Multi-Domain Adversarial Feature Generalization for Person Re-Identification

Shan Lin, *Member, IEEE*, Chang-Tsun Li, *Senior Member, IEEE*, and Alex C. Kot, *Fellow, IEEE*

***Abstract*— With the assistance of sophisticated training methods applied to single labeled datasets, the performance of fully-supervised person re-identification (Person Re-ID) has been improved significantly in recent years. However, these models trained on a single dataset usually suffer from considerable performance degradation when applied to videos of a different camera network. To make Person Re-ID systems more practical and scalable, several cross-dataset domain adaptation methods have been proposed, which achieve high performance without the labeled data from the target domain. However, these approaches still require the unlabeled data of the target domain during the training process, making them impractical. A practical Person Re-ID system pre-trained on other datasets should start running immediately after deployment on a new site without having to wait until sufficient images or videos are collected and the pre-trained model is tuned. To serve this purpose, in this paper, we reformulate person re-identification as a multi-dataset domain generalization problem. We propose a multi-dataset feature generalization network (MMFA-AAE), which is capable of learning a universal domain-invariant feature representation from multiple labeled datasets and generalizing it to 'unseen' camera systems. The network is based on an adversarial auto-encoder to learn a generalized domain-invariant latent feature representation with the Maximum Mean Discrepancy (MMD) measure to align the distributions across multiple domains. Extensive experiments demonstrate the effectiveness of the proposed method. Our MMFA-AAE approach not only outperforms most of the domain generalization Person Re-ID methods, but also surpasses many state-of-the-art supervised methods and unsupervised domain adaptation methods by a large margin.**

***Index Terms*— Person re-identification, domain generalization, video surveillance, adversarial feature learning.**

I. Introduction

RE-IDENTIFYING a person in CCTV surveillance systems, also known as Person Re-ID, is a critical but also labor-intensive task. In recent years, the computer vision community has proposed various methods to automatically identify re-appearing people in multi-camera surveillance systems. Most of these proposed approaches are modeled, trained, and tested on the same dataset collected from a very small camera network [1]–[8]. However, a real-world CCTV system usually consists of tens to hundreds of cameras. Detecting, extracting and annotating thousands of people across hundreds of cameras is an extremely challenging task. Hence, using the actual annotated data collected from every target surveillance camera to train a fully-supervised Person Re-ID model is not a practical approach. Besides, most conventional supervised single-dataset models often over-fit to specific datasets (camera networks). Once these supervised models are trained on a given dataset, they usually suffer from considerable performance degradation when applied to a different camera network. Table I illustrates the performance of a simple supervised model tested on the same dataset and a different dataset. This model uses the ResNet50 network [9] as the feature extraction backbone. In Table I, the model trained on the Market-1501 dataset [10] can achieve 91.6 % Rank 1 accuracy with 78.7% mAP score when tested on the same dataset. However, it can only achieve 37.6% Rank 1 and 22.6% mAP when tested on the DukeMTMC-reID dataset [11]. The performance of the model trained on the DukeMTMC-reID dataset also drops from 83.4% Rank-1 with 66.6% mAP to only 48.2% Rank 1 with 21.6% mAP when tested on the Market-1501 dataset. This suggests that models trained on a single dataset are prone to over-fitting and have poor generalization performance.

TABLE I

The Performance Degradation of the Single-Dataset-Trained Baseline Model (ResNet50) When Tested on a Different Dataset

Training Dataset	**Testing Dataset**			
	Market1501		DukeMTMC-reID	
	Rank 1	mAP	Rank 1	mAP
Market-1501	91.6%	78.7%	37.6%	22.6%
DukeMTMC-reID	48.2 %	21.6%	83.4%	66.6%

The weak generalization capacity and poor scalability in most single-dataset-trained models severely hinder the real-world deployment of Person Re-ID systems. Different datasets are often collected in very different environments (e.g., indoors/outdoors, summer/winter, daytime/nighttime). If we consider each dataset (camera system) as a domain, there are often large domain gaps between datasets. Hence, recent researches focus on unsupervised cross-dataset *domain adaptation* (DA) for Person Re-ID [12]–[15] to obviate the need for annotating the images from new camera systems. These cross-dataset DA methods aim to adapt a model trained on

Manuscript received April 30, 2020; revised September 2, 2020, October 24, 2020, and December 17, 2020; accepted December 17, 2020. Date of publication December 31, 2020; date of current version January 7, 2021. This work was supported in part by the EU Horizon 2020 Marie Sklodowska-Curie Actions through the project entitled Computer Vision Enabled Multimedia Forensics and People Identification (Project No. 690907, Acronym: IDENTITY); and in part by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG-100E-2018-018) with the Defence Science and Technology Agency, Singapore, under the project agreement No. DST000ECI19300431. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Liang Wang. *(Corresponding author: Chang-Tsun Li.)*

Shan Lin is with the Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K.

Chang-Tsun Li is with the Department of Information Technology, Deakin University, Waurn Ponds, VIC 3216, Australia (e-mail: changtsun.li@deakin.edu.au).

Alex C. Kot is with the School of EEE, Nanyang Technological University, Singapore 639798.

Digital Object Identifier 10.1109/TIP.2020.3046864

an annotated source dataset to an unlabeled target dataset by image translation, feature alignment, or multi-task learning. By transferring the domain-specific knowledge, the cross-dataset domain adaptation methods do not require labeled (i.e., annotated) data of the target domain. However, for the DA approaches to be effective, the following two issues are yet to be resolved.

1) *Generalization Issue*: Cross-dataset domain adaptation methods require a large amount of unlabeled data from the target network prior to model adaptation training. However, it may not be known in advance where the model would be deployed. The unlabeled data collection also takes time even if the target site is known, especially when images/videos of all four seasons are required. The additional data collection process will delay the system deployment.
2) *Scalability Issue*: Cross-dataset domain adaptation methods require the training and fine-tuning of a bespoke model for every new camera network. The training or fine-tuning for a new model may take from hours to days, depending on the system scale. Besides, the scales and configurations of CCTV systems may not always be constant. More new cameras may be added to the system to meet the ever-changing demands. Such changes require the model to be re-trained in order to accommodate the new cameras.

In this paper, we address the generalization and scalability issues of Person Re-ID from a different perspective. Since no single dataset can cover all possible backgrounds and imaging conditions, we decide to learn a universal feature representation from multiple datasets. In recent years, many large-scale Person Re-ID datasets such as CUHK02 [16], CUHK03 [17], Market-1501 [10], DukeMTMC-reID [11], MSMT17 [18], RAP [19], and CUHK-SYSU [20] have been collected. They cover a wide variety of visual scenes with various camera settings. Each dataset can be considered as a different surveillance system representing a different domain. Therefore, we reformulate Person Re-ID as a *domain generalization* (DG) problem, in which we train a model from multiple existing datasets without any prior knowledge of the target system (i.e., no domain adaptation). We aim to develop a domain generalization model that can leverage the labeled images from multiple datasets to learn a domain-invariant feature representation. Domain generalization applied to the feature learning on these datasets helps learn a representation that can be relatively well generalized to any unseen surveillance system. This setting simulates the real-world scenario, in which a strong feature learner only needs to be trained on multiple datasets once and can be deployed to new camera networks without further data collection or adaptation training.

However, due to its challenging nature, few methods have attempted the domain generalization setting [21]–[23]. The recent DIMN [22] sets a standard training and evaluation procedure for the multi-dataset domain generalization for Person Re-ID. The DIMM method is based on a complicated meta-learning procedure. However, the dynamic model synthesis during the testing process makes the DIMN model relatively slow and cumbersome. The DualNorm method [23] uses a domain style normalization by performing instance normalization (IN) in the early layers of the feature extractor networks such as MobileNet and ResNet. The DualNorm method is efficient and can be integrated into most of the existing Person Re-ID methods. However, it does not fully utilize the domain label for training.

In this paper, we proposed a novel framework for domain generalization, which aims to learn a universal representation via domain-based adversarial learning while aligning the distribution of mid-level features between them. Our proposed framework can be considered as an extension of our **M**ulti-task **M**id-level **F**eature **A**lignment (MMFA) network [14] in a multiple domain learning setting. We called it MMFA with **A**dversarial **A**uto-**E**ncoder (MMFA-AAE). Our MMFA-AAE can simultaneously minimize the losses of data reconstruction, identity, and triplet loss. It alleviates the domain difference via adversarial training and also matches the distribution of mid-level features across multiple datasets. Our contributions can be summarized as follows.

