

Part-aware Progressive Unsupervised Domain Adaptation for Person Re-Identification

Fan Yang, Ke Yan, Shijian Lu, Huizhu Jia, Don Xie, Zongqiao Yu, Xiaowei Guo, Feiyue Huang, and Wen Gao, *Fellow, IEEE*

Abstract—Unsupervised domain adaptation (UDA) aims to mitigate the domain shift that occurs when transferring knowledge from a labeled source domain to an unlabeled target domain. While it has been studied for application in unsupervised person re-identification (ReID), the relations of feature distribution across the source and target domains remain underexplored, as they either ignore the local relations or omit the in-depth consideration of negative transfer when two domains do not share identical label spaces. In light of the above, this paper presents an innovative part-aware progressive adaptation network (PPAN) that exploits global and local relations for UDA-based ReID across domains. A multi-branch network is developed that explicitly learns discriminative feature representation from both whole-body images and body-part images under the supervision of a labeled source domain. Within each network branch, an independent UDA constraint is designed that aligns the global and local feature distributions from a labeled source domain with those of an unlabeled target domain. In addition, a novel progressive adaptation strategy (PAS) is designed that effectively alleviates the negative influence of outlier source identities. The proposed unsupervised ReID model is evaluated on five widely used datasets (Market-1501, DukeMTMC-reID, CUHK03, VIPeR and PRID), and experimental results demonstrate its superior robustness and effectiveness relative to state-of-the-art approaches.

Index Terms—Unsupervised domain adaptation, Person re-identification, Part-aware, Feature alignment, Progressive adaptation.

I. INTRODUCTION

PERSON re-identification (ReID) aims to identify the same person from non-overlapping cameras. It has attracted increasing attention in recent years in the multimedia community [1–8] due to its importance in various applications including cross-camera tracking [9–11] and image/video retrieval [12, 13]. While leveraging large-scale labeled person ReID datasets such as Market-1501 [14] and DukeMTMC-reID [15], most existing ReID systems adopt a supervised learning approach such as metric learning based methods [1–3] and deep learning based methods [4, 5, 7, 8, 16–26]. These efforts have achieved great success; for example, the Rank-1

This work is partially supported by the Beijing Major Science and Technology Project under contract No. Z191100010618003 and Project 2019BD004 supported by PKU-Baidu Fund, and also supported by grants from NVIDIA NVAIL program.

F. Yang and K. Yan contributed equally to this work.

Corresponding author: Huizhu Jia (email: hzjia@pku.edu.cn).

F. Yang, H. Jia, D. Xie and W. Gao are with the National Engineering Laboratory for Video Technology, Peking University, Beijing, China.

S. Lu is with the Nanyang Technological University, Singapore.

K. Yan, Z. Yu, X. Guo and F. Huang are with the YouTu Lab, Tencent, Shanghai.

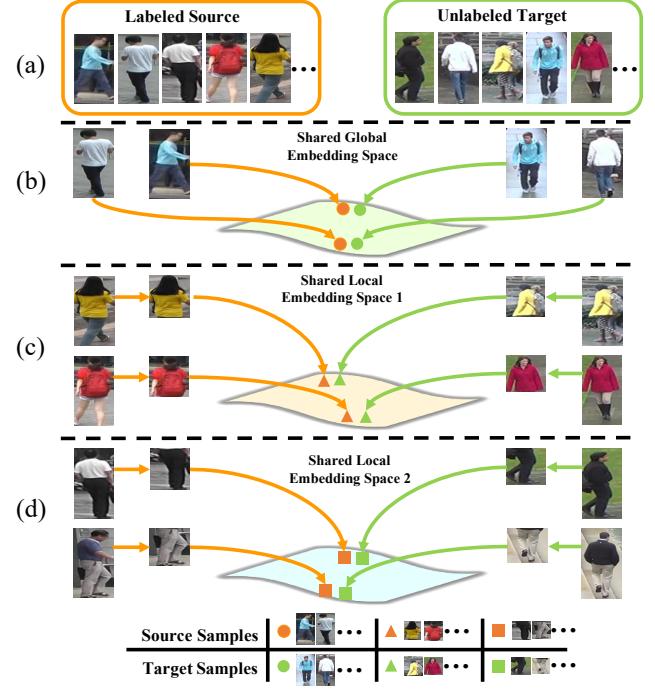


Fig. 1. Motivations and concepts behind the proposed part-aware unsupervised domain adaptation in person ReID: Persons from different domains share many common characteristics, including genders, hairstyle and outfit color and texture. These shared characteristics could be either global, as illustrated in 1b, or local, as illustrated in 1c and 1d. Boxes and arrows in orange highlight image samples in the labeled source domain, while those in green highlight image samples in the unlabeled target domain.

accuracy for the Market-1501 and DukeMTMC-reID datasets has been improved from 43.79% [14] and 30.75% [27] to 94.99% [26] and 86.04% [26], respectively. On the other hand, existing person ReID systems continue to face two major constraints in practical deployments. First, collecting and manually annotating a large number of training images is extremely expensive and time-consuming. Second, a ReID model trained on one dataset often exhibits low generalizability when applied to another dataset due to the impact of domain bias or shift [28]. This scalability issue severely hinders supervised ReID in many real-world tasks, where domain-specific labels are often unavailable.

Designing unsupervised algorithms is a common method adopted to improve the generalization capacity of person ReID models. Most traditional unsupervised methods extract hand-crafted features that are invariant even under different cameras [27, 29–31]. Once these invariant features are extracted, naive

feature matching or unsupervised distance metric learning [32–37] is applied to measure the image similarity. However, these approaches use shallow hand-crafted features that suffer from representational limitations. In recent years, several works [38–43] have used pseudo-labels to learn target representations; however, these works are affected by two major constraints. First, label estimation is susceptible to view-specific interference, while images from the same view (i.e. rather than those of the same person from different views) are likely to be labeled as the same person. Second, falsely pseudo-labeled samples cause the learned model to be vulnerable to error accumulation. Without correctly labeled pairwise samples, it is very challenging to learn discriminative feature representations due to the large ‘intra-class’ variation and relatively small ‘inter-class’ variation [26].

Unsupervised domain adaptation (UDA) [44] is able to transfer knowledge containing view-invariant discriminative information from a labeled source domain to a related, unlabeled target domain. The challenge here is to mitigate the domain shift resulting from domain-specific data characteristics. Vanilla UDA [45, 46] assumes that the source and target domains share an identical label space; however, this is invalid in person ReID, where the source and target domains usually contain entirely different identities. To address this problem, several works [12, 47–50] have exploited attribute information under the assumption that visual attributes are commonly shared between different persons across domains. However, sparse and low-dimensional representations of attributes are often insufficient for identity discrimination. Very recently, several attempts [51–53] have used CycleGAN [54] to translate the labeled source-domain images to the target domain. Using generated images that inherit the identity labels of the source domain and the image styles of the target domain, deep models can learn feature representations directly under the supervised learning paradigm. On the other hand, these works only consider general shifts, such as illumination, but ignore the bias of imaging conditions, such as viewing angles, human poses and background clutter which tends to lead to sub-optimal ReID matching.

We can observe three distinct feature patterns across images in different domains that could be exploited to address the constraints of state-of-the-art methods. First, rich high-level features such as gender, hairstyle and outfit color and texture are commonly shared across domains that have non-overlapping label spaces. These commonly shared features lie within a specific latent space that can be exploited for better UDA in person ReID. Second, persons from different domains share multi-granular latent features; for example, some people share global features, as illustrated in Fig. 1b, while some people share either upper-body features (but different lower-body features) or lower-body features (but different upper-body features), as illustrated in Figs. 1c and 1d. Third, certain outlier source identities are located far away from the target domain, which may cause the positive transfer from the source domain to the target domains in the shared embedding space to deteriorate. When aiming to alleviate the influence of such negative transfer in UDA, it is more helpful to select positive source identities in order to align the global and local feature

distributions across domains.

Based on the aforementioned observations, we propose an innovative Part-aware Progressive Adaptation Network (PPAN), which explicitly enforces the feature alignment across domains in a progressive manner. More specifically, a Part-aware Adaptation Network (PAN) is designed to learn view-invariant feature representations from the labeled source domain and perform domain adaptation to the unlabeled target domain by independently aligning global and local feature distributions. Moreover, to suppress the negative influence of outlier identities in the source domain, a Progressive Adaptation Strategy (PAS) is designed that progressively eliminates outlier source identities according to cross-domain similarity measures. The proposed PPAN makes four major contributions:

- It designs a novel UDA framework that learns multi-granular domain-invariant ReID features;
- It designs a novel feature alignment that regularizes feature representations concurrently via the unsupervised domain-adaptive constraint and feature norm constraint;
- It designs a PAS that alleviates the negative transfer problem by automatically evaluating the transferability of source examples based on their similarity to the target domain;
- It develops an end-to-end trainable deep network that achieves superior person ReID performance across a number of widely used benchmarking datasets.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of related works. Section 3 presents our proposed method in detail. Implementation and experimental results are presented in Section 4. Finally, several concluding remarks are drawn in Section 5.

II. RELATED WORK

Person ReID has been extensively studied and a large number of person ReID systems have been reported in recent years. From the perspective of feature representation extraction, which is most relevant to our work, existing methods can be broadly grouped into three categories: 1) supervised learning based [16–26, 55, 56], 2) unsupervised learning based [32–43] and 3) UDA based [47–53, 57–59].

A. Supervised Person ReID

Early examples of deep person ReID learned global representation from whole-body images only. Different methods have been reported that learn feature representations and distance metrics with different losses, such as identity classification loss, pair-wise verification loss, triplet ranking and quadruplet ranking loss, etc. [16–19, 21, 56]. For example, [17] trains a classification model by treating images of each individual person as a specific category, while [18] includes pose-normalized images to train classification models. Different networks have also been studied to measure image similarity, including Siamese networks [60, 61], triplet networks [56], and quadruplet deep networks [19]. However, while these methods can effectively learn global person representation,

