Pixel and Feature Transfer Fusion for Unsupervised Cross-Dataset Person Re-Identification

Yang Yang *Member, IEEE,* Guan'an Wang, Prayag Tiwari, Hari Mohan Pandey, and Zhen Lei *Senior Member, IEEE*

***Abstract*—Recently, unsupervised cross-dataset person re-identification (Re-ID) has attracted more and more attention, which aims to transfer knowledge of a labeled source domain to an unlabeled target domain. There are two common frameworks: one is pixel-alignment of transferring low-level knowledge and the other is feature-alignment of transferring high-level knowledge.**

In this paper, we propose a novel Recurrent Auto-Encoder (RAE) framework to unify these two kinds of methods and inherit their merits. Specifically, the proposed RAE includes three modules, *i.e.* a feature-transfer module, a pixel-transfer module and a fusion module. The feature-transfer module utilizes an encoder to map source and target images to a shared feature space. In the space, not only features are identity-discriminative, but also the gap between source and target features is reduced. The pixel-transfer module takes a decoder to reconstruct original images with its features. Here, we hope the images reconstructed from target features are in source-style. Thus, the low-level knowledge can be propagated to the target domain. After transferring both high- and low-level knowledge with the two proposed module above, we design another bilinear pooling layer to fuse both kinds of knowledge. Extensive experiments on Market-1501, DukeMTMC-ReID and MSMT17 datasets show that our method significantly outperforms either pixel-alignment or feature-alignment Re-ID methods, and achieves new state-of-the-art results.

***Index Terms*—Person Re-Identification, Unsupervised Learning, Generate Adversarial Nets, Feature Fusion**

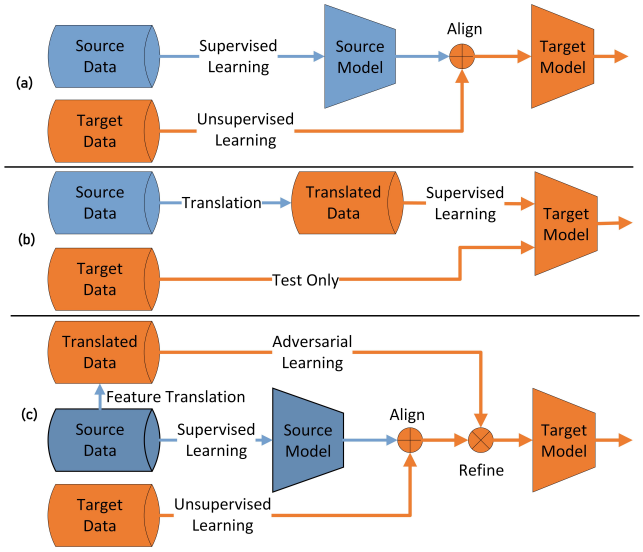


Fig. 1. Brief illustration of existing unsupervised cross-dataset Re-IDs and our proposed approach. (a) Feature-transfer Re-ID methods try to transfer high-level semantic knowledge in a pixel space by reducing the divergence between source and target dataset. (b) Pixel-transfer Re-ID methods transfer low-level knowledge in a pixel space by translating source images to be target-style. (c) Our approach unifies the two types of transfer learning in a framework. With it, we can not only simultaneously transfer both high- and low-level knowledge, but also encourage the two knowledge enhance each other.

I. Introduction

Person re-identification (Re-ID) [1] has attracted more and more attention in recent years for its wide applications in video surveillance, smart city, public security, etc. The goal of Re-ID is to match pedestrian images under dis-joint cameras. For any query pedestrian image in one camera, all pedestrian images with the same identity in other cameras need to be found. However, due to dramatic intra-class variation caused by the view, pose, illumination, occlusion, and small inter-class similarity, Re-ID is still a challenging and unsolved problem.

Traditional Re-ID methods can be grouped into feature learning based methods [2], [3] and metric learning based methods [4], [5]. However, those methods are limited by less-semantic hand-crafted features or weakly-discriminative linear metric functions. Inspired by recent progress on deep learning [6], [7], deep supervised Re-ID methods [8], [9] have been proposed by simultaneously learning feature representation and distance metric with deep neural networks. However, most of those methods are trained in a supervised way, which requires a large number of accurately labeled data from each camera. Considering that pedestrian images under different scenes are dramatically different, which is known as domain shift, directly using the model trained on one scene to another usually performs poorly. Labeling the massive online pedestrian images to support supervised learning is expensive and impractical. Those weaknesses seriously limit the scalability of supervised person Re-ID methods. For better scalability, unsupervised Re-ID methods [10], [11], [12] are proposed, which learn feature representation with unlabelled data. However, for

Yang Yang and Guan'an Wang are co-first authors. Hari Mohan Pandey and Zhen Lei are the corresponding author.

Yang Yang and Guan'an Wang are with National Laboratory of Pattern Recognition (NLPR), Institute of Automation Chinese Academy of Sciences (CASIA), Beijing, China. Email: yang.yang@nlpr.ia.ac.cn, wangguanan2015@ia.ac.cn.

Prayag Tiwari is with Department of Computer Science, Aalto University, Espoo, Finland. Email: prayag.tiwari@aalto.fi.

Hari Mohan Pandey is with Department of Computer Science, Edge Hill University. Email: pandeyh@edgehill.ac.uk.

Zhen Lei is with the National Laboratory of Pattern Recognition (NLPR), Center for Biometrics and Security Research (CBSR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong. Email: zlei@nlpr.ia.ac.cn.

Manuscript received ***; revised ***.

lacking in knowledge about how the visual appearance of identical objects changes cross-cameras, unsupervised methods typically offer weaker performances compared with supervised counterparts. Further, semi-supervised Re-ID methods [13] is also proposed to improve matching accuracy by learning with both labeled and unlabeled data. Yet, those semi-supervised manners still require massive labeled data, which is difficult to obtain in large-scale Re-ID applications.

Recently, Re-ID community focus on unsupervised cross-dataset Re-ID, which tries to adapt the knowledge about cross-camera identity information from an existing labeled dataset (source domain) to unlabeled datasets (target domain). Existing unsupervised cross-dataset Re-ID methods can be divided into two groups, *i.e.* pixel-transfer and feature-transfer Re-ID methods. As shown in Fig. 1(a), pixel-transfer Re-ID methods first translate labeled source data to target-like (source2target) data via a generation model (such as GAN [14]), then use the target-like labeled data to train a target model in a supervised way. Feature-transfer Re-ID methods are displayed in Fig. 1(b), which first train a source model with the source data in a supervised way, then adapt the source model to the target data by pull source and target features with distribution distance metrics such as KL divergence. However, both types of solutions suffer from their weaknesses. On the one hand, feature-transfer Re-ID methods pull domain-level source- and target-features, which may lead to semantic misalignment and harm performance. On the other hand, for pixel-transfer Re-ID methods, the unexpected low-level characters of resolutions, backgrounds, and illuminations will be transferred to the target domain.

This paper unifies both pixel-transfer and feature-transfer learning for cross-dataset unsupervised Re-ID, which enjoys their advantages, meanwhile overcoming shortcomings. Specifically, we propose a novel recurrent auto-encoder (RAE). It includes the encoder of feature-transfer learning and the decoder of pixel-transfer learning. The encoder's task is to transfer feature knowledge from source to target domain by learning a consistent feature space with a GAN loss. Meanwhile, the decoder is to transfer pixel knowledge from source to target domain by encouraging a feature of the target domain to reconstruct its original images but with source-style. Thus, the pixel knowledge can be transferred while avoiding encoding low-level noises such as resolution, background, and so on. Besides, to take full advantage of the source pixel knowledge, we extract the features of the reconstructed source-style image with the encoder and use a bilinear pooling layer to fuse it with the feature of the original target image. Because bilinear pooling uses second-order statistics and can interactively model the pairwise features, both pixel-transfer, and feature-transfer learning information is enhanced.

The main contributions of this work are summarized as below:

(1) We explore the problem of unsupervised cross-dataset Re-ID by unifying pixel and feature-transfer learning. To the best of our knowledge, this is the first work towards this target in the person Re-ID community.

(2) We propose a novel recurrent auto-encoder (RAE) which simultaneously performs pixel- and feature-transfer learning. In the framework, both types of transfer learning can be jointly performed, meanwhile they can constrain and enhance each other for a better transfer.