1) We propose an effective feature generalization mechanism utilizing domain-based adversarial learning. We introduce an additional feature distribution alignment (i.e., Maximum Mean Discrepancy [24]) to regularize the feature learning process. By integrating the adversarial auto-encoder [25] and Maximum Mean Discrepancy (MMD) alignment, our MMFA-AAE architecture is capable of extracting domain-invariant features from multiple source datasets and generalize the features to unseen target domains (datasets).
2) The proposed MMFA-AAE method not only demonstrates the state-of-the-art performance on the multi-dataset domain generalization setting but also surpasses many domain adaptation Person Re-ID methods.
3) Unlike the DIMN [22] and DualNorm methods [23], our MMFA-AAE reduces the dimension of the feature vectors to only 512 without affecting the overall performance. It can significantly shorten the subject retrieval time and reduce the storage requirement for saving the processed features.
4) Our domain-based adversarial learning sub-network can be easily integrated into most existing Person Re-ID methods. It can help to boost the generalization capacity of the existing Person Re-ID models.

II. Related Work

A. Single-Dataset Person Re-ID

In recent years, Person Re-ID methods are often based on deep convolutional neural networks. Early deep learning based approaches are developed based the Siamese architecture [7], [17], [26], [27] to learn the the corresponding regions matching between two input images. The recent methods [5], [28], [29] are usually consisted of both softmax classification loss and triplet verification loss. The latest approaches, such as DCC [8] and DuATM [30], utilize the attention mechanism to further boost the Person Re-ID performance. Most of these methods are trained and tested on a single dataset to evaluate their performance. However, different datasets are collected from different cameras under different imaging conditions. It has been noted that these supervised single-dataset methods often over-fit to the training dataset and generalize poorly when tested on other unseen datasets. Collecting a labeled dataset for a new camera system is an expensive and time-consuming task. Hence, many recent methods are focusing on cross-dataset domain adaptation (DA) learning [12]–[15], [31]–[33].

B. Cross-Dataset Domain Adaption Person Re-ID

DA approaches assume that there exists a massive amount of unlabeled data obtained from the target camera system (also known as the *target domain*). These approaches utilize the information extracted from the unlabeled target domain data to help the models trained on the source domain to adapt to the target domain. Early proposed cross-dataset DA approaches rely on weak label information in the target dataset [31], [32]. Therefore, these methods can only be considered as semi-supervised or weakly-supervised learning. Recent cross-dataset works, such as UMDL [33], SPGAN [12], TJ-AIDL [13], MMFA [14], do not require any labeled 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 the source dataset and the target dataset. The SPGAN approach uses the generative adversarial network (GAN) to generate a 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 transferred dataset can automatically adapt to the target domain. The TJ-AIDL approach individually trains two models: an identity classification model and an attribute recognition model. Domain adaptation in TJ-AIDL is achieved by minimizing the distance between the inferred attributes from the identity classification model and the predicted attributes from the attribute recognition model. Similar to TJ-AIDL, the MMFA network is jointly optimized through people identity classification and attribute learning with cross-dataset mid-level feature alignment regularization. In this way, the learned feature representation can be better generalized from one dataset to another. In [15], the Hetero-Homogeneous Learning (HHL) method improves the capability of generalization to target datasets by achieving camera invariance and domain connectedness simultaneously. The BUC [34] and the PurifyNet [35], on the other hand, try to estimate labels for the target domain dataset. Compared to previous unsupervised single dataset approaches, recent unsupervised cross-dataset domain adaptation methods yield much better performance. Although DA approaches do not require labeled data from the target domain, they do require a large amount of unlabeled data from the target domain to facilitate the adaptation. They require additional time for data collection and model adaptation, which will delay the system deployment. Compared to DA methods, domain generalization (DG) approaches are relatively more practical in real-world applications.

C. Multi-Dataset Domain Generalization Person Re-ID

Multi-dataset domain generalization (DG) methods aim to learn a universal domain-invariant feature representation that is robust to various domain-shift across different datasets (camera systems). As a result, a domain generalization model can be incorporated into a new surveillance system without fine-tuning and adaptation. In the Person Re-ID research community, only a few works focus on multi-domain generalization [21]–[23]. The Domain Guided Dropout (DGD) method [21] is the first multi-dataset domain generalization work. By removing the domain-specific neurons, the DGD method achieves multi-domain generalization by only utilizing the neurons that are effective across all domains. The DGD method is only trained on several small Person Re-ID datasets such as CUHK01 [36] and CUHK03 [17]. It only performs the evaluation on the same dataset without considering the cross-dataset situation. DIMN [22] proposed recently is trained on 5 large datasets (CUHK02 [16], CUHK03 [17], Market-1501 [10], DukeMTMC-reID [11], and CUHK-SYSU [20]) and tested on 4 small benchmarks (VIPeR [37], PRID [38], GRID [39],and i-LIDS [40]). The DIMN method [22] follows the meta-learning approach [20]. Different from the common way of using feature distances for matching scores, DIMN generates classifier weights from gallery images and then takes the inner product of the classifier weights and probe image features to calculate matching scores. This meta-learning pipeline makes the model domain-invariant, but the complicated meta-learning procedure makes optimization difficult. In addition, classifier weight generation during testing slows down the speed of model inference. Considering these drawbacks, a simpler approach, DualNorm [23], that utilizes normalization was proposed. Unlike DIMN, the DualNorm method focuses on learning domain-invariant features. It regards the style and content variations as the cause of domain bias and suppresses them by inserting instance normalization (IN) [41] in the early layers and a batch normalization (BN) [42] to a feature extraction layer. Both DIMN and DualNorm only use the person identity labels during the model training. They do not fully utilize domain labels (dataset labels). In our proposed MMFA-AAE method, we use the MMD-based [24] adversarial domain learning to suppress the domain-specific information.

III. The Proposed Methodology

Domain Aggregation Baseline: To evaluate the effectiveness of the proposed MMFA-AAE model, we first build a Person Re-ID model to serve as the baseline reference. This baseline model uses MobileNetV2 [43] and ResNet50 [9] as the backbone. We keep the default structure of the backbone and only change the dimension of the last classification layer (fully connected layer) to the total number of identities. Similar to [22], [23], the baseline model is trained on labeled images aggregated from multiple source domains. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be the extracted feature vectors (feature embeddings) from the backbone network with batch size n and $\mathbf{Y} = \left[\mathbf{y}_1, \ldots, \mathbf{y}_n\right]$ the corresponding person identity label set of $\mathbf{X}$. The mini-batch contains samples randomly selected from all source domains. The baseline model is pre-trained on ImageNet [44] and jointly optimized with the cross-entropy loss $\mathcal{L}_{id}$ for identity classification and the triplet loss $\mathcal{L}_{tri}$ for people verification. $p_{id}(\mathbf{x}_i, \mathbf{y}_i)$ denotes the predicted probability that feature vector $\mathbf{x}_i$ belongs to person identity $\mathbf{y}_i$.

The identity loss $\mathcal{L}_{id}$ can be expressed as

$$\mathcal{L}_{id} = \frac{1}{n}\sum_{i=1}^{n} log(p_{id}(\mathbf{x}_i, \mathbf{y}_i)). \quad (1)$$

In every mini-batch, the images can be divided into three groups, anchor images, positive pairs to the anchor and negative pairs to the anchor. The feature embeddings $\mathbf{X}_a$ of the images of a person are used as the *anchor* of the triplet. $\mathbf{X}_p$ denotes the different feature embeddings of the same person of the anchor image (positive pairs to the anchor image). $\mathbf{X}_n$ denotes the feature embeddings of different people (negative pairs to the anchor image). The training process encourages the model to make the l_2 distance between the positive pair $d_{ap} = d(\mathbf{X}_a, \mathbf{X}_p)$ smaller than the negative pair $d_{an} = d(\mathbf{X}_a, \mathbf{X}_n)$ by a margin α_1. The triplet loss function $\mathcal{L}_{tri}$ of one triplet can be defined as

$$\begin{aligned}\mathcal{L}_{tri} &= \max\left\{0, d_{ap} - d_{an} + \alpha_1\right\}\\ &= \max\left\{0, d(\mathbf{X}_a, \mathbf{X}_p) - d(\mathbf{X}_a, \mathbf{X}_n) + \alpha_1\right\}. \quad (2)\end{aligned}$$

Our baseline model follows the same settings of most triplet-based models with the distance margin α set to 0.3. The overall loss for the baseline $\mathcal{L}_{baseline}$ will be the summation of the cross-entropy loss and the triplet loss:

$$\mathcal{L}_{baseline} = \mathcal{L}_{id} + \mathcal{L}_{tri} \quad (3)$$

We use the Euclidean distance of the extracted feature vectors from the baseline network to perform person retrieval evaluation. The performance of the baseline model (ResNet50) on a single dataset setting is shown in Table I.

