PROXY TASK LEARNING FOR CROSS-DOMAIN PERSON RE-IDENTIFICATION

Houjing Huang [1,2], *Xiaotang Chen* [1,2*] *and Kaiqi Huang* [1,2,3]

[1] CRISE, CASIA [2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{houjing.huang, xtchen, kaiqi.huang}@nlpr.ia.ac.cn

ABSTRACT

Person re-identification (ReID) has achieved rapid improvement recently. However, exploiting the model in a new scene is always faced with huge performance drop. The cause lies in distribution discrepancy between domains, including both low-level (*e.g*. image quality) and high-level (*e.g*. pedestrian attribute) variance. To alleviate the problem of domain shift, we propose a novel framework Proxy Task Learning (PTL), which performs body perception tasks on target-domain images while training source-domain ReID, in a multi-task manner. The backbone is shared between tasks and domains, hence both low- and high-level distributions are deeply aligned. We experimentally verify two proxy tasks, *i.e*. human parsing and attribute recognition, that prominently enhance generalization of the model. When integrating our method into an existing cross-domain pipeline, we achieve state-of-the-art performance on large-scale benchmarks.

Index Terms— Person Re-identification, Cross-domain, Multi-task, Human Parsing, Attribute Recognition

1. INTRODUCTION

Person re-identification (ReID) aims to associate images of some given person across cameras, which has wide application in video surveillance, *e.g*. cross-camera tracking and pedestrian retrieval, *etc*. Just as with many other tasks, ReID is confronted with the problem of domain shift. That is, when the model trained on one domain is exploited to another, it tends to have a huge performance drop. In ReID, one domain usually refers to images captured from a group of nearby cameras. This is really frustrating if the model is confined to the scene where the training images come from.

The cause of insufficient cross-domain generalization lies in both low- and high-level distribution discrepancy between domains. In terms of low-level variance, lighting condition and image quality are frequently observed. As illustrated in Fig. 1, CUHK03 has lighting much dimmer than the other two datasets. The former seems captured near corridors of teaching buildings, while the latter are taken from outdoor scenes.

*Corresponding author

(a) Market1501 [1] (b) CUHK03 [2] (c) Duke [3]

Fig. 1: Distribution discrepancy between scenes raises an obstacle for ReID model to generalize across domains.

In terms of high-level distinction, environment composition and pedestrian attribute are factors easy to identify. For example, due to different seasons, Market1501 contains pedestrians usually wearing shorts, while DukeMTMC-reID shows frequent trousers and coats.

To alleviate the problem of domain shift, researchers propose Unsupervised Domain Adaptation (UDA) to adjust the model using target-domain images, which do not have identity labels. Three mainstream groups of approaches exist in the literature. 1) *GAN Based Methods.* The first group align the two domains in image space [4, 5, 6]. They typically utilize Generative Adversarial Networks (GAN) to transfer source-domain images into the style of target domain before training a normal ReID model. These GAN based methods consider low-level factors like image quality and lighting, as well as high-level factors like background composition. However, attribute distribution of pedestrians is not addressed. Moreover, designing constraints for adversarial training requires heavy efforts. 2) *MMD Methods.* By contrast, the second group align distributions in feature space. It is also able to tackle both low- and high-level elements, because the whole network would be updated in back propagation. Maximum Mean Discrepancy (MMD) is widely used to pull close two domains [7]. Nonetheless, it only takes low-order statistics into account, making it inadequate to encompass the underlying gap. 3) *Clustering-and-Finetuning Methods.* The final group put domain discrepancy aside, concentrating on mining pseudo identity labels in target domain [8, 9, 10, 11]. The representative pipeline is to iterate between label mining and

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on September 26,2020 at 15:12:10 UTC from IEEE Xplore. Restrictions apply.

model finetuning till convergence. The initial performance of the source model on target images determines neighborhood purity and thus is highly related to the final performance it can achieve.

