

Multi-task Mid-level Feature Alignment Network for Unsupervised Cross-Dataset Person Re-Identification

Shan Lin¹
shan.lin@warwick.ac.uk

Haoliang Li²
lihaoliang@ntu.edu.sg

Chang-Tsun Li³
chli@csu.edu.au

Alex Chichung Kot²
eackot@ntu.edu.sg

¹ Department of Computer Science
University of Warwick
United Kingdom

² Rapid-Rich Object Search Lab
Nanyang Technological University
Singapore

³ School of Computing & Mathematics
Charles Sturt University
Australia

Abstract

Most existing person re-identification (Re-ID) approaches follow a supervised learning framework, in which a large number of labelled matching pairs are required for training. Such a setting severely limits their scalability in real-world applications where no labelled samples are available during the training phase. To overcome this limitation, we develop a novel unsupervised Multi-task Mid-level Feature Alignment (MMFA) network for the unsupervised cross-dataset person re-identification task. Under the assumption that the source and target datasets share the same set of mid-level semantic attributes, our proposed model can be jointly optimised under the person's identity classification and the attribute learning task with a cross-dataset mid-level feature alignment regularisation term. In this way, the learned feature representation can be better generalised from one dataset to another which further improve the person re-identification accuracy. Experimental results on four benchmark datasets demonstrate that our proposed method outperforms the state-of-the-art baselines.

1 Introduction

Person Re-identification (Re-ID) is the problem of identifying the re-appearing person in a non-overlapping multi-camera surveillance system. Two primary tasks in person Re-ID are learning the subjects' features and developing new similarity measurements, which should be invariant to the viewpoint, pose, illumination and occlusion. Due to its potential applications in security and surveillance, person Re-ID has received substantial attention from both academia and industry. As a result, the person Re-ID performance on existing datasets has been significantly improved in the recent years. For example, the Rank-1 accuracy of a single query search on the Market1501 dataset [45] has been pushed from 44.4% [45] to 91.2% [47]. The Rank-1 accuracy of the DukeMTMC-reID dataset [46] released in 2017 has been

quickly improved from 30.8% [19] to 81.8% [80]. However, most of these approaches follow supervised learning frameworks which required a large number of manually labelled images. In real-world person Re-ID deployment, typical video surveillance systems usually consist of over one hundred cameras. Manual labelling all those cameras is a prohibitively expensive job. The limited scalability severely hinders the applicability of existing supervised Re-ID approaches in the real-world scenarios.

One solution to make a person Re-ID model scalable is designing an unsupervised algorithm for the unlabelled data. In recent years, some unsupervised methods have been proposed to extract view-invariant features and measure the similarity of images without label information [12, 52, 55, 40]. These approaches only analyse the unlabelled datasets and generally yield poor person Re-ID performance due to the lack of strong supervised tuning and optimisation. Another approach to solve the scalability issue of Re-ID is unsupervised transfer learning via domain adaptation strategy. The unsupervised domain adaptation methods leverage labelled data in one or more related source datasets (also known as source domains) to learn models for unlabelled data in a target domain. However, most domain adaptation frameworks [23, 24] assume that the source domain and target domain contain the same set of class labels. Such assumption does not hold for person Re-ID because different Re-ID datasets usually contain completely different sets of persons (classes). Therefore, most unsupervised cross-dataset Re-ID methods proposed in recent years [3, 27, 36] did not use conventional domain adaptation mechanisms. For example, [9] uses image-to-image translation to transfer the style of images in the target domain to the source domain images for generating a new training dataset. These newly generated samples which inherit the identity labels from the source domain and the image style of the target domain can be used for supervised person Re-ID learning. [56] trains two individual models: identity classification and attribute recognition and performs the domain adaptation between two models.

In our work, we rethink the assumption made for the unsupervised cross-dataset Re-ID. Although the identity labels of the source and target datasets are non-overlapping, many of the mid-level semantic features of different people such as genders, age-groups or colour and texture of the outfits are commonly shared between different people across different datasets. Hence, these mid-level visual attributes of the people can be considered as the common labels between different datasets. If we assume these mid-level semantic features are shared between the different domains, we can then treat the unsupervised cross-dataset person Re-ID as a domain adaptation transfer learning based on the mid-level semantic features from the source domain to the target domain. Therefore, we propose a **Multi-task Mid-level Feature Alignment** network (MMFA) which can simultaneously learn the feature representation from the source dataset and perform domain adaptation to the target dataset via aligning the distributions of the mid-level features. The contributions of our MMFA model are summarized below:

