Cross-Entropy Adversarial View Adaptation for Person Re-identification

Lin Wu, Richang Hong, Yang Wang*, Meng Wang

Abstract—Person re-identification (re-ID) is a task of matching pedestrians under disjoint camera views. To recognise paired snapshots, it has to cope with large cross-view variations caused by the camera view shift. Supervised deep neural networks are effective in producing a set of non-linear projections that can transform cross-view images into a common feature space. However, they typically impose a symmetric architecture, yielding the network ill-conditioned on its optimisation. In this paper, we learn view-invariant subspace for person re-ID, and its corresponding similarity metric using an adversarial view adaptation approach. The main contribution is to learn coupled asymmetric mappings regarding view characteristics which are adversarially trained to address the view discrepancy by optimising the cross-entropy view confusion objective. To determine the similarity value, the network is empowered with a similarity discriminator to promote features that are highly discriminant in distinguishing positive and negative pairs. The other contribution includes an adaptive weighing on the most difficult samples to address the imbalance of within/between-identity pairs. Our approach achieves notable improved performance in comparison to state-of-the-arts on benchmark datasets.

Index Terms—Person re-identification, View adaptation, Adversarial learning, Entropy regularisation.

I. INTRODUCTION

PERSON re-identification (re-ID) is a challenging problem specialising on pedestrian matching across a network of cameras. It has not been solved yet principally because of the significant visual changes caused by colour, background, camera viewpoints and human poses. Recent state-of-the-arts are developed in the basis of supervised deep neural networks [1]–[9] to learn robust and discriminative representations against visual variations. However, training deep architectures requires a large number of labeled image pairs across multiple camera views, which is prohibitively expensive and not scalable to real-world scenarios. To combat that challenge, a number of semi/un-supervised methods have been developed [10]–[17]. Some of them attempt to seek feature invariance by

Lin Wu is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230000, China.

Richang Hong is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230000, China. Email: hongrc.hfut@gmail.com.

Yang Wang (Corresponding author) is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230000, China; Dalian University of Technology, Dalian 116024, China. Email: yang-wang@hfut.edu.cn. Yang Wang was supported by National Natural Science Foundation of China, under Grant No 61806035.

Meng Wang is with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230000, China. Email: eric.mengwang@gmail.com. Meng Wang was supported by he National Key Research and Development Program of China under grant 2018YFB0804200; NSFC 61725203, 61732008.
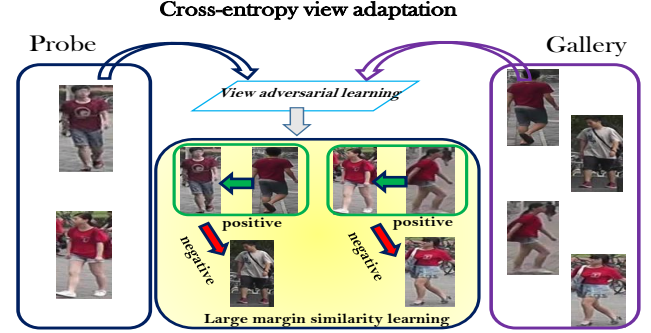

Fig. 1: Cross-entropy adversarial view adaptation for person re-ID. Given paired images across views (the probe and the gallery), the task is to jointly learn view-invariant feature space and its corresponding similarity metric. We propose to optimise asymmetric mappings regarding view specificity, and conditioned on each other through view-adversarial training (cross-entropy loss). The similarity metric is jointly learned by a discriminator network to identify positive pairs against negatives. See the text for details.

designing robust hand-crafted features [10]–[12]. However, without the supervision of labeled data, the discrimination and specificity apt to camera-pair changes are not captured. Also, unsupervised methods treat samples from different views indiscriminately, and the effect of view-specific inference is not considered. On the other hand, some unsupervised methods introduce graph structure or clustering centroid [14]–[17] to keep visually similar people close in the projected space. Nonetheless, it is insufficient to explore the discriminative space as the learning of view-specific projections into a shared subspace is optimised independently.

Matching pedestrian snapshots across camera views (probe and gallery) can be achieved by seeking a common subspace therein, and jointly optimising a measure for each pair of cross-view images. Siamese networks with deep convolutions are demonstrated to hold promise in person re-ID [3], [19], [20] by learning a set of nonlinear transformations that align the correlation of layer activations in deep neural networks. However, Siamese networks have layer-wise equality constraints on deep layered representations, which are commonly imposed within convolutional networks through weight sharing. The idea of Siamese networks is to enforce the exact consistency between the probe and the gallery mapping, where the learning of symmetric transformations can reduce the number of parameters in the deep model. Unfortunately, this may induce the optimisation poorly conditioned because the same network must handle images from two disjoint distributions.