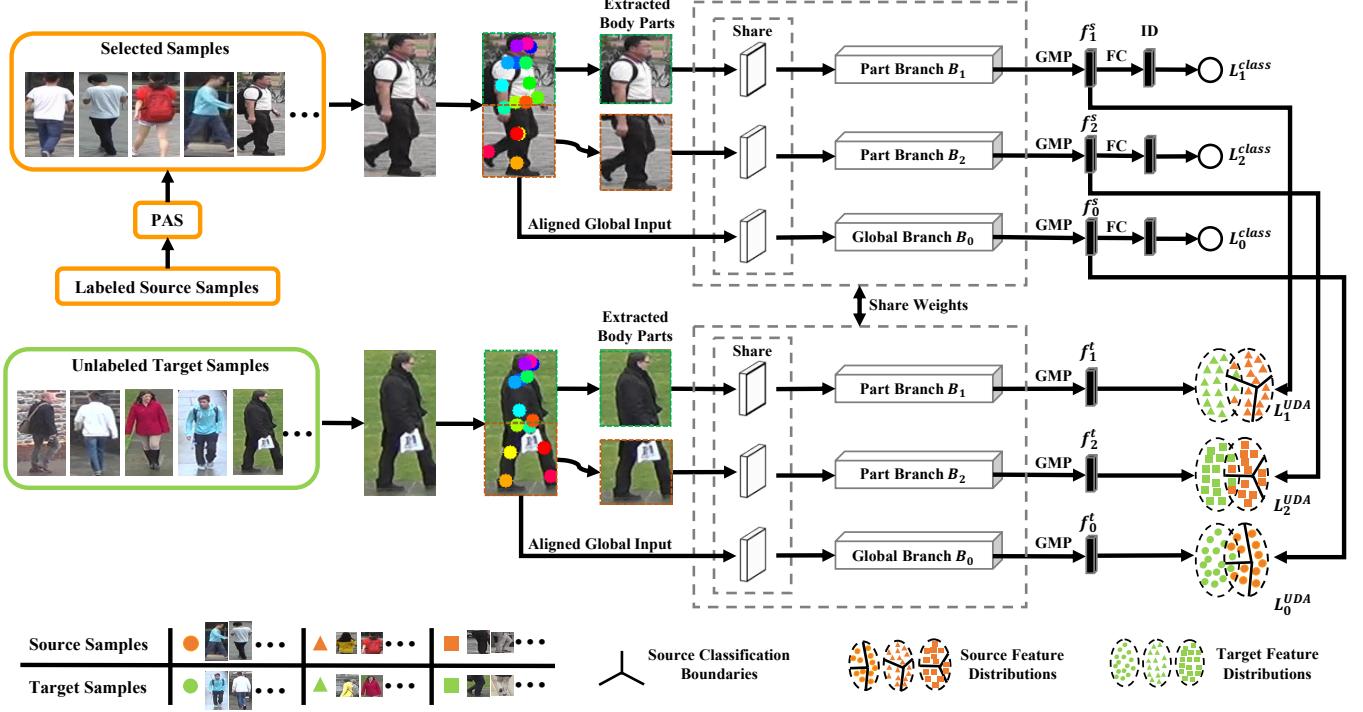


Fig. 2. The framework of the proposed part-aware progressive adaptation network (PPAN): Given images from the source and target domains, a whole-body image and two body-part images are first determined via pose estimation. Three feature extraction branches with global max pooling (GMP) layers then map the input image to a discriminative feature embedding space as supervised by the independent softmax classification loss of the source domain. Three unsupervised domain adaptation (UDA) constraints are introduced to align the distribution of global and local features between the source and target domains. In addition, a progressive adaptation strategy (PAS) is designed to select positive source samples according to cross-domain similarity.

they often produce sub-optimal ReID performance as they ignore the very informative details concerning body parts.

Several systems [21, 22] have been designed to learn part representations from pre-defined fixed-height horizontal strips, with the aim of capturing richer and finer visual cues from body-part images. To address the misalignment problem, [23] uses off-the-shelf pose estimation models to detect body parts, then iteratively fuses these part representations with global representations. [24] proposes a Harmonious Attention (HA) model that simultaneously locates body parts from whole-body images and learns multi-scale feature maps. [26] designs a multi-branch intra-attention network to learn feature representations from attentive global and local regions and adaptively fuses the output of intra-attention networks by means of an inter-attention network. However, while supervised person ReID methods achieve good accuracy, they also suffer from poor scalability due to the requirement that labeled training samples be available.

Although a number of part-based supervised person ReID methods [21–24, 26] have been reported, exploiting parts to learn discriminative local features from unlabeled data/domain remains an open research challenge. For our part, by observing that persons from different domains often share local features, we design a part-aware UDA framework that learns multi-granular domain-invariant features for domain-adaptive person ReID. From this aspect, we incorporate part-aware multi-granular feature learning, UDA constraint, feature norm constraint and PAS and make them learn in a collaborative

and complementary way. To the best of our knowledge, this is the first domain-adaptive person ReID work to model multi-granular UDA in a progressive manner.

B. Unsupervised Person ReID

With the goal of addressing the scalability issue, unsupervised ReID directly extracts or learns cross-view feature representations from unlabeled images. Existing works can be broadly classified into three categories according to the utilized feature extraction scheme. Methods in the first category involve creating hand-crafted features [27, 29–33, 35] such as color histograms, Histogram of Oriented Gradient (HOG) and SIFT. For example, [32, 33] exploit salient image patches that are informative for ReID matching. [35] extracts weighted histograms of overlapping stripe features and applies iterative sparse ranking to generate the ranked list of gallery individuals. The challenge associated with this approach is that of how to design view-invariant features that are both robust and tolerant to different camera views.

Methods in the second category learn a dictionary with atoms corresponding to specific semantic elements. Using the learned dictionary, feature representations are generated by minimizing the reconstruction error. For example, [34] formulates unsupervised ReID as a sparse dictionary learning problem. To regularize the learned dictionary, graph Laplacian regularization is exploited to iteratively update the graph Laplacian matrix. Moreover, in order to make the graph Laplacian regularization tolerant to outliers, [36] introduces

a L_1 -norm to enable joint learning of the graph and the dictionary. However, the issue with unsupervised methods is that their accuracy is usually much lower than supervised methods due to the lack of supervision.

Methods in the third category jointly estimate labels for unlabeled domains and learn feature representations using pseudo-labels [38–43]. For example, [38] learns a specific projection for each view based on asymmetric clustering on cross-view person images. [39] constructs a graph for samples in each camera and introduces a dynamic graph matching (DGM) process for cross-camera labelling association. [43] utilizes k -means clustering to generate pseudo-labels used in fine-tuning the deep ReID model. [42] uses k -reciprocal nearest neighbors to obtain proximate label information for unsupervised video ReID. While the use of pseudo-labels results in a certain level of success, pseudo-labels are often noisy and lead to poor view-invariant and discriminative representations when there are large discrepancies between person images captured under different conditions.

C. Unsupervised Domain-Adaptive Person ReID

UDA [44] transfers view-invariant and discriminative feature representations from the labeled source domain to the unlabeled target domain. The main strategy involves guiding feature learning by minimizing the discrepancy between the source and target feature distributions [45, 46, 54, 62–70]. For example, [45] and [62] use Maximum Mean Discrepancy (MMD) to compute the difference between two domains in order to reduce the MMD distance. [63, 64] match the mean and covariance of the two distributions with the deep Correlation Alignment (CORAL), while [65–67] use the features extracted by encoder networks to reconstruct the images in both domains. Moreover, due to the great success of generative adversarial networks (GAN) [71], adversarial domain adaptation has received increasing attention in the learning of domain-invariant representations in recent years [46, 54, 68–70]. Most existing UDA methods assume that class labels remain the same across domains; however, this assumption is invalid in domain adaptation in the person ReID context, where different datasets have non-overlapping identities.

Over the past few years, several unsupervised domain-adaptive ReID models have been proposed to learn feature representation for the target domain. For example, [47–50, 58] use visual semantic attributes as a common space in order to bridge the gap between the source and target domains. [49] designs a joint attribute-identity transfer learning scheme that can adapt the knowledge from the source domain to the target domain. [50] regularizes networks by aligning the distribution of predefined attributes, while [51–53] use style-translated images with inherited identity labels to train ReID models in a supervised manner. [57] searches for a common factorized space shared between the source and target domains and exploits both an unsupervised factorization loss and a graph-based loss in training.

Most existing UDA works focus on matching the entire source and target data distribution in order to learn global domain-invariant representations. On the other hand, mapping

the source and target domain images to a single global embedding space might result in sub-optimal distribution matching in cases where cross-domain persons share multi-granular characteristics. At the same time, however, aligning the entire source and target domains might make the model vulnerable to negative transfer, as the source and target domains do not share an identical label space. The proposed Part-aware Progressive Adaptation Network (PPAN) addresses these constraints by independently aligning the complementary distributions of global and local feature embedding. In addition, it also exploits identity selection that progressively selects positive source identities according to their similarity to the target samples (more details will be provided in the following sections).

While MMD loss has been used in existing works, we here adapt it for the challenging domain-adaptive person ReID task. More specifically, we explore the usability of MMD in multi-granular domain-invariant ReID feature learning, which helps in overcoming the sub-optimal alignment of global feature distribution. The feature norm is widely used for information measurement in the statistics and machine learning communities [72]; here, we exploit it to guide the global branch to facilitate the learning of informative and complementary global domain-invariant features (rather than treating it as an information measurement tool).

In addition, we propose a novel Progressive Adaptation Strategy (PAS), which evaluates the transferability of source examples based on their similarities to the target domain. Given that source and target domains usually contain entirely different identities, certain outlier identities in the source domain that are located far away from the target domain may deteriorate the positive transfer in the shared embedding space. PAS addresses this issue by estimating a transferability that computes the distance between each identity in the source domain and all samples in the target domain. Using this estimated transferability, we transfer the relevant source identities, thus promoting positive transfer, and ignore the irrelevant ones to mitigate negative transfer.

III. METHODOLOGY

Problem Definition For person ReID utilizing UDA, there is a source domain $D^s = \{I_i^s, c_i^s\}_{i=1}^{N^s}$ ($I_i^s \in I^s$, $c_i^s \in [1, \dots, Q]$) that has N^s labeled samples I^s with corresponding identity labels c_i^s that belong to Q distinct persons. There is also a target domain $D^t = \{I_j^t\}_{j=1}^{N^t}$ ($I_j^t \in I^t$) that has N^t unlabeled samples I^t for domain adaptation. The assumption here is that the source and target domains contain entirely different identities, while the objective is to learn view-invariant feature representations from the labeled D^s and adapt them to the unlabeled D^t under certain UDA constraints.

In this paper, we develop a Part-aware Adaptation Network (PAN) that learns multi-granular view-invariant and domain-invariant feature representations from both whole-body images and body-part images, as illustrated in Fig. 2. Moreover, we also design a Progressive Adaptation Strategy (PAS) that selects positive source identities according to the cross-domain similarity measures. The below subsections will describe the network design and implementation (including body part detection, PAN and PAS) in detail.

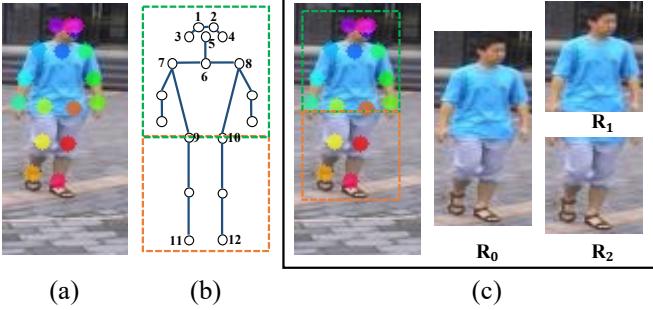


Fig. 3. Illustration of human part detection: (a) 18 major human body joints detected by pose estimation; (b) definition of three body parts by using 12 of the 18 detected major body joints; (c) detection of three body regions based on the definition in Fig. 3b.