- We propose a novel unsupervised cross-dataset domain adaptation framework for person Re-ID by minimising the distribution variation of the source’s and the target’s mid-level features based on the Maximum Mean Discrepancy (MMD) distance [8]. Due to the low dimensionality of attribute annotations, we also include mid-level feature maps in our deep neuron network as additional latent attributes to capture a more completed representation of mid-level features of each domain. In our experiments, the proposed MMFA method surpasses other state-of-the-art unsupervised models on four popular unsupervised benchmarks datasets.
- The existing unsupervised domain adaptation Re-ID approaches based on deep learn-

ing [4, 36] require two-stage learning processes: supervised feature learning and unsupervised domain adaptation. Different from those methods, Our MMFA model introduce a new jointly training structure which simultaneously learns the feature representation from the source domain and adapts the feature to the target domain in a single training process. Because our model does not require two-step training procedure, the training time for our method is much less than many other unsupervised deep learning person Re-ID approaches.

2 Related Work

Most existing Re-ID models are supervised approaches focusing on features engineering [4, 19, 37, 43], distance metrics development [12, 19, 26, 38] or creating new deep learning architectures [4, 12, 20]. However, in real-world person Re-ID deployment, supervised methods suffer from poor scalability due to the lack of subject's identity for each camera pair. Therefore, some unsupervised person Re-ID methods have been developed based on hand-crafted features learned from a single unlabelled dataset [12, 34, 42]. However, due to the absence of the pairwise identity labels, these unsupervised methods cannot learn robust cross-view discriminative features and usually yield much weaker performance compared to the supervised learning approaches.

Because of the poor person Re-ID performance of the single dataset unsupervised learning, many of recent works are focusing on developing cross-dataset domain adaptation methods for a scalable person Re-ID system [15, 25, 27, 36]. These approaches leverage the pre-trained supervised Re-ID models and adapt these models to the target dataset. Early proposed cross-dataset person Re-ID domain adaptation approaches rely on weak label information in target dataset [15, 25]. Therefore, these methods can only be considered as semi-supervised or weakly-supervised learning. The recent cross-dataset works such as UMDL [27], SPGAN [8] and TJ-AIDL [36] do not require any labelled information from the target dataset and can be considered as fully unsupervised cross-dataset domain adaptation learning. The UMDL method tries to transfer the view-invariant feature representation via multi-task dictionary learning on both source and target datasets. The SPGAN approach uses the generative adversarial network (GAN) to generate new training dataset by transferring the image style from the target dataset to the source dataset while preserving the source identity information. Hence, the supervised training on the new translated dataset can be automatically adapted to the target domain. The TJ-AIDL approach individually trains two models: an identity classification model and an attribute recognition model. The domain adaptation in TJ-AIDL is achieved by minimising the distance between inferred attributes from the identity classification model and the predicted attributes from the attribute recognition model. Compared to the previous single dataset unsupervised approaches, the recent cross-dataset unsupervised domain adaptation methods yield much better performance. Our work improved upon these cross-dataset unsupervised methods by introducing a more intuitive domain adaptation mechanism and proposing a novel jointly training framework for simultaneous feature learning and domain adaptation.



Figure 1: In each of these three pairs of images, the one on the left-hand side is randomly selected from the Market1501 dataset while the other one shows the attention regions from highest activated feature maps (1749_{th} , 511_{th} and 1091_{th}) of the last convolutional layer. These feature maps highlight distinctive semantic features such as green shorts, red backpack, red T-shirt. Best view in colour.

3 The Proposed Methodology

One basic assumption behind domain adaptation is that there exists a feature space which underlying both the source and the target domain. Although high-level information like person’s identity is not shared between different Re-ID datasets, the mid-level features such as visual attributes can be overlapped between persons. For example, the people in dataset A and dataset B can be different, but some of mid-level semantic information like genders, age-groups, colour of clothes or accessories may be the same. Hence, in our proposed method MMFA, we assume that the source and target datasets contain the same set of mid-level attribute labels. As a result, the unsupervised cross-dataset person Re-ID can be transformed into an unsupervised domain adaptation problem by regularising the distribution variance of the attribute feature space between the source domain and the target domain.

