

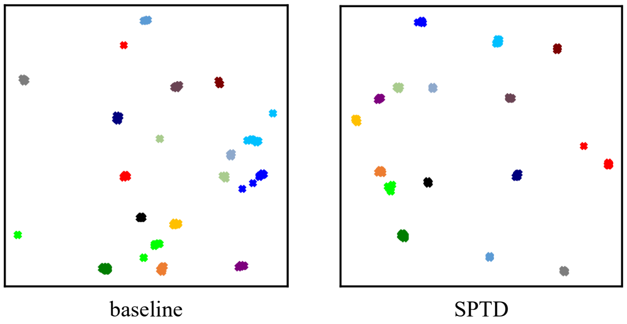


Figure 5: The visualisation of feature distributions of baseline and SPTD on DukeMTMC-ReID dataset. Compared with the proposed method, image features with the same identity extracted by baseline are incompact. Different colors denote different identities. (Best viewed in color)

Retrieval Results. To verify the superiority of our pre-trained model, we visualise the retrieval results of baseline and the proposed method on the DukeMTMC-ReID. Results are shown in Fig. 6. It can be seen that some hard samples, which have similar appearances to the query image but with different identities, are at the top of the retrieval results of baseline. While the retrieval results of our SPTD are more accurate, due to introducing the specific discriminative knowledge from the target domain in the pre-training stage. For example, in the first row, baseline only focuses on the information of clothes and ignores more discriminative features, i.e. the discriminative information of hat. While our pre-trained model with the prior knowledge from the target domain can capture the discriminative features and obtain a more robust pedestrian representation. Compared with baseline, the retrieval results of our SPTD are more accurate. In the second and third rows, due to the illumination change, retrieved images in the rank-list of baseline almost have different identities compared with query. However, with the spcific discriminative knowledge from the target domain, our pre-trained model is robust when images have illumination change, which also promises the accuracy of the retrieval results. In the fourth row, the $4th$ image of the rank-list is occluded by a car, which leads to the appearance of pedestrian change. In this case, baseline has made a false decision. While our pre-trained model is benefits from the specific discriminative knowledge from the target domain, which has corrected the false sample and improve the retrieval results more accurate. Compared with pre-trained model of baseline, our pre-trained model is more robust to the hard samples on the target domain, which also demonstrate that the specific knowledge from target domain is important to cross domain person re-id tasks.

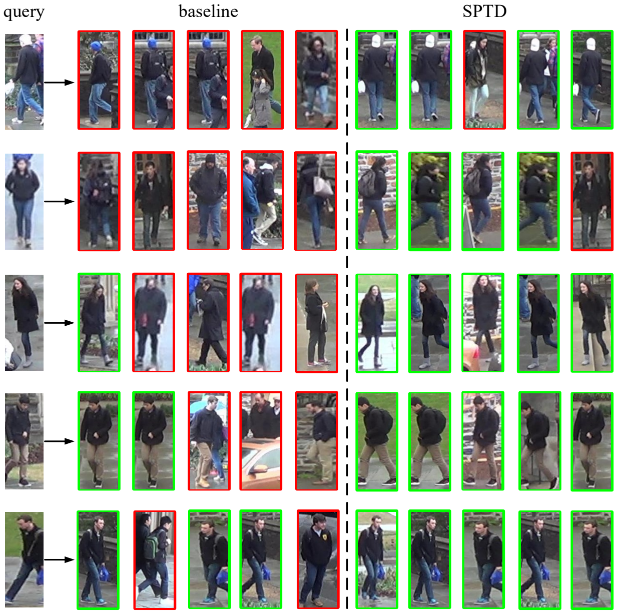


Figure 6: Visual comparison of the retrieval results on DukeMTMC-ReID. Five query images are used to evaluate. The green and red boxes represent the positive and negative images with the query image, respectively. (Best viewd in color)

5 CONCLUSION

In this paper, we propose a novel Self-Supervised Pre-training on the Target Domain(SPTD) to introduce specific knowledge from the target domain to the pre-trained model for cluster-based cross-domain re-id methods. Specifically, in the pre-training stage, we use different data augmentation manners to simulate different intra-class variations for the samples from the target domain and constrain the consistency between the augmented data distribution and the original data distribution in a shared representation space. Consequently, some prior knowledge from the target domain can be added to the pre-trained model, which is beneficial to the following optimization. Comprehensive experiments on three widely used datatsets including Market1501, DukeMTMC-ReID and MSMT17 demonstrate that the proposed SPTD can estimate better pseudo labels and improve the performance of cluster-based cross-domain re-id methods in both pre-training and fine-tuning stages.