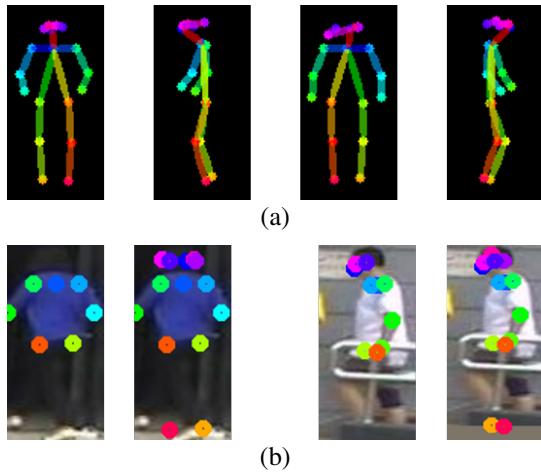


Fig. 4. Illustration of improvement in joint detection: (a) four estimated canonical poses on the dataset Market-1501, including front-view, left-side-view, back-view and right-side-view from left to right; (b) visualization of the improved joint detection—the graph on the left of each pair shows the original detection, while the one on the right shows the improved detection.

A. Body Part Detection

Human images, especially those collected using automatic detectors, often suffer from various types of misalignment, either because they include too much background clutter or are missing certain body parts. On the other hand, positive transfer in UDA often requires good alignment of the human body and body parts. Based on the observation that body parts can generally be localized using body joints, we first employ pose estimation to localize body joints, then use the localized body joints for human alignment. More specifically, we adopt an off-the-shelf pose estimator [73] that directly produces the 2D locations of 18 major body joints. We use 12 of the 18 key points $K_a(a = 1, \dots, 12)$ for alignment, as illustrated in Fig. 3a and 3b.

It should be noted here that joint detection may suffer from detection errors when dealing with low-quality images due to occlusions, poor lighting, etc. We deal with this issue by leveraging a set of canonical human poses that represent a list of typical human body configurations as exhibited on public surveillance cameras. Canonical human poses define the relative position relations between body joints from a

specific viewpoint. Following [74], we predict the poses of all training images in a dataset, then split the poses with high detection confidence into four groups: front-view, left-side-view, back-view and right-side-view. The mean pose images of these four groups are designated as the canonical poses. With the angle between vector $\vec{K_6K_9}$ and vector $\vec{K_6K_{10}}$ (K_6 and K_{10} denoting the sixth and tenth body joints, as illustrated in Fig. 3b), the poses can be split into two groups, namely {front-view, back-view} and {left-side-view, right-side-view}. The first group can be further divided based on the positions of the left and right shoulders, while the second group is split according to the orientation of the knees. The mean pose images of the four groups are shown in Fig. 4a (on dataset Market-1501). Joint detection can thus be improved by replacing the missing joints (or joints identified with ultra-low confidence) according to the estimated canonical poses and existing joints identified with confidence. Fig. 4b illustrates how joint detection can be improved using the estimated canonical poses; here, the image on the left of each pair shows the original detection, while the one on the right shows the improved detection with canonical pose.

Using the 12 landmark body joints illustrated in Fig. 3b, we first divide the human images into three regions, as illustrated in Fig. 3c: namely, the whole body region $P_0 = \{K_1, \dots, K_{12}\}$, the upper-body region $P_1 = \{K_1, \dots, K_{10}\}$, and the lower-body region $P_2 = \{K_9, \dots, K_{12}\}$. Let (x_a, y_a) denote the coordinates of the 12 major joints; accordingly, the height $H_i(i = 1, 2)$ of the upper-body region P_1 and lower-body region P_2 can be computed as follows:

$$H_i = \max_{a \in P_i}(y_a) - \min_{a \in P_i}(y_a) \quad (1)$$

The vertical boundary (y_t, y_b) of the whole body region P_0 can be determined by:

$$\begin{aligned} y_t &= \min_{a \in P_1}(y_a) - \frac{1}{3}H_1 \\ y_b &= \max_{a \in P_2}(y_a) + \frac{1}{6}H_2 \end{aligned} \quad (2)$$

where $\frac{1}{3}H_1$ and $\frac{1}{6}H_2$ denote the expansions that aim to enlarge the vertical boundary so that it includes the head and foot regions completely. The key parameters of the global human region can be computed as follows:

$$\begin{aligned} H_0 &= y_b - y_t \\ (x_0, y_0) &= (\bar{x}_a, \bar{y}_a) \quad a \in P_0 \\ x_l &= \min(x_0 - H_0/4, \min_{a \in P_0}(x_a)) \\ x_r &= \max(x_0 + H_0/4, \max_{a \in P_0}(x_a)) \end{aligned} \quad (3)$$

where (x_l, x_r) , (x_0, y_0) and H_0 denote the horizontal boundary, the center and the height of the whole body region, respectively. Moreover, the bar symbol in \bar{x}_a and \bar{y}_a denotes a mean operator. For each body region $P_b(b = 0, 1, 2)$, the corresponding bounding box $R_b(b = 0, 1, 2)$ can be determined by:

$$\begin{aligned} R_0 &= (x_l, y_t, x_r, y_b) \\ R_1 &= (x_l, y_t, x_r, \max_{a \in P_1}(y_a)) \\ R_2 &= (x_l, \min_{a \in P_2}(y_a), x_r, y_b) \end{aligned} \quad (4)$$

For the sample image in Fig. 3a, Fig. 3c shows the three identified human regions.

B. Multi-granular Feature Representation Learning

We utilize ResNet50 [75], pre-trained on ImageNet [76], for the basic building elements and design a multi-branch network for multi-granular feature representation learning. As illustrated in Fig. 2, the designed network consists of three branches: 1) one global branch, B_0 , that aims to learn global features from whole-body images, and 2) two independent local branches (B_1 and B_2) that aim to learn local features from two body-part images. To reduce the number of model parameters and over-fitting risks, the global and local branches share the first CONV layer; this is because the lower CONV layers capture low-level features that are common to all patterns in the same image. We also attempted to share more residual units [75], but this resulted in a marked performance drop, largely because of the limited representation capacity of the shared higher CONV layers. For each branch, we removed the global average pooling layer and the following fully connected layers. Additionally, the last down-sampling operation in each branch is also removed to increase the granularity of the learned features.

Given a whole-body region R_0 or body-part region R_b ($b = 1, 2$), a feature map M_b is first extracted by the corresponding branch B_b ($b = 0, 1, 2$), as follows:

$$M_b = \mathcal{F}(R_b, \theta_b) \quad (5)$$

where R_b and M_b are the input and output of the feature extraction branch B_b , while \mathcal{F} is the feature extraction operation with parameters θ_b . The output M_b is a 3D tensor $M_b \in \mathbb{R}^{h \times w \times c}$, where h , w , and c denote the height, width, and channel number of the feature map M_b , respectively. A Global Max Pooling (GMP) layer is then applied to convert the feature map M_b to a feature representation f_b , as follows:

$$f_b = GMP(M_b) \quad (6)$$

Global Average Pooling (GAP) is a traditional pooling strategy that works well in classification networks. When applied to person ReID, however, the GAP operation could overlook discriminative cues that are critical to the ReID performance. Let us take some small but discriminative body part (surrounded by the background) as an example: in this case, the average pooling will suppress its strong response and lead to misdetection. By contrast, Global Max Pooling (GMP) can preserve the largest response from the human body region without being affected by the responses from the surrounding background.

Since the source and target domains share the same feature extraction branch parameters, the feature representations extracted from the source and target domain images can be denoted by f_b^s ($b = 0, 1, 2$) and f_b^t ($b = 0, 1, 2$), respectively.

Given the ground-truth of source sample pairs (I_i^s, c_i^s) , the three feature extraction branches map the respective input image to view-invariant feature representation as supervised by the following independent softmax classification loss:

$$\mathcal{L}_b^{\text{Class}} = -\frac{1}{N^s} \sum_{i=1}^{N^s} \log\left(\frac{\exp(W_{c_i^s}^b \mathcal{BN}(f_b^s))}{\sum_{q=1}^Q \exp(W_q^b \mathcal{BN}(f_b^s))}\right) \quad (7)$$

where N^s is the number of source training images, Q is the number of identities, W_q^b denotes the parameter of the prediction function for the training identity q in the b -th branch, and \mathcal{BN} denotes the batch normalization operation.

The overall source cross-entropy classification loss can thus be defined as follows:

$$\mathcal{L}^{\text{Class}} = \mathcal{L}_0^{\text{Class}} + \lambda \sum_{b=1}^2 \mathcal{L}_b^{\text{Class}} \quad (8)$$

where λ controls the relative weights of the global-branch and local-branch loss.

C. Part-aware Adaptation

1) *UDA Constraints*: Supervised multi-granular feature representation usually contains rich domain-specific information, such as background clusters and human poses, which restricts its generalization capacity (see Section 4 for more detail). As described in the previous subsection, cross-domain persons share similar characteristics at different granularity levels. From a feature distribution perspective, persons from different domains should be closer together in the embedding space if they have similar appearance (and, thus, similar visual features). Inspired by this observation, three independent UDA constraints are introduced to align feature distribution at multiple granularity levels (i.e. global and local), with the aim of transferring the learned view-invariant representations and eliminating domain-specific features comprehensively. More specifically, the training in source identity classification acts as a builder that initializes and maintains the embedding space, while the UDA constraint acts as a regulator that refines the space.

For each branch B_b ($b = 0, 1, 2$) as shown in Fig. 2, an independent UDA constraint processes every output feature representation (i.e. f_b^s and f_b^t , extracted from the source and target domains) to minimize the feature distribution distance between the source and target domains, as follows:

$$\mathcal{L}_b^{\text{UDA}} = \mathcal{D}(\mathbb{P}(f_b^s), \mathbb{P}(f_b^t)) \quad (9)$$

where $\mathbb{P}(f_b^s)$ and $\mathbb{P}(f_b^t)$ denote the feature distributions in the source and target domains, respectively. $\mathcal{D}(\cdot, \cdot)$ indicates the distance metric between the two distributions, and we adopt the Maximum Mean Discrepancy (MMD) [62] in implementation. More specifically, MMD defines the difference between two feature distributions based on their mean embedding in the reproducing kernel Hilbert space (RKHS):

$$\mathcal{L}_b^{\text{UDA}} = \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \phi(f_{i,b}^s) - \frac{1}{n^t} \sum_{j=1}^{n^t} \phi(f_{j,b}^t) \right\|_{\mathcal{H}}^2 \quad (10)$$

where n^s and n^t denote the mini-batch source and target data sampled from D^s and D^t , $f_{i,b}^s$ and $f_{j,b}^t$ denote the feature representation extracted by the b -th branch from the i -th source sample and j -th target sample, and ϕ denotes a mapping operation that maps a feature distribution into the RKHS, where arbitrary probability distribution is uniquely represented by an element.