Currently, there are a few attribute annotations available for some Re-ID datasets. However, the number of these attribute labels are limited. There are 27 attribute labels for the Market1501 dataset and 23 for the DukeMTMC-reID dataset [17]. The features from 27 or 23 user-defined attributes alone cannot give a good representation of the overall mid-level semantic features for both source and target datasets. There may exist many shared mid-level visual clues between domains which cannot fully captured by those 27/23 user-defined annotations. To obtain more attributes for representing the shared mid-level features, we start to consider the feature-maps generated from the different convolutional layers. In our experiment, we observed that most feature maps from the last convolutional layer of an attribute-identity multi-task classification model are able to capture many distinctive semantic features of a person, see Figure 1 for example. Hence, we treat those feature maps as the attribute-like mid-level deep features in our proposed MMFA model.

3.1 Architecture

Our model is optimised using stochastic gradient descent (SGD) method on mini-batches. Each mini-batch consists n_S number of labelled images $[\mathbf{I}_{S,1}, \mathbf{I}_{S,2}, \dots, \mathbf{I}_{S,n_S}]$ from a source dataset S and n_T number of unlabelled images $[\mathbf{I}_{T,1}, \mathbf{I}_{T,2}, \dots, \mathbf{I}_{T,n_T}]$ from a target dataset T . Each labelled image $\mathbf{I}_{S,i}$ is associated with an identity label $y_{S,i}$ and a set of M attributes $\mathbf{A}_{S,i} = [a_{S,i}^1, a_{S,i}^2, \dots, a_{S,i}^M]$. Our model consists of one pre-trained ResNet50-based backbone network [18] as the feature extractor with 1 fully connected layer for identity classification

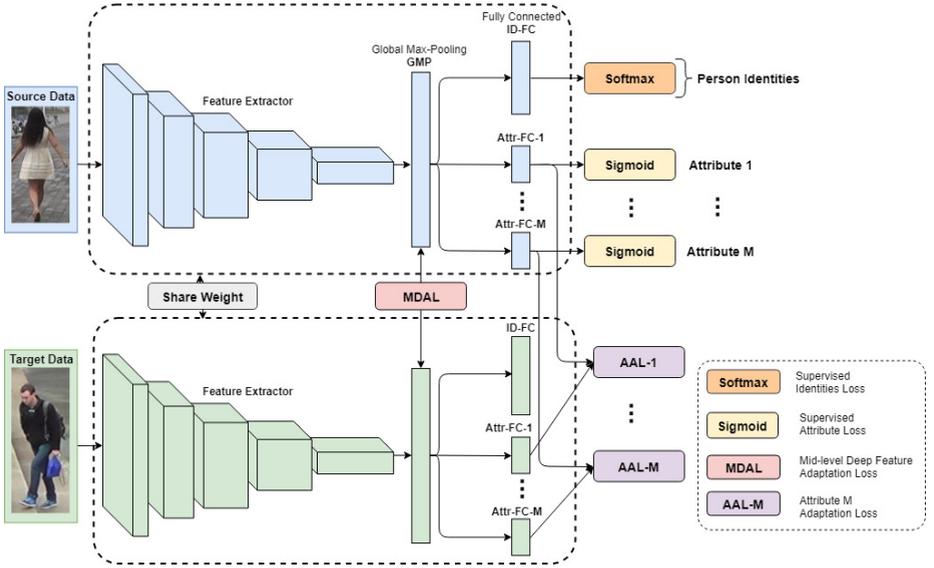


Figure 2: In the proposed MMFA model, the source and target images will undergo two networks with shared weights. The global max-pooling will extract the most sensitive feature maps from the last convolutional layer and feed them into each independent softmax classifier for classifying the identity or attributes of the person. In order to generalise the feature representation to the target dataset, we also regularised our network by aligning the distribution of the pre-defined attributes and mid-level deep features from the source to the target domain.

and M individual fully connected layers for single attribute recognition. The overview of our architecture is shown in Figure 2. We change the last average pooling layer from ResNet50 to a global max-pooling (GMP) layer to emphasise the semantic regions from the feature maps of the last convolutional layer.