(3) We design a bilinear pooling layer to fuse features of a target image and its reconstructed source-style image and enhance the final transferring information from the source domain to the target domain.

(4) Extensive experiments are conducted on three pedestrian datasets Market-1501, DukeMTMC-reID and MSMT-17. Our proposed framework significantly outperforms either pixel- or feature-transfer Re-IDs and achieves state-of-the-art performance, which verifies the effectiveness of our proposed framework.

II. Related Work

A. Supervised Re-ID

Existing methods can be grouped into hand-crafted descriptors [15], [16], [17], metric learning methods [18], [19], [2] and deep learning algorithms [1], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. The goal of hand-crafted descriptors is to design robust features. For example, Ma *et al.* [15] handle the background and illumination variations by combining biologically inspired features with covariance descriptors. Yang *et al.* [16] explore color information by using salient color names. In [17], Liao *et al.* propose an effective feature representation called local maximal occurrence, which can analyze the horizontal occurrence of local features and maximize the occurrence to make a stable representation against viewpoint changes. Metric learning methods are designed to make a pair of true matches have a relatively smaller distance than that of a wrong match pair in a discriminant manner. Zheng *et al.* [18] formulate person RE-ID as a relative distance comparison learning problem in order to learn the optimal similarity measure between a pair of person images. The model is formulated to maximize the likelihood of a pair of true matches having a relatively smaller distance than that of a wrong match pair in a soft discriminant manner. Deep learning algorithms adopt deep neural networks to learn robust and discriminative features in an end-to-end manner straightly. For example, Zheng *et al.* [1] learn identity-discriminative features by fine-tuning a pre-trained CNN to minimize a classification loss. In [20], Hermans *et al.* show that using a variant of the triplet loss outperforms most other published methods by a large margin. In [21], a network named Part-based Convolutional Baseline (PCB) is proposed to learn fine-grained part-level features with a uniform partition strategy. Most of the existing methods are designed for supervised Re-ID task, and can not be adapted to unsupervised Re-ID task. This limits the applicability in practical surveillance scenarios.

B. Unsupervised and Semi-Supervised Re-ID

To alleviate the above limitations, researchers also focus on unsupervised person Re-ID methods using unlabeled data. For example, Fan *et al.* [12] applies techniques of data clustering, instance selection, and fine-tuning methods to obtain pseudo labels for the unlabeled data. Nevertheless, due to the lack

of the knowledge about identity information across cameras, unsupervised Re-ID methods typically cannot achieve comparable results as the supervised ones do. The low retrieval accuracy limits the practicability of unsupervised Re-ID methods. To achieve a balance between scalability and practicability, semi-supervised Re-ID methods are proposed to learn with both labelled and unlabeled data. For example, Liu *et al.* [13] propose a coupled dictionaries model, where a dictionary learns labelled data to carry the relationship between features from different cameras, and another one learns unlabeled data to exploit the geometry of the marginal distribution. Nevertheless, semi-supervised manners still require massive labelled data, which is difficult to obtain in a large-scale Re-ID system.

C. Cross-dataset Unsupervised Re-ID

For better scalability and higher retrieval accuracy, cross-dataset unsupervised Re-ID methods are proposed, where beside a target unlabeled dataset used to both train and test, an extra lableled source dataset is import to enhance performance during training. Its main idea is to transfer knowledge of labeled source dataset to the unlabeled target dataset, so that a high test accuracy can be achieved in the target dataset.

There are two different kinds of solutions for cross-dataset unsupervised Re-ID, *i.e.* pixel-transfer and feature-transfer Re-ID. Pixel-transfer cross-dataset unsupervised Re-IDs aims to transfer knowledge in the pixel space. It usually includes two stages, where firstly translate source images to fit target images' style (such as color, illumination, view), then uses the adapted images to train a target model in a supervised manner. For example, Deng *et al.* [33] propose a novel SPGAN to translate source pedestrian images to target style with preserved self-similarity and domain-dissimilarity, and use the translated images to train a model for the target domain. Wang *et al.* [34] propose a novel identity-preserving GAN named PTGAN to transfer source data to the target domain, and use the translated images to train a model for the target domain. Bake *et al.* [35] propose to translate pedestrian images to target domain from synthesis images of the game engine with labeled with illumination information to solve illumination bias.

Feature-transfer cross-dataset unsupervised Re-IDs try to transfer knowledge in the feature space. Its key point is learning a common feature space shared by both source and target datasets. For example, Peng *et al.* [36] adopt a multi-task dictionary learning method to learn a dataset-shared but target biased feature representation. Wang *et al* [37] adopt deep architecture to learn an transferable attribute-semantic and identity-discriminative feature, Lv *et al.* [38] propose to transfer spatial-temporal patterns from source domain to target domain. Li *et al.* [39] propose to use source classification and ranking loss to learn identity-discriminative features and use the orthogonal constraint to learn domain-invariant feature for both source and target data.

The pixel-transfer manners can significantly reduce the low-level divergence in the image space such as illumination and color, but may not deal with the high-level variation such as age, carrying, pose. Complementarily, feature-transfer manners are good at reducing high-level variation but often neglect low-level divergence. Different from either pixel- or feature-transfer cross-dataset unsupervised Re-ID datasets, our proposed approach inherent advantages of both methods and overcome their shortcomings by unifying the two kinds of manners. Besides, in our framework, the two types of transfer learning enhance other, thus get better performance than that of simply cascading them.

D. Unsupervised Domain Adaptation

Our work relates to unsupervised domain adaptation (UDA). Recent trend consists of learning domain-invariant features and image-level domain translation. For example, [40] aim to learn a mapping between source and target distributions. [41] use an adversarial approach to learn a transformation in the pixel space from the source domain to the target domain. [42] focus on learning domain-invariant feature space. Most UDA methods assume that class labels are the same across domains, while identities of different pedestrian datasets are non-overlapping. Therefore, those UDA methods mentioned above cannot be utilized for domain-adaptation in Re-ID.

E. Re-ID with GAN

Recently, many methods attempt to utilize GAN to generate training samples for improving Re-ID. Zheng *et al.* [43] use a GAN model to generate unlabeled images as data augmentation. Huang *et al.* [44] first assign pseudo labels to generated pedestrian images and then learn them in a supervision manner. Zhong *et al.* [45], [46], [47] translate images to different camera styles with CycleGAN [48], and then use both real and generated images to reduce inter-camera variation. Ma *et al.* [49], [50] use a cGAN [51] to generate pedestrian images with different poses to learn features free of influences of pose variation. Zheng *et al* [52] propose joint learning framework that end-to-end couples re-id learning and image generation in a unified network. All those methods focus on supervised Re-ID. Different from them, our method utilise GAN as an objective function to reduce the gap between source and target domains for cross-dataset unsupervised Re-ID.

III. Recurrent Auto-Encoder

Suppose there are an annotated pedestrian dataset (source domain) $\{I_i^s\}_{i=1}^{N_s}$ labelled with identities $\{y_i^s\}_{i=1}^{N_s}$, and an unlabelled pedestrian dataset (target domain) $\{I_i^t\}_{i=1}^{N_t}$, where N_s and N_t denote total images in source domain and target domain, respectively. Our primary goal is to learn a Re-ID model that generalizes well in the target domain by leveraging labeled samples in the source domain and unlabeled samples in the target domain. The key idea is to transfer pixel and feature knowledge of how to re-identify pedestrians from the source domain to the target one, and then fuse both knowledge.