In practice, MMD is estimated via empirical kernel mean embedding, which modifies the distance $\mathcal{L}_b^{\text{UDA}}$ as follows:

$$\begin{aligned}\mathcal{L}_b^{\text{UDA}} = & \frac{1}{(n^s)^2} \sum_{i=1}^{n^s} \sum_{i'=1}^{n^s} \mathcal{K}(f_{i,b}^s, f_{i',b}^s) \\ & + \frac{1}{(n^t)^2} \sum_{j=1}^{n^t} \sum_{j'=1}^{n^t} \mathcal{K}(f_{j,b}^t, f_{j',b}^t) \\ & - \frac{2}{n^s n^t} \sum_{i=1}^{n^s} \sum_{j=1}^{n^t} \mathcal{K}(f_{i,b}^s, f_{j,b}^t)\end{aligned}\quad (11)$$

where \mathcal{K} denotes the selected kernel function; moreover, we use the RBF characteristic kernel as follows:

$$\mathcal{K}(f_{i,b}^s, f_{j,b}^t) = \exp\left(-\frac{1}{\alpha}\|f_{i,b}^s - f_{j,b}^t\|^2\right) \quad (12)$$

where α is the bandwidth that is empirically set to the dimension of feature representation.

For each branch B_b ($b = 0, 1, 2$), the objective of feature adaptation is to minimize the distance $\mathcal{L}_b^{\text{UDA}}$ between the feature distributions of the source and target domains. The overall feature adaptation loss can thus be determined as follows:

$$\mathcal{L}^{\text{UDA}} = L_0^{\text{UDA}} + \lambda \sum_{b=1}^2 L_b^{\text{UDA}} \quad (13)$$

where λ is the same as in Eq. (8).

2) *Feature Norm Adaptation*: The smaller-norm-less-informative assumption is that parameters or features with smaller norms play a less informative role during inference [72]. This assumption has been widely accepted in the statistics and machine learning communities and become an important measurement of information. In our proposed ReID model, the feature extraction branches attempt to exploit features from the whole body and the body parts in a supervised manner. After being regularized by the UDA constraint, the feature alignment module can be interpreted as a process of selecting the domain-invariant features across domains. To investigate the ways in which the UDA constraint increases feature representation information for the target domain, we compute target-domain L_2 -norm histograms of global and local feature representations with and without UDA constraint. Fig. 5 presents the histograms, from which three observations can be derived. First, global features from whole-body images have higher norms than local features from any individual body-parts, as they contain more informative cues. Second, local features tend to aggregate in higher norm regions following feature alignment, demonstrating that the UDA constraint helps to exploit more discriminative and domain-invariant local features. Third, global features tend to aggregate in lower norm regions with feature alignment, which makes them less

informative and leads to sub-optimal ReID matching. The intuition behind this phenomenon is that the global branch may pay attention to local regions and thus fail to fully explore global features when regularized with the UDA constraint.

Based on the above observations, we design an adaptive feature norm constraint to bridge the gap between the original feature norm and aligned feature norm, and increase the possibility of global features being transferred, as follows:

$$\mathcal{L}^{\text{Norm}} = \left(\frac{1}{n^s} \sum_{i=1}^{n^s} \|f_{i,0}^s\|_2 - R\right)^2 + \left(\frac{1}{n^t} \sum_{j=1}^{n^t} \|f_{j,0}^t\|_2 - R\right)^2 \quad (14)$$

where $\|\cdot\|_2$ denotes the L_2 -norm of a feature vector, which can be interpreted as the radius from the hypersphere origin to the vector point, while R is the mean feature norm of the direct transferred ReID model. The convergence of the mean feature norms of the adapted ReID model to R results in a much higher proportion of information being retained following alignment with the target domain.

Algorithm 1 PPAN Learning Algorithm

Input:

Source domain $D^s = \{I_i^s, c_i^s\}_{i=1}^{N^s}$, Target domain $D^t = \{I_j^t\}_{j=1}^{N^t}$, Multi-branch feature extraction network G .

Output:

Optimized G

- 1: **Stage-1:**
 - 2: $m \leftarrow 0$
 - 3: Train G on $D^s = \{I_i^s, c_i^s\}_{i=1}^{N^s}$ using Eq. (8).
 - 4: **Stage-2:**
 - 5: $\hat{D}^s \leftarrow \emptyset$
 - 6: $m \leftarrow \text{pre_trained_epochs}$
 - 7: **while** not converge **do**
 - 8: Run the PAS based on G , output: \hat{D}^s
 - 9: **for** $e = 1$ **to** max_iter **do**
 - 10: Randomly sample $B^s = \{(I_i^s, c_i^s)\}_{i=1}^{n^s}$, $B^t = \{x_j^t\}_{j=1}^{n^t}$ from \hat{D}^s and D^t respectively
 - 11: Train G on B^s and B^t using Eq. (17).
 - 12: **end for**
 - 13: $\hat{D}^s \leftarrow \emptyset$
 - 14: $m \leftarrow m + 1$
 - 15: **end while**
 - 16: **return** G
-

D. Progressive Adaptation Strategy

In unsupervised domain-adaptive person ReID, the persons in the target domain are completely different from those in the source domain. Matching the whole source-domain distribution and target-domain distribution will thus lead to negative transfer due to the existence of outlier source identities. To alleviate the negative transfer, outlier source identities and the associated source samples therefore need to be identified while performing domain adaptation. This is not an easy task, as the target and source domains have different label spaces. Fortunately, the shared embedding space characterizes the similarity between target sample and each source identity.

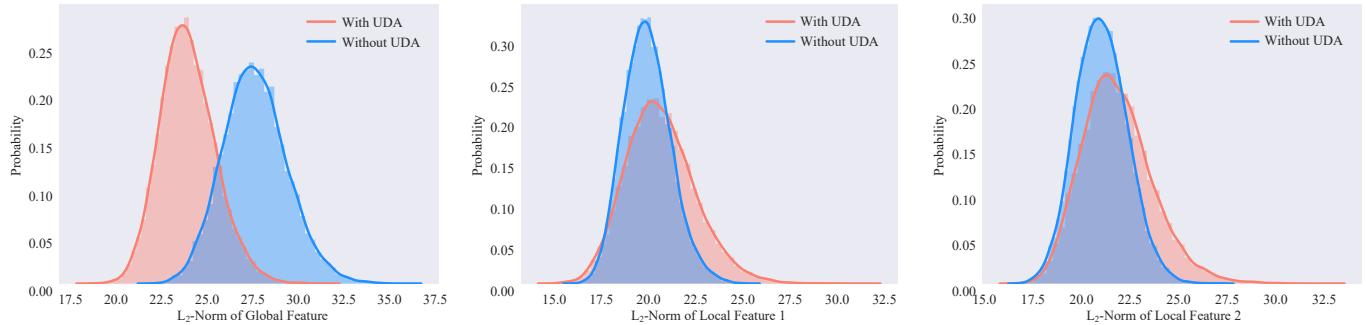


Fig. 5. The L_2 -norm histograms of global and local feature representations extracted from the target domain with and without UDA

We design a PAS that selects source identities progressively and adaptively depending on the cross-domain similarity. For each identity in the source domain, the PAS first computes a prototype $E_q^s (q = 1, \dots, Q)$ using the mean vector of the embedded source samples associated with each identity, as follows:

$$f_i^s = [f_{i,0}^s, f_{i,1}^s, f_{i,2}^s] \\ E_q^s = \frac{1}{N_q^s} \sum_{i \in D_q^s} f_i^s \quad (15)$$

where D_q^s denotes the set of samples labeled with identity class q in the source domain, while N_q^s is the number of corresponding samples; moreover, $f_{i,b}^s (b = 1, 2, 3)$ denotes the feature embedding (defined in Eq. (6)) of the i -th sample in D_q^s and $[,]$ denotes concatenation. A set of prototypes $\{E_q^s\}_{q=1}^Q$ can thus be obtained for Q distinct persons in the source domain. As the embedded target samples should be located close to the source prototypes in the latent feature space, the similarity between the j -th unlabeled target sample and the Q prototypes can thus be measured by using cosine distance, as follows:

$$f_j^t = [f_{j,0}^t, f_{j,1}^t, f_{j,2}^t] \\ d_j^q = \mathcal{C}(f_j^t, E_q^s) \quad q = 1, 2, \dots, Q \quad (16)$$

where f_j^t denotes the feature embedding of the j -th target sample and $\mathcal{C}(,)$ is the cosine distance between two feature representations. The Q distinct source identities are then ranked according to d_j^q , after which the top V identities are added into the selected training dataset \hat{D}^s .

E. Joint Loss Function and Optimization

The overall objective function of the proposed PPAN can be derived by combining the source classification loss $\mathcal{L}^{\text{Class}}$ and the learning constraints \mathcal{L}^{UDA} and $\mathcal{L}^{\text{Norm}}$, as follows:

$$\mathcal{L} = \mathcal{L}^{\text{Class}} + \beta_1 \mathcal{L}^{\text{UDA}} + \beta_2 \mathcal{L}^{\text{Norm}} \quad (17)$$

where β_1 and β_2 control the relative weights of the classification loss of the UDA constraints and feature norm constraints. The PPAN thus enables end-to-end training with the Stochastic Gradient Descent algorithm, where the PAN and the PAS learn alternatively and iteratively, as described in Algorithm 1.

TABLE I
SETTINGS OF THE FIVE PERSON REID DATASETS USED IN OUR EXPERIMENTS

| Dataset | Cams | IDs | Train IDs | Test IDs | Images |
|---------------|------|-------|-----------|----------|--------|
| Market-1501 | 6 | 1,501 | 751 | 750 | 32,668 |
| DukeMTMC-reID | 8 | 1,404 | 702 | 702 | 36,411 |
| CUHK03 | 6 | 1,467 | 1,367 | 100 | 14,097 |
| VIPeR | 2 | 632 | 316 | 316 | 1,264 |
| PRID | 2 | 934 | 100 | 649 | 1,134 |

IV. EXPERIMENTS

A. Datasets and Settings

1) *Datasets*: Our proposed PPAN is evaluated on five widely used benchmarking datasets: namely, Market-1501 [14], DukeMTMC-reID [15], CUHK03 [16], VIPeR [77] and PRID [78] (see Table I). Market-1501 was collected using six cameras, and consists of 32,668 images of 1,501 identities, as generated by an automatic detector. We follow the training and evaluation protocol in [14], which splits images into a training set (of 12,936 images) and a testing set (19,732 images). DukeMTMC-reID is a subset of the DukeMTMC dataset that was collected using eight cameras for the study of cross-camera tracking. It consists of images of 1,404 identities, half of which are used for training while the other half are used for testing; specifically, there are 2,228 queries, 17,661 galleries, and 16,522 training images, respectively. We follow the protocol outlined in [15] for evaluations on this dataset. Moreover, CUHK03 consists of 14,097 images of 1,467 different identities, where six campus cameras were deployed for image collection and each identity has been captured by two campus cameras. This dataset provides two types of annotations: one employing manually labeled bounding boxes and the other using bounding boxes produced by an automatic detector. We adopted a challenging setup with detected person images, as reported in [16], where the training and testing sets consist of 1,367 identities and 100 identities respectively. The VIPeR [77] dataset was collected using two cameras and consists of 632 identities, each with two images. We follow the training and evaluation protocol in [77] that randomly splits the entire population into two halves to create the training/test sets. The PRID [78] was collected from two camera views, A and B , each capturing 385 and 749 identities respectively; however, only 200 identities appear in both views. The single-shot version of the dataset is used in our experiments, as in

[49]. In each data split, 100 identities (with one image from each view) are randomly chosen from the 200 present in both camera views to form the training set, while the remaining 100 of View *A* are used as the query set and the remaining 649 of View *B* are used as the gallery set. For the VIPeR and PRID datasets, we repeat the random split 10 times and report the average result.