In this paper, we propose a novel framework named Proxy Task Learning (PTL) to overcome the discrepancy between domains. Specifically, we perform body perception tasks on target-domain images while training ReID on source domain, in a multi-task manner. Perception tasks facilitate comprehension of the model on target pedestrians, working as a proxy between domains. Compared with GAN based methods, our strategy pays attention to both background and human body, with an easy-to-implement framework. Compared with MMD, ours is not limited to low-order statistics. When considering clustering-and-finetuning methods, the proposed approach is able to obtain an initial model more compatible with target domain and facilitate further finetuning. Our experiment shows that human parsing and attribute recognition are two effective proxy tasks, which prominently improves generalization of the model. The parsing and attribute labels on target domain are obtained by inferring target images using models trained on corresponding public datasets, *i.e.* COCO Densepose [12] and PETA [13]. As a consequence, our method comes with no extra labeling cost.

Our contribution is summarized as follows. First, we propose the Proxy Task Learning (PTL) framework to mitigate the discrepancy between source and target domains. To the best of our knowledge, we are the first to perform perception tasks on target images as domain adaptation. Second, we verify human parsing and attribute recognition as effective proxy tasks, through extensive experiments. Finally, when integrated with existing clustering-and-finetuning pipeline, our model achieves state-of-the-art performance on large-scale benchmarks.

2. RELATED WORK

Person Re-identification. Feature representation is a critical factor to distinguish between identities. Some researchers devise specific backbones [14], taking into account lightweight implementation and multi-scale features. In order to obtain fine-grained representations, a line of works extract features from multiple uniformly divided regions on the feature map [15]. Part annotations, *e.g.* keypoints and segmentation masks, have also proven to be effective for extracting fine-grained features or part alignment [16]. Ranking loss functions and sampling strategies are also proposed to effectively optimize the feature space [17].

Cross-domain Person Re-identification. Researchers propose to adapt ReID model using unlabeled target-domain images, also known as Unsupervised Domain Adaptation (UDA). We categorize these approaches into groups. 1) *GAN Based Methods.* SPGAN [4] employs CycleGAN with carefully designed generator constraints to transfer source images into style of target images. Then a usual ReID model is trained on these translated images for a model suitable for the target domain. CR-GAN [5] synthesizes images by augmenting each source pedestrian with various contextual images from target domain. 2) *MMD Methods.* Lin *et al*. [7] propose to align distributions of both identity features and attribute features between domains, using Maximum Mean Discrepancy (MMD). 3) *Clustering-and-Finetuning Methods.* Fan *et al*. [8] use the model trained on source domain to extract features for target images, perform clustering, assign pseudo labels, and finally finetune the model in a supervised manner. The process in target domain is iterated till convergence. Fu *et al*. [10] conduct clustering for upper-, lower- and whole-body features independently to obtain three label sets, and finetune the model with three loss functions simultaneously. 4) *Other methods* mainly focus on self-supervised constraints. TJ-AIDL [18] formulates relationship between identity and attribute. ECN [19] utilizes the assumption that there is no identity duplication inside a batch of randomly selected target images. CASCL [20] emphasizes on the premise of camera invariance.

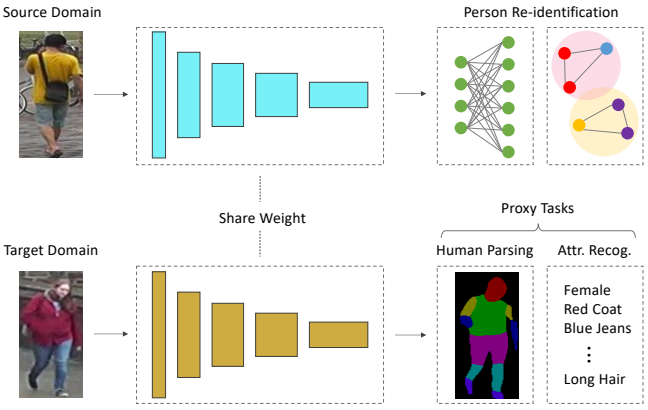

Fig. 2: Overview of our framework. We perform body perception on target-domain pedestrians while training ReID on source domain, sharing the same backbone.