MMFA-AAE Network: Most domain generalization methods assume that there exists a common feature space that is able to span both seen source domains and unseen target domains. If a model can extract features from this common feature space, it is able to generalize well to other unseen domains. In order to find this feature space, we extend our previous work **M**ulti-task **M**id-level **F**eature **A**lignment network (MMFA) with an additional **A**dversarial **A**uto-**E**ncoder (AAE) [25] to the multi-domain setting. We call it **MMFA** with **A**dversarial **A**uto-**E**ncoder (MMFA-AAE). The proposed method aims to learn a model from multiple labeled datasets and removes the domain-specific information via domain-based adversarial learning. The proposed network also minimizes the mid-level feature distribution variance based on the MMD distance [24]. In this section, we describe how the proposed MMFA-AAE network is designed for domain generalization.

A. Architecture

The architecture of the proposed MMFA-AAE network is shown in Figure 1. In this model, we add several components to the baseline proposed at the beginning of this section. The images from multiple domains will be the inputs for the same backbone networks (MobileNetV2 [43] or ResNet50 [9]) with shared weights. The feature vectors extracted from the backbone network will then be passed on to an adversarial auto-encoder [25]. The auto-encoder [45] aims to map the feature vectors $\mathbf{X}$ from different datasets to a common latent space (hidden codes $\mathbf{H}$). The hidden codes from the auto-encoder serve as new compressed feature vectors, which are used for supervised feature training and domain discrimination. The domain discriminator determines the dataset from which the feature vector is drawn. By using the strong domain discriminator to train the feature extractor in an adversarial manner, the MMFA-AAE network aims to produce a domain-invariant latent space among multiple domains (multiple datasets). In order to further generalize the feature representation across multiple domains, we follow our previous MMFA method [14], which uses Maximum Mean Discrepancy (MMD) [24] regularization to align the distribution of the extracted deep features between different domains. In the following section, we will describe how the proposed MMFA-AAE network generalizes the feature representation from multiple domains.

B. Instant Normalization

From recent studies on the generative adversarial network (GAN), especially in the style transformation area [46], [47], it is observed that some image style information can be encoded in the mean and variance of the convolutional feature maps inside the network [46]. Hence, the instance normalization (IN) [41], which performs the normalization on a single image across all channels, can potentially eliminate the appearance divergence caused by style variation [47]. Therefore, the IBN-Net was proposed to enhance the generalization capability of the network for various computer vision tasks [47]. The DualNorm method [23] applied this technique to the Person Re-ID problem and boosted the identification performance in the multi-dataset domain generalization setting. Hence, our MMFA-AAE network adopts the same setting as in [23] and applies the IN in the first 6 blocks in MobileNetV2 and the first 4 blocks in ResNet50.

C. MMD-Regularized Adversarial Auto-Encoder

1) Reconstruction Loss: In the domain adversarial auto-encoder of our MMFA-AAE network, we use an encoder $Q(\mathbf{X})$ to map the feature embeddings $\mathbf{X}$ to the hidden codes $\mathbf{H}$ (i.e., $\mathbf{H} = Q(\mathbf{X})$) and a decoder $P(\mathbf{H})$ to reconstruct the feature embeddings $\mathbf{X}$ from the hidden codes $\mathbf{H}$. The encoder-decoder pair is shared across all the domains. The reconstructed feature embedding is denoted as $P(Q(\mathbf{X}))$ and the reconstruction loss of the auto-encoder is defined as

$$\mathcal{L}_{\text{rec}} = \|\mathbf{X} - P(Q(\mathbf{X}))\|_2^2 \quad (4)$$

2) Identity Loss: In our MMFA-AAE network, $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$ is a set of corresponding hidden codes of $\mathbf{X}$., serves as new compressed feature vectors for supervised feature training. Hence, the identity loss $\mathcal{L}_{id}$ for the network (cf. Eq. 1) can be expressed as:

$$\begin{aligned}\mathcal{L}_{id} &= \frac{1}{n}\sum_{i=1}^{n} log(p_{id}(\mathbf{h}_i, \mathbf{y}_i))\\ &= \frac{1}{n}\sum_{i=1}^{n} log(p_{id}(Q(\mathbf{x}_i), \mathbf{y}_i)). \quad (5)\end{aligned}$$

3) Triplet Loss: Let $\mathbf{H}_a$, $\mathbf{H}_p$, and $\mathbf{H}_n$ denote the hidden codes of $\mathbf{X}_a$, $\mathbf{X}_p$, and $\mathbf{X}_n$, respectively. The triplet verification loss $\mathcal{L}_{tri}$ for our MMFA-AAE network (cf. Eq. 2) can be

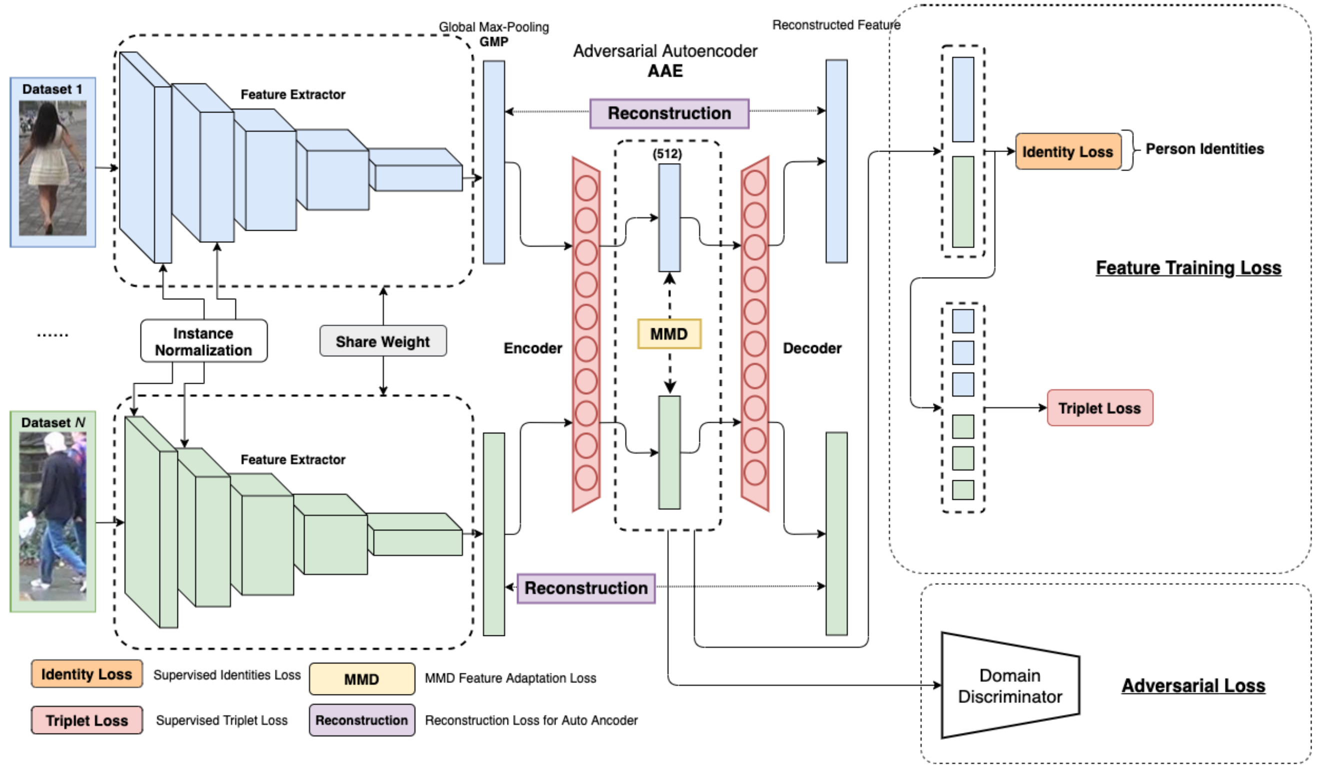


Fig. 1. An overview of MMFA-AAE framework for Person Re-ID multi-domain generalization.