2) *Evaluation Protocol*: The performance of person ReID is evaluated using the widely used cumulative matching characteristic (CMC) [79] across all five datasets. CMC is a widely used evaluation metric in person ReID studies. Let us take the single-gallery-shot setting (each gallery identity has only one instance) as an example: for each query, all gallery samples are ranked according to their distances from the query, and the CMC top-k accuracy is evaluated as follows

$$Acc_k = \begin{cases} 1 & \text{if top-}k \text{ ranked gallery samples} \\ & \text{contain the query identity} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

This is in fact a shifted step function, where the CMC curve can be derived by averaging the shifted step functions over all queries. Due to space limitations, and also to facilitate direct comparison with state-of-the-art methods, we only report the CMC accuracy at selected ranks. The mean Average Precision (mAP) [14] is also reported for Market-1501 and DukeMTMC-reID. For CUHK03, VIPeR and PRID, the mAP is not reported [16, 49, 50] because the gallery of the three datasets contains only one image for each identity. All experiments adopt the single-query evaluation mode, and no re-ranking is performed for any of the evaluated methods.

3) *Implementation Details*: The proposed PPAN is implemented on the PyTorch framework. It is first pre-trained on ImageNet [76], then fine-tuned on a pair of ReID datasets; here, one dataset is used as the source (with identity labels) and the other used as the target (without labels). The VIPeR and PRID datasets are used only as the target, as they are too small for deep network training. We use the Stochastic Gradient Descent (SGD) optimizer with a batch size of 32 for both the source and target domain images, and set the Nesterov momentum to 0.999 with weight decay of 5×10^{-4} . All training and test images are rescaled to 384×192 and each of the two body parts has a fixed size of 192×192 . Each training image is first normalized by subtracting its channel means, then fed to the network for training in a random order. The network is trained on the source domain for 10 epochs using the classification loss in Eq. (8) ($\lambda = 0.5$) in *Stage-1*, then trained on both source and target domains for 90 epochs with joint loss in Eq. 17 ($\lambda = 0.8$, $\beta_1 = 1$ and $\beta_2 = 0.001$) in *Stage-2*. The learning rate is initially set to 0.01 and then divided by 10 after every 30 epochs.

B. Comparison with State-of-the-Art Methods

The proposed PPAN is evaluated and compared with state-of-the-art methods on five widely used datasets: Market-1501, DukeMTMC-reID, CUHK03, VIPeR and PRID.

For Market-1501, PPAN is compared with 12 state-of-the-art methods, comprising three unsupervised methods and nine

TABLE II
COMPARISON WITH STATE-OF-THE-ART UNSUPERVISED METHODS ON THE MARKET-1501 DATASET: ‘-’ DENOTES ‘NO RESULTS REPORTED’, ‘SRC’ DENOTES ‘SOURCE DOMAIN’, ‘**’ DENOTES ‘NO AUXILIARY SOURCE DOMAIN USED’, AND ‘†’ DENOTES ‘MULTIPLE SOURCE DOMAINS USED’.

| Methods | Src | R1 | R5 | R10 | R20 | mAP |
|---------------|--------|--------------|--------------|--------------|--------------|--------------|
| ISR [35] | * | 40.30 | - | - | - | 14.30 |
| UNCA [80] | * | 45.20 | - | - | - | 18.90 |
| CAMEL [38] | * | 54.50 | - | - | - | 26.30 |
| SSDAL [48] | † | 39.40 | - | - | - | 19.60 |
| PTGAN [51] | CUHK03 | 31.50 | - | 60.20 | - | - |
| PTGAN [51] | Duke | 38.60 | - | 66.10 | - | - |
| PUL [43] | Duke | 44.70 | 59.10 | 65.60 | 71.70 | 20.10 |
| MMFA [50] | Duke | 56.70 | 75.00 | 81.80 | - | 27.40 |
| SPGAN [52] | Duke | 51.50 | 70.10 | 76.80 | 82.40 | 22.80 |
| TJ-AIDL [49] | Duke | 58.20 | 74.80 | 81.10 | 86.50 | 26.50 |
| CamStyle [53] | Duke | 58.80 | 78.20 | 84.30 | 88.80 | 27.40 |
| Tfusion [59] | CUHK01 | 60.75 | 74.44 | 79.25 | - | - |
| CFSM [57] | Duke | 61.20 | - | - | - | 28.30 |
| PPAN | Duke | 62.71 | 77.29 | 82.54 | 87.08 | 30.20 |
| PPAN | CUHK03 | 63.69 | 77.34 | 83.50 | 87.93 | 32.84 |

TABLE III
COMPARISON WITH STATE-OF-THE-ART UNSUPERVISED METHODS ON THE DUKEMTMC-REID DATASET: ‘-’ DENOTES ‘NO RESULTS REPORTED’, ‘SRC’ DENOTES ‘SOURCE DOMAIN’.

| Methods | Src | R1 | R5 | R10 | R20 | mAP |
|---------------|--------|--------------|--------------|--------------|--------------|--------------|
| PTGAN [51] | CUHK03 | 27.40 | - | 50.70 | - | - |
| PTGAN [51] | Market | 27.40 | - | 50.70 | - | - |
| PN-GAN [74] | Market | 29.94 | 51.62 | 50.70 | - | 15.77 |
| PUL [43] | Market | 30.40 | 44.50 | 50.70 | 56.00 | 16.40 |
| SPGAN [52] | Market | 41.10 | 56.60 | 63.00 | 69.60 | 22.30 |
| MMFA [50] | Market | 45.30 | 59.80 | 66.30 | - | 24.70 |
| TJ-AIDL [49] | Market | 44.30 | 59.60 | 65.00 | 70.00 | 23.00 |
| CamStyle [53] | Market | 48.40 | 62.50 | 68.90 | 74.40 | 25.10 |
| CFSM [57] | Market | 49.80 | - | - | - | 27.30 |
| PPAN | CUHK03 | 52.24 | 64.90 | 69.75 | 73.92 | 30.54 |
| PPAN | Market | 55.66 | 68.17 | 73.16 | 77.38 | 34.01 |

UDA methods (leverage labeled source dataset), as shown in Table II. It can be observed that PPAN achieves superior ReID accuracy and outperforms state-of-the-art methods by 2.49% in Rank-1 (63.69% versus 61.20%) and 4.54% in mAP (32.84% versus 28.30%). More specifically, PPAN outperforms CAMEL (without using a labeled source dataset) by 9.19% in Rank-1 and 6.54% in mAP. The experimental results clearly demonstrate the importance of leveraging a related label-rich source domain in feature learning. In addition, PPAN improves Rank-1 by 5.49% and 6.99% and mAP by 6.34% and 5.44% compared with TJ-AIDL and MMFA (which align global features only). Note that TJ-AIDL and MMFA exploit attribute labels to help adaptation, whereas PPAN automatically discovers latent multi-granular characteristics with no additional annotation. These evident performance gains can largely be attributed to our comprehensive feature distribution alignment and progressive adaptation, which help to alleviate the negative transfer problem.

On the larger and more recent DukeMTMC-reID, PPAN is compared with eight state-of-the-art methods. As shown in Table III, PPAN obtains superior accuracy on this dataset, outperforming the state-of-the-art approaches by 5.86% in Rank-1 (55.66% versus 49.80% by CFSM) and 6.71% in mAP (34.01% versus 27.30%), respectively. This further verifies the advantages of PPAN due to its utilization of PAN and PAS to guide feature learning and feature adaptation simultaneously. Note that lower accuracy is obtained on DukeMTMC-reID

TABLE IV
COMPARISON WITH STATE-OF-THE-ART UNSUPERVISED METHODS ON THE CUHK03 DATASET: ‘-’ DENOTES ‘NO RESULTS REPORTED’, ‘SRC’ DENOTES ‘SOURCE DOMAIN’, ‘*’ DENOTES ‘NO AUXILIARY SOURCE DOMAIN USED’.

| Methods | Src | R1 | R5 | R10 | R20 |
|--------------|--------|--------------|--------------|--------------|--------------|
| GL [36] | * | 30.40 | - | - | - |
| CAMEL [38] | * | 31.90 | - | - | - |
| DECAMEL [40] | * | 38.27 | 51.62 | - | - |
| PN-GAN [74] | Market | 16.85 | 39.05 | 53.32 | - |
| PTGAN [51] | Duke | 24.80 | - | 66.20 | - |
| PTGAN [51] | Market | 26.90 | - | 72.10 | - |
| PPAN | Duke | 42.01 | 64.95 | 75.61 | 85.04 |
| PPAN | Market | 46.08 | 70.44 | 80.43 | 88.58 |

TABLE V
COMPARISON WITH STATE-OF-THE-ART UNSUPERVISED METHODS ON THE VIPER AND PRID DATASETS: ‘-’ DENOTES ‘NO RESULTS REPORTED’, ‘SRC’ DENOTES ‘SOURCE DOMAIN’, ‘*’ DENOTES ‘NO AUXILIARY SOURCE DOMAIN USED’, AND ‘†’ DENOTES ‘MULTIPLE SOURCE DOMAINS USED’.

| Methods | Src | VIPeR | PRID |
|--------------|--------|--------------|--------------|
| | | R1 | R1 |
| SDALF [29] | * | 19.90 | 16.30 |
| GTS [33] | * | 25.15 | - |
| SDC [37] | * | 25.76 | - |
| ISR [35] | * | 27.00 | 17.00 |
| DLLR [34] | * | 29.60 | 21.10 |
| GL [36] | * | 33.50 | 25.00 |
| UsNCA [80] | * | 24.30 | - |
| CAMEL [38] | * | 30.90 | - |
| DECAMEL [40] | * | 34.15 | - |
| UDML [58] | † | 31.50 | - |
| JSLAM [47] | † | 34.60 | - |
| SSDAL [48] | † | 37.90 | 20.10 |
| TJ-AIDL [49] | Duke | 35.10 | 34.80 |
| TJ-AIDL [49] | Market | 38.50 | 26.80 |
| MMFA [50] | Duke | 36.30 | 34.50 |
| MMFA [50] | Market | 39.10 | 35.10 |
| PPAN | Duke | 40.82 | 39.10 |
| PPAN | Market | 43.40 | 41.50 |

than Market-1501, as images in the former contain more variations in background and scene layout. However, after taking a second look, it can be observed that PPAN performs more consistently following transfer from Market-1501 to DukeMTMC-reID, whereas most state-of-the-art methods undergo much larger performance drops while adapting to a more difficult dataset.

For CUHK03, PPAN is compared with five state-of-the-art methods, including three unsupervised methods and two UDA methods (see Table IV). PPAN achieves superior ReID accuracy when compared to all five state-of-the-art methods. Specifically, it outperforms state-of-the-art unsupervised methods by 7.81% in Rank-1 (46.08% versus 38.27%), as well as state-of-the-art UDA methods by 19.18% in Rank-1 (for adapting from Market to CUHK03) and 17.21% (adapting from Duke to CUHK03). These outstanding performance results demonstrate the superiority and robustness of our proposed method.