3. METHODOLOGY

Our framework is illustrated in Fig. 2. It performs body perception tasks on target pedestrians at the same time of training source domain ReID, sharing the same backbone. Intuitively, there is strong correlation between body analysis and ReID. For example, human parsing distinguishes foreground from background and localizes each part of the body, which should be an underlying capability of ReID to extract features over body for discrimination. Attribute recognition not only learns features of clothes style, texture, and color, but also involves age, gender, body shape, *etc*. These are useful information for recognizing an identity. We resort to these related body tasks, with the assumption that if the model behaves well in perceiving pedestrians of target domain, it would be benefi-

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on September 26,2020 at 15:12:10 UTC from IEEE Xplore. Restrictions apply.

cial for generalizing the ReID model. These tasks work as a bridge between two domains. Consequently, we term them as *proxy tasks* and our framework as Proxy Task Learning (PTL). By sharing all layers of the backbone between domains and tasks, both low- and high-level variance of domains are tackled. We obtain human parsing labels and attributes for target images with models trained on COCO [12] and PETA [13] respectively. Since these two datasets are publicly available, our method does not introduce extra labeling cost.

3.1. Source-domain Identification

We denote the source training set as $\{(\mathcal{I}_i^s, y_i | i = 1, 2, ..., N^s\}$, where N^s is the number of source images, and $\mathcal{I}_i^s$ is the i-th image with identity label y_i. There are C identities in total and $y_i \in \{1, 2, \ldots, C\}$. For training ReID, we adopt the strong baseline proposed by Luo *et al*. [21]. Concretely, each image is first transformed into feature maps by the backbone (Conv1∼Conv5). Then we perform Global Max Pooling (GMP) to obtain a feature vector $f_i \in \mathbb{R}^D$. Batch Normalization (BN) is further applied upon the result to obtain feature $g_i \in \mathbb{R}^D$. Treating each identity as a class, we adopt a multi-class classifier consisting of cascaded Fully Connected (FC) layer and Softmax layer. We denote the classifier as $\psi : \mathbb{R}^D \to \mathbb{R}^C$, which predicts a probability distribution $p_i = \psi(g_i)$ for image $\mathcal{I}_i^s$, where $p_i \in \mathbb{R}^C$. The identification loss is negative log likelihood of the output node corresponding to ground truth, and the loss over a batch is computed as

$$\mathcal{L}_{ide}^s = -\frac{1}{N_b^s} \sum_{i=1}^{N_b^s} \log(p_{i,y_i}), \tag{1}$$

in which N_b^s is number of images in a batch.

To learn discriminative features, triplet loss [17] is also utilized, which directly optimizes distance in feature space. Following Luo *et al*. [21], the triplet loss is applied to features before BN. Consider a triplet $(\mathcal{I}_{i1}^s, \mathcal{I}_{i2}^s, \mathcal{I}_{i3}^s)$, where $(\mathcal{I}_{i1}^s, \mathcal{I}_{i2}^s)$ are from the same person and $(\mathcal{I}_{i1}^s, \mathcal{I}_{i3}^s)$ from different identities, which are named positive and negative pairs respectively. The loss resulting from this triplet is

$$\mathcal{L}_{tri}^s(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3}) = [\delta + d(f_{i1}, f_{i2}) - d(f_{i1}, f_{i3})]_+, \tag{2}$$

in which $\delta = 0.3$ is a margin and $d(\cdot, \cdot)$ is euclidean distance. The triplet loss inside a batch is thus calculated as

$$\mathcal{L}_{tri}^s = \frac{1}{N_b^s} \sum_{i1,i2,i3} \mathcal{L}_{tri}^s(\mathcal{I}_{i1}, \mathcal{I}_{i2}, \mathcal{I}_{i3}). \tag{3}$$

According to BatchHard [17] sampling strategy, the number of triplets in a batch is the same as batch size N_b^s.

3.2. Target-domain Proxy Tasks

We denote the target training set as $\{(\mathcal{I}_i^t, \mathcal{S}_i, \mathcal{A}_i | i = 1, 2, \ldots, N^t\}$, where N^t is the number of images, and $\mathcal{I}_i^t$ is the i-th image with its parsing label being a 2-dim map $\mathcal{S}_i \in \{1, \ldots, K\}^{H \times W}$ and attribute label being $\mathcal{A}_i$. $\mathcal{A}_i$ is a list of probability distributions, and its j-th distribution $\mathcal{A}_{i,j} \in \mathbb{R}^{M_j}$ is the soft labels for the j-th attribute, where M_j is the number of classes of the attribute, and $\sum_j \mathcal{A}_{i,j} = 1$. There are N_{attr} attributes in total, and $j \in \{1, 2, \ldots, N_{attr}\}$. Here our soft labels for each attribute are obtained from the corresponding Softmax layer of the model trained on public attribute dataset. Note that we do not discretize them into one-hot labels. As discovered in Knowledge Distillation, soft supervision maintains the underlying structure of label space and would be advantageous for optimization. Examples of parsing and attribute labels are illustrated in Fig. 3.