expressed as

$$\begin{aligned}\mathcal{L}_{tri} &= \max\left\{0, d(\mathbf{H}_a, \mathbf{H}_p) - d(\mathbf{H}_a, \mathbf{H}_n) + \alpha_1\right\}\\ &= \max\left\{0, d(Q(\mathbf{X}_a), Q(\mathbf{X}_p)) - d(Q(\mathbf{X}_a), Q(\mathbf{X}_n)) + \alpha_1\right\}\end{aligned} \tag{6}$$

4) Adversarial Loss: The hidden codes can create a common latent feature space for multiple domains. Although the IN helps remove the domain style information, the extracted feature vectors may still contain other kinds of domain-specific information. Hence, there is a risk that certain hidden codes could be over-fitted to the training datasets. Therefore, we impose a domain discriminator D to determine the dataset from which the feature vector is to be drawn. Suppose we have K different datasets in total (i.e., K domains). Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n], \mathbf{z}_i \in \{1, 2, \ldots, K\}$ denote the corresponding domain labels of $\mathbf{X}$. Thus, the domain discriminator D can be optimized by minimizing the domain classification loss defined as

$$\mathcal{L}_{\mathrm{D}}(D, Q) = \sum_{i=l}^{n} log(D(Q(\mathbf{x}_i), \mathbf{z}_i)) \tag{7}$$

where $D(\cdot)$ denotes the predicted probability that the hidden code $Q(\mathbf{x}_i)$ belongs to domain $\mathbf{z}_i$. After the training, the domain discriminator can capture the hidden domain-specific information, which is useful for determining the source domain of the feature vector. We can then eliminate the domain information from the network via adversarial learning using the domain discriminator. The overall adversarial learning process is a mini-max optimization problem:

$$arg \min_{Q} \max_{D} \mathcal{L}_{\mathrm{D}}(D, Q) \tag{8}$$

Q can be minimized using Eq. 5 and Eq. 6. The network can learn the feature vector by mapping it to the corresponding person identity via the identity loss $\mathcal{L}_{id}$ and the triplet verification loss $\mathcal{L}_{tri}$. D, on the other hand, needs to be maximized in order to suppress the domain-related information. To simplify the training process, we convert the mini-max optimization problem to a full minimization process by negating the domain classification loss $\mathcal{L}_{\mathrm{D}}$. The domain adversarial loss $\mathcal{L}_{\mathrm{adv}}$ is defined as

$$\mathcal{L}_{\mathrm{adv}} = -\mathcal{L}_{\mathrm{D}}(D, Q). \tag{9}$$

Minimizing the domain adversarial loss $\mathcal{L}_{\mathrm{adv}}$ is equivalent to maximizing the domain classification loss $\mathcal{L}_{\mathrm{D}}$. By minimizing the domain adversarial loss $\mathcal{L}_{\mathrm{adv}}$ in the feature training process, it can guide the feature extractor to produce features that are difficult for the domain discriminator to predict the corresponding domain labels. This mechanism encourages the network to focus less on the domain-specific visual information, but more on the domain-invariant features.

5) MMD-Based Regularization: To further enhance the domain invariance of the hidden codes, we adopt our previous MMFA architecture and incorporate the Maximum Mean Discrepancy (MMD) [24] regularization to align the distributions among different training datasets. Let $\mathbf{H}_l = [\mathbf{h}_{l,1}, \mathbf{h}_{l,2}, \ldots, \mathbf{h}_{l,n_l}]$ and $\mathbf{H}_t = [\mathbf{h}_{t,1}, \mathbf{h}_{t,2}, \ldots, \mathbf{h}_{t,n_t}]$ with batch sizes n_l and n_t be the hidden codes extracted from the encoder of two domains, l and t. Also let $\phi(\cdot)$ denote a mapping operation that projects the distributions onto a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ [48]. The MMD distance between domains l and t can be measured according to the following equation:

$$MMD(\mathbf{H}_l, \mathbf{H}_t)^2 = \left\| \frac{1}{n_l} \sum_{i=1}^{n_l} \phi(\mathbf{h}_{l,i}) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi(\mathbf{h}_{t,j}) \right\|_{\mathcal{H}}^2 \tag{10}$$

The arbitrary distribution of the hidden codes of different domains can be represented by using the kernel embedding technique [49], [50]. If the kernel $k(\cdot,\cdot)$ is characteristic, the mapping to the RKHS $\mathcal{H}$ is injective [51]. The injectivity indicates that the arbitrary distribution is uniquely represented by an element in RKHS. Therefore, we have a kernel function $k(\mathbf{h}_{l,i}, \mathbf{h}_{t,j}) = \phi(\mathbf{h}_{l,i})\phi(\mathbf{h}_{t,j})^\intercal$ induced by $\phi(\cdot)$. The MMD distance formulated in Eq. 10 can therefore be expressed as

$$\begin{aligned} MMD(\mathbf{H}_l, \mathbf{H}_t)^2 = & \frac{1}{(n_l)^2} \sum_{i=1}^{n_l} \sum_{i'=1}^{n_l} k(\mathbf{h}_{l,i}, \mathbf{h}_{l,i'}) \\ & + \frac{1}{(n_t)^2} \sum_{j=1}^{n_t} \sum_{j'=1}^{n_t} k(\mathbf{h}_{t,j}, \mathbf{h}_{t,j'}) \\ & - \frac{2}{n_l \cdot n_t} \sum_{i=1}^{n_l} \sum_{j=1}^{n_t} k(\mathbf{h}_{l,i}, \mathbf{h}_{t,j}) \end{aligned} \tag{11}$$

We follow the same setting as that of our previous domain adaptation MMFA model [14], which uses the RBF characteristic kernel with bandwidth $\alpha_2 = 1; 5; 10$ to compute the MMD distance:

$$k(\mathbf{h}_{l,i}, \mathbf{h}_{t,j}) = exp(-\frac{1}{2\alpha_2} \left\| \mathbf{h}_{l,i} - \mathbf{h}_{t,j} \right\|^2) \tag{12}$$

Since the MMFA-AAE network focuses on the feature generalization of multiple domains (K domains), the overall MMD regularization term $\mathcal{L}_{MMD}$ on the hidden codes is expressed as

$$\mathcal{L}_{\text{MMD}}\left(\mathbf{H}_1, \ldots, \mathbf{H}_K\right) = \frac{1}{K^2} \sum_{1 \le i,j \le K} \text{MMD}\left(\mathbf{H}_i, \mathbf{H}_j\right) \tag{13}$$

D. Training Procedure

The learning procedure of MMFA-AAE is similar to training an AAE network [25]. Unlike AAE, which only aims to minimize the reconstruction loss, our MMFA-AAE aims to jointly minimize the losses of identification, verification (triplet), reconstruction as well as MMD regularization on hidden codes. In our MMFA-AAE, the MMD-based adversarial auto-encoder with the early layer instance normalization enhances the feature generalization among different dataset domains. However, in order to learn a robust feature representation, the network also needs to incorporate the person identity loss and triplet loss. Our MMFA-AAE network uses the same network structure as our domain aggregation baseline proposed at the beginning of Section III. We use the same equations to compute the identity loss $\mathcal{L}_{\text{id}}$ and the triplet loss $\mathcal{L}_{\text{tri}}$ as formulated in Eq. 5 and Eq. 6, respectively. Unlike our baseline method, the MMFA-AAE model makes use of three additional loss functions. The reconstruction loss $\mathcal{L}_{\text{rec}}$ is used to preserve the content information of the feature vectors while performing latent space projection during the dimension reduction. The MMD regularization loss $\mathcal{L}_{\text{MMD}}$ helps align the distribution between different domains. The adversarial loss $\mathcal{L}_{\text{adv}}$ is computed according to Eq. 9. By maximizing the domain classification loss $\mathcal{L}_{\text{D}}$ as defined in Eq. 7 (i.e., minimizing $\mathcal{L}_{\text{adv}}$ as defined in Eq. 9), the network is guided to suppress the domain-specific information encoded in the extracted feature vectors. Similar to training other adversarial learning models, the training procedures for the MMFA-AAE model can be divided into two phases:

1) Freezing the feature extractor while using the feature vectors extracted from the network to train and update the domain discriminator D by minimizing $\mathcal{L}_{\text{D}}$. The domain discriminator D aims to predict the dataset from which a feature map is extracted. The domain classification loss can be computed with Eq. 7. We repeat the same process five times in a single iteration to minimize the domain classification loss for a relatively accurate domain prediction.
2) Freezing the domain discriminator D while training the feature extractor using the identity loss $\mathcal{L}_{\text{id}}$ and triplet loss $\mathcal{L}_{\text{tri}}$ to predict the identity labels and minimize the triplet distance, respectively. Meanwhile, the reconstruction loss $\mathcal{L}_{\text{rec}}$, the MMD domain distance loss $\mathcal{L}_{\text{MMD}}$ and adversarial loss $\mathcal{L}_{\text{adv}}$ help to remove the domain-specific information. The feature extractor training loss $\mathcal{L}$ can thus be formulated as a weighted sum of all these losses:

$$\mathcal{L} = \mathcal{L}_{\text{id}} + \lambda_1 \mathcal{L}_{\text{tri}} + \lambda_2 \mathcal{L}_{\text{rec}} + \lambda_3 \mathcal{L}_{\text{MMD}} + \lambda_4 \mathcal{L}_{\text{adv}} \tag{14}$$

Let E^*, Q^*, P^* and D^* denotes the parameters for feature extractor, encoder, decoder and domain discriminator, respectively. The overall algorithm of MMFA-AAE is illustrated in Algorithm 1.

IV. Experiments

A. Datasets and Settings

To evaluate our method, we follow the experiment settings in the DIMN method [22], which were also adopted by DualNorm [23]. In these settings, multiple large-scale benchmark datasets are combined to train a model. Small-scale datasets are individually used to evaluate the domain generalization ability of the model. In the experiments, the CUHK02 [16], CUHK03 [17], Market-1501 [10], DukeMTMC-reID [11] and CUHK-SYSU [20] datasets are selected for training. All these datasets have more than one thousand identities and thousands of images. We use all the images in this combined dataset to train our model, regardless of their original training/testing splits. All Person Re-ID models involved in the comparisons are trained with 121, 765 images from 18, 530 identities. The statistics of the training dataset are shown in Table II. We test the models on the VIPeR [37], PRID [38], GRID [39] and i-LIDS [40] datasets. However, these datasets are relatively small and have no more than one thousand identities. To evaluate the models in a more realistic manner, we also include the currently largest dataset, MSMT17 [18], in the experiments. The overall statistics of the testing datasets are

Algorithm 1 MMFA-AAE Network Training

```
Input: Multiple Dataset Domains $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$
Output: Learned parameters $E^*$, $Q^*$, $P^*$ and $D^*$.
for $t = 1$ to max iteration do
    Sample a mini-batch of images with the corresponding person identity labels $\mathbf{Y}$ and dataset labels (domain labels) $\mathbf{Z}$
    for $i = 1$ to 5 do
        Freezing $E^*$, $Q^*$, $P^*$ parameters and sample hidden codes $\mathbf{H}$ from the feature extractor and the encoder.
        Compute the gradient of Eq. 7 with respect to $D$ on hidden codes $\mathbf{H}$ and the corresponding domain labels $\mathbf{Z}$.
        Use the gradient to update $D^*$ by minimizing the objective of Eq. 7.
    end for
    Freezing $D^*$ parameters and sample hidden codes $\mathbf{H}$ from the feature extractor and the encoder.
    Compute the gradient of Eq. 14 with respect to $E^*$, $Q^*$, $P^*$ on $\mathbf{H}$ and the corresponding person identity labels $\mathbf{Y}$.
    Use the gradient to update $E^*$, $Q^*$, $P^*$ by minimizing the objective of Eq. 14.
end for
```

TABLE II
THE STATISTICS OF THE TRAINING DATASETS

Dataset	Total IDs	Total Images
CUHK02 [16]	1,816	7.264
CUHK03 [17]	1,467	14,097
Market-1501 [10]	1,501	29,419
DukeMTMC-reID [11]	1,812	36,411
CUHK-SYSU [20]	11,934	34,574
Total	18,530	121,765

shown in Table III. The evaluation of the performance of our domain generalization method follows the same settings as in [22], [23].

1) Evaluation Protocols: We follow the evaluation protocols used in [37] for VIPeR, [38] for PRID, [39] for GRID, and [40] for i-LIDS. Because we have to use the same testing in order to compare to other methods, we randomly select half of the VIPeR dataset for testing. For the PRID dataset, we follow the same single-shot experiments as in [52]. Since the VIPeR and PRID datasets contain only two images per person, the mean average precision (mAP) metric cannot be used. On GRID, we follow the standard testing split recommended in [39]. On i-LIDS, two images per identity are randomly selected as the probe image and the gallery image, respectively. For all the testing datasets mentioned above, the average results over 10 random selection of the testing sets are reported. The MSMT17 dataset has already been split into training, query, and gallery set. We follow the single-query retrieval setting for the MSMT17 dataset evaluation.

TABLE III
THE STATISTICS OF TESTING DATASETS

Dataset	#Test IDs		# Test Images	
	Probe	Gallery	Probe	Gallery
VIPeR [37]	316	316	316	316
PRID [38]	100	649	100	649
GRID [39]	125	900	125	1025
i-LIDS [40]	60	60	60	60
MSMT17 [18]	3,060	3,060	9,716	82,161

The cumulative matching characteristics (CMC) curve is used for our performance evaluation, as it is the most common metric used for evaluating Person Re-ID performance. This metric is adopted since Person Re-ID is intuitively posed as a ranking problem, where each image in the gallery is ranked based on its comparison to the probe. The probability that the correct match in the ranking equal to or less than a particular value is plotted against the size of the gallery set [37]. To make the comparison concise, we simplify the CMC curve by only comparing Rank 1, Rank 5, and Rank 10 successful retrieval rates. The CMC curve evaluation is valid when only one ground truth matches each given query image. The MSMT17 dataset contains multiple ground-truth images for the same person. Therefore, we use the mean average precision (mAP) proposed in [10] as an additional new evaluation metric. For each query image, the average precision (AP) is calculated as the area under its precision-recall curve. The mean value of the average precision (mAP) will reflect the overall recall of the person Re-ID algorithms.

2) Implementation Details: For the auto-encoder sub-network, we follow the same setting as that reported in [53], which uses a single hidden layer with a size of 512 neurons. The value of the hidden layer is used as an input for both the adversarial and classification sub-networks. Both sub-networks are composed of two fully-connected (FC) layers. The size of one FC layer is set to the same size as the hidden layer and while the size of the other is made the same as the identity labels. The weights for the identity and triplet losses are made equal, *i.e*, $\lambda_1 = 1$. Through various testings, it is observed that the parameters $\lambda_2 = 10, \lambda_3 = 0.2, \lambda_4 = 0.5$ yield the best performance. The Adam optimizer [54] is used for all experiments. The initial learning rate is set to 0.00035 with the warm-up training technique [55] and is decreased by 10% at the 40th epoch and 70th epoch, respectively. Totally, there are 120 training epochs with a batch size of 64. We implement our model in PyTorch and train it on a single Titan X GPU. The extracted features are l_2 normalized before matching scores are calculated.

B. Comparison Against State-of-the-art Methods

To demonstrate the superiority of our method, we compare it with various state-of-the-art methods under three different experimental conditions: fully supervised, unsupervised domain adaptation, and domain generalization. In Table IV, the *DG* methods are the multi-dataset domain generalization approaches. The AGG methods in the *DG* category are the domain aggregation baselines trained without any domain generalization layer or sub-network. *S* denotes a fully supervised method trained using images and labels from the corre-

TABLE IV

COMPARISON AGAINST STATE-OF-THE-ART METHODS. (R: RANK, S: SUPERVISED TRAINING WITH A TARGET DATASET, DA: DOMAIN ADAPTATION, DG: DOMAIN GENERALIZATION, -: NO REPORT)