$\mathbf{H}_S = [\mathbf{h}_{S,1}, \mathbf{h}_{S,2}, \dots, \mathbf{h}_{S,n_S}]$ and $\mathbf{H}_T = [\mathbf{h}_{T,1}, \mathbf{h}_{T,2}, \dots, \mathbf{h}_{T,n_T}]$ are the mid-level deep features of the source domain and the target domain obtained after the GMP layer, respectively. The identity features $\mathbf{H}_S^{id} = [\mathbf{h}_{S,1}^{id}, \mathbf{h}_{S,2}^{id}, \dots, \mathbf{h}_{S,n_S}^{id}]$ and $\mathbf{H}_T^{id} = [\mathbf{h}_{T,1}^{id}, \mathbf{h}_{T,2}^{id}, \dots, \mathbf{h}_{T,n_T}^{id}]$ are the outputs from the fully connected layer with \mathbf{H}_S and \mathbf{H}_T as input for identity classification (shown as *ID-FC* in Figure 2). For a specific m -th attribute where $m \in M$, the m -th attribute features $\mathbf{H}_S^{attr_m} = [\mathbf{h}_{S,1}^{attr_m}, \mathbf{h}_{S,2}^{attr_m}, \dots, \mathbf{h}_{S,n_S}^{attr_m}]$, $\mathbf{H}_T^{attr_m} = [\mathbf{h}_{T,1}^{attr_m}, \mathbf{h}_{T,2}^{attr_m}, \dots, \mathbf{h}_{T,n_T}^{attr_m}]$ can be obtained from its corresponding fully connected layer with \mathbf{H}_S and \mathbf{H}_T as input (shown as *Attr-FC- m* in Figure 2). Our model can be jointly trained in a multi-task manner: two supervised classification losses for identity classification and attribute recognition, one adaptation losses based on the attribute features and another adaptation loss based on the mid-level deep features.

3.2 Multi-task Supervised Classification for Feature Learning

The view-invariant feature representations are learned from a multi-task identity and attribute classification training. The additional attribute annotations provide further regularisation and additional supervision to the feature learning process

Identity Loss: We denote that $p_{id}(\mathbf{h}_{S,i}^{id}, y_{S,i})$ is the predicted probability on the identity feature $\mathbf{h}_{S,i}^{id}$ with the ground-truth label $y_{S,i}$. The identity loss is computed by softmax cross entropy function:

$$L_{id} = -\frac{1}{n_S} \sum_{i=1}^{n_S} \log(p_{id}(\mathbf{h}_{S,i}^{id}, y_{S,i})) \quad (1)$$

Attribute Loss: We denote that $p_{attr}(\mathbf{h}_{S,i}^{attr_m}, m)$ is the predicted probability for the m -th attribute feature $\mathbf{h}_{S,i}^{attr_m}$ with ground-truth label $a_{S,i}^m$. The overall attributes loss can be expressed as the average of sigmoid cross entropy loss of each attribute:

$$L_{attr} = -\frac{1}{M} \frac{1}{n_S} \sum_{m=1}^M \sum_{i=1}^{n_S} (a_{S,i}^m \cdot \log(p_{attr}(\mathbf{h}_{S,i}^{attr_m}, m)) + (1 - a_{S,i}^m) \cdot \log(1 - p_{attr}(\mathbf{h}_{S,i}^{attr_m}, m))) \quad (2)$$

3.3 MMD-based Regularisation for Mid-level Feature Alignment

As we make a shared mid-level latent space assumption in our MMFA model, the domain adaptation can be achieved by reducing the distribution distance of attribute features between the source domain and the target domain. Based on the attribute features $\{\mathbf{H}_S^{attr_1}, \dots, \mathbf{H}_S^{attr_M}\}$ and $\{\mathbf{H}_T^{attr_1}, \dots, \mathbf{H}_T^{attr_M}\}$ obtained from the supervised classification learning, we use the Maximum Mean Discrepancy (MMD) measure [9] to calculate the feature distribution distance of each attribute. The overall attribute distribution distance is the the mean MMD distance of all attributes:

$$L_{AAL} = \frac{1}{M} \sum_{m=1}^M MMD(\mathbf{H}_S^{attr_m}, \mathbf{H}_T^{attr_m})^2 = \frac{1}{M} \sum_{m=1}^M \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(\mathbf{h}_{S,i}^{attr_m}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \phi(\mathbf{h}_{T,j}^{attr_m}) \right\|_{\mathcal{H}}^2 \quad (3)$$