	False	True	False	True	False	True	False	True
hairLong	0.71	0.29	0.99	0.01	0.15	0.85	0.97	0.03
personalMale	0.79	0.21	0.03	0.97	0.80	0.20	0.02	0.98
upperBodyRed	0.62	0.38	0.60	0.40	0.98	0.02	0.92	0.08
lowerBodyJeans	0.33	0.67	0.56	0.44	0.91	0.09	0.64	0.36

Fig. 3: Examples of human parsing and soft attribute labels predicted by models trained on COCO [12] and PETA [13], respectively. Each attribute has two classes *True* and *False*.

For a target image $\mathcal{I}_i^t$, we represent the output of backbone as $\mathcal{E}_i \in \mathbb{R}^{H/2 \times W/2}$. Upon the backbone, we connect a human parsing head and N_{attr} attribute recognition heads to perceive body information of the target person. The human parsing head contains ⟨Deconv, BN, ReLU, Conv, Softmax⟩ layers, denoted by ϕ. It predicts the part label of each pixel on the feature map, with result $\mathcal{G}_i = \phi(\mathcal{E}_i)$, $\mathcal{G}_i \in \mathbb{R}^{K \times H \times W}$. Here $K = 8$ is number of parts plus one, since *background* is viewed as one class as well. For clarity, we denote the vector at spatial location (m, n) of $\mathcal{G}_i$ as $q \in \mathbb{R}^K$, and the corresponding label on $\mathcal{S}_i$ as $r \in \{1, 2, \ldots, K\}$. The human parsing loss for image I_i^t at this location is negative log likelihood $\mathcal{L}_{hp}^t(i, m, n) = -\log(q_r)$. Hence the parsing loss over the whole batch is

$$\mathcal{L}_{hp}^t = \frac{1}{N_b^t HW} \sum_{i=1}^{N_b^t} \sum_{m=1}^{H} \sum_{n=1}^{W} \mathcal{L}_{hp}^t(i, m, n). \tag{4}$$

For perceiving attributes of the target person, Global Average Pooling (GAP) is first conducted upon $\mathcal{E}_i$, obtaining $e_i \in \mathbb{R}^D$. Each attribute has a head consisting of ⟨FC, BN, ReLU, FC, Softmax⟩ layers. Denoting prediction of the j-th head for image I_i^t as $h_i^j \in \mathbb{R}^{M_j}$, the corresponding attribute loss is $\mathcal{L}_{attr}^t(i, j) = -\sum_{l=1}^{M_j} \mathcal{A}_{i,j,l} \log(h_{i,l}^j)$, *i.e*. the weighted sum of negative log likelihoods. The overall attribute loss in

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on September 26,2020 at 15:12:10 UTC from IEEE Xplore. Restrictions apply.

the batch is thus computed as

$$\mathcal{L}^t_{attr} = \frac{1}{N^t_b N_{attr}} \sum_{i=1}^{N^t_b} \sum_{j=1}^{N_{attr}} \mathcal{L}^t_{attr}(i,j). \quad (5)$$

3.3. Multi-task Learning

We share the backbone between source and target domains, and tasks of identification and body perception. In each iteration, a batch of source images are fed to the network to calculate identification loss $\mathcal{L}^s_{ide}$ and triplet loss $\mathcal{L}^s_{tri}$. Then a batch of target images are processed, with human parsing loss $\mathcal{L}^t_{hp}$ and attribute loss $\mathcal{L}^t_{attr}$ computed. All these loss functions are summed up as

$$\mathcal{L} = \lambda^s_{ide}\mathcal{L}^s_{ide} + \lambda^s_{tri}\mathcal{L}^s_{tri} + \lambda^t_{hp}\mathcal{L}^t_{hp} + \lambda^t_{attr}\mathcal{L}^t_{attr} \quad (6)$$

where λ^s_{ide}, λ^s_{tri}, λ^t_{hp} and λ^t_{attr} are hyper parameters to balance between different tasks, which are set to 1, 1, 0.1 and 1 by default. The gradient of final loss w.r.t. network parameters is then back propagated. Eventually the network is optimized to not only distinguish between source identities, but also to perceive body structure and appearance of target persons. As a result, the model would be compatible with target distribution and generalize better. When training is finished, human parsing and attribute heads are removed.