Type	Method	VIPeR			PRID			GRID			i-LIDS		
		R 1	R 5	R 10	R 1	R 5	R 10	R 1	R 5	R 10	R 1	R 5	R 10
S	Ensemble [28]	45.9	77.5	88.9	17.9	40.0	50.0	-	-	-	50.3	72.0	82.5
S	DNS [52]	42.3	71.5	82.9	29.8	52.9	66.0	-	-	-	-	-	-
S	ImpTrpLoss [56]	47.8	74.4	84.8	22.0	-	47.0	-	-	-	60.4	82.7	90.7
S	GOG [1]	49.7	79.7	88.7	-	-	-	24.7	47.0	58.4	-	-	-
S	MTDnet [57]	47.5	73.1	82.6	32.0	51.0	62.0	-	-	-	58.4	80.4	87.3
S	OneShot [58]	34.3	-	-	41.4	-	-	-	-	-	51.2	-	-
S	SpindleNet [2]	53.8	74.1	83.2	67.0	89.0	89.0	-	-	-	66.3	86.6	91.8
S	SSM [59]	53.7	-	91.5	-	-	-	27.2	-	61.2	-	-	-
S	JLML [60]	50.2	74.2	84.3	-	-	-	37.5	61.4	69.4	-	-	-
DA	MMFA(Market-1501) [14]	39.1	-	-	35.1	-	-	-	-	-	-	-	-
DA	MMFA(DukeMTMC-reID) [14]	36.3	-	-	34.5	-	-	-	-	-	-	-	-
DA	TJ-AIDL(Market-1501) [13]	38.5	-	-	26.8	-	-	-	-	-	-	-	-
DA	TJ-AIDL(DukeMTMC-reID) [13]	35.1	-	-	34.8	-	-	-	-	-	-	-	-
DA	SyRI [61]	43.0	-	-	43.0	-	-	-	-	-	56.5	-	-
DG	AGG(DIMN)	42.9	61.3	68.9	38.9	63.5	75.0	29.7	51.1	60.2	69.2	84.2	88.8
DG	AGG(DualNorm)	42.1	-	-	27.2	-	-	28.6	-	-	66.3	-	-
DG	AGG(MMFA-AAE)	48.1	-	-	27.7	-	-	32.6	-	-	67.3	-	-
DG	DIMN [22]	51.2	70.2	76.0	39.2	67.0	76.7	29.3	53.3	65.8	70.2	89.7	94.5
DG	DualNorm [23]	53.9	-	-	**60.4**	-	-	41.4	-	-	74.8	-	-
DG	**MMFA-AAE**	**58.4**	-	-	57.2	-	-	**47.4**	-	-	**84.8**	-	-

sponding target dataset. The *DA* methods utilize unsupervised domain adaptation techniques. It is important to note that the *DA* and *S* methods are advantaged in the comparison in the sense that they have information about the target domain while our MMFA-AAE does not. We include them not as direct competitors, but to contextualize our results.

1) Comparison With Domain Generalization Methods: As discussed earlier, domain generalization (DG) is the most practical approach to the Person Re-ID problem. It assumes that a target dataset cannot be seen during training. Because of this challenge, domain generalization methods have to learn a domain-invariant feature representation from other datasets. However, there are only few prior studies [22], [23] on domain generalization for the Person Re-ID task. To make a fair comparison with these methods, we use the same MobileNetV2 [43] feature extractor backbone and follow the same evaluation protocol and experiment settings as those adopted in [22] and [23]. The lower part of Table IV shows the benchmark results of the methods. Our AGG baseline is slightly higher because of the additional triplet loss used during the supervised training. The MMFA-AAE network attains a 10% to 30% increase in terms of Rank 1 retrieval accuracy for all four datasets. Our MMFA-AAE method outperforms the DIMN and DualNorm on VIPeR, GRID and i-LIDS by a large margin. MMFA-AAE only falls behind DualNorm by 3% in Rank 1 accuracy when tested on the PRID dataset but still performs nearly 20% higher than the DIMN method.

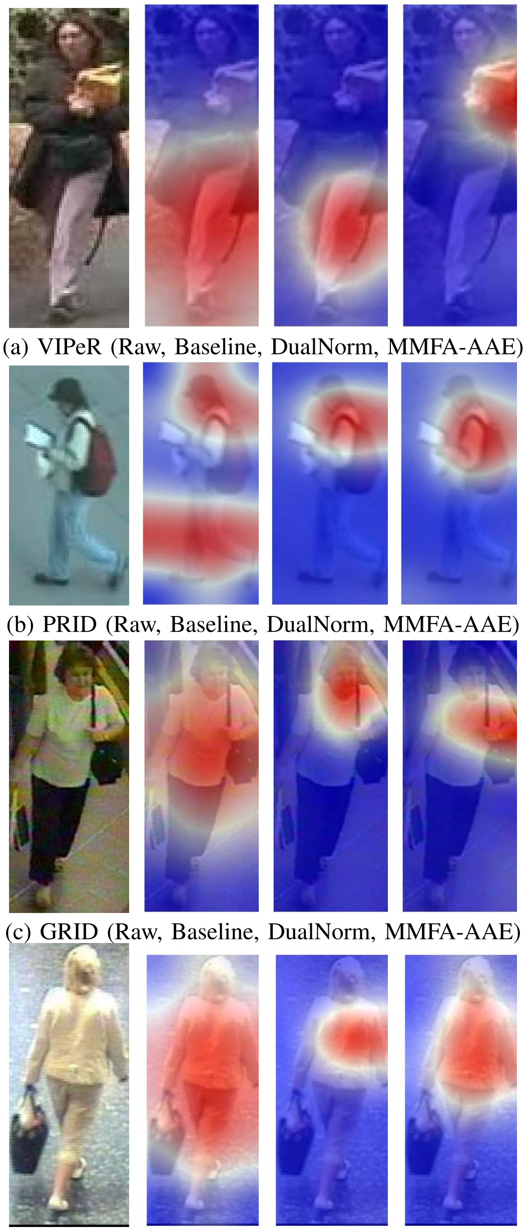
To further demonstrate the proposed MMFA-AAE's superiority to other methods, we also conduct the experiments on the largest Person Re-ID benchmark: MSMT17. Table V provides a performance comparison of our domain aggregation baseline, the DualNorm method and our MMFA-AAE network. All three methods use the same ResNet50 backbone to allow a fair comparison. The domain aggregation baseline without any domain generalization capability can only achieve 14.8% Rank 1 accuracy and 5.9% mAP score. Both DualNorm and our MMFA-AAE outperform the baseline method by a large margin in both Rank 1 and mAP scores. Our MMFA-AAE consistently surpasses the DualNorm by 3 to 4% in terms of Rank 1, Rank 5, and Rank 10 accuracy. Overall, our MMFA-AAE yields a much better performance most of the time without any additional data collection and domain adaptation process.

TABLE V

COMPARISON BETWEEN DUALNORM AND MMFA-AAE WITH RESNET50 BACKBONE ON THE MSMT17 DATASET

Model	MSMT17			
	Rank 1	Rank 5	Rank 10	mAP
AGG(MMFA-AAE)	14.8	27.8	37.6	5.9
DualNorm	42.6	55.9	61.8	19.6
MMFA-AAE	**46.0**	**59.5**	**64.2**	**20.7**

2) Comparison With Domain Adaptation Methods: We also compare our MMFA-AAE with other unsupervised domain adaptation methods. Multi-dataset domain generalization approaches focus on learning the universal feature representation from multiple datasets and assume the model can learn well-generalized features for any unseen camera network. Domain adaptation (DA) approaches focus on analyzing the characteristics between the images from labeled datasets and unlabeled images obtained from the new camera systems. Note, as discussed earlier, the DA methods' requirement for the unlabeled images from the new camera systems makes them impractical. Although the training and experimentation



(a) VIPeR (Raw, Baseline, DualNorm, MMFA-AAE)

(b) PRID (Raw, Baseline, DualNorm, MMFA-AAE)

(c) GRID (Raw, Baseline, DualNorm, MMFA-AAE)

(c) i-LIDS (Raw, Baseline, DualNorm, MMFA-AAE)

Fig. 2. Most activated feature maps produced by three different models on the same raw image. The images on the left-most column are the raw images while the other one shows the attention regions from the most activated feature maps of the last residual block. These feature maps highlight distinctive semantic features obtained from each model (Baseline, DualNorm and MMFA-AAE from left to right).

setting is different for DA and DG models, our MMFA-AAE model without using any target domain image surpasses the latest unsupervised domain adaptation approaches such as TJ-AIDL [13], MMFA [14], and SyRI [61]. The performance of the DA methods is shown in the middle section of Table IV. MMFA-AAE outperforms all of these DA methods on all the benchmark datasets without using any image from the target dataset and does not use additional adaptation. This means that our method can effectively use the domain-invariant feature learned from multiple large-scale datasets.