Our proposed approach includes three modules, namely feature-transfer module, pixel-transfer module and fusion module. The feature-transfer module aims to learn a shared feature space for both source and target domain with an

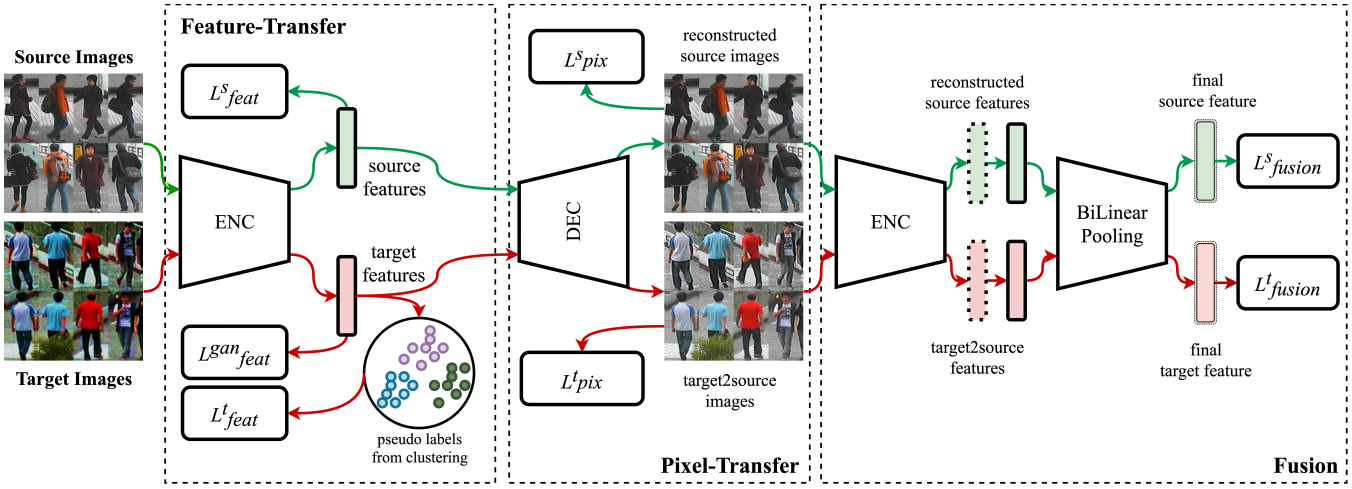


Fig. 2. Overview of our proposed framework. The proposed framework includes a feature-transfer module, a pixel-transfer module and a fusion module.

Encoder Enc. It first trained using labeled images of source domain with cross-entropy and triplet loss [20]. In the target domain, unlabeled images are represented by features by encoder Enc, then a clustering algorithm learns cluster labels of samples. With the pseudo-labels, the encoder can be refined by target images. Besides, a GAN loss [14] is used to reduced the domain gap between source and target features. The pixel-transfer module try to transfer the low-level information from source to target domain with Decoder Dec. It takes features as inputs and outputs images. The Dec is first train by source images with a reconstructed loss and GAN loss. Then backward source knowledge by force the generated target images to be source-style. Finally, the fusion module fuses feature-transfer and pixel-transfer module with a bilinear pooling layer.

A. Feature-Transfer Module

The feature-transfer learning module aims to learn a shared feature space for both source and target domain with an encoder Enc as in Eq.(1). I mean images, F mean corresponding feature maps. We first extract identity-discriminative knowledge from labeled source data in a supervised way and unlabeled target data in a self-supervised (clustering) learning way. Then to better adapt the knowledge to the target domain, the source and target features are pulled with KL divergence.

$$F = Enc(I) \tag{1}$$

Learning source images in a supervised way. Following existing supervised Re-ID models [22], [9], we first train the encoder Enc with labeled source images in cross-entropy loss and triplet loss [20]. They can be formulated as below:

$$\mathcal{L}^s_{feat} = \mathcal{L}^s_{cls} + \lambda_{tri}\mathcal{L}^s_{tri} \tag{2}$$

$$\mathcal{L}^s_{cls} = -\frac{1}{N_s}\sum_{i=1}^{N_s} logp(y_{s,i}|I_{s,i}) \tag{3}$$

$$\mathcal{L}^s_{tri} = \frac{1}{N_s}\sum_{i=1}^{N_s}[m + ||F_i - F_{i,p})||_2 - ||F_i - F_{i,n}||_2] \tag{4}$$

where N_s denotes the number of source images, $p(y_{s,i}|I_{s,i})$ is the predicted probability of image $I_{s,i}$ belonging to $y_{s,i}$, F_i is the feature of image I_i which can be computed by Eq.(1), $F_{i,p}$ and $F_{i,n}$ are the postive and negative features of F_i, m is a margin parameter.

Learning target images in self-supervised way. Since the target images are unlabeled, following [53], we utilise clustering algorithm to learn their pseudo-labels, then using the pseudo-labels to train the encoder in a self-supervised way. The clustering procedure includes three steps: after every epoch, 1) extracting features for all target images, 2) computing a distance matrix with k-reciprocal encoding for all features and then performing density-based clustering to assign samples to different groups, 3) assigning group-index to every images as their pseudo-labels y_t. Thus, the overall loss of this paper can be described as below:

$$\mathcal{L}^t_{feat} = \mathcal{L}^t_{cls} + \lambda_{tri}\mathcal{L}^t_{tri} \tag{5}$$

$$\mathcal{L}^t_{cls} = -\frac{1}{N_t}\sum_{i=1}^{N_t} logp(y_{t,i}|I_{t,i}) \tag{6}$$

$$\mathcal{L}^t_{tri} = \frac{1}{N_t}\sum_{i=1}^{N_t}[m + ||F_i - F_{i,p})||_2 - ||F_i - F_{i,n}||_2] \tag{7}$$

where N_t denotes the number of target images, $p(y_{t,i}|I_{t,i})$ is the predicted probability of image $I_{t,i}$ belonging to $y_{t,i}$, F_i is the feature of image I_i which can be computed by Eq.(1), $F_{i,p}$ and $F_{i,n}$ are the positive and negative features of F_i, m is a margin parameter.

Adapt to the target domain. Considering the gap between source and target domains and the learned encoder may be more biased to the source images (the encoder Enc is trained with confident source ground truth labels but noisy target pseudo labels), it is necessary to adapt the Enc to the target domain. Here, we propose to adapt by pulling features from source and target features with a GAN loss [14]. GAN loss

can reduce the KL divergence between two distribution. Its formulation is shown in below:

$$\begin{aligned}\min \mathcal{L}_{feat}^{gan} &\equiv \textit{iter}(\min_{Enc} \mathcal{L}_{enc},\ \min_{Dis} \mathcal{L}_{dis}),\\ \min_{Enc} \mathcal{L}_{enc} &= \frac{1}{N_t}\sum_{i}^{N_t}[log\ Dis(F_i^t)],\\ \min_{Dis} \mathcal{L}_{dis} &= \frac{1}{N_s}\sum_{i}^{N_s}[log\ Dis(F_i^s)]\\ &\quad + \frac{1}{N_t}\sum_{i}^{N_t}[log(1-Dis(F_i^t))],\end{aligned} \tag{8}$$

where Dis is a discriminator to distinguish target features from source features, $\textit{iter}(\cdot,\cdot)$ means iteratively optimize the two items.

Overall Feature-Transfer loss. The overall loss of feature-transfer module is summarised as below:

$$\mathcal{L}_{feat} = \mathcal{L}_{feat}^{s} + \mathcal{L}_{feat}^{t} + \lambda_{feat}^{gan}\mathcal{L}_{feat}^{gan} \tag{9}$$

B. Pixel-Transfer Module

Although the feature-transfer module learns a shared feature space for source and target domains, it only considers high-level semantic knowledge but fails to deal with some low-level information such as color and illumination. Recent pixel-transfer unsupervised methods [33], [34], [35] have shown their powerful performance for unsupervised cross-dataset person Re-ID. They first translates target images to be source-style, then extract corresponding features. Thus low-level information of source domain can be transferred to target domain. However, due to large gap and trivial variation of raw images, the generated images is not necessary to preserve identity information. As a consequence, those generated samples have a serious side effect on feature learning. What's more, existing transfer methods are two-stage procedure (first generate images, then extract their features), which is cumbersome and time-consuming, and difficult to deal with large-scale dataset. Different from all existing pixel-transfer unsupervised Re-ID methods in a forward, we choose to backward the information, which naturally avoids the noise from unperfect generated images.