ACKNOWLEDGEMENTS

The work described in this paper was partially supported by the National Natural Science Foundation of China (Grant no.61772093), Graduate Research and Innovation Foundation of Chongqing, China (Grant no. CYS20074), and the Fundamental Research Funds for the Central Universities (Grant no. 2021CDJQY-018).

REFERENCES

[1] Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M Hospedales. 2019. Disjoint label space transfer learning with common factorised space. In *Proceedings of the Artificial Intelligence*. 3288–3295.

[2] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. 2020. Dual Distribution Alignment Network for Generalizable Person Re-Identification. *arXiv:2007.13249*.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*. 1597–1607.

[4] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 994–1003.

[5] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the International Conference on Computer Vision*. 1422–1430.

[6] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. 2018. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 1–18.

[7] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. 2019. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the International Conference on Computer Vision*. 6112–6121.

[8] Yixiao Ge, Dapeng Chen, and Hongsheng Li. 2020. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. In *Proceedings of the International Conference on Learning Representations*.

[9] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *arXiv preprint arXiv:2006.02713*.

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv:1803.07728*.

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the Computer Vision and Pattern Recognition*. 9729–9738.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the Computer Vision and Pattern Recognition*. 770–778.

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

[14] Yangru Huang, Peixi Peng, Yi Jin, Junliang Xing, Congyan Lang, and Songhe Feng. 2019. Domain adaptive attention model for unsupervised cross-domain person re-identification. *arXiv:1905.10529*.

[15] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. 2020. Global distance-distributions separation for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision*. 735–751.

[16] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. 2019. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *Proceedings of the International Conference on Computer Vision*. 7919–7929.

[17] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the Computer Vision and Pattern Recognition*. 2197–2206.

[18] Shengcai Liao and Ling Shao. 2019. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. *arXiv:1904.10424*.

[19] Yutian Lin, Lingxi Xie, Yu Wu, Chenggang Yan, and Qi Tian. 2020. Unsupervised person re-identification via softened similarity learning. In *Proceedings of the Computer Vision and Pattern Recognition*. 3390–3399.

[20] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. 2019. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 7202–7211.

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Workshops*.

[22] Bingpeng Ma, Yu Su, and Frederic Jurie. [n.d.]. Covariance Descriptor based on Bio-inspired Features for Person Re-identification and Face Verification. *Image & Vision Computing*.

[23] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*. 69–84.

[24] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. 2016. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 1306–1315.

[25] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision*. 17–35.

[26] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling. In *Proceedings of the European Conference on Computer Vision*. 480–496.

[27] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*.

[28] Dongkai Wang and Shiliang Zhang. 2020. Unsupervised person re-identification via multi-label classification. In *Proceedings of the Computer Vision and Pattern Recognition*. 10981–10990.

[29] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. 2020. Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 10568–10577.

[30] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. 2018. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 2275–2284.

[31] Menglin Wang, Baisheng Lai, Jianqiang Huang, Xiaojin Gong, and Xian-Sheng Hua. 2020. Camera-aware Proxies for Unsupervised Person Re-Identification. *arXiv:2012.10674*.

[32] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 79–88.

[33] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. 2019. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the Computer Vision and Pattern Recognition*. 1187–1196.

[34] Guile Wu, Xiatian Zhu, and Shaogang Gong. 2020. Tracklet self-supervised learning for unsupervised person re-identification. In *Proceedings of the Artificial Intelligence*. 12362–12369.

[35] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. 2020. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *Proceedings of the Artificial Intelligence*. 12597–12604.

[36] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 2019. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 1389–1398.

[37] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. 2020. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 13657–13665.

[38] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. 2020. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 9021–9030.

[39] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. 2020. Multiple expert brainstorming for domain adaptive person re-identification. *arXiv:2007.01546*.

[40] Xinyu Zhang, Jiewei Cao, Chunhua Shen, and Mingyu You. 2019. Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In *Proceedings of the International Conference on Computer Vision*. 8222–8231.

[41] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-aware global attention for person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 3186–3195.

[42] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. 2020. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *Proceedings of the European Conference on Computer Vision*. 526–544.

[43] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Zhizheng Zhan, and Zheng-Jun Zha. 2020. Exploiting Sample Uncertainty for Domain Adaptive Person Re-Identification. *arXiv:2012.08733*.

[44] Kecheng Zheng, Wu Liu, Lingxiao He, Tao Mei, Jiebo Luo, and Zheng-Jun Zha. 2021. Group-aware Label Transfer for Domain Adaptive Person Re-identification. *arXiv:2103.12366* (2021).

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the International Conference on Computer Vision*. 1116–1124.

[46] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. 2018. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision*. 172–188.

[47] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. 2019. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition*. 598–607.

[48] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. 2020. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *Proceedings of the European Conference on Computer Vision*. 140–157.