A. Motivations

This paper is motivated towards person re-ID by presenting a deep view adaptation approach in the sense that the non-linear transformations into a common feature space corresponding to paired observations should be asymmetric. This asymmetric architecture is necessary to characterise the view-specific entailments, and the optimisation regarding asymmetric mappings should be conditioned on each other to capture the identity interplay between the probe and gallery views. Also, a re-ID system proceeds paired images and requires a comparable metric to determine the similarity for each pair. Towards the above practices, in this paper we present a deep feature learning approach to optimise a feature space such that the invariance between the probe and gallery distribution is maximum (*cross-view invariance*), and we jointly learn similarity metrics for paired images. Our approach is appealing in the ability to learn asymmetric mappings characterising cross-view images without enforcing any sharing constraints. The minimisation on view discrepancy is achieved by performing adversarial learning with entropy regularisation to operate cross-entropy minimisation over cross-view samples. To jointly learn a comparable metric, the adversarial framework is empowered with a discriminator network to distinguish positive pairs against negatives. More importantly, the network training does not require a large number of training samples as opposed to existing deep learning methods [5], [19], [21] because we introduce adaptive weighting into the paired inputs which would emphasise the most difficult ones by assigning batch-based adaptive weights into positive/negative pairs.

It is noted that our framework is different from the study on domain adaptation to person re-ID [1], [22]–[24]. First, this line typically reuses pre-trained models from a closely related dataset with a large amount of samples (source), and then design the training towards the much smaller dataset of interest (target). Here instead, we are interested in adapting adversarial learning into cross-view invariant feature learning a.k.a *adversarial view adaptation*, to effectively address the view discrepancy in re-identifying persons. Secondly, a common problem of existing domain adaptation approaches is that a principled alignment between the source and target is missing, and thus they are unable to penalise the correlated domain misalignment in practical terms. In contrast, our method explicitly minimises the view discrepancy through the proposed view-adversarial objective. Our method is also distinct from existing methods based on adversarial losses [22]. For instance, SPGAN [22] is composed of GAN loss to update the target domain w.r.t the source, our method instead uses cross-entropy loss to optimise the view confusion objective.

B. Our Approach and Contributions

Our approach is designed based on the view adaptation scheme to learn asymmetric deep neural transformations in order to map view-specific distributions into a common feature space. In this sense, we introduce adversarial learning [25] into *view discriminator* which is optimised through cross-entropy based *view confusion objective*. This objective is to confuse the view discriminator that will perceive the two distributions identically so as to minimise the cross-view discrepancy. Specifically, we develop an adaptive learning framework to produce asymmetric mappings over two views through a view-adversarial training. When the view discriminator cannot determine if a pair is from the probe or the gallery view, the feature inference becomes optimal in terms of creating a view-invariant space. In this adversarial learning regime, view adaptation is seen as a generative adversarial network but there is not necessary to generate samples. In fact, the discriminator is confused by a view confusion objective and cannot determine the samples are from the probe or the gallery distribution when the network is optimal in creating a latent space conditioned on two views.

To address the similarity learning, we additionally enforce the semantic similarity by learning a distance metric jointly with the feature learning. This similarity is pertained in the discriminative base model of adversarial networks by using a contrastive loss (i.e., *similarity discriminator*), which pulls the images in positive pairs closer while pushes the negative pairs away from positives. Thus, our network is end-to-end trainable to process paired samples and outputs its similarity value to determine whether the pair is from the same identity or not. The overview of our framework is shown in Fig. 1. However, training paired input would raise an imbalance issue between the within-identity and between-identity samples. Hence, we particular introduce adaptive weighing into the most difficult positive/negative ones, which leads to optimised re-ID rank loss and quick convergence [26].

The major contributions of this paper can be summarised as follows.

- We propose a principled adversarial feature learning approach to person re-ID to jointly produce a latent view-invariant feature space and its corresponding distance metric which maintains high discrimination on positive pairs from negatives.
- Our method is differentiated from the literature in conceptualising cross-view matching through **asymmetric mappings** followed by explicit view adaptation. This is achieved by presenting view-adversarial learning to train cross-view embedding whilst confusing a view discriminator in a cross-entropy objective function.
- We provide insights into adaptive weighing which assigns larger weights to difficult samples such that positive/negative class imbalance is effectively addressed.
- Extensive validations of the proposed method against the state-of-the-art are performed to demonstrate the competence of our model.