For VIPeR and PRID, PPAN is compared with 14 state-of-the-art methods, including nine unsupervised methods and five UDA methods (see Table V). For VIPeR and PRID, PPAN outperforms the comparison methods by 4.30% (43.40% versus 39.10% by MMFA) and 6.40% (41.50% versus 35.10% by MMFA) in Rank-1, respectively. This further demonstrates the superior performance of our PPAN. Note that PPAN performs

more consistently for transfer from large-scale to small-scale domains. As Table V shows, PPAN experiences a 1.90% performance drop in Rank-1 while transferring from Market-1501 to VIPeR compared with transferring from Market-1501 to PRID. By contrast, state-of-the-art methods experience much larger performance drops (e.g. 39.10% versus 35.10% by MMFA, 38.50% versus 26.80% by TJ-AIDL, etc).

C. Ablation Study

Our proposed method learns discriminative features by using both global whole-body images and local body-part images. To facilitate domain adaptation to the unlabeled target domain, part-aware adaptation is employed in order to align the global and local feature distributions. Furthermore, a Progressive Adaptation Strategy (PAS) is designed that eliminates outlier source identities according to cross-domain similarity measures. To determine how each of these innovative components help to achieve the outstanding person ReID performance (see Tables II-V), we develop five models for ablation analysis, as follows: 1) a **Global** model that implements the base global branch network; 2) a **Global+Local** model that implements the global and local branch networks; 3) a **Global+Local+UDA** model that includes the part-aware adaptation beyond the **Global+Local** model; 4) a **Global+Local+PAS** model that includes the progressive adaptation strategy beyond the **Global+Local** model; and 5) a **Global+Local+UDA+PAS** model that includes the PAS beyond the **Global+Local+UDA** model.

Table VI presents results for the performance of each of the five models on the two datasets. As can be seen from the table, the inclusion of local-branch networks, part-aware adaptation and progressive adaptation strategy all clearly help to improve the person ReID performance. More specifically, **Global+Local** consistently outperform **Global**, which is largely due to the incorporation of the complementary local features. In addition, both **Global+Local+UDA** and **Global+Local+PAS** consistently outperform **Global+Local** across the two datasets, demonstrating the effectiveness of using the part-aware adaptation and progressive adaptation strategies in unsupervised domain adaptive ReID. Moreover, **Global+Local+UDA+PAS** outperforms either part-aware adaptation or progressive adaptation strategy alone, demonstrating the complementary nature of the two proposed adaptation schemes.

D. Discussion

In addition to the ablation study, we also study five factors that could potentially affect the ReID: namely, the contributions of individual human regions with and without using the proposed UDA and PAS, the feature norm constraint of the global-branch network, the pooling strategies, and the hyper-parameters.

1) *UDA Constraint Analysis:* We evaluate the ways in which the global and local features affect the ReID with and without using UDA. Table VII presents the experimental results on DukeMTMC-reID. As can be seen from the table,

TABLE VI

ABLATION STUDY ON MARKET-1501 AND DUKEMTMC-REID DATASETS. * HIGHLIGHTS THE METHODS THAT DIRECTLY APPLY THE SOURCE-TRAINED MODEL TO THE TARGET DOMAIN.

| Models | Market-1501→DukeMTMC-reID | | | | | DukeMTMC-reID→Market-1501 | | | | |
|----------------------|---------------------------|--------------|--------------|--------------|--------------|---------------------------|--------------|--------------|--------------|--------------|
| | R1 | R5 | R10 | R20 | mAP | R1 | R5 | R10 | R20 | mAP |
| *Global | 43.63 | 58.3 | 64.18 | 69.16 | 24.33 | 53.56 | 70.48 | 75.83 | 82.09 | 23.88 |
| *Global+Local | 48.20 | 63.89 | 68.46 | 72.31 | 28.15 | 58.25 | 73.57 | 79.27 | 84.02 | 25.76 |
| Global+Local+UDA | 53.30 | 67.14 | 72.35 | 76.88 | 32.93 | 61.31 | 75.53 | 81.47 | 86.81 | 28.56 |
| Global+Local+PAS | 50.63 | 64.49 | 69.29 | 73.74 | 30.54 | 59.97 | 74.70 | 79.93 | 85.06 | 27.58 |
| Global+Local+UDA+PAS | 55.66 | 68.17 | 73.16 | 77.38 | 34.01 | 62.71 | 77.29 | 82.54 | 87.08 | 30.20 |

TABLE VII

EVALUATIONS OF HOW INFORMATION FROM DIFFERENT HUMAN PARTS CONTRIBUTES TO PERSON REID WITH AND WITHOUT THE USE OF UNSUPERVISED DOMAIN ADAPTATION (UDA) CONSTRAINTS (MARKET-1501→DUKEMTMC-REID).

| Models | Without UDA constraint | | | | | With UDA constraint | | | | |
|-------------------|------------------------|--------------|--------------|--------------|--------------|---------------------|--------------|--------------|--------------|--------------|
| | R1 | R5 | R10 | R20 | mAP | R1 | R5 | R10 | R20 | mAP |
| Global | 43.63 | 58.30 | 64.18 | 69.16 | 24.33 | 49.65 | 64.54 | 69.61 | 73.65 | 29.56 |
| Upper-body | 38.86 | 53.37 | 59.96 | 65.66 | 19.73 | 42.32 | 55.52 | 61.98 | 67.36 | 21.29 |
| Lower-body | 18.81 | 31.55 | 37.31 | 43.98 | 8.80 | 23.25 | 36.59 | 42.95 | 49.51 | 11.30 |
| Part Fusion | 42.37 | 56.59 | 63.18 | 68.85 | 23.77 | 50.40 | 64.54 | 68.48 | 73.83 | 28.42 |
| Global+Upper-body | 47.26 | 61.80 | 67.63 | 71.45 | 27.14 | 52.78 | 65.66 | 70.91 | 75.13 | 31.92 |
| Global+Lower-body | 44.17 | 58.63 | 64.60 | 69.49 | 24.66 | 50.52 | 64.90 | 69.34 | 74.32 | 30.04 |
| Global+Part | 48.20 | 63.89 | 68.46 | 72.31 | 28.15 | 53.30 | 67.14 | 72.35 | 76.88 | 32.93 |

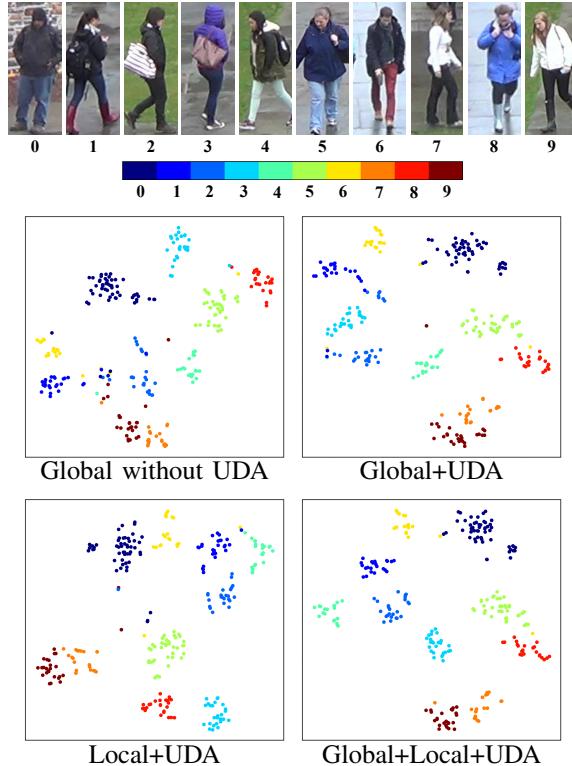


Fig. 6. Feature distributions of 10 random test identities in four transferred feature spaces (Market-1501 → DukeMTMC-ReID) as visualized by t-SNE with color-coded identity classes

the ReID performance is greatly improved for both global branch and local-branch networks when UDA is included; ‘Global’ outperforms both ‘Upper-body’ and ‘Lower-body’, as it captures more comprehensive information from the entire body. In addition, ‘Upper-body’ clearly outperforms ‘Lower-body’, and the fusion of the two parts outperforms the use of either part alone. Furthermore, the inclusion of ‘Global’ clearly improves ReID for ‘Upper-body’, ‘Lower-body’ and

‘Part Fusion’, demonstrating the complementary nature of global and local features in person ReID.

Fig. 6 illustrates the importance of using the part-aware adaptation network to capture the complementary domain-invariant global and local features. In the figure, we plot four feature distributions of 10 randomly selected test identities of DukeMTMC-ReID (transferred from Market-1501). As the figure shows, adapting either global or local features alone does not lead to compact per-identity clusters, although adapting local features forms more compact clusters than adapting global features. By comparison, adapting both global and local features assists in the formation of far more compact and well-separated feature clusters, demonstrating the necessity and effectiveness of aligning global and local feature distributions simultaneously.

TABLE VIII

EVALUATIONS OF HOW GLOBAL FEATURES CONTRIBUTE TO PERSON REID WITH AND WITHOUT THE USE OF UDA AND FN CONSTRAINT (MARKET-1501→DUKEMTMC-REID): ‘G’ DENOTES ‘GLOBAL BRANCH WITHOUT UDA’; ‘GU’ DENOTES ‘GLOBAL BRANCH WITH UDA’; ‘PU’ DENOTES ‘LOCAL BRANCHES WITH UDA’.

| Models | R1 | R5 | R10 | R20 | mAP |
|----------|--------------|--------------|--------------|--------------|--------------|
| G+PU | 51.91 | 66.05 | 71.54 | 75.18 | 31.95 |
| GU | 47.76 | 63.38 | 67.68 | 72.84 | 27.68 |
| GU+FN | 49.65 | 64.54 | 69.61 | 73.65 | 29.56 |
| GU+PU | 52.42 | 66.33 | 71.63 | 75.41 | 32.09 |
| GU+PU+FN | 53.30 | 67.14 | 72.35 | 76.88 | 32.93 |

2) *Feature Norm Constraint Analysis:* When the local branches employ UDA, the inclusion of UDA in the global branch has no evident impact on the ReID, as shown in Table VIII. One major reason for this is that a global branch with UDA learns a large number of redundant local features rather than global domain-invariant features, since persons from different domains are more likely to share local characteristics. We therefore design a Feature Norm (FN) constraint for global feature distribution alignment. As can be seen from Table VIII, FN consistently improves the ReID performance when it works with either the global branch alone (with UDA) or

TABLE IX
EVALUATIONS OF HOW INFORMATION FROM DIFFERENT HUMAN PARTS CONTRIBUTES TO PERSON REID WITH AND WITHOUT THE USE OF PROGRESSIVE ADAPTATION STRATEGY (PAS) (MARKET-1501→DUKEMTMC-REID).