3) Comparison With Supervised Methods: Although many fully supervised methods are reported to have high performance on large-scale datasets such as Market-1501 and DukeMTMC-reID, their performance is still low when trained on small-scale datasets. Many methods have been proposed to address this issue [1], [2], [13], [28], [52], [56]–[60]. We have selected several supervised methods (labeled as *S* in Table IV) with reports on at least one of the four benchmark datasets. These methods are Ensemble [28], DNS [52], ImpTriplet [56], GOG [1], MTDnet [57], OneShot [13], SpindleNet [2], SSM [59], and JLML [60]. They follow conventional single-dataset training and testing procedures. It is not a fair comparison for MMFA-AAE method, which operates under the more challenging cross-dataset generalization setting. However, we use their results as references to illustrate the generalization capability of our MMFA-AAE model. Our MMFA-AAE method shows competitive or even better results on all four benchmarks.

Overall, our proposed MMFA-AAE network demonstrates state-of-the-art performance. It can effectively reduce the influence of domain-specific features by using the adversarial training method and learn a more general feature representation.

C. Ablation Study

1) Feature Heat-Map Visualization: To evaluate the effectiveness of the feature generated from our MMFA-AAE model, we randomly select images from each testing dataset and plot the most activated feature-maps obtained from the backbone network, as shown in Figure 2. We observed that the feature maps obtained from the domain aggregate baseline model could only focus on a vague global region. The domain generalization models such as DualNorm and MMFA-AAE can focus more on the local region with semantic meaning. In comparison with the DualNorm approach, the proposed MMFA-AAE can concentrate on the more meaningful areas like laptop or handbag, as shown in Figure 2 (a) and (c). For images from the PRID and the i-LIDS dataset, the MMFA-AAE and DualNorm also focus on similar regions. However, the MMFA-AAE still shows superior semantic region coverage. For example, the i-LIDS image, MMFA-AAE are focusing on the entire upper torso while the DualNorm can only focus on the shoulder region.

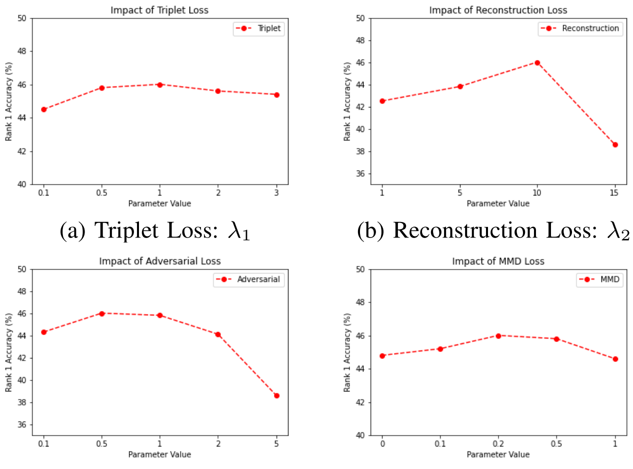
2) Components Analysis: There are four important components in the MMFA-AAE framework: Instance Normalization (IN), Triplet Loss, Adversarial Auto-Encoder (AAE), and Maximum Mean Discrepancy (MMD). To evaluate the contribution of each component, we incrementally adding one component into our baseline method and compare the performance in Table VI. The baseline we use in the experiment uses batch normalization after global average pooling. The baseline is trained with identity loss only first. We then introduce the instance normalization into the lower convolutional layer like DualNorm. The triplet loss will further enhance the performance by 1% to 2% on VIPeR, GRID, and i-LIDS. The domain-based adversarial auto-encoder gives a significant 3% to 8% boost for all the datasets. The final MMD alignment helps further boost the overall performance by 1% to 2%.

MMFA-AAE has four hyper-parameters that affect the re-ID accuracy: λ_1, λ_2, λ_3 and λ_4. We conduct experiments

TABLE VI

ABLATION STUDY ON THE IMPACT OF DIFFERENT COMPONENTS FOR MMFA-AAE NETWORKS

Method	VIPeR	PRID	GRID	i-LIDS
	R-1	R-1	R-1	R-1
Baseline (ResNet50)	42.9	38.9	29.7	69.2
Baseline + IN (DualNorm)	54.4	68.6	43.7	72.2
Baseline + IN + Triplet	55.9	61.6	43.0	74.8
Baseline + IN + Triplet + AAE	57	**67.6**	46.3	82.3
Baseline + IN + Triplet + AAE + MMD (MMFA-AAE)	**58.4**	65.7	**47.4**	**84.8**



(a) Triplet Loss: λ_1

(b) Reconstruction Loss: λ_2

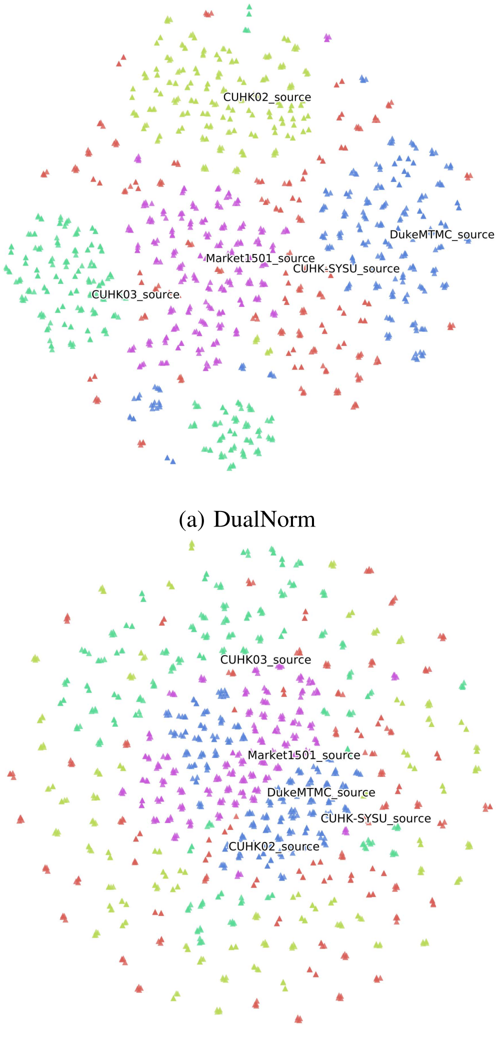
(c) Adversarial Loss: λ_3

(d) MMD Loss: λ_4

Fig. 3. The impact of the hyper-parameters of MMFA on the Re-ID Rank 1 accuracy of the MSMT17 dataset.

to analyze the impact of these hyper-parameters. For easier comparison, we only select the largest and the most complex MSMT17 dataset for evaluation and use Rank 1 accuracy. The results are shown in Figure 3. As shown Figure 3, each loss function can contribute 1% to 2% increase to the overall performance. However, adversarial loss λ_2 and re-construction loss λ_3 need to be carefully tuned, otherwise it may even degrading the performance of the Re-ID model.

3) t-SNE Visualization: We also visualize the 2D point cloud of the feature vectors extracted from the DualNorm network and our MMFA-AAE method using t-SNE [62], as shown in Figure 4a and 4b. We used a random sample of 6000 images from all five training datasets with a perplexity of 5000 for this visualization. As shown in Figure 4a, the DualNorm network can merge 5 different datasets well with low domain gaps between different datasets. However, the datasets are still clustered into several groups based on the property of the extracted feature vectors. On the other hand, our MMFA-AAE introduced the additional Adversarial-Auto-encoder (AAE) to mix up the feature vectors of different domains and alleviate the domain information. Figure 4b depicts our feature-point clouds extracted from the MMFA-AAE network. We can easily see that the overlap between different feature domains is more prominent in the case of the MMFA-AAE network.



(a) DualNorm

(b) MMFA-AAE

Fig. 4. The t-SNE visualization of the feature vectors from the DualNorm network and our MMFA-AAE network. Different color points indicate the training dataset domains.