II. RELATED WORK

A. Person Re-identification

Most of existing re-ID models are developed in a supervised manner to learn discriminative features [2], [5], [7], [27], [28] or learn distance metrics [8], [21], [29]. However, these models commonly rely on substantial labeled training data, which would hinder the application of them in large networked cameras. Semi-supervised and unsupervised methods are presented to overcome the scalability issue by using limited number

of labeled samples or without using label information. These techniques often focus on designing handcrafted features (e.g., colour, texture) [10]–[12] that should be robust to visual changes in imaging conditions. However, low-level features are not expressive in view-invariance because features are not learned to be apt to view-specific bias. On the other hand, transfer learning has been applied into re-ID [10], [13], [15], [30], [31], and these methods learn the model using large labeled datasets and then transfer the discriminative knowledge to the unlabelled target pairs. For example, they can learn a cross-view metric either by asymmetric clustering on person images [15] or by transferable colour metric from a single pair of images [10]. However, there is still a considerable performance gap relative to supervised learning approaches because it is not principled to fully explore the discriminative space in the context of source and domain image translation. In contrast to existing approaches that derive a metric independently from images of people, we learn a deep metric jointly with feature learning from few labeled training pairs in an adversarial manner.

B. Generative Adversarial Networks

The Generative Adversarial Networks (GANs) [25] consist of a generator G and a discriminator D to compete the learning where the generator is learned to map samples from a latent distribution to confuse D by producing samples close to real data, while the discriminator tries to distinguish between real and generated samples. The most popular variation of GAN is the Deep Convolutional GAN (DCGAN) introduced by Radford et al. [32]. DCGAN improved the overall quality of generated images by adapting Convolutional Neural Network (CNN) into GAN architecture. Then, GANs have been extensively studied and widely used in several applications including realistic image generation [33], image-image translation [34], domain adaptation [35], and cross-modal retrieval [36].

Recently, GANs are adopted into person re-ID community by Zhong *et al* [14] which is to introduce a semi-supervised pipeline that integrates GAN-generated samples into the CNN learning. Following up the work in [14], Deng *et al* [22] present an unsupervised domain adaptation method (SPGAN) to preserve the similarity after translation and then train re-ID models with the translated images using supervised feature learning methods. In this work, we do not focus on generating samples for person re-ID as opposed to [14]. Instead, we formulate adversarial networks into the feature learning process to produce a view-invariant subspace and jointly learning a similarity metric. Our method is also different from SPGAN [22] in the sense that we learn asymmetric transformations regarding view discrepancy rather than addressing a target domain into matching the source domain.

C. Adversarial Adaptation Methods

Deep convolutional neural networks trained on large-scale datasets can learn representations which are generically useful across different tasks and visual domains [37], [38]. However, due to the domain shift/bias, generalising the well-trained recognition models to novel tasks typically require fine-tuning these networks. While it is difficult to obtain enough labelled data to properly fine-tune the large number of parameters in deep networks, and thus recent deep adaptation methods attempt to mitigate the difficulty by learning deep neural transformations that map both domains into a common feature space. This can be generally achieved by optimising the representations to align the source and target sets [39]–[41]. For instance, several methods use the maximum mean discrepancy loss to measure the difference between the source and the target feature distributions [39], [42], [43]. Inspired by the idea of adapting higher order statistics of the two distributions [44]–[47], some methods propose a transformation to minimise the distance between the covariance representations of source and target datasets to ultimately achieve the correlation alignment [40], [41]. These approaches are unsupervised domain adaptation that do not need any target data labels, but they require large amounts of target training samples, which may not be available always. Also, semantic alignment of classes is difficult without a shared feature space which can be sought by creating positive and negative pairs using the source and target data [48]–[52].

Our framework is closely related to the adversarial adaptive methods [53]–[55] particularly in the employment of view confusion loss (*i.e.,* cross-entropy loss). However, these works on domain adaptation are in the case of unlabelled target domains, and their ultimate goal is to regulate the learning of the source and target mappings so as to minimise the distance between the empirical mapping distributions. They chose an adversarial loss to minimise domain shift, learning a representation that is simultaneously discriminative of source labels while being able to distinguish between domains. Our method is not designed to match the target distribution to the source through an adversarial loss. In stead, we allow individual mappings which are not enforced to have weight sharing or any consistency to characterise view shifts and the adaptation is achieved through a view-adversarial loss.