$\phi(\cdot)$ is a map operation which project the attribute distribution into a reproducing kernel Hilbert space (RKHS) \mathcal{H} [9]. n_S and n_T are the batch sizes of the source domain images and target domain images. The arbitrary distribution of the attribute features can be represented by using the kernel embedding technique [60]. It has been proven that if the kernel $k(\cdot, \cdot)$ is characteristic, then the mapping to the RKHS \mathcal{H} is injective [62]. The injectivity indicates that the arbitrary probability distribution is uniquely represented by an element in RKHS. Therefore, we have a kernel function $k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{T,j}^{attr_m}) = \phi(\mathbf{h}_{S,i}^{attr_m}) \phi(\mathbf{h}_{T,j}^{attr_m})^\top$ induced by $\phi(\cdot)$. Now, the average MMD distance between the source domain's and the target domain's attribute distributions can be re-expressed as:

$$L_{AAL} = \frac{1}{M} \sum_{m=1}^M \left[\frac{1}{(n_S)^2} \sum_{i=1}^{n_S} \sum_{i'=1}^{n_S} k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{S,i'}^{attr_m}) + \frac{1}{(n_T)^2} \sum_{j=1}^{n_T} \sum_{j'=1}^{n_T} k(\mathbf{h}_{T,j}^{attr_m}, \mathbf{h}_{T,j'}^{attr_m}) - \frac{2}{n_S \cdot n_T} \sum_{i=1}^{n_S} \sum_{j=1}^{n_T} k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{T,j}^{attr_m}) \right] \quad (4)$$

In our MMFA model, we decided to use the well-know RBF characteristic kernel with bandwidth α :

$$k(\mathbf{h}_{S,i}^{attr_m}, \mathbf{h}_{T,j}^{attr_m}) = \exp\left(-\frac{1}{2\alpha} \left\| \mathbf{h}_{S,i}^{attr_m} - \mathbf{h}_{T,j}^{attr_m} \right\|^2\right) \quad (5)$$

Due to the limited size of available attribute annotations, these attributes alone cannot give a good representation of all domain shared mid-level features. By assuming the last feature maps after the feature extractor are attribute-like mid-level features, we introduce the additional mid-level deep feature alignment to our model. The mid-level deep features adaptation loss L_{MDAL} is the MMD distance between the source and target mid-level deep feature $\mathbf{H}_S, \mathbf{H}_T$, similar to our attributes features adaptation loss:

$$L_{MDAL} = MMD(\mathbf{H}_S, \mathbf{H}_T)^2 = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \phi(\mathbf{h}_{S,i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \phi(\mathbf{h}_{T,j}) \right\|_{\mathcal{H}}^2 \quad (6)$$

Finally, we formulate the overall loss function by incorporating the weighted summation of above components L_{id} , L_{attr} , L_{AAL} and L_{MDAL} :

$$L_{all} = L_{id} + \lambda_1 L_{attr} + \lambda_2 L_{AAL} + \lambda_3 L_{MDAL} \quad (7)$$

4 Experiments

4.1 Datasets and Settings

Person Re-ID Datasets: Four widely used person Re-ID benchmarks are chosen for experimental evaluations: Market1501, DukeMTMC-reID, VIPeR and PRID. The Market1501 dataset [45] contains 32,668 images of 1,501 pedestrian. 751 identities are selected for training and 750 remaining identities are for testing. Each identity was captured by at most 6 non-overlapping cameras. The DukeMTMC-reID dataset [46] is the redesign version of pedestrian tracking dataset DukeMTMC [49] for person Re-ID task. It contains 34,183 image of 1,404 pedestrians. 702 identities are used for training and the remaining 702 are for testing. Each identity was captured by 8 non-overlapping cameras. The VIPeR dataset [9] is one of the oldest person Re-ID dataset. It contains 632 identities, but only two images for each identity. Due to its low resolution and large variation in illumination and viewpoints, the VIPeR dataset is still a very challenging dataset. The PRID dataset [10] consists of 934 identities from two camera views. There are 385 identities in View A and 749 identities in View B, but only 200 identities appear in both views.

Evaluation Protocol: We follow the proposed single-query evaluation protocols for Market1501 and DukeMTMC-reID [45, 46]. For the VIPeR dataset, we randomly half-split the dataset into training and testing sets. The overall performance on VIPeR is the average results from 10 randomly 50/50 split testing. For the PRID dataset evaluation, we follow the same single-shot experiments as [10]. Similar to the VIPeR dataset setting, the final performance is the average of the experimental results based on 10 random split testing. Since the VIPeR and PRID datasets are too small for training the deep learning network, our MMFA model trains on the entire Market1501 or the DukeMTMC-reID datasets. We adopted the commonly used Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP) as performance metrics.