| Models | Without PAS | | | | | With PAS | | | | |
|-------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | R1 | R5 | R10 | R20 | mAP | R1 | R5 | R10 | R20 | mAP |
| Global | 43.63 | 58.30 | 64.18 | 69.16 | 24.33 | 46.05 | 60.68 | 65.97 | 70.73 | 26.66 |
| Upper-body | 38.86 | 53.37 | 59.96 | 65.66 | 19.73 | 40.22 | 55.83 | 62.07 | 67.24 | 20.21 |
| Lower-body | 18.81 | 31.55 | 37.31 | 43.98 | 8.80 | 21.77 | 35.36 | 41.61 | 47.98 | 10.07 |
| Part Fusion | 42.37 | 56.59 | 63.18 | 68.85 | 23.77 | 46.14 | 61.18 | 67.19 | 71.90 | 25.79 |
| Global+Upper-body | 47.26 | 61.80 | 67.63 | 71.45 | 27.14 | 49.28 | 63.59 | 68.26 | 72.53 | 29.39 |
| Global+Lower-body | 44.17 | 58.63 | 64.60 | 69.49 | 24.66 | 46.83 | 61.25 | 67.17 | 72.11 | 27.08 |
| Global+Part | 48.20 | 63.89 | 68.46 | 72.31 | 28.15 | 50.63 | 64.49 | 69.29 | 73.74 | 30.54 |

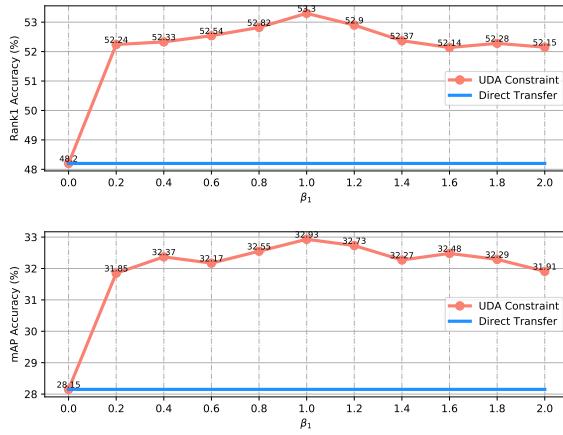


Fig. 7. Evaluation on the influence of parameter β_1 (Market-1501→DukeMTMC-reID) in person ReID with and without UDA

the global branch plus local branches (both with UDA). These experimental results allow us to conclude that FN effectively guides the global branch to learn domain-invariant features from the whole body effectively.

3) *PAS Analysis*: We investigate how the proposed Progressive Adaptation Strategy (PAS) affects the ReID performance, in cases where the models are trained on the source domain with and without PAS and directly applied on the target domain. As Table IX shows, PAS clearly helps with improving the ReID performance for both global-branch and local-branch networks. In addition, including PAS beyond **Global+Local+UDA** greatly improves the ReID, as shown in Table VI. This validates the rationale of our PAS design, in that selecting positive source identities is beneficial to unsupervised domain adaptive ReID.

TABLE X
EVALUATION OF THE INFLUENCE OF POOLING STRATEGIES (MARKET-1501→DUKEMTMC-REID).

| Models | R1 | R5 | R10 | R20 | mAP |
|-------------------|--------------|--------------|--------------|--------------|--------------|
| Baseline with GAP | 40.31 | 54.71 | 60.37 | 65.89 | 22.32 |
| Baseline with GMP | 48.20 | 63.89 | 68.46 | 72.31 | 28.15 |
| PPAN with GAP | 46.32 | 61.36 | 66.20 | 71.32 | 26.93 |
| PPAN with GMP | 55.66 | 68.17 | 73.16 | 77.38 | 34.01 |

4) *Pooling Strategies*: We also evaluate the impact of two pooling strategies: namely, traditional Global Average Pooling (GAP) and our adopted Global Max Pooling (GMP). While GAP achieved great success in the classification task, it could result in the loss of discriminative cues that are critical to

TABLE XI
EVALUATION OF THE INFLUENCE OF PARAMETER λ (MARKET-1501→DUKEMTMC-REID).

| λ | Direct transfer | | UDA constraint | | PAS | |
|-----------|-----------------|--------------|----------------|--------------|--------------|--------------|
| | R1 | mAP | R1 | mAP | R1 | mAP |
| 0 | 43.63 | 24.33 | 49.65 | 29.56 | 46.05 | 26.66 |
| 0.1 | 44.03 | 24.64 | 50.34 | 29.75 | 46.31 | 26.93 |
| 0.2 | 45.28 | 25.42 | 50.66 | 30.19 | 47.35 | 27.66 |
| 0.3 | 46.36 | 26.43 | 50.76 | 30.28 | 48.61 | 28.64 |
| 0.4 | 47.53 | 27.38 | 51.63 | 31.59 | 49.41 | 29.53 |
| 0.5 | 48.20 | 28.15 | 52.42 | 32.03 | 50.13 | 30.15 |
| 0.6 | 48.19 | 28.07 | 52.63 | 32.61 | 50.62 | 30.48 |
| 0.7 | 48.02 | 27.84 | 52.90 | 32.85 | 50.63 | 30.54 |
| 0.8 | 47.80 | 27.62 | 53.30 | 32.93 | 50.53 | 30.45 |
| 0.9 | 47.53 | 27.41 | 52.99 | 32.82 | 60.26 | 30.21 |
| 1.0 | 47.39 | 27.16 | 52.53 | 32.59 | 49.82 | 29.96 |

TABLE XII
EVALUATION OF THE INFLUENCE OF PARAMETER V (MARKET-1501→DUKEMTMC-REID).

| V | Selected IDs | R1 | R5 | R10 | R20 | mAP |
|-----|--------------|--------------|--------------|--------------|--------------|--------------|
| all | 751 | 48.2 | 63.89 | 68.46 | 72.31 | 28.15 |
| 10 | 668 | 48.56 | 62.38 | 67.91 | 71.90 | 28.76 |
| 9 | 644 | 49.01 | 63.64 | 68.12 | 72.98 | 29.17 |
| 8 | 629 | 49.41 | 62.83 | 68.71 | 73.15 | 29.29 |
| 7 | 625 | 49.55 | 63.86 | 68.35 | 72.84 | 29.80 |
| 6 | 615 | 50.09 | 64.36 | 69.12 | 73.47 | 29.82 |
| 5 | 604 | 50.63 | 64.49 | 69.29 | 73.74 | 30.54 |
| 4 | 576 | 49.64 | 63.37 | 68.53 | 73.29 | 29.56 |
| 3 | 527 | 49.41 | 62.79 | 67.90 | 72.71 | 29.33 |
| 2 | 491 | 48.74 | 62.51 | 67.76 | 72.66 | 28.74 |
| 1 | 411 | 46.36 | 61.75 | 67.10 | 71.36 | 27.34 |

the ReID. By contrast, unlike GAP, GMP is able to preserve the peak response from the human body region without being affected by the responses from the surrounding background. Experimental results in Table X show that GMP achieves far better performance for both baseline network and PPAN.

5) *Parameter Analysis*: We first evaluate the impact of λ (defined in Eq. (8) and (13)), which controls the relative weights of the global-branch and local-branch loss. As Table XI shows, a moderate λ helps to enhance the feature discriminability learned by the global-branch and local-branch networks. When the model is trained with only classification loss, the global and local features have comparable importance for the ReID matching; on the other hand, when classification loss is combined with UDA constraint or PAS, the local-branch networks fully explore domain-invariant features (as shown in Table VII and Table IX). A larger value of λ therefore aids the performance.

We also evaluate the impact of β_1 , which controls the weight of the UDA constraint. Fig. 7 presents the experimental results with different values of β_1 . It can be observed that

the proposed PAN achieves consistently better performance when the UDA constraint is included. The best person ReID performance is obtained when $\beta_1 = 1.0$.

In addition, we also evaluate the impact of V as defined in Section 3.4. Table XII shows experimental results where the second column lists the average number of source training identities when selecting the corresponding top V source identities for each target training sample. As can be seen from Table XII, the Rank-1 accuracy and mAP improve as V decreases, and the best results are achieved when V is set to 5 while applying PAS on source identities.

V. CONCLUSION

In this paper, we propose an end-to-end trainable network, PPAN, for unsupervised domain adaptive person ReID. Unlike most existing domain adaptive ReID methods, which either ignore the local feature distribution in feature alignment or match whole source and target distributions, the proposed part-aware domain adaptive network independently adapts global and local feature representations by incorporating UDA constraints and a feature norm constraint, as well as a three-branch CNN architecture. In addition, a novel progressive adaptation strategy is designed that progressively selects positive source identities in order to alleviate the negative transfer problem. Experiments on five widely used benchmarking datasets demonstrate that the proposed PPAN achieves superior person ReID performance compared to state-of-the-art approaches. Extensive ablation studies were also performed so as to provide more insight into our proposed network.

In our future work, we will continue to investigate fine-grained unsupervised domain adaptive ReID by incorporating trainable pose estimation. It may also be possible to study semantic human part parsing by incorporating our unique exploitation of multi-granular domain-invariant features with UDA constraints. Beyond that, we will also go on to explore body-part based sample selection schemes.

REFERENCES

- [1] N. O'Hare and A. F. Smeaton, "Context-aware person identification in personal photo collections," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 220–228, 2009.
- [2] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 260–272, 2016.
- [3] S. Sunderrajan and B. S. Manjunath, "Context-aware hypergraph modeling for re-identification and summarization," *IEEE Transactions on Multimedia*, vol. 18, no. 1, pp. 51–63, 2016.
- [4] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, and N. Zheng, "Large margin learning in set-to-set similarity comparison for person re-identification," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 593–604, 2018.
- [5] Z. Wang, J. Jiang, Y. Yu, and S. Satoh, "Incremental re-identification by cross-direction and cross-ranking adaption," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2376–2386, 2019.
- [6] G. Ding, S. Zhang, S. Khan, Z. Tang, J. Zhang, and F. Porikli, "Feature affinity based pseudo labeling for semi-supervised person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2891–2902, 2019.
- [7] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for scalable person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 986–999, 2019.
- [8] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, vol. 21, no. 6, pp. 1412–1424, 2019.
- [9] N. Liang, G. Wu, W. Kang, Z. Wang, and D. D. Feng, "Real-time long-term tracking with prediction-detection-correction," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2289–2302, 2018.
- [10] Z. Liu, Z. Lin, X. Wei, and S. Chan, "A new model-based method for multi-view human body tracking and its application to view transfer in image-based rendering," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1321–1334, 2018.
- [11] K. Lee and J. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1429–1438, 2015.
- [12] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2035–2045, 2013.
- [13] Z. Zhang, Y. Xie, W. Zhang, and Q. Tian, "Effective image retrieval via multilinear multi-index fusion," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2878–2890, 2019.
- [14] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, 2015, pp. 1116–1124.
- [15] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proc. ICCV*, 2017, pp. 3774–3782.
- [16] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, 2014, pp. 152–159.
- [17] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. CVPR*, 2016, pp. 1249–1258.
- [18] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. ICCV*, 2017, pp. 3980–3989.
- [19] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. CVPR*, 2017, pp. 1320–1329.
- [20] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. CVPR*, 2017, pp. 3356–3365.
- [21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proc. CVPR*, 2016, pp. 1335–1344.
- [22] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. IJCAI*, 2017, pp. 2194–2200.
- [23] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. CVPR*, 2017, pp. 907–915.
- [24] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. CVPR*, 2018, pp. 2285–2294.
- [25] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. CVPR*, 2018, pp. 5363–5372.
- [26] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.
- [27] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, 2015, pp. 2197–2206.
- [28] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, 2011, pp. 1521–1528.
- [29] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. CVPR*, 2010, pp. 2360–2367.
- [30] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Proc. CVPR*, 2016, pp. 1363–1372.
- [31] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6, pp. 379–390, 2014.
- [32] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. CVPR*, 2013, pp. 3586–3593.
- [33] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *Proc. BMVC*, 2014, pp. 3586–3593.
- [34] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification," in *Proc. BMVC*, 2015, pp. 44.1–44.12.
- [35] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Transac-*