III. OUR APPROACH

A. Problem Formulation

We consider the training task where $\{X_s, X_t\}$ is the input space with X_s and X_t containing person images captured by two disjoint cameras, namely probe view (source) and gallery view (target). Specifically, the model is trained on labeled pairs in correspondence (x_s^i, x_t^i) where x_s^i and x_t^i are examples of the same person i across camera views. To address the view variance, we formulate it into adversarial adaptive manner: the main goal is to regularise the learning of the source and target mappings, M_s and M_t, so as to minimise the distance between the empirical source and target mapping distributions: $M_s(x_s^i)$ and $M_t(x_t^i)$. Under this setting, the similarity discriminator is to learn to directly determine the input pair (x_s^i, x_t^i) belongs to the same person or not, eliminating the cross-view variance.

The standard generative adversarial learning pits two networks against each other: a discriminator and a generator. The generator is in principle trained to produce images in a way that confuses the discriminator, which in turn tries to distinguish them from real image examples. In our case
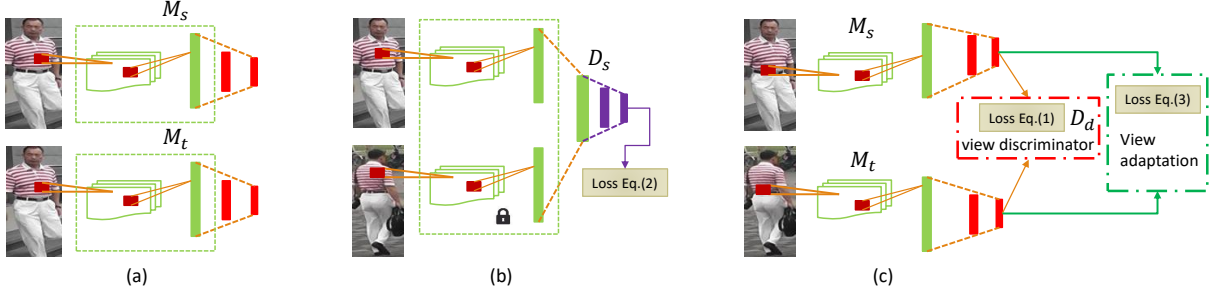
Fig. 2: The overview of adversarial adaptation learning for cross-view person re-ID. The network is trained with stage-wise updating on parameters (See text for details). Stage (a): The two mappings $\{M_s, M_t\}$ corresponding to different views are initialised by two CNNs. Stage (b): The similarity discriminator D_s is trained by optimising the loss function Eq.(2). Stage (c): The view discriminator D_d is trained by optimising the Eq.(1), and the mappings are updated by minimising the Eq.(3) to achieve view adaptation.

of cross-view adaptation for matching persons, this principle is employed to ensure that the networks cannot distinguish between the distribution of its probe view (X_s) and gallery view (X_t) [53], [55], [56]. In other words, a view discriminator (D_d) is adopted to classify whether an example is from the source or the target view. However, the generator is not needed in our network because generative modelling of input image distributions is not necessary, as the ultimate task is to learn discriminative representations regarding identities. On the other hand, asymmetric mappings can better model the differences of camera views than symmetric ones. Therefore, we first learn a couple of asymmetric mappings conditioned on each other through the view-adversarial training to produce view-invariant feature space. Then, a similarity estimator (D_s) with a margin-based separability is optimised on the Euclidean distances of positive/negative pairs to learn the effective similarity metrics.

B. Discriminator Networks

In our full adversarial adaptation framework, we have a **view discriminator** D_d, which classifies whether a data point is drawn from the probe or the gallery domain. Thus, D_d can be optimised according to a supervised loss, and the label indicates the origin domain. Herein, D_d is defined below:

$$\min_{D_d} \mathcal{L}_{D_d}(X_s, X_t, M_s, M_t) = -E_{x_s \sim X_s} \left[\log D_d(M_s(x_s)) \right]$$
$$- E_{x_t \sim X_t} \left[\log(1 - D_d(M_t(x_t))) \right], \tag{1}$$

where we design the individual probe and gallery mappings M_s and M_t. It is clear that the two mappings are both parameterised in the supervised training with their asymmetric structures. This strategy is different from existing discriminative domain adaptation approaches [54] which generally consider a separate adaptation: the probe mapping is first learned through supervised losses, and then target mappings are initialised while adapting with the probe. By contrast, we aim to ensure the distance minimisation between the probe and gallery domains under their respective mappings, while crucially maintaining both mappings semantically discriminative. To effectively minimise the view discrepancy, we design the view-adversarial mapping loss (as defined in Eq.(3)) which suits the case where we initially use independent mappings

and then the galley mapping (M_t) is updated to adversarially to match the probe (M_s).