Implementation Details: The input images are randomly cropped and resized to (256,128,3). All the fully-connected layers after global max-pooling layer are equipped with batch normalization, the dropout rate of 0.5 and the leaky RELU activation function. λ_1 , λ_2 and λ_3 in the final loss function (Equation 7) are empirically fixed to 0.1, 1, 1. For all the adaptation losses, we adopted the mixture kernel strategy [46, 48] by averaging the RBF kernels

Dataset	VIPeR	PRID	Market1501		DukeMCMT-reID	
Metric (%)	Rank-1	Rank-1	Rank-1	mAP	Rank-1	mAP
SDALF [9]	19.9	16.3	-	-	-	-
CPS [0]	22.0	-	-	-	-	-
DLLR [10]	29.6	21.1	-	-	-	-
GL [13]	33.5	25.0	-	-	-	-
ISR [12]	27.0	17.0	40.3	14.3	-	-
GTS [52]	25.2	-	-	-	-	-
SDC [42]	25.8	-	-	-	-	-
AML [39]	23.1	-	44.7	18.4	-	-
UsNCA [28]	24.3	-	45.2	18.9	-	-
CAMEL [40]	30.9	-	54.5	26.3	-	-
PUL [9]	-	-	44.7	20.1	30.4	16.4
AdaRSVM [25]	10.9	4.9	-	-	-	-
UDML [27]	31.5	24.2	-	-	-	-
SSDAL [53]	37.9	20.1	39.4	19.6	-	-
TJ-AIDL ^{Duke} [56]	35.1	<u>34.8</u>	58.2	26.5	-	-
SPGAN ^{Duke} [9]	-	-	51.1	22.8	-	-
TJ-AIDL ^{Market} [56]	<u>38.5</u>	26.8	-	-	<u>44.3</u>	<u>23.0</u>
SPGAN ^{Market} [9]	-	-	-	-	41.1	22.3
MMFA ^{Duke}	36.3	34.5	<u>56.7</u>	27.4	-	-
MMFA ^{Market}	39.1	35.1	-	-	45.3	24.7

Table 1: Performance comparisons with state-of-the-art unsupervised person Re-ID methods. The best and second best results are highlighted by bold and underline respectively. The superscripts: *Duke* and *Market* indicate the source dataset which the model is trained on.

with the bandwidth $\alpha = 1, 5, 10$. We use the stochastic gradient descent (SGD) optimizer with batch size 32 for both source domain images and the target domain images. We set the learning rate to 0.01 and the nesterov momentum to 0.9 with the weight decay of 5×10^{-4} . The learning rate will decrease by 10 after the 20-th epoch. The person Re-ID evaluation of the target domain is measured by the L_2 distance of the 2048-D mid-level deep features H_T after the global max-pooling layer.

4.2 Comparisons with State-of-the-art Methods

The performance of our proposed MMFA model is extensively compared with 16 state-of-the-art unsupervised person Re-ID methods as shown in Table 1. These methods include: view-invariant feature learning methods SDALF [9] and CPS [0], graph learning method GL [13], sparse ranking method ISR [12], saliency learning methods GTS [52] and SDC [42], neighbourhood clustering methods AML [39], UsNCA [28], CAMEL [40] and PUL [9], ranking SVM method AdaRSVM [25], attribute co-training method SSDAL [53], dictionary learning method DLLR [10] and UDML [27], id-to-attribute transfer method TJ-AIDL [56] and image style transfer method SPGAN [9]. These methods can be categorised into three groups:

1. hand-craft features approaches: SDALF, CPS, DLLR, GL, ISR, GTS, SDC
2. clustering approaches: AML, UsNCA, CAMEL, PUL
3. domain adaptation approaches: AdaRSVM, UDML, SSDAL, TJ-AIDL, SPGAN

Our MMFA method outperforms most existing state-of-the-art models on VIPeR, PRID, Market1501 and DukeMTMC-reID datasets. the Rank-1 accuracy increases from 38.5% to

39.1% in VIPeR, from 34.8% to 35.1% in PRID and from 44.3% to 45.3% in DukeMTMC-reID. The mAP performance of our approach surpasses all exiting methods by a good margin from 23.0% to 24.7% and 26.5% to 27.4% in DukeMTMC-reID and Market1501 receptively. Although, the Rank-1 accuracy of our MMFA model on the Maket1501 dataset did not surpass the TJ-AIDL method, our mAP score and the overall performance (Rank-5 to Rank-10 accuracy) are better than TJ-AIDL. The complete comparisons with TH-AIDL and SPGAN are shown in Table 2. It is worth noting that the performance of our MMFA is achieved in one single end-to-end training session with only 25 epochs. Our performance can be further improved by implementing any pre- and post-processing techniques such as part-based local max pooling (LMP), attention mechanisms or re-ranking. For fair comparisons, the performance results shown the Table 1 and Table 2 are all based on the basic models without any pre or post-processing.