Specifically, the pixel-transfer module utilise a Decoder Dec to learn low-level information of source domain, and then transfer it to target domain by making the reconstructed target images to be source-style. The decoder Dec can be formluated as in Eq.(10).

$$\hat{I} = Dec(F) \tag{10}$$

Learn low-level knowledge from source domain. We first learn low-level information from source domain with a Decoder Dec, The Dec is trained with source data by reconstructing source images with source features. Its loss is shown in Eq.(11)

$$\mathcal{L}_{pix}^{s} = \mathcal{L}_{reconst}^{s} + \lambda_{pix}^{gan}\mathcal{L}_{gan}^{s} \tag{11}$$

$$\mathcal{L}_{reconst}^{s} = \frac{1}{N_s}\sum_{i=1}^{N_s}[||I_{i,s} - \hat{I}_{i,s}||_2] \tag{12}$$

$$\begin{aligned}\min \mathcal{L}_{gan}^{s} &\equiv \textit{iter}(\min_{Dec} \mathcal{L}_{dec},\ \min_{Enc} \mathcal{L}_{enc})\\ \min_{Dec} \mathcal{L}_{dec} &= \frac{1}{N_s}\sum_{i=1}^{N_s}[log\ Dis(\hat{I}_{i,s})]\\ \min_{Dis} \mathcal{L}_{dis} &= \frac{1}{N_s}\sum_{i=1}^{N_s}[log\ Dis(I_{i,s})\\ &\quad + log(1-Dis(\hat{I}_{i,s}))]\end{aligned} \tag{13}$$

$\mathcal{L}_{pix}^{s}$ consists of two parts including a L2 loss and a GAN loss. The former makes Dec easy to be optimized and the latter encourages the reconstructed images to be less blurry and more realistic. The two losses has been proved to be effective by many works [54]. $\hat{I}$ is the reconstructed images with Eq.(10), Dis is a discriminator to distinguish reconstructed source images $\hat{I}_s$ from real images I_s. Please note that the discriminators Dis in Eq.(13) and Eq.(8) do not share weights.

Transfer low-level pixel knowledge to target domain. After trained by source data, the Decoder Dec contains low-level source knowledge. Then, we transfer the low-level pixel knowledge to target domain by forcing the target features to reconstruct images with source-style. Thus, the low-level source knowledge can be contained in the feature space, meanwhile avoid encoding noisy generated images.

$$\mathcal{L}_{pix}^{t} = \mathcal{L}_{reconst}^{t} + \lambda_{pix}^{gan}\mathcal{L}_{gan}^{t} \tag{14}$$

$$\mathcal{L}_{reconst}^{t} = \frac{1}{N_t}\sum_{i=1}^{N_t}[||I_{i,t} - \hat{I}_{i,t}||_2] \tag{15}$$

$$\mathcal{L}_{gan}^{t} \equiv \min_{Dec} \frac{1}{N_t}\sum_{i=1}^{N_t}[log\ Dis(\hat{I}_{i,t})], \tag{16}$$

Overall pixel-transfer loss. The overall loss of the pixel-transfer module can be summarised as below:

$$\mathcal{L}_{pix} = \mathcal{L}_{pix}^{s} + \mathcal{L}_{pix}^{t} \tag{17}$$

C. Fusion

Review bilinear pooling. Factorized bilinear pooling has been proven to be effective for feature fusion [55]. Suppose a feature map $F_1/F_2 \in R^{h\times w\times c}$ with height h, width w and channels c, we denote a $h \times w$ dimensional descriptor at a channel location on F as $F = [f_1, f_2, ..., f_c]^T$. Then the full bilinear model is defined by

$$z_i = F_1^T W_i F_2 = F_1^T U_i V_i^T F_2 = U_i^T F_1 \circ V_i^T F_2 \tag{18}$$

Here, $W_i \in R^{c\times c}$ is a projection matrix, z_i is the output of the bilinear model. The $W \in R^{c\times c\times o}$ is a trainable parameters to obtain a o dimenional output z. According to matrix factorization, the projection matrix W_i can be factorized in to two one-rank vectors. $U_i \in R^c$ and $V_i \in R^c$. Thus the output feature $z \in R^o$ is given by

$$z = \mathcal{P}(F_1, F_2) = Q^T(U^T F_1 \circ V^T F_2) \tag{19}$$

where $U \in R^d$ and $V \in R^{c\times d}$ are projection matrices, $Q \in R^{d\times o}$ is the classification matrix, $\circ$ is Hadamard product

Algorithm 1 Training Procedure

Input: Source domain dataset **S**, target domain dataset **T**
Output: Encoder Enc, Decoder Dec, Fusion $\mathcal{P}$
While unit converge:
1. predict pseudo-labels with clustering algorithm
2. optimize feature-transfer module by minimizing Eq.(9)
3. optimize pixel-transfer module by minimizing Eq.(17)
4. optimize fusion module by minimizing Eq.(21)

Algorithm 2 Inference Procedure

Input: target test images $\{I^t\}_{i=1}^n$ **T**, trained Enc, Dec and $\mathcal{P}$
Output: features $\{F_i^{pt}\}_{i=1}^n$
For every target image I^t:
1. feature transfer $F^t = Enc(I^t)$ with Eq.(1)
2. pixel transfer $I^{t2s} = Dec(F^t)$ with Eq.(10)
3. repeat feature transfer $F^{t2s} = Enc(I^{t2s})$ with Eq.(1)
4. fuse feature $F^{pt} = \mathcal{P}(F^t, F^{t2s})$ with Eq.(20)
5. final feature $F^{pt} = gap(F^{pt})$, gap is global average pooling

and d is a hyperparameter deciding the dimension of joint embeddings.

Fuse feature-transfer and pixel-transfer modules. The feature-transfer and pixel-transfer modules can well transfer high-level and low-level knowledge of source domain to the target domain. To further enhance the performance, explicitly fuse the two modules with a bilinear pooling. The final fused feature with bilinear pooling F^p can be formulated as below:

$$\begin{aligned} F^{pt} &= \mathcal{P}(F^t, F^{t2s}) \\ F^{ps} &= \mathcal{P}(F^s, F^{s2s}) \end{aligned} \tag{20}$$

where F^t is the feature map of a target image I^t, F^s is the feature map of a source image I^s, F^{t2s} is the feature map of reconstructed images I^{t2s} from F^t, F^{s2s} is the feature map of reconstructed images I^{s2s} from F^s.

Loss. We train the fusion module with classification loss and triplet loss, Specifically,

$$\mathcal{L}_{fusion} = \mathcal{L}_{fusion}^s + \mathcal{L}_{fusion}^t \tag{21}$$

$$\begin{aligned} \mathcal{L}_{fusion}^s = -\frac{1}{N_s}\sum_{i=1}^{N_s}\{log\ p(y_{s,i}|F_i^{ps}) \\ + \lambda_{tri}[m + ||F_i^{ps} - F_{i,p}^{ps})||_2 - ||F_i^{ps} - F_{i,n}^{ps}||_2]\} \end{aligned} \tag{22}$$

$$\begin{aligned} \mathcal{L}_{fusion}^t = -\frac{1}{N_t}\sum_{i=1}^{N_t}\{log\ p(y_{t,i}|F_i^{ts}) \\ + \lambda_{tri}[m + ||F_i^{pt} - F_{i,p}^{pt})||_2 - ||F_i^{pt} - F_{i,n}^{pt}||_2]\} \end{aligned} \tag{23}$$

where N_s denotes the number of source images, $p(y_{s,i}|I_{s,i})$ is the predicted probability of image $I_{s,i}$ belonging to $y_{s,i}$, F_i is the feature of image I_i which can be computed by Eq.(1), $F_{i,p}$ and $F_{i,n}$ are the positive and negative features of F_i, m is a margin parameter.

D. Optimization and Inference

The overall optimization and inference procedures are summarised in Algorithm 1 and Algorithm 2, respectively.

IV. Experiment

Fig. 3. Examples from Market-1501, DukeMTMC-reID and MSMT17 datasets. In each row, the left/right 5 images contain the same pedestrian.

A. Datasets and Evaluation Procotols

We choose three widely used large-scale pedestrian benchmarks Market-1501 [56], DukeMTMC-ReID [57] and MSMT17 [34] to evaluate our proposed approach. Following [33] we only adopt identity labels of source domain as supervised information, and no attribute labels is used. Their statistic and illustration are shown in Table I and Fig. 3, respectively.

Market-1501. Market-1501 contains $32,668$ images of $1,501$ pedestrians collected from 6 camera at an university campus. The bounding box was cropped by DPM [58]. Therefore more background clutter and misalignment problem are characterized. The dataset is split into two non-over-lapping subsets, 751 identities for training and 750 identities for testing. During the test stage, $3,368$ images are used as probes to retrieve gallery set with $19,732$ images.