An effective re-ID system requires a metric to estimate the similarity for the paired pedestrian snapshots. This is amount to learning discriminative representations that are able to distinguish positive pairs against negative ones. Thus, to empower the view-adaptation framework with discriminative capability, we propose to optimise a view-invariant feature space such that data examples with the same identity are closer than those with different identities. In this work, we are interested in performing an end-to-end training for each paired images in their RGB values and optimising the view-discrepancy jointly with their similarity metrics. As a result, we can simply estimate similarity values for persons by directly computing the Euclidean distances of their embeddings.

To generate the embedding for each pair (x_s, x_t) [1], and the corresponding similarity metric, we adopt the **similarity discriminator** network $D_s(\cdot)$ that aims to map semantically similar examples onto metrically close while simultaneously map semantically different examples onto metrically distant points in the embedding space. Hence, we formulate the similarity discriminator to minimise the following loss:

$$\min_{M_s, M_t, D_s} \mathcal{L}_{D_s}(X_s, X_t, Y) =$$
$$- E_{(x_s, x_t) \sim (X_s, X_t)} \left[\sum_{(x_s, x_t)} y \log D_s(M_s(x_s), M_t(x_t)) \right]$$
$$- \gamma E_{(x_s, x_t) \sim (X_s, X_t)} \left[\sum_{(x_s, x_t)} (1-y)\{\max(0, m - d(x_s, x_t))\}^2 + y d(x_s, x_t)^2 \right] \tag{2}$$

where $y \in Y$ is the binary label assigned to the pair (x_s, x_t), and $y = 1$ if the pair is positive and $y = 0$ otherwise. Y denotes the number of identities in training. $d(x_s, x_t)$ denotes the Euclidean distance between two input vectors: $d(x_s, x_t) = ||M_s(x_s) - M_t(x_t)||_2$. m is the margin that defines the separability in the embedding space and γ is the parameter to control the relative importance of two losses. In our experiments, m is empirically set to be $m = 2$ and γ is set to be $\gamma = 2.5$ (see the empirical evaluations in Section IV). The scheme of similarity discriminator is illustrated in Fig.3.

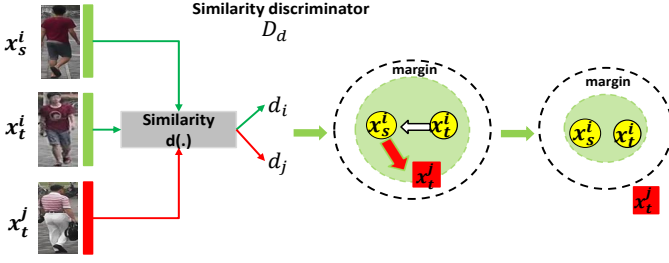[1] We omit the superscript i for the notation simplification.

Fig. 3: The similarity discriminator learns to embed images into their Euclidean distances ($d(\cdot)$) whereby the loss is to minimise the distances between positive examples, and maximise the distances between negative ones. This is motivated by the nearest-neigh classification by enforcing a margin between positive and negative pairs.

C. Adversarial View Adaptation with Cross-Entropy Loss

In our framework, we need to minimise the distances between the probe and gallery representations through alternating the minimisation between two functions. Thereby, the probe and gallery mappings should be optimised according to a constrained adversarial objective [53], which can be formulated as:

$$
\min_{M_s, M_t} \mathcal{L}_M(X_s, X_t, D_d) =
$$
$$
- \sum_{d \in \{s,t\}} E_{x_d \sim X_d} \left[\frac{1}{2} \log D_d(M_d(x_d)) + \frac{1}{2} \log(1 - D_d(M_d(x_d))) \right].
$$
(3)

Intuitively, the loss function in Eq.(3) is a view confusion objective, under which the mapping can be trained using a **cross-entropy** loss function against a uniform distribution. This loss is to ensure the adversarial discriminator will view the two domains identically. Finally, the full objective function is formulated to be the unconstrained optimisation as follow:

$$
\min_{D_d} \mathcal{L}_{D_d}(X_s, X_t, M_s, M_t);
$$
$$
\min_{M_s, M_t} \mathcal{L}_M(X_s, X_t, D_d); \qquad (4)
$$
$$
\min_{M_s, M_t, D_s} \mathcal{L}_{D_s}(X_s, X_t, Y).
$$

The components of the objective function Eq.(4) can be interpreted as:

- $\min_{D_d} \mathcal{L}_{D_d}(X_s, X_t, M_s, M_t)$: We allow independent view mappings without enforcing weight sharing ($M_s \neq M_t$). This introduces a more flexible learning paradigm that allows view specific feature extractions to be learned. Siamese-like networks in person re-ID [3], [8], [21] have layer-wise equality constraint, thus enforcing exact probe and target mapping consistency. Indeed, learning a symmetric transformation reduces the number of parameters in the model, and ensures the mapping is view-invariant when the optimisation is converged. However, this may render the optimisation poorly conditioned because the same network is demanded to deal with images from two separate distributions.
- $\min_{M_s, M_t} \mathcal{L}_M(X_s, X_t, D_d)$: In the setting where both the mappings are changing, the standard GAN loss cannot be applied because in the GAN setting the source distribution remains fixed while the target distribution is learned to match it. Thus, we aim to optimise the view

---

Algorithm 1 CROSS-ENTROPY adversarial view adaptation learning for person re-ID.