Discussion. Previous works APR [22] and TJ-AIDL [18] also implement multi-task learning of identification and attribute recognition. However, the two tasks are carried out on source images, without explicit attribute learning on target images. As demonstrated in later experiments, our semantic comprehension of target persons plays an important role in domain adaptation.

4. EXPERIMENT

Implementation Details. Our implementation is based on Pytorch. ResNet-50 is adopted as the backbone, and Adam as the optimizer. Learning rate is set to 0.00035 and reduced with a factor of 10 at epochs 160 and 280 respectively. The total training epochs is 480. Warmup [21] is applied at the early stage of training as well. A source batch contains 64 images from 16 identities, each with 4 images sampled; Target batch adopts random sampling with batch size 32. Image resolution for network input is $width \times height = 128 \times 256$. Random flipping and low-level image processing, *e.g.* brightness and contrast perturbation, are used as data augmentation during training. **Datasets and Evaluation Metrics.** We perform cross-domain experiments between three large-scale datasets, Market1501 [1], CUHK03 [2] and DukeMTMC-reID [3]. For CUHK03, we adopt the *detected* subset and the new protocol proposed by Zhong *et al*. [23]. The dataset statistics are summarized in Table 1. Two common evaluation metrics are used, mean Average Precision (mAP) [1] and Cumulative Match Characteristic (CMC) [24]. For CMC, we report the Rank-1, -5 and -10 accuracy.

Dataset	Training	Testing	
		Query	Gallery
Market1501 [1]	751 / 12,936	750 / 3,368	750 / 15,913
CUHK03 [2]	767 / 7,365	700 / 1,400	700 / 5,332
DukeMTMC-reID [3]	702 / 16,522	702 / 2,228	1,110 / 17,661

Table 1: Statistics (#Identities / #Images) of ReID datasets.

	Market→Duke				Duke→Market			
	mAP	R1	R5	R10	mAP	R1	R5	R10
BL	29.1	49.9	63.8	69.9	27.8	58.2	75.2	81.1
BL+HP^t	32.7	54.1	66.8	72.3	**34.5**	65.6	**80.9**	85.7
BL+Attr^t	35.6	56.8	70.2	75.1	32.1	63.3	78.7	84.1
BL+HP^t+Attr^t	**36.2**	**57.4**	**71.0**	**75.8**	34.4	**66.1**	**80.9**	**85.8**
BL+SSG†	45.7	64.7	78.3	81.7	37.9	66.2	81.9	86.9
PTL+SSG†	**52.6**	**71.4**	**82.8**	**86.7**	**46.0**	**74.1**	**87.2**	**91.5**

Table 2: Effectiveness of proxy task learning. Market→Duke means Market1501 is source dataset and DukeMTMC-reID is target dataset. PTL is equivalent to BL+HP^t+Attr^t.

4.1. Substantial Improvement of PTL over Baseline

We first implement a baseline which is only trained on source images, *i.e.* training with only $\mathcal{L}^s_{ide}$ and $\mathcal{L}^s_{tri}$ in Equation 6, denoted by BL in Table 2. Then we train the baseline with various combinations of proxy tasks on target domain, either with human parsing (BL+HP^t), attribute recognition (BL+Attr^t), or both (BL+HP^t+Attr^t). We have following observations. First, both BL+HP^t and BL+Attr^t have substantial improvement over BL. For example, BL+HP^t increases mAP by 3.6% and 6.7% for Market→Duke and Duke→Market respectively, while BL+Attr^t increases 6.5% and 4.3%. Second, BL+Attr^t is superior to BL+HP^t for Market→Duke but worse for Duke→Market. Finally, combining two proxy tasks achieves best performance, indicating that both body structure and appearance are crucial for feature learning. The performance boost of BL+HP^t+Attr^t over BL is 7.1% mAP (7.5% Rank-1) under Market→Duke, and 6.4% mAP (7.9% Rank-1) under Duke→Market. The significant improvement verifies the efficacy of our proxy task learning framework.