DukeMTMC-reID. DukeMTMC-reID contains $36,441$ images of $1,404$ identities collected from 8 cameras. When DukeTMTC-reID is as target dataset, we split all identities into two halves $702/702$ for training and testing. In test stage, $2,228$ images are used as probes to query gallery set with $17,661$ images.

MSMT17. MSMT17 [34] is a newly released person ReID dataset. It is composed of 126,411 person images from 4,101 identities collected by 15 cameras. The dataset suffers from substantial variations of scene and lighting, and is more challenging than the other two datasets.

We utilise mean average precision (mAP) to represent overall precision and recall rates, Cumulative Matching Characteristic (CMC) curve to reflect the top-k retrieval precision.

B. Implementation Details

Encoder. For a fair comparison and following most Re-ID methods, we adopt ResNet-50 [7] pre-trained by ImageNet [6]

Dataset	Train Nums (ID / Image)	Testing Nums (ID / Image)	
		Gallery	Query
Market-1501	751 / 12,936	750 / 19,732	750 / 3,368
DukeMTMC-reID	702 / 16,522	1,110 / 17,661	702 / 2,228
MSMT17	1,041 / 32,621	3,060 / 82,161	3,060 / 11,659

TABLE I
DATASET DETAILS. WE EVALUATE OUR METHOD ON 3 PUBLIC RE-ID DATASETS, INCLUDING MARKET-1501, DUKEMTMC-REID AND MSMT17.

Methods	Year	Target:n DukeMTMC-reID					Target: Market-1501				
		Source	Rank1	Rank5	Rank10	mAP	Source	Rank1	Rank5	Rank10	mAP
BOW	ICCV'15	duke	25.1	-	-	12.2	market	44.4	-	-	20.8
LDNS	CVPR'16	duke	-	-	-	-	market	61.0	-	-	35.7
TriNet	ArXiv'17	duke	72.4	-	-	53.5	market	84.9	-	-	69.1
DuATM	CVPR'18	duke	81.8	-	-	64.6	market	91.4	-	-	76.6
PCB	ECCV'18	duke	83.3	90.5	92.5	69.2	market	93.8	97.5	98.5	81.6
BOW	ICCV'15	none	17.1	28.8	34.9	8.3	none	35.8	52.4	60.3	14.8
LOMO	CVPR'15	none	12.3	21.3	26.6	4.8	none	27.2	41.6	49.1	8.0
PUL	TOMM'18	market	30.0	43.4	48.5	16.4	duke	45.5	60.7	66.7	20.5
CAMEL	CVPR'17	-	-	-	-	-	duke	54.5	-	-	26.3
TJ-AIDL	CVPR'18	market	44.3	59.6	65.0	23.0	duke	58.2	74.8	81.1	26.5
SyRI	ECCV'18	-	-	-	-	-	mix	65.7	-	-	-
SPGAN	CVPR'18	market	46.9	62.6	68.5	26.4	duke	58.1	76.0	82.7	26.9
HHL	ECCV'18	market	46.9	61.0	66.7	27.2	duke	62.2	78.8	84.0	31.4
ECN	CVPR'19	market	63.3	75.8	80.4	40.4	duke	75.1	87.6	91.6	43.0
PDA-Net	ICCV'19	market	63.2	77.0	82.5	45.1	duke	75.2	86.3	90.2	47.6
UDAP	CVPR'20	market	68.4	80.1	83.5	49.0	duke	75.8	89.5	93.2	53.7
ECN++	TPAMI'20	market	74.0	83.7	87.4	54.4	duke	84.1	92.8	95.4	63.8
BUC	AAAI'19	none	40.4	52.5	58.2	40.4	none	61.9	73.5	78.2	29.6
CR-GAN	ICCV'19	market	68.9	80.2	84.7	48.6	duke	77.7	89.7	92.7	54.0
PCB-PAST	ICCV'19	market	72.4	-	-	54.3	duke	78.4	-	-	54.6
SSG	ICCV'19	market	73.0	80.6	83.2	53.4	duke	80.0	90.0	92.4	58.3
MMCL	CVPR'20	market	72.4	82.9	85.0	51.4	duke	84.4	92.8	95.0	60.4
AD-Cluster	CVPR'20	market	72.6	82.5	85.5	54.1	duke	86.7	94.4	96.5	68.3
SADA	CVPR'20	market	74.5	85.3	88.7	55.8	duke	83.0	91.8	94.1	59.8
ADTC	ECCV'20	market	71.9	84.1	87.5	52.5	duke	79.3	90.8	94.1	59.7
D-MMD	ECCV'20	market	63.5	78.8	83.9	46.0	duke	70.6	87.0	91.5	48.8
GDS-H	ECCV'20	market	73.1	-	-	55.1	duke	81.1	-	-	61.2
JVTC	ECCV'20	market	75.0	85.1	88.2	56.2	duke	83.8	93.0	95.2	61.1
RAE (Ours)	-	market	72.4	84.4	87.7	53.0	duke	85.7	94.2	**96.7**	69.8
MMCL + RAE (Ours)	-	market	75.6	84.9	87.6	54.4	duke	86.1	93.9	95.9	**69.9**
JVTC + RAE (Ours)	-	market	**77.9**	**86.3**	**88.7**	**58.4**	duke	**86.9**	**94.3**	96.1	69.0

TABLE II
COMPARISON WITH STATE-OF-THE-ART ON MARKET-1501 AND DUKEMTMC-REID. SOURCE DATASET IS USED AS LABELED ONE AND TARGET AS UNLABELED ONE. OURS PERFORMS BEST COMPARED WITH EXISTING UNSUPERVISED CROSS-DATASET RE-ID METHODS.

as the CNN backbone. Specifically, we discard the last layer to extract feature map, and apply dropout with 0.5 on *pool5* layer for better generalization. Following [22], to get a high-resolution feature map, the stride of the last layer is set 1.

Decoder. The decoder Dec consists of three fractional-strided convolutional layers. Each layer but the last one utilizes the batch normalization technique and activated by Rectified Linear Units (RelU), and the last layer is projected to the range $[-1, 1]$ by $tanh$.

Discriminator. The discriminators in feature-transfer and pixel-transfer modules consist of three strided convolutional layers activated by Leaky ReLU and a FC layer to one logit. Please note that the two kinds of discriminators do not share weights.

We implement our approach with Pytorch framework and trained using 4 NVIDIA TITANXP GPUs (each with 12GB VRAM). We use the SGD algorithm with momentum 0.9 and weight decay $5e^{-5}$ as optimizer. Considering that the Encoder Enc has well initialized with ImageNet [6], we set the learning rate of Enc relatively small value 0.05, and the other part are set 0.5. Following [20], we set batch size $72 = 18 \times 4$ for both source and target data set, where 18 means total identities in each batch and 4 means total images of each identity. We train the whole model by $20,000$ iterations, and the learning rates decay by 0.1 after every $8,000$ iterations. We have three hyper-parameters: λ_{tri}, λ_{feat}^{gan} and λ_{pix}^{gan}. We set $\lambda_{tri} = 1.0$, $\lambda_{feat}^{gan} = 10.0$ and $\lambda_{pix}^{gan} = 10.0$ via cross-validation. More analysis can be seen in Section IV-E (Parameters Analysis).

C. Comparisons with the State-of-the-arts

State-of-the-art methods. In this section, we compare our method with state-of-the-art supervised and unsupervised Re-

ID methods on Market1501 and DukeMTMC-reID datasets. We first compare our results with two hand-crafted features Bag-of-Words (BOW) [56] and local maximal occurrence (LOMO) [17]. Those two hand-crafted features are directly applied on the target dataset without any training process. We then compare existing unsupervised cross-dataset methods including feature-transfer and pixel-transfer methods. The feature-transfer cross-dataset unsupervised Re-ID methods includes PUL [12], CAMEL[11], TJ-AIDL[37]. The pixel-transfer cross-dataset unsupervised Re-ID methods includes SPGAN[33], SyRI[35], HHL[46], ECN [47], PDA-Net [59] and UDAP [60], ECN++ [61]. Note that, different from common setting, SyRI uses synthesis images of a 3D game engine together with DukeMTMC and CUHK datasets as source data, which contain more label information. Besides the transfer-based unsupervised Re-ID methods, we also compare some cluster-based ones, which utilise and improve clustering algorithms for better robust and reliable pesudo-labels. They are BUC [62], UCDA [63], CR-GAN [64], PCB-PAST [65], SSG [66], MMCL [67], AD-Cluster [68], SADA [69], ADTC [70], D-MMD [71], JVTC [72]. What's more, we also introduce some supervised methods, which are straightly trained with target data in a supervised way. They are LDNS [73]. TriNet [20], DuATM [74] and PCB [9]. The performance evaluation of those methods are shown in Table II.