V. CONCLUSION

In this paper, we propose a novel framework, MMFA-AAE, for multi-dataset feature generalization. Our MMFA-AAE network enables a Person Re-ID model to be deployed out-of-the-box for new camera networks. The main objective of our MMFA architecture is to learn a domain-invariant feature representation by jointly optimizing an adversarial auto-encoder with an MMD distance regularization. The adversarial auto-encoder is designed to learn a latent feature space among different Person Re-ID datasets via domain-based adversarial learning. The MMD-based regularization further enhances the domain-invariant features by aligning the distributions among different domains. In this way, the learned feature embedding is supposed to be universal to the seen training datasets and is expected to generalize well to unseen datasets. Extensive experiments demonstrate that our proposed MMFA-AAE is able to learn domain-invariant features, which lead to state-of-the-art performance on many datasets that it has never seen before. The proposed MMFA-AAE also out-performs most of the cross-dataset domain adaptation approaches and many fully-supervised methods. In conclusion, our MMFA-AAE approach addresses the scalability and generalization issues facing many existing Person Re-ID methods by providing a practical multi-dataset feature generalization strategy. With promising results, our MMFA-AAE approach paves the way for further research into the use of domain generalization within Person Re-ID and beyond.

References

[1] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical Gaussian Descriptor for Person Re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1363–1372.

[2] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 907–915.

[3] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for Person re-identification," Mar. 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737

[4] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: http://arxiv.org/abs/1711.08184

[5] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.

[6] Z. Dai, M. Chen, S. Zhu, and P. Tan, "Batch feature erasing for Person re-identification and beyond," 2018, *arXiv:1811.07130*. [Online]. Available: https://arxiv.org/abs/1811.07130

[7] L. Wu, Y. Wang, X. Li, and J. Gao, "What-and-where to match: Deep spatially multiplicative integration networks for person re-identification," *Pattern Recognit.*, vol. 76, pp. 727–738, Apr. 2018.

[8] L. Wu, Y. Wang, J. Gao, M. Wang, Z.-J. Zha, and D. Tao, "Deep coattention-based comparator for relative representation learning in person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 2, 2020, doi: 10.1109/TNNLS.2020.2979190.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[11] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3774–3782.

[12] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.

[13] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.

[14] S. Lin, H. Li, C.-T. Li, and A. C. Kot, "Multi-task mid-level feature alignment network for unsupervised cross-dataset Person re-identification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–13.

[15] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a Person retrieval model hetero- and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 176–192.

[16] W. Li and X. Wang, "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.

[17] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1288–1296.

[18] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[19] D. Li, Z. Zhang, X. Chen, and K. Huang, "A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1575–1590, Apr. 2019.

[20] S. Qiao, C. Liu, W. Shen, and A. Yuille, "Few-shot image recognition by predicting parameters from activations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7229–7238.

[21] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1249–1258.

[22] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 719–728.

[23] J. Jia, Q. Ruan, and T. M. Hospedales, "Frustratingly easy Person re-identification: Generalizing Person re-ID in practice," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–14.

[24] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. K. Sriperumbudur, "A fast, consistent Kernel two-sample test," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 673–681.

[25] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," in *Proc. Int. Conf. Learn. Represent. Workshop (ICLRW)*, 2015, pp. 1–16.

[26] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1391–1403, Mar. 2019.

[27] S. Lin and C.-T. Li, "End-to-end correspondence and relationship learning of mid-level deep features for person re-identification," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2017, pp. 1–6.

[28] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1846–1855.

[29] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.

[30] J. Si *et al.*, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.

[31] R. Layne, T. M. Hospedales, and S. Gong, "Domain transfer for person re-identification," in *Proc. 4th ACM/IEEE Int. Workshop Anal. Retr. Tracked Events Motion Imag. Stream - ARTEMIS*, 2013, pp. 25–32.

[32] A. J. Ma, J. Li, P. C. Yuen, and P. Li, "Cross-domain person reidentification using domain adaptation ranking SVMs," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1599–1613, May 2015.

[33] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.

[34] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised Person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8738–8745.

[35] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2655–2666, 2020.

[36] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2012, pp. 31–44.

[37] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. Int. Workshop Perform. Eval. Tracking Surveill. (PETS)*, 2007, pp. 1–7.

[38] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandin. Conf. Image Anal. (SCIA)*, 2011, pp. 91–102.

[39] C. Change Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1988–1995.

[40] W.-S. Zheng, S. Gong, and T. Xiang, "Associating groups of people," in *Proc. Brit. Mach. Vis. Conf.*, 2009, p. 23-1.

[41] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–11.

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[45] D. E. Rumelhart, G. E. Hinton, and R. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. Cambridge, MA, USA: MIT Press, 1987, ch. 8, pp. 399–421.

[46] H. Nam and H. E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 2558–2567.

[47] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-Net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 484–500.

[48] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample problem," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 513–520.

[49] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007, pp. 13–31.

[50] J. Hu, J. Lu, Y.-P. Tan, and J. Zhou, "Deep transfer metric learning," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5576–5588, Dec. 2016.

[51] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, G. R. Lanckriet, and B. Schölkopf, "Kernel choice and classifiability for RKHS embeddings of probability distributions," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 1750–1758.

[52] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.

[53] M. Ghifary, W. B. Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2551–2559.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–15.

[55] P. Goyal *et al.*, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," 2017, *arXiv:1706.02677*. [Online]. Available: http://arxiv.org/abs/1706.02677

[56] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[57] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for Person re-identification," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 1–8.

[58] S. Bak and P. Carr, "One-shot metric learning for person re-identificatio," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[59] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3356–3365.

[60] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2194–2200.

[61] S. Bąk, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised Person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 193–209.

[62] D. Graham-Rowe, "Visualizing data Using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Shan Lin (Member, IEEE) received the B.Sc. and Ph.D. degrees from the University of Warwick, U.K, in 2015 and 2020, respectively. He is currently a Research Fellow with the ROSE Lab, Nanyang Technological University, Singapore. His current research interests are in the area of person re-identification, computer vision, and deep learning. His studies are funded by the European Union EU H2020 project IDENTITY and National Research Foundation, Singapore under AI Singapore program. He has published several technical papers in these areas.

Chang-Tsun Li (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from National Defense University, Taiwan, the M.Sc. degree in computer science from the U.S. Naval Postgraduate School, USA, and the Ph.D. degree in computer science from the University of Warwick, U.K. He is currently a Professor of cyber security with Deakin University. He has had over 20 years of research experience in multimedia forensics and security, biometrics, machine learning, data analytics, computer vision, image processing, pattern recognition, bioinformatics, and content-based image retrieval. The outcomes of his research have been translated into award-winning commercial products protected by a series of international patents and have been used by a number of law enforcement agencies, national security institutions and companies around the world, including INTERPOL (Lyon, France), UK Home Office, Metropolitan Police Service (U.K.), Sussex Police Service (U.K.), Guildford Crown Court (U.K.), Barclays Bank PLC, US Department of Homeland Security. In addition to his active contribution to the advancement of his field of research through publication, he is also enthusiastically serving the international cyber security community. He is currently a Vice Chair of Computational Forensics Technical Committee of the International Association of Pattern Recognition (IAPR), a Member of the IEEE Information Forensics and Security Technical Committee, an Associate Editor of the IEEE ACCESS, *EURASIP Journal of Image and Video Processing and IET Biometrics*.

Alex C. Kot (Fellow, IEEE) has been with Nanyang Technological University (NTU), Singapore, since 1991. He headed the Division of Information Engineering, School of Electrical and Electronic Engineering (EEE) for eight years. He was the Vice Dean Research and Associate Chair (Research) of the School of EEE for three years, overseeing the research activities for the School with over 200 faculty members. He was the Associate Dean (Graduate Studies) of the College of Engineering (COE) for eight years. He is currently the Director of ROSE Lab [Rapid(Rich) Object SEearch Lab) and the Director of NTU-PKU Joint Research Institute. He has published extensively with over 300 technical papers in the areas of signal processing for communication, biometrics recognition, authentication, image forensics, machine learning, and AI. He served as an Associate Editor for a number of IEEE transactions, including IEEE TSP, IMM, TCSVT, TCAS-I, TCAS-II, TIP, SPM, SPL, JSTSP, JASP, and TIFS. He was a TC member for several IEEE Technical Committee in SPS and CASS. He has served the IEEE in various capacities such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing (ICIP) and area/track chairs for several IEEE flagship conferences. He also served as the IEEE Signal Processing Society Distinguished Lecturer Program Coordinator and the Chapters Chair for IEEE Signal Processing Chapters worldwide. He received the Best Teacher of The Year Award at NTU, the Microsoft MSRA Award, and as a coauthor for several award papers. He was elected as the IEEE CAS Distinguished Lecturer, in 2005. He was a Vice President in the Signal Processing Society and an IEEE Signal Processing Society Distinguished Lecturer. He is now a Fellow of the Academy of Engineering, Singapore and a Fellow of IES.