Input: Paired person images in cross-view $\{X_s, X_t\}$ where pairs are labeled in correspondence.
Output: A similarity discriminator D_s and cross-view mappings M_s and M_t.

1 Initialise two mappings M_s and M_t using M-Net and D-Net.
2 Initialise D_s using VGG pre-trained on ImageNet and fine-tuned on a soft-max function.
3 Uniformly sample (x_s, x_t) from $\{X_s, X_t\}$.
4 Train D_s using Eq.(2).
5 **while** *not convergent* **do**
6   Train D_d by minimising Eq.(1).
7   Update M_s, M_t by minimising Eq.(3).
8 **end**
9 **return** D_s, M_s, M_t.

---

confusion objective, and the mappings are updated using cross-entropy loss against a uniform distribution.
- $\min_{M_s, M_t, D_s} \mathcal{L}_{D_s}(X_s, X_t, Y)$: We choose $\mathcal{L}_{D_s}(\cdot)$ to be a discriminative base model, as most prior adversarial adaptive methods suggest a generative model is not necessary while optimisation can be performed directly in a discriminative space for this purpose [35], [54].

D. Network Training

We optimise the objective function Eq.(4) in stages. The overall network has three components to be trained: a similarity discriminator D_s, a view discriminator D_d, and mappings across views $\{M_s, M_t\}$. First, M_s and M_t are initialised by two deep models: M-Net [57] and D-Net [58], which are effective in independent feature detection and extraction [3], [59]. Then, the similarity discriminator D_s is modelled by stacked fully-connected layers: 1024 hidden units, 2048 hidden units, and the final similarity output. With the exception of the similarity output layer, these fully-connected layers are using a ReLU activation function. However, there would be a severe imbalance between the number of within-identity pairs and the much greater number of between-identity pairs because the model requires the access to all pairs as input. Thus, our first improvement is to introduce an adaptive weighted loss into the similarity discriminator for the sake of imbalance.

1) Adaptive Weighted Loss: The challenge of learning effective features during training with a balanced model is to assign larger weights to difficult positive and negative samples [26]. We improve the similarity loss in Eq. (2) by introducing adaptive weight distribution on the positive/negative class. Thus, Eq.(2) can be rewritten as:

$$
\min \mathcal{L}_{D_s}^* = \sum_{x_p \in P(x_s)} \left[\log D_s(M_s(x_s), M_t(x_p)) + w_p d(x_s, x_p)^2 \right]
$$
$$
- \gamma \sum_{x_n \in N(x_s)} \left[\{\max(0, m - w_n d(x_s, x_n))\}^2 \right],
$$
(5)

where the gallery sample $x_t \in X_t$ is positive to x_s, i.e. $x_p \in P(x_s)$ or negative to x_s, i.e., $x_n \in N(x_s)$. w_p and w_n denote the weights assigned to the positive and negative pairs, respectively. Through this adaptive weight loss, the

Fig. 4: Examples from person re-ID datasets. From left to right: VIPeR, CUHK03, Market-1501, and DukeMTMC-reID. Columns indicate the same identities.

positive/negative class imbalance is alleviated by the explicit reflection on weight distribution. Apparently, the advantage of this adaptive weighing on positive/negative samples is to pertain the contribution of hard samples whilst the original loss using the uniform weights can eliminate the effect of hard samples, and thus very likely to get into the local minima as driven by easy samples. In our implementation, w_p and w_n are defined by using the soft-max/min weight distributions as:

$$w_p = \frac{\exp^{d(x_s,x_p)}}{\sum_{x_p \in P(x_s)} \exp^{d(x_s,x_p)}}; w_n = \frac{\exp^{-d(x_s,x_n)}}{\sum_{x_n \in N(x_s)} \exp^{d(x_s,x_n)}}. \quad (6)$$