- tions on Pattern Analysis and Machine Intelligence, vol. 37, no. 8, pp. 1629–1642, 2015.
- [36] E. Kadirov, T. Xiang, Z. Fu, and S. Gong, “Person re-identification by unsupervised 11 graph learning,” in Proc. ECCV, 2016, pp. 178–195.
- [37] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by saliency learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 2, pp. 356–370, 2017.
- [38] H. Yu, A. Wu, and W. Zheng, “Cross-view asymmetric metric learning for unsupervised person re-identification,” in Proc. ICCV, 2017, pp. 994–1002.
- [39] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, “Dynamic label graph matching for unsupervised video re-identification,” in Proc. ICCV, 2017, pp. 5152–5160.
- [40] H. Yu, A. Wu, and W. Zheng, “Unsupervised person re-identification by deep asymmetric metric embedding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 956–973, 2020.
- [41] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, “Dynamic graph co-matching for unsupervised video-based person re-identification,” IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2976–2990, 2019.
- [42] Z. Liu, D. Wang, and H. Lu, “Stepwise metric promotion for unsupervised video person re-identification,” in Proc. ICCV, 2017, pp. 2448–2457.
- [43] H. Fan, L. Zheng, C. Yan, and Y. Yang, “Unsupervised person re-identification: Clustering and fine-tuning,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 14, no. 4, pp. 83:1–83:18, 2018.
- [44] M. Wang and W. Deng, “Deep visual domain adaptation: A survey,” Neurocomputing, vol. 312, pp. 135–153, 2018.
- [45] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, “Deep domain confusion: Maximizing for domain invariance,” in CoRR, abs/1412.3474, 2014.
- [46] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” Journal of Machine Learning Research, vol. 17, no. 1, pp. 2096–2030, 2016.
- [47] P. Peng, Y. Tian, T. Xiang, Y. Wang, M. Pontil, and T. Huang, “Joint semantic and latent attribute modelling for cross-class transfer learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 7, pp. 1625–1638, 2018.
- [48] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Deep attributes driven multi-camera person re-identification,” in Proc. ECCV, 2016, pp. 475–491.
- [49] J. Wang, X. Zhu, S. Gong, and W. Li, “Transferable joint attribute-identity deep learning for unsupervised person re-identification,” in Proc. CVPR, 2018, pp. 2275–2284.
- [50] S. Lin, H. Li, C. Li, and A. C. Kot, “Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification,” in Proc. BMVC, 2018.
- [51] L. Wei, S. Zhang, W. Gao, and Q. Tian, “Person transfer gan to bridge domain gap for person re-identification,” in Proc. CVPR, 2018, pp. 79–88.
- [52] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, “Image-image domain adaptation with preserved selfsimilarity and domain-dissimilarity for person reidentification,” in Proc. CVPR, 2018, pp. 994–1003.
- [53] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, “Camstyle: A novel data augmentation method for person re-identification,” IEEE Transactions on Image Processing, vol. 28, no. 3, pp. 1176–1190, 2019.
- [54] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. ICCV, 2017, pp. 2242–2251.
- [55] F. Xiong, M. Gou, O. Camps, and M. Sznaier, “Person reidentification using kernel-based metric learning methods,” in Proc. ECCV, 2014, pp. 1–16.
- [56] Y. Zhang, Q. Zhong, L. Ma, D. Xie, and S. Pu, “Learning incremental triplet margin for person re-identification,” in Proc. AAAI, 2019, pp. 9243–9250.
- [57] X. Chang, Y. Yang, T. Xiang, and T. Hospedales, “Disjoint label space transfer learning with common factorised space,” in Proc. AAAI, 2019, pp. 3288–3295.
- [58] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, “Unsupervised cross-dataset transfer learning for person re-identification,” in Proc. CVPR, 2016, pp. 1306–1315.
- [59] J. Lv, W. Chen, Q. Li, and C. Yang, “Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns,” in Proc. CVPR, 2018, pp. 7948–7956.
- [60] Z. Zheng, L. Zheng, and Y. Yang, “A discriminatively learned cnn embedding for person re-identification,” ACM Multimedia Computing, Communications, and Applications, vol. 14, no. 1, 2017.
- [61] X. Qian, Y. Fu, Y. Jiang, T. Xiang, and X. Xue, “Multi-scale deep learning architectures for person re-identification,” in Proc. ICCV, 2017, pp. 5409–5418.
- [62] M. Long and J. Wang, “Learning transferable features with deep adaptation networks,” in Proc. ICML, 2015, pp. 97–105.
- [63] B. Sun, J. Feng, and K. Saenko, “Return of frustratingly easy domain adaptation,” in Proc. AAAI, 2016, pp. 2058–2065.
- [64] B. Sun and K. Saenko, “Deep coral: Correlation alignment for deep domain adaptation,” in Proc. ECCV, 2016, pp. 443–450.
- [65] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, “Deep reconstruction-classification networks for unsupervised domain adaptation,” in Proc. ECCV, 2016, pp. 597–613.
- [66] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, “Domain separation networks,” in Proc. NIPS, 2016, pp. 343–351.
- [67] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim, “Image to image translation for domain adaptation,” in Proc. CVPR, 2018, pp. 4500–4509.
- [68] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in Proc. ICML, 2015, p. 1180–1189.
- [69] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in Proc. CVPR, 2017, pp. 2962–2971.
- [70] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, “Unsupervised pixel-level domain adaptation with generative adversarial networks,” in Proc. CVPR, 2017, pp. 1063–6919.
- [71] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. NIPS, 2014, pp. 2672–2680.
- [72] J. Ye, X. Lu, Z. Lin, and J. Z. Wang, “Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers,” in Proc. ICLR, 2018.
- [73] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in Proc. CVPR, 2017, pp. 1302–1310.
- [74] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y. Jiang, and X. Xue, “Pose-normalized image generation for person re-identification,” in Proc. ECCV, 2018, pp. 661–678.
- [75] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, 2016, pp. 770–778.
- [76] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei, “Imagenet: A large-scale hierarchical image database,” in Proc. CVPR, 2009, pp. 248–255.
- [77] D. Gray and H. Tao, “Viewpoint invariant pedestrian recognition with an ensemble of localized features,” in Proc. ECCV, 2008, pp. 262–275.
- [78] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, “Person re-identification by descriptive and discriminative classification,” in Proc. SCIA, 2011, pp. 91–102.
- [79] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, pp. 523–536, 2018.
- [80] C. Qin, S. Song, G. Huang, and L. Zhu, “Unsupervised neighborhood component analysis for clustering,” Neurocomputing, vol. 168, no. C, pp. 609–617, 2015.



Fan Yang is currently working toward the Ph.D. degree in the National Engineering Laboratory for Video Technology, Peking University, China. He received his B.Eng. degree in computer science in 2014 from Harbin Institute of Technology, Weihai, China. His research interests include computer vision, object detection, object tracking and person re-identification.



Ke Yan Ke Yan is currently a senior researcher of Tencent YouTu Lab, Shanghai, China. He received his M.Sc. in computer science in 2017 from Peking University, Beijing, China, and his B.Eng. degree in software engineering in 2014 from Wuhan University, Wuhan, China. His research interests include computer vision, image recognition and search.



Xiaowei Guo graduated from the School of Mathematics and Computational Science, Sun Yat-sen University, currently working at Tencent YouTu Lab, a technical expert. His research interests include computer vision, deep learning, machine learning, etc.



Shijian Lu received his PhD in electrical and computer engineering from the National University of Singapore. He is an Assistant Professor with School of Computer Science & Engineering, the Nanyang Technological University, Singapore. His major research interests include image and video analytics, visual intelligence, and machine learning. He has published more than 100 international refereed journal and conference papers and co-authored over 10 patents in these research areas. He is currently an Associate Editor for the journal Pattern Recognition (PR). He has also served in the program committee of a number of conferences, e.g. the Area Chair of the International Conference on Document Analysis and Recognition (ICDAR) 2017, the Senior Program Committee of the International Joint Conferences on Artificial Intelligence (IJCAI) 2018, etc.



Feiyue Huang is the director of Tencent YouTu Lab and expert researcher. He graduated from Tsinghua University with a Ph.D. in computer science and joined Tencent in 2008. Since 2012, Huang Feiyue founded YouTu Team and he is responsible for research projects of YouTu Lab till now, including computer vision, pattern recognition and machine learning.



Huizhu Jia received his Ph.D. degree in electrical engineering from the Chinese Academy of Sciences, Beijing, China, in 2007. He is currently with the National Engineering Laboratory for Video Technology at Peking University, China. Before joining Peking University, he engaged in chip design in Grandview semiconductor Corp. and Spreadtrum communications Corp. from 2004 to 2010. He has successfully developed a variety of video codec chip. His research interests include: ASIC design, image processing, multimedia data compression and communication.



Wen Gao received the Ph.D. degree in electronics engineering from the University of Tokyo, Tokyo, Japan, in 1991. He was a professor of computer science in the Harbin Institute of Technology, Harbin, China, from 1991 to 1995, and a professor in the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. He is currently a professor of computer science with Peking University, Beijing. He has authored extensively, including five books and more than 600 technical articles in refereed journals and conference proceedings in image processing, video coding and communication, pattern recognition, multimedia information retrieval, multimodal interface, and bioinformatics. He served or serves on the editorial board for several journals, such as the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Multimedia, the IEEE Transactions on Image Processing, the IEEE Transactions on Autonomous Mental Development, the EURASIP Journal of Image Communications, and the Journal of Visual Communication and Image Representation. He chaired a number of prestigious international conferences on multimedia and video signal processing, such as the IEEE International Conference on Multimedia & Expo and ACM Multimedia, and also served on the advisory and technical committees of numerous professional organizations. He is a fellow of the IEEE and ACM.



Don Xie received his Ph.D. degree in Electrical Engineering, University of Rochester, USA. He was a Senior Scientist at Eastman Kodak Company, New York, USA, from 1994 to 1997; a Principal Scientist at Broadcom Corporation, California, USA, from 1997 to 2009. He is currently a Professor with the Department of Electrical Engineering, Peking University, Beijing, China. His research interests include multimedia SoC design, embedded system. He holds 24 U.S. Patents.



Zongqiao Yu received the BEng degree from the University of Science and Technology of China, Hefei, in 2010 and the MEng degree from Nanjing University, Jiangsu, China, in 2013. He is currently a staff researcher at Tencent Inc., Shanghai. His research interests include machine learning and computer vision.