4.2. Integration with Clustering-and-Finetuning

As previously discussed, the final performance of clustering-and-finetuning (CFT) methods is highly related to the initial state of the model in target domain. To verify this assumption, we integrate either baseline or PTL (BL+HP^t+Attr^t) with an existing CFT pipeline and compare the results. The CFT method we adopt is SSG [10]. In order to speed up the process of SSG, here we omit the neighbor reranking step and feature extraction of source images. This simplified version of SSG is denoted by SSG†. The results are recorded in Table 2. We first observe that BL+SSG† largely improves upon BL by clustering and finetuning, which demonstrates the benefit of rectifying neighborhood relationship in target domain. Further, comparing PTL+SSG† with BL+SSG†, we see that our method achieves improvement with a large margin. In mAP (Rank-1), the superiority brought by PTL is 6.9% (6.7%) for

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on September 26,2020 at 15:12:10 UTC from IEEE Xplore. Restrictions apply.

	Market→Duke			
	mAP	R1	R5	R10
BL	29.1	49.9	63.8	69.9
BL+HP^s	30.6	52.0	66.1	71.1
BL+HP^t	32.7	54.1	66.8	72.3
BL+HP^s+HP^t	34.3	55.4	69.3	73.9
BL+Attr^s	30.7	51.5	65.3	70.9
BL+Attr^t	35.6	56.8	70.2	75.1
BL+Attr^s+Attr^t	35.4	56.1	70.3	75.0

Table 3: Relationship with source-domain body perception.

	Market→CUHK			
	mAP	R1	R5	R10
BL	13.3	15.4	27.6	34.7
BL+Attr$^t_{\text{PETA, soft}}$	18.9	20.5	37.3	46.1
BL+Attr$^t_{\text{PETA, one-hot}}$	18.3	20.5	35.9	45.2
BL+Attr$^t_{\text{RAP, soft}}$	17.2	18.4	33.3	43.0

Table 4: Component analysis for attribute recognition.

Market→Duke and 8.1% (7.9%) for Duke→Market. The results implies that, by performing perception tasks on target pedestrians, our method results in a model more compatible with target domain and facilitates the following finetuning.

4.3. Relationship with Source-domain Body Perception

Considering the close relation between ReID and body perception tasks, we wonder whether training these tasks on source domain has an influence on cross-domain generalization, and whether it is still beneficial to keep training these tasks on target domain. To answer these questions, we involve source-domain human parsing and attribute recognition in training. Specifically, for human parsing, we train it either only on source images (BL+HP^s), only on target images (BL+HP^t), or on both (BL+HP^s+HP^t). Similar experiments are conducted for attribute recognition as well. The results are reported in Table 3. Comparing BL+HP^s and BL+Attr^s with BL, we notice that source-domain perception tasks indeed benefit cross-domain testing. However, we can observe much more improvement brought by target-domain tasks, by comparing BL+HP^t with BL+HP^s, and BL+Attr^t with BL+Attr^s. The indispensable role of target-domain tasks is to perceive and adapt to distribution of target pedestrians, which could not be accomplished with only source images.

4.4. Component Analysis for Attribute Recognition

In other sections of the paper, we use model trained on PETA to predict soft attribute labels for ReID images. Here we experiment with discretized one-hot labels. Besides, we also try an alternative attribute dataset RAP [25]. The results are shown in Table 4. We first notice that all three variants improve upon baseline considerably. In addition, the superiority of BL+Attr$^t_{\text{PETA, soft}}$ over BL+Attr$^t_{\text{PETA, one-hot}}$ verifies the rationality of choosing soft label as supervision. Finally, comparing BL+Attr$^t_{\text{RAP, soft}}$ and BL+Attr$^t_{\text{PETA, soft}}$, we conclude that