In the training, the VGG-16 network [58] pre-trained on ImageNet [60] is used as the base feature architecture. Following the conventional fine-tuning strategy [14], the last fully-connected layer is modified to have K neuron to predict the K-classes, where K is the number of training persons. Once fine-tuning is done, the convolutional layers of VGG architecture are used to be the non-linear transformations for the two mappings. As the network takes paired inputs, the two mappings are not applied with weight-sharing to ensure the network asymmetric. The outputs of each pair is concatenated before passing into the similarity discriminator [61]. Once the D_s and $\{M_s, M_t\}$ are trained, the next step is training the view discriminator D_d by classifying the images into X_s or X_t. We model D_d by using two fully-connected layers with a soft-max activation in the last layer to optimise the loss function of Eq. (1). This is implemented by freezing the D_s, $\{M_s, M_t\}$ and updating the parameters of D_d. Then, the network is trained to confuse D_d in which the cross-entropy loss is computed by optimising Eq. (3). The training process is illustrated in Fig.2, and the procedure is summarised in Algorithm 1.

To address the imbalance on training pairs, we improve the optimisation on the similarity discriminator D_s by introducing the adaptive weighted loss. Thus, to construct a batch during training and calculate adaptive weights, we follow MTMCT [26] to construct $\mathcal{SP}$ batches. In specific, during a training epoch each identity is selected into its batch, and the remaining $\mathcal{P} - 1$ batch identities are selected at random. And $\mathcal{S}$ samples for each identity are also selected at random.

IV. EXPERIMENTS

A. Datasets and Evaluation

We perform extensive experiments and comparative studies to evaluate our approach over four benchmark datasets: VIPeR [62], CUHK03 [19], Market-1501 [63], and DukeMTMC-reID [14], [64]. Example images are shown in Fig.4.

TABLE I: The division on train/validation/test set of each dataset.

Dataset	$\sharp C_{train}$	$\sharp C_{valid}$	$\sharp C_{test}$
VIPeR	216	100	316
CUHK03	1160	100	100
Market-1501	650	100	751
DukeMTMC-reID	602	100	702

- The **VIPeR** dataset [62] contains 632 individuals taken from two cameras with arbitrary viewpoints and varying illumination conditions. The 632 person images are randomly divided into two equal halves, one for training and the other for testing.
- The **CUHK03** dataset [19] includes 13,164 images of 1360 pedestrians. The whole dataset is captured with six surveillance camera. Each identity is observed by two disjoint camera views, yielding an average 4.8 images in each view. This dataset provides both manually labeled and detected pedestrian bounding boxes. Our experiments report results on the labeled dataset.
- The **Market-1501** dataset [63] contains 32,668 fully annotated boxes of 1501 pedestrians. Each identity is captured by at most six cameras and boxes of person are obtained by running a state-of-the-art detector, the Deformable Part Model (DPM) [65]. The dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively.
- The **DukeMTMC-reID** dataset is a re-ID version of the DukeMTMC dataset [64]. It contains 34,183 image boxes of 1,404 identities of which 702 are used for training and the remaining 702 for testing. The probe and gallery images are 2,228 and 17,661, respectively.

We evaluate all the approaches with Cumulative Matching Characteristic (CMC) results by the single-shot setting. The CMC curve can characterise a ranking result for every image in the gallery given the probe image. We also use mean Average Precision (mAP) as performance measure on CUHK03, Market-1501, and DukeMTMC-reID.

B. Settings

Considering the training of our network is accessible to few examples from each person because we do not perform data augmentation, it is necessary to perform cross-validation on hyper-parameters to improve the generalisation on unseen observations. We therefore construct two disjoint sets of classes to be C_{train} and C_{valid} on each train set of each dataset. For example, on VIPeR dataset, the three subsets are randomly divided to be C_{train} (216 persons), C_{valid} (100 persons), and C_{test} (316 persons). The details of the train/validation/test division on four datasets are given in Table I. There are up to six cameras for CUHK03 and Market-1501, and thus for each person we randomly select two cameras to be the probe and gallery views. Then, each person's images across views are selected to be samples in pairs.

We use the VGG architecture and its variants M-Net and D-Net as the feature bases which are initialised from weights pre-trained on ImageNet and fine-tuned on target C_{train} of each re-ID dataset. Once fine-tuning is done, the convolution layers of each network are used as M_s (M_t), and a three-layer fully connection with ReLU as activation function is

used as the similarity discriminator D_s. The hidden layers in D_s have the dimensionality of 1,024 and 2,048, respectively. The learning rate starts with 0.001 and is divided by 10 every 10 epoches. The network uses a batch size of 128 images. The training is stopped when the loss stops decreasing during the validation on C_{valid}.