Overall results. Our proposed framework clearly outperforms existing state-of-the-art unsupervised Re-ID methods, improving mAP scores by at least 4% and 10% on DukeMTMC-reID and Market-1501 respectively. This demonstrates the overall performance advantages of our proposed approach in capability of simultaneous increment and translation learning for cross-dataset unsupervised Re-ID

Comparison with hand-crafted features. When comparing with unsupervised hand-crafted Re-ID methods BOW [56] and LOMO [17], the performance margins are even much larger, e.g. *Ours* outperforms BOW/LOMO $24.7\%/28.2\%$ and $25.0\%/31.8\%$ on DukeMTMC-reID and Market-1501 datasets respectively. This indicates that hand-crafted features are not sufficient to describe dramatic intra-class variation.

Comparison with cross-dataset unsupervised Re-ID. When comparing with feature-transfer or pixel-transfer unsupervised Re-ID methods, several phenomenons can be observed. Firstly, we can find that both feature- and pixel-transfer methods significantly outperform the hand-crafted method, *e.g.* PUL outperform BOW by 8.1% and 7.7% on DukeMTMC-reID and Market-1501 datasets. This verifies that those cross-dataset unsupervised Re-ID methods can transfer the knowledge about identity cross-camera to target domain from the source domain and improve the matching accuracy. Secondly, we can find feature-transfer (TJ-AIDL) and pixel-transfer (SPGAN) methods achieve similar performance, *e.g.* only rank1 gap of 2.6% and 0.1% between TJ-AIDL and SPGAN on two datasets. But when the source dataset are augmented by more labelled data, pixel-transfer method SyRI outperform alignment TJ-AIDL by 8.5% on DukeMTMC-reID in terms of Rank1, which indicates the potential of translation Re-IDs for more labelled data. Finally, by unifying both feature- and pixel-transfer in an end-to-end framework, our approach significantly outperforms either feature-transfer or pixel-transfer unsupervised cross-dataset Re-IDs, and achieves the best performance among unsupervised cross-dataset Re-ID methods, which demonstrates the effectiveness of our proposed framework.

Comparison with clustering-based unsupervised Re-ID. In the pixel-transfer module, we utilise a clustering algorithm to learn pseudo-labels. Here, we also compare with recently proposed clustering-based unsupervised Re-ID. Several phenomenons can be observed. Firstly, most cluster-based unsupervised Re-ID methods achieve better accuracy than transfer-based ones. For example, JVTC get 75% Rank-1 score in market2duke setting, outperforming ECN++ by 1%. This demonstrates the effectiveness of pseudo-labels. Secondly, by improving clustering algorithms, such as iterative trick (AD-Cluster), metric learning (D-MMD), mutual-learning (MMCL), memory (JVTC), pseudo-labels can be significantly improved to be more reliable and less noisy, thus contribute to better accuracy. Specifically, in market2duke setting, Rank-1 score increases from 40% of BUC (AAAI'19) to 75% of JVTC (ECCV'20). Thirdly, since our proposed method mainly contributes to a combination framework of feature-transfer and pixel-transfer learning, any clustering algorithm can be used to learn pseudo-labels, *i.e.* ours is compatibility of any clustering algorithms. Using state-of-the-art clustering algorithms such as MMCL or JVTC, we can further achieve better accuracy. For example, powered by the proposed RAE, JVTC+RAE (Ours) improves Rank-1 score from 75.0% to 77.9%.

Comparison with supervised Re-ID. We also compare with supervised Re-ID methods. As we can see in Table II, Supervised hand-crafted Re-ID (BOW) exceeds its unsupervised version by $3.9\%/6.0\%$, which verifies the importance of label information. Our approach outperforms supervised hand-crafted Re-ID BOW by $20.8\%/19.0\%$, achieve comparable performance with LDNS, but is still behind most deep supervised Re-IDs by $20.5\% - 36.2\%$ and $29.3\% - 41.8\%$, which shows the potentiality of unsupervised cross-dataset Re-ID.

D. Results on Large-Scale datasets

We also conduct experiments on MSMT17, a larger and more challenging dataset. A limited number of works report performance on MSMT17, i.e., PTGAN [34], ECN [47], and SSG [66], ECN++ [61], MMCL [67], D-MMD [71], NRMT [75] and JVTC [72]. The experimental results are shown in Table III. As we can see, the overall performance is much lower than that on Market-1501 and DukeMTMC-reID, showing that the MSMT17 is much more difficult than the others. Our approach outperforms existing methods by large margins under both unsupervised and transfer learning settings. For example, our method achieves 43.6%/40.8% rank-1 accuracy respectively. This outperforms SSG by 11.4% in rank-1 accuracy. When using state-of-the-art clustering methods (such as MMCL and JVTC), ours can achieve better accuracy. The above experiments demonstrate the effectiveness of our proposed method on complex dataset.

Methods	Year	Target: MSMT17					Target: MSMT17				
		Source	Rank1	Rank5	Rank10	mAP	Source	Rank1	Rank5	Rank10	mAP
PTGAN	CVPR'18	market	10.2	-	24.4	2.9	duke	11.8	-	27.4	3.3
ECN	ECCV'18	market	25.3	36.3	42.1	8.5	duke	30.2	41.5	46.8	10.2
SSG	ICCV'19	market	31.6	-	49.6	13.2	duke	32.2	-	51.2	13.3
ECN++	TPAMI'20	market	40.4	53.1	58.7	15.2	duke	42.5	55.9	61.5	16.0
MMCL	CVPR'20	market	40.8	51.8	56.7	15.1	duke	43.6	54.3	58.9	16.2
D-MMD	ECCV'20	market	29.1	46.3	54.1	13.5	duke	34.4	51.1	58.5	15.3
NRMT	ECCV'20	market	43.7	56.5	62.2	19.8	duke	45.2	57.8	63.3	20.6
JVTC	ECCV'20	market	45.4	58.4	64.3	20.3	duke	42.1	53.4	58.9	19.0
RAE (Ours)	-	market	40.1	50.2	55.2	14.8	duke	43.3	55.0	57.9	16.2
MMCL + RAE (Ours)	-	market	43.4	52.2	57.0	17.8	duke	46.6	57.2	59.2	19.2
JVTC + RAE (Ours)	-	market	47.7	60.1	66.0	22.9	duke	45.4	56.9	59.3	22.0

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART ON MSMT17. WE UTILISE MARKET-1501 OR DUKEMTMC-REID 90 AS SOURCE DATASET. OUR PROPOSED METHOD PERFORM BEST ON LARGE-SCALE RE-ID DATASET.

Methods	Target: DukeMTMC-reID					Target: Market-1501				
	Source	Rank1	Rank5	Rank10	mAP	Source	Rank1	Rank5	Rank10	mAP
supervised (upper bound)	duke	82.8	92.2	94.5	70.2	market	92.1	96.8	98.4	81.4
direct-transfer	market	28.1	43.9	48.4	14.6	duke	46.0	64.1	71.0	20.9
baseline	market	60.8	75.4	79.6	45.1	duke	70.4	83.4	87.1	51.0
baseline + FT	market	67.1	79.4	81.2	48.0	duke	79.1	90.1	93.5	62.7
baseline + PT	market	64.5	77.9	80.6	46.8	duke	74.3	88.2	92.1	48.6
baseline + FT + PT	market	69.1	81.7	82.9	50.1	duke	82.7	92.2	94.9	66.1
baseline + FT + PT + F	market	**72.4**	**84.4**	**87.7**	**53.0**	duke	**85.7**	**94.2**	**96.7**	**69.8**
baseline + FT	market	67.1	79.4	81.2	48.0	duke	79.1	90.1	93.5	62.7
baseline + $\hat{PT}$	market	63.0	76.0	88.9	44.8	duke	73.5	87.1	91.5	47.2
baseline + FT + $\hat{PT}$	market	67.0	79.7	82.9	49.1	duke	78.7	91.0	93.9	65.3
baseline + FT + $\hat{PT}$ + F	market	68.2	80.4	81.8	49.5	duke	80.9	91.8	94.7	66.8

TABLE IV
COMPONENT ANALYSIS. 'FT' DENOTES FEATURE-TRANSFER MODULE. "PT" MEANS (POST)-PIXEL-TRANSFER MODULE. "F" IS FUSION MODEL. EXPERIMENTAL RESULTS SHOW THE EFFECTIVENESS OF EVERY MODULE. "$\hat{PT}$" MEANS PRE-PIXEL-TRANSFER MODULE, PLEASE VIEW TEXT FOR MORE DETAILS.