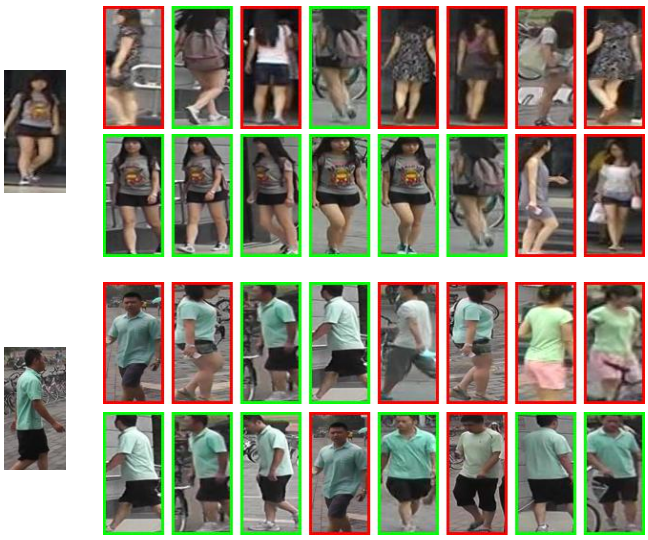

Fig. 4: Two ranking examples under Duke→Market setting. The query image is on the left, and top-8 gallery images are listed on the right. In each case, the first row is returned by BL, and second row by BL+HP^t+Attr^t. Green and red surrounding boxes denote having same and different identity with query, respectively.

PETA is a better choice. We utilize similar number of attributes of PETA and RAP, *i.e*. 105 *vs*. 96. Nonetheless, PETA consists of ten different scenes, while RAP mainly involves limited indoor scenarios. The huge diversity of PETA images may result in a more robust attribute model, which could predict attributes on ReID images with higher quality and in turn facilitates our PTL training.

4.5. Qualitative Analysis of Ranking Result

To qualitatively demonstrate the benefits of our method, we illustrate two ranking cases where BL+HP^t+Attr^t shows superiority over BL, as in Fig. 4. In the first case, the baseline model fails to retrieve most of correct results, even if these images are with similar body pose and thus appearance as query. This implies undesirable model collapse caused by domain shift. Through adaptation, our method BL+HP^t+Attr^t successfully returns much more target images. In the second case, the baseline mistakenly matches with multiple pedestrians wearing similar T-shirts. With the perception of body structure and attributes of target-domain persons, our method manages to reject those images with different gender or with obvious color discrepancy.

4.6. Comparison with State of the Art

To compare with state-of-the-art methods, we combine PTL with SSG [10], denoted by PTL+SSG. The results under various settings are reported in Table 5. The approaches we compare include GAN based methods (SPGAN [4], SBSGAN [6], CR-GAN [5]), MMD methods (MMFA [7]), clustering-and-finetuning methods (PUL [8],

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on September 26,2020 at 15:12:10 UTC from IEEE Xplore. Restrictions apply.

	Publication	Market→Duke		Duke→Market	
		mAP	R1	mAP	R1
PUL [8]	TOMM18	16.4	30.0	20.5	45.5
TJ-AIDL [18]	CVPR18	23.0	44.3	26.5	58.2
MMFA [7]	BMVC18	24.7	45.3	27.4	56.7
SPGAN [4]	CVPR18	26.2	46.4	26.7	57.7
CASCL [20]	ICCV19	30.5	51.5	35.6	64.7
SBSGAN [6]	ICCV19	30.8	53.5	27.3	58.5
ECN [19]	CVPR19	40.4	63.3	43.0	75.1
MAR [9]	CVPR19	48.0	67.1	40.0	67.7
CR-GAN [5]	ICCV19	48.6	68.9	54.0	77.7
SSG [10]	ICCV19	53.4	73.0	58.3	80.0
PAST [11]	ICCV19	54.3	72.4	54.6	78.4
PTL+SSG (Ours)		60.7	76.2	69.0	87.3

	Publication	CUHK→Market		CUHK→Duke	
		mAP	R1	mAP	R1
PUL [8]	TOMM18	18.0	41.9	12.0	23.0
SPGAN [4]	CVPR18	19.0	42.3	-	-
SBSGAN [6]	ICCV19	28.5	57.6	27.8	47.7
CR-GAN [5]	ICCV19	56.0	78.3	47.7	67.7
PAST [11]	ICCV19	57.3	79.5	51.8	69.9
PTL+SSG (Ours)		73.1	88.9	58.5	75.6

Table 5: Comparison with state-of-the-art methods under various settings. The 1st, 2nd and 3rd highest scores in each column are marked by red, green and blue, repectively.