C. Experimental Results

In this section, we compared the proposed method with recent un/semi-supervised and supervised models on four datasets. The comparison results measured by rank-R accuracies of CMC are shown in Fig.5. And respective rank-R ($R = 1, 5, 20$) values on four datasets are given in Table II, Table III, Table IV, and Table V. We also conduct self-ablation evaluations on parameter sensitivity and network architecture.

a) Comparison to Un/semi-supervised Methods: We compared our method with several unsupervised re-ID models, including local salience learning based models (GST [11] and eSDC [12]), transfer-learning based models (t-LRDC [13], PUL [31], and UMDL [30]), metric learning methods (OSML [10], CAMEL [15], OL-MANS [16]), and a semi-supervised method of LSRO [14].

On the VIPeR dataset, Table II shows that our method outperforms other models in the case when there is only one example for each person in each view. For example, our method achieves rank-1=51.3, which is noticeably improved performance compared to OL-MANS [16] with rank-1=44.9. The main reason is that the assumptions without supervision cannot provide the view-specific inference, and thus impedes these unsupervised methods from achieving higher accuracies. In contrast, the proposed method is based on adversarial learning which is able to effectively minimise the view discrepancy without requiring large numbers of labeled training examples. Moreover, the improved variant of our approach (denoted as $Ours^*$) with adaptive weighted loss can emphasise the most difficult samples in a batch, an thus outperforms the state-of-the-art SpindleNet [4] at rank-1 value.

On the CUHK03 dataset, in Table III it can be seen that our method outperforms the state-of-the-art by large margins. For instance, the rank-1 value is improved by 25% compared to OL-MANS [16]. The reality is the illumination changes in CUHK03 are extremely severe and even human beings may find difficulty in identifying the persons across views. Without the aid of supervision, unsupervised methods cannot retrain the appearance robustness against visual variations. As a comparison, our approach is able to address this issue by training a discriminative distance metric jointly with the view-invariant feature learning. This leads to better performance of the proposed method. Also, the performance of our method with adaptive weighting outperforms the state-of-the-art SpindleNet [4] which builds the discriminative representations by extensive body region decompositions.

Table IV and Table V report the comparison results on the Market-1501 and DukeMTMC-reID datasets, respectively. Our method has achieved notable performance gain on the two datasets in comparison with these un/semi-supervised methods. These empirical evaluations on different benchmark

	Method	R=1	R=5	R=20
	Ours	51.3	77.0	96.1
	Ours *	**55.9**	**79.2**	**97.9**
Un/s-supervised	GTS [11]	25.2	44.8	71.0
	eSDC [12]	26.3	46.6	72.8
	t-LRDC [13]	27.4	46.0	75.1
	OSML [10]	34.3	-	-
	CAMEL [15]	30.9	52.0	72.5
	OL-MANS [16]	44.9	74.4	93.6
Supervised	DCSL [66]	44.6	73.4	82.6
	JSTL [1]	20.9	-	-
	Deep-Embed [2]	49.0	77.1	96.2
	SpindleNet [4]	53.8	74.1	92.1
	Part-Aligned [5]	48.7	74.7	93.0
	DNSL [29]	42.3	71.5	92.1
	SI-CI [8]	35.8	72.3	97.1
	PIE [7]	18.1	25.3	49.4

TABLE II: Comparison results with state-of-the-arts on VIPeR. Un/s-supervised stands for unsupervised/semi-supervised. The best results are in bold.

datasets demonstrate the effectiveness of our model in cross-view person re-ID owing to the effective view adaptation while learning discriminative metrics in the context of view-aligned feature space.

b) Comparison to Supervised Methods: We compared the proposed method against recent state-of-the-art supervised models: DCSL [66], JSTL [1], DNSL [29], Deep-Embed [2], SpindleNet [4], Part-Aligned [5], MSCAN [6], SI-CI [8], PIE [7], JLML [21], MTMCT [26], SPReID [67], SVDNet [68], and DPFL [69]. Comparison results on three datasets are reported in Table II, Table III, Table IV, and Table V respectively. It can be noticed that our method achieves better results compared to these supervised methods, and can outperform them when the adaptive weighting is applied. In Table II, we obtain rank-1=51.3 (55.9 from $Ours^*$) on the VIPeR dataset which has gained the recognition improvement over the SpindleNet [4] by 2.1% in rank-1 value. And in Table III, we obtain rank-1=86.6 (88.9 after weighted adaptation) as opposed to SpindleNet [4] with rank-1=88.5. We remark that SpindleNet [4] is a fully-supervised method that needs annotations on each body region to focus/extract these local features to describe each person. The process of annotating each body region is very cumbersome and not scalable in large networked cameras. On Market-1501, comparison results in Table IV show that our method greatly improves the rank-1 accuracy for this task. For example, in comparison with MSCAN [6], a state-of-the