E. Model Analysis

Components analysis. To further analyze each component of our proposed approach, we compare the whole framework with several variants. Firstly, as a baseline, a *direct-transfer* variant is conducted, where only a Encoder Enc is trained with source dataset and directly tested on target dataset. Secondly, *Baseline* means training the Encoder Enc with source dataset in a supervised way and target dataset in a self-supervised way (clustering). Besides, we analyze the "FT", "PT" and "F" modules, *i.e.* feature-transfer, pixel-transfer and fusion modules, respectively.

The mean Average Precision (mAP) scores and cumulative Matching Curve (CMC) are shown in Table IV. Firstly, the *direct transfer* setting performs only $20.1\%/22.4\%$, which indicates that the model trained on one domain cannot be straightly adapted in another one without any adaptation strategy. Secondly, the *baseline* significantly outperforms *direct transfer* by more than 30%, showing the effectiveness of the self-supervised learning in the target domain. Thirdly, when using "FT", *baseline + FT*, mAP scores improves about 3% and 10% on DukeMTMC-reID and Market-1501, respectively. This shows that GAN loss reduces the gap between source and target features. Besides, when using "PT", *baseline + PT*, mAP scores improves about 3% and 10% on DukeMTMC-reID and Market-1501, respectively. This shows that GAN loss reduces the gap between source and target features. What's more, combining feature transfer ("FT") and pixel-transfer ("PT") modules outperform either FT or PT by at least 2% and 4% mAP scores on DukeMTMC-reID and Market-1501. This verifies the complementary between "FT" and "PT". Finally, the whole framework achieves the best performance under the Fusion Module. The experimental results show the effectiveness of each component and the complementary of "FT" + "PT".

Post- *v.s.* pre-pixel-transfer learning. Existing pixel-transfer based methods first translate source images to target-style with a GAN model (e.g. CycleGAN [48], StarGAN [54]), then perform feature-transfer learning. As discussed in the Section 1, such procedure may import unexpected low-level characters of resolutions, backgrounds, and illuminations into the target domain. Different from them, to avoid such low-level noises, we first perform feature-transfer learning, then do the pixel-transfer learning in a format of decoding. We call the former as pre-pixel-transfer learning, and the latter (ours) as post-pixel-transfer learning. To verify the effectiveness of the post-pixel-transfer learning, we also report the results of pre-pixel-transfer as contrast experiments. We name the pre-pixel-transfer learning as "$\hat{PT}$". Experimental

Fig. 4. Examples of satisfactory results. The first column is an query while the rest lists the ranking results (the yellow number represents its similarity) based on our method. Green rectangles means correct matching and red ones are wrong.

Fig. 5. Examples of unsatisfactory results. The first column is an query while the rest lists the ranking results (the yellow number represents its similarity) based on our method. Green rectangles means correct matching and red ones are wrong.

results are shown in Table IV. As we can see, in all cases, replacing post-pixel-transfer learning ("PT") with pre-pixel-transfer learning ("$\hat{PT}$") significantly harm performance. For Example, in market2duke setting, $baseline + FT + \hat{PT} + F$ perform worse than $baseline + FT + PT + F$ by 4% Rank-1 score.

Parameters Analysis. There are three hyper-parameters, including λ_{tri} in Eq.(2)/Eq.(5), λ_{feat}^{gan} in Eq.(8) and λ_{pix}^{gan} in Eq.(11)/Eq.(14). We carefully analyse their effects under different values. The experimental results of are shown in Fig. 6, where the source and target datasets are DukeMTMC-reID and Market-1501, respectively. As we can see, λ_{tri} is stable to different values and reaches the best accuracy at 1.0, this satisfies the conclusion in most Re-ID works [22], [21]. λ_{feat}^{gan} and λ_{pix}^{gan} is robust when their values are smaller than 10 but perform worse when larger than 10. Thus the two parameters should be carefully searched. The results are consistent with popular GAN works [54].

We note that on all the 4 settings including Duke2Market, Market2Duke, Duke2MSMT and Market2MSMT, the values of λ_{tri}, λ_{feat}^{gan} and λ_{pix}^{gan} are unchanged, i.e., set to 1, 10 and 10, respectively. Thus, our proposed method is relatively robust to hyper-parameters.

Satisfactory & unsatisfactory results. In Fig. 4 and 5, we present some satisfactory and unsatisfactory results based on our method, respectively. From Fig. 4, we can see that nearly all listing results (including wrong results) own similar clothing styles and the same belongings with the query image. It demonstrates that our proposed method can both translate low-level information (color, illumination, view, etc) and high-level information (identity). In Fig. 5, although most listing results are wrong, their low-level information (color, illumination, view, etc) is still translated well. The reason of failure in translating identity information may be because (1) the appearance of listing images is too similar with the query to recognize them correctly and (2) there are some interference factors like occlusion, similar backgrounds, etc.

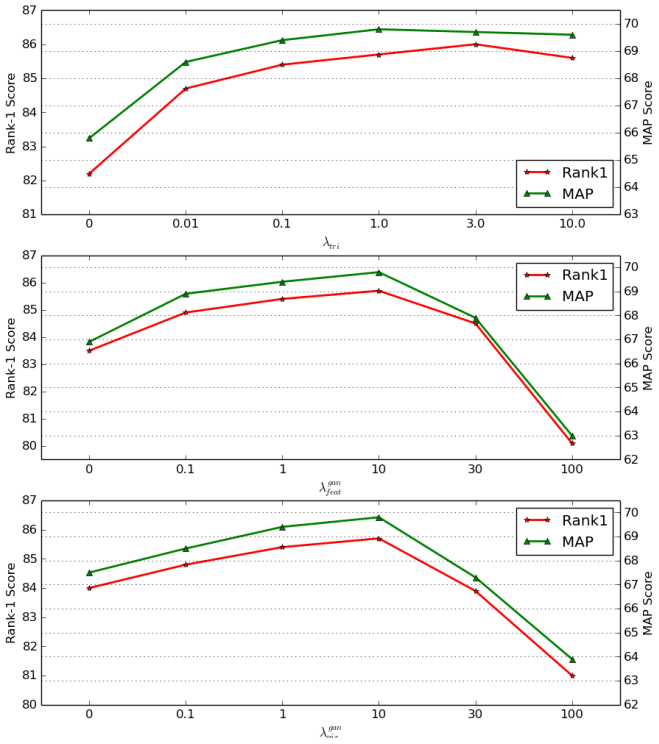


Fig. 6. Hyper-parameters analysis.

V. Conclusion

In this paper, we propose a novel idea for unsupervised cross-dataset Re-ID which unifies pixel-transfer and feature transfer learning by bilinear pooling. It performs translation learning of low-level and high-level knowledge from source domain to target domain. The source images' styles are learned and preserved in the pixel-transfer module while the identity-discriminative features are achieved in the feature-transfer module. These two kinds of domain translation information are finally enhanced in the bilinear pooling layer. Our proposed method is an end-to-end framework and simultaneously performs above-mentioned operations via an adversary strategy and Hadamard product. Extensive experiments on Market-1501, DukeMTMC-ReID and MSMT17 datasets verify the effectiveness of our method.

Acknowledgement

This work was supported in part by the National Key Research & Development Program (No. 2020YFC2003901), Chinese National Natural Science Foundation Projects #61806203, #62106264, #61872367, #61976229, #61876178. This work was also supported in part by the Academy of Finland (grants 336033, 315896), Business Finland (grant 884/31/2018), and EU H2020 (grant 101016775).

References

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[2] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3685–3693.

[3] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3200–3208.

[4] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2288–2295.

[5] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916.