MAR [9], SSG [10], PAST [11]), as well as those devising novel cross-domain constraints (TJ-AIDL [18], CASCL [20], ECN [19]). Our method achieves highest performance under all these cross-domain settings, surpassing previous methods by a large margin. Concretely, in terms of mAP, the advantage over 2nd highest method reaches 6.4% (60.7 *vs*. 54.3) for Market→Duke, 11.7% (69.0 *vs*. 58.3) for Duke→Market, 15.8% (73.1 *vs*. 57.3) for CUHK→Market, and 6.7% (58.5 *vs*. 51.8) for CUHK→Duke. In terms of Rank-1, the corresponding boosts are 3.2% (76.2 *vs*. 73.0), 7.3% (87.3 *vs*. 80.0), 9.4% (88.9 *vs*. 79.5), and 5.7% (75.6 *vs*. 69.9), respectively. The significant improvement verifies the efficacy of our proxy task learning framework.

5. CONCLUSION

In order to increase the generalization ability of ReID model from source to target domain, we propose a Proxy Task Learning framework that performs body perception on target-domain pedestrians while training ReID on source domain, sharing the same backbone in a multi-task manner. We choose human parsing and attribute recognition as two proxy tasks, considering their strong relation to ReID. Extensive experiments are conducted to verify the efficacy of our framework, demonstrating the important role of perceiving target pedestrians to minimize domain gap. Our final model achieves state-of-the-art performance on large-scale benchmarks.

6. ACKNOWLEDGEMENT

This work is supported in part by the National Key Research and Development Program of China (Grant No. 2016YFB1001005), the National Natural Science Foundation of China (Grant No. 61673375 and 61721004), and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006).

7. REFERENCES

[1] Zheng et al., "Scalable person re-identification: A benchmark," in *ICCV*, 2015.
[2] Li et al., "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.
[3] Zheng et al., "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.
[4] Deng et al., "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification," in *CVPR*, 2018.
[5] Chen et al., "Instance-guided context rendering for cross-domain person re-identification," in *ICCV*, 2019.
[6] Huang et al., "Sbsgan: Suppression of inter-domain background shift for person re-identification," in *ICCV*, 2019.
[7] Lin et al., "Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification," in *BMVC*, 2018.
[8] Fan et al., "Unsupervised person re-identification: Clustering and fine-tuning," *TOMM*, 2018.
[9] Yu et al., "Unsupervised person re-identification by soft multilabel learning," in *CVPR*, 2019.
[10] Fu et al., "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *ICCV*, 2019.
[11] Zhang et al., "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *ICCV*, 2019.
[12] Güler et al., "Densepose: Dense human pose estimation in the wild," in *CVPR*, 2018.
[13] Deng et al., "Pedestrian attribute recognition at far distance," in *ACM MM*, 2014.
[14] Zhou et al., "Omni-scale feature learning for person re-identification," in *CVPR*, 2019.
[15] Sun et al., "Beyond part models: Person retrieval with refined part pooling," in *ECCV*, 2018.
[16] Zhang et al., "Densely semantically aligned person re-identification," in *CVPR*, 2019.
[17] Hermans et al., "In defense of the triplet loss for person re-identification," *arXiv*, 2017.
[18] Wang et al., "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *CVPR*, 2018.
[19] Zhong et al., "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *CVPR*, 2019.
[20] Wu et al., "Unsupervised person re-identification by camera-aware similarity consistency learning," in *ICCV*, 2019.
[21] Luo et al., "Bag of tricks and a strong baseline for deep person re-identification," in *CVPR Workshops*, 2019.
[22] Lin et al., "Improving person re-identification by attribute and identity learning," *PR*, 2019.
[23] Zhong et al., "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.
[24] Gray et al., "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS Workshop*, 2007.
[25] Li et al., "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *TIP*, 2018.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on September 26,2020 at 15:12:10 UTC from IEEE Xplore. Restrictions apply.