Source → Target	Market1501 → DukeMTMC-reID				DukeMTMC-reID → Market1501			
Metric (%)	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
SPGAN	41.1	56.6	63.0	22.3	51.5	70.1	76.8	22.8
TJ-AIDL	44.3	59.6	65.0	23.0	58.2	74.8	81.1	26.5
MMFA	45.3	59.8	66.3	24.7	<u>56.7</u>	75.0	81.8	27.4

Table 2: Detail Comparison with SPGAN and TJ-AIDL

4.3 Component Analysis and Evaluation

We also analysed each component of our MMFA model based on their contributions to the cross-domain feature learning. The first set of experiments is the unsupervised performance based on the feature representation learned from the source domain attributes or identity, without any domain adaptation. In the top section of Table 3, the attribute annotations alone cannot give a good representation of a person due to its low dimensionality, only 6.4% and 19.2% Rank1 accuracy achieved. The features from identity labels on the other hand yield much better performance compared to attributes. When attribute and identity information are jointly trained as a multi-objective learning task, the feature representations show a better generalization-ability. This experiment shows that the attribute annotations do provide extra information to the system which serves as additional supervision for learning more generalised cross-dataset features.

Source → Target	Market1501 → DukeMTMC-reID				DukeMTMC-reID → Market1501			
Metric (%)	Rank1	Rank5	Rank10	mAP	Rank1	Rank5	Rank10	mAP
Attribute Only	6.4	14.4	18.6	2.3	19.2	34.8	45.1	6.2
ID Only	37.6	54.9	61.6	20.6	48.2	66.1	73.3	21.6
Attribute+ID Only	41.7	57.5	63.6	23.3	52.2	69.1	75.7	23.5
Attribute with Attribute Feature Adaptation	15.8	26.0	48.2	5.7	35.5	55.3	64.0	12.7
ID with Mid-level Deep Feature Adaptation	42.1	57.7	63.9	24.3	53.4	70.2	76.4	25.2
Mid-level Deep Feature + Attribute Adaptation	45.3	59.8	66.3	24.7	56.7	75.0	81.8	27.4

Table 3: Adaptation performance on each model components

The lower section of Table 3 shows the unsupervised re-id performances after aligning the mid-level feature distribution. After aligning the source and target distributes of attributes features, mid-level features or both, we can see a large performance increase when compared with previously non-adapted features. It shows that the proposed mid-level feature distribution alignment strategy is a feasible approach for the unsupervised person Re-ID task.

5 Conclusion

In this paper, we presented a novel unsupervised cross-dataset feature learning and domain adaptation framework MMFA for person Re-ID task. We utilised the multi-supervision identity and attribute classifications to learn a discriminative feature for person Re-ID on the labelled source dataset. With a shared mid-level feature space assumption, we proposed the mid-level feature alignment domain adaptation strategy to reduce the MMD distance based on the source domain's and the target domain's mid-level feature distributions. In contrast to most existing learn-then-adapt unsupervised cross-dataset approaches, our MMFA is a one-step learn-and-adapt method which can simultaneously learn the feature representation and adapt to the target domain in a single end-to-end training procedure. Meanwhile, our proposed method is still able to outperform a wide range of state-of-the-art unsupervised Re-ID methods.

Acknowledgement: This work is supported by EU Horizon 2020 project, entitled Computer Vision Enable Multimedia Forensics and People Identification (acronym:IDENTITY, Project ID:690907) and carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore.

References

- [1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An Improved Deep Learning Architecture for Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [2] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conference (BMVC)*, 2011.
- [3] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [4] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised Person Re-identification: Clustering and Fine-tuning. *arXiv preprint*, 2017.
- [5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person Re-Identification by Symmetry-Driven Accumulation of Local Features. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [6] Doug Gray, Shane Brennan, and Hai Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 41–47, 2007.
- [7] Douglas Gray and Hai Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *European Conference on Computer Vision (ECCV)*, 2008.