[9] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," *arXiv preprint arXiv:1711.09349*, 2017.

[10] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 591–606, 2016.

[11] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 994–1002.

[12] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning." *arXiv preprint arXiv:1705.10444*, 2017.

[13] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3550–3557.

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks." *arXiv preprint arXiv:1406.2661*, 2014.

[15] 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, pp. 379–390, 2014.

[16] Y. Yang, J. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European Conference on Computer Vision*, 2014, pp. 536–551.

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2197–2206.

[18] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[19] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2288–2295.

[20] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[21] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[22] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[23] J. Wu, S. Liao, Z. Lei, X. Wang, Y. Yang, and S. Z. Li, "Clustering and dynamic sampling for unsupervised domain adaptation in person re-identification," in *ICME 2019*. IEEE, 2019.

[24] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li, "Unsupervised graph association for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8321–8330.

[25] G. Wang, Y. Yang, J. Cheng, J. Wang, and Z. Hou, "Color-sensitive person re-identification," in *IJCAI'19 Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 933–939.

[26] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, "Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3622–3631.

[27] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, "High-order information matters: Learning relation and topology for occluded person re-identification." *arXiv preprint arXiv:2003.08177*, 2020.

[28] G. Wang, T. Zhang, Y. Yang, J. Cheng, J. Chang, and Z. Hou, "Cross-modality paired-images generation for rgb-infrared person re-identification," in *AAAI 2020 : The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[29] G. Wang, Y. Yang, T. Zhang, J. Cheng, Z. Hou, P. Tiwari, H. M. Pandey *et al.*, "Cross-modality paired-images generation and augmentation for rgb-infrared person re-identification," *Neural Networks*, vol. 128, pp. 294–304, 2020.

[30] G. Wang, Q. Hu, Y. Yang, J. Cheng, and Z.-G. Hou, "Adversarial binary mutual learning for semi-supervised deep hashing," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[31] G. Wang, Q. Hu, J. Cheng, and Z. Hou, "Semi-supervised generative adversarial hashing for image retrieval," in *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 469–485.

[32] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification." *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[33] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," *computer vision and pattern recognition*, pp. 994–1003, 2018.

[34] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," *computer vision and pattern recognition*, pp. 79–88, 2018.

[35] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification." *arXiv preprint arXiv:1804.10094*, 2018.

[36] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," *computer vision and pattern recognition*, pp. 2275–2284, 2018.

[38] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," *computer vision and pattern recognition*, 2018.

[39] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. F. Wang, "Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification." *arXiv preprint arXiv:1804.09347*, 2018.

[40] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," *national conference on artificial intelligence*, pp. 2058–2065, 2016.

[41] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," *neural information processing systems*, pp. 469–477, 2016.

[42] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[43] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, 2017. [Online]. Available: https://academic.microsoft.com/paper/2949257576

[44] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated samples in person re-identification," *arXiv preprint arXiv:1801.06742*, 2018. [Online]. Available: https://academic.microsoft.com/paper/2785286326

[45] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5157–5166. [Online]. Available: https://academic.microsoft.com/paper/2963289251

[46] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero- and homogeneously," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 172–188.

[47] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[48] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[49] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool, "Pose guided person image generation," in *31st Annual Conference on Neural Information Processing Systems*, 2017, pp. 406–416. [Online]. Available: https://academic.microsoft.com/paper/2962819541

[50] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 99–108. [Online]. Available: https://academic.microsoft.com/paper/2771558241

[51] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[52] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147. [Online]. Available: https://academic.microsoft.com/paper/2963049565

[53] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," *AAAI 2019 : Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 8738–8745, 2019.

[54] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[55] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *ICLR 2017 : International Conference on Learning Representations 2017*, 2017.

[56] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[57] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *european conference on computer vision*, pp. 17–35, 2016.

[58] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[59] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7919–7929.

[60] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognition*, vol. 102, p. 107173, 2020.

[61] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Learning to adapt invariance in memory for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[62] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 2, 2019, pp. 1–8.

[63] H. Wang, G. Wang, Y. Li, D. Zhang, and L. Lin, "Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[64] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "Cr-gan: Learning complete representations for multi-view generation," *arXiv preprint arXiv:1806.11191*, 2018.

[65] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8222–8231.

[66] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6111–6120.

[67] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 981–10 990.

[68] Y. Zhai, S. Lu, Q. Ye, X. Shan, J. Chen, R. Ji, and Y. Tian, "Ad-cluster: Augmented discriminative clustering for domain adaptive person

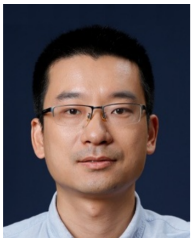
re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9021–9030.
[69] G. Wang, J.-H. Lai, W. Liang, and G. Wang, "Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[70] Z. Ji, X. Zou, X. Lin, L. Xiao, H. Tiejun, and S. Wu, "An attention-driven two-stage clustering method for unsupervised person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
[71] D. Mekhazni, A. Bhuiyan, G. Ekladious, and E. Granger, "Unsupervised domain adaptation in the dissimilarity space for person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 159–174.
[72] J. Li and S. Zhang, "Joint visual and temporal consistency for unsupervised domain adaptive person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 483–499.
[73] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1239–1248.
[74] J. Si, H. Zhang, C.-G. 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5363–5372.
[75] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, and L. Shao, "Unsupervised domain adaptation with noise resistible mutual-training for person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 526–544.

Yang Yang (Member, IEEE) received his B.S. degree and M.S. degree from Xidian University in 2009 and 2013, respectively, and the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2016, where he is currently an associate professor. His research interests are in computer vision, image processing, and machine learning, and particularly in person re-identification, attribute analysis, and face recognition. He has published more than 30 papers in international journals and conferences. He serves as a reviewer for several international journals and conferences including IEEE TPAMI, IEEE CVPR, etc. He attended the Tutorial in ECCV 2018, ICPR 2018, CCCV 2017.

Guan'an Wang received his BS degree in automation from the Central South University (CSU), China, in 2015, and his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (CASIA), China, in 2021. From October, 2019 to October, 2020, he was funded by CSC and visited to Queen Mary University of London (advised by Prof. Shaogang Gong), UK. His research interests are in computer vision, pattern recognition and person re-identification. He has published more than 10 papers in international journals and conferences, including CVPR/ICCV/ECCV/AAAI/IJCAI/TNNLS and so on. He serves as a reviewer for several international journals and conferences including IJCV/TMM/TCSVT/CVPR/ICCV and so on. More information can be found on his homepage: https://wangguanan.github.io/.

Prayag Tiwari (Member, IEEE) received his Master Degree from the National University of Science and Technology "MISiS", Moscow, Russia, and Ph.D Degree from the University of Padova, Italy. He is currently working as a Postdoctoral Researcher at the Aalto University. Previously, he was working as a Marie Curie Researcher at the University of Padova, Italy. He also worked as a research assistant at the NUST "MISiS", Moscow, Russia. He has several publications in top journals and conferences including NN, IPM, FGCS, JCP, ASOC, NCAA, IJCV, IEEE TFS, IEEE TII, IEEE IOTJ, ACM TOIT, IJIS, CIKM, SIGIR, AAAI, etc. His research interests include Machine Learning, Deep Learning, Quantum-Inspired Machine Learning, Information Retrieval, Health Informatics, and IoT.

Hari Mohan Pandey (Senior Member, IEEE) received the B.Tech. degree from Uttar Pradesh Technical University, India, the M.Tech. degree from the Narsee Monjee Institute of Management Studies, India, and the Ph.D. degree computer science and engineering from the Amity University, India. He worked as a Postdoctoral Research Fellow with the Middlesex University, London, U.K. He also worked on a European Commission project – Dream4car under H2020. He is a Senior Lecturer with the Department of Computer Science, Edge Hill University, Lancashire, U.K. He is 966 specialized in computer science and engineering and his research area includes, artificial intelligence, soft computing, natural language processing, language acquisition, machine learning, and deep learning.



Zhen Lei (Senior Member, IEEE) received the BS degree in automation from the University of Science and Technology of China, in 2005, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently a professor. He has published more than 190 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an area chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015. He was awarded 2019 IAPR YOUNG BIOMETRICS INVESTIGATOR AWARD. He is a senior member of the IEEE.