- [8] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur. A Fast, Consistent Kernel Two-Sample Test. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.
- [9] Arthur Gretton, Karsten Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alexander J. Smola. A Kernel Method for the Two-Sample Problem. *Journal of Machine Learning Research (JMLR)*, 2008.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [11] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person Re-identification by Descriptive and Discriminative Classification. In *Scandinavian Conference on Image Analysis (SCIA)*, 2011.
- [12] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Dictionary Learning with Iterative Laplacian Regularisation for Unsupervised Person Re-identification. In *British Machine Vision Conference (BMVC)*, 2015.
- [13] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Person Re-identification by Unsupervised L1 Graph Learning. In *European Conference on Computer Vision (ECCV)*, 2016.
- [14] Martin Kostinger, Martin Hirzer, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Large Scale Metric Learning from Equivalence Constraints. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [15] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Domain Transfer for Person Re-identification. In *International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream (ARTEMIS)*, 2013.
- [16] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain Generalization with Adversarial Feature Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious Attention Network for Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [18] Yujia Li, Kevin Swersky, and Richard Zemel. Generative Moment Matching Networks. *International Conference for Machine Learning (ICML)*, 2015.
- [19] Shengcai Liao, Yang Hu, Xiangyu Zhu, Stan Z. Li, Xiangyu Zhu, and Stan Z. Li. Person Re-identification by Local Maximal Occurrence Representation and Metric Learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [20] Shan Lin and Chang-tsun Li. End-to-End Correspondence and Relationship Learning of Mid-Level Deep Features for Person Re-Identification. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017.
- [21] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving Person Re-identification by Attribute and Identity Learning. In *arXiv preprint*, 2017.

- [22] Giuseppe Lisanti, Iacopo Masi, Andrew D. Bagdanov, and Alberto Del Bimbo. Person Re-Identification by Iterative Re-Weighted Sparse Ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.
- [23] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *International Conference for Machine Learning (ICML)*, 37, 2015.
- [24] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep Transfer Learning with Joint Adaptation Networks. In *International Conference for Machine Learning (ICML)*, 2017.
- [25] Andy J Ma, Jiawei Li, Pong C Yuen, and Ping Li. Cross-Domain Person Reidentification Using Domain Adaptation Ranking SVMs. *IEEE Transactions on Image Processing (TIP)*, 5 2015.
- [26] Sakrapee Paisitkiangkrai, Chunhua Shen, and Anton van den Hengel. Learning to Rank in Person Re-identification with Metric Ensembles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [27] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised Cross-Dataset Transfer Learning for Person Re-identification. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [28] Chen Qin, Shiji Song, Gao Huang, and Lei Zhu. Unsupervised Neighborhood Component Analysis for Clustering. *Neurocomputing*, 168:609–617, 2015.
- [29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision Workshops (ECCVW)*, 2016.
- [30] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [31] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert Space Embedding for Distributions. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.
- [32] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R. G. Lanckriet, and Bernhard Schölkopf. Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.
- [33] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep Attributes Driven Multi-camera Person Re-identification. In *European Conference on Computer Vision (ECCV)*, 2016.
- [34] Hanxiao Wang, Shaogang Gong, and Tao Xiang. Unsupervised Learning of Generative Topic Saliency for Person Re-identification. In *British Machine Vision Conference (BMVC)*, 2014.

- [35] Hanxiao Wang, Xiatian Zhu, Tao Xiang, and Shaogang Gong. Towards Unsupervised Open-Set Person Re-identification. In *International Conference on Image Processing (ICIP)*, 2016.
- [36] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-Identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [37] Ke Yan, Yaowei Wang, Dawei Liang, Tiejun Huang, and Yonghong Tian. CNN vs. SIFT for Image Retrieval: Alternative or Complementary? In *ACM International Conference on Multimedia (ACM MM)*, 2016.
- [38] Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Exploiting Multi-grain Ranking Constraints for Precisely Searching Visually-similar Vehicles. In *International Conference on Computer Vision (ICCV)*, 2017.
- [39] Jieping Ye, Zheng Zhao, and Huan Liu. Adaptive Distance Metric Learning for Clustering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [40] Hong-Xing Xing Yu, Ancong Wu, and Wei-Shi Shi Zheng. Cross-View Asymmetric Metric Learning for Unsupervised Person Re-Identification. In *International Conference on Computer Vision (ICCV)*, 2017.
- [41] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a Discriminative Null Space for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [42] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised Saliency Learning for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [43] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning Mid-level Filters for Person Re-identification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [44] Rui Zhao, Wanli Ouyang, Xiaogang Wang, Wanli Ouyang, and Xiaogang Wang. Person Re-Identification by Saliency Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(2):356–370, 12 2017.
- [45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *International Conference on Computer Vision (ICCV)*, 2015.
- [46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. In *International Conference on Computer Vision (ICCV)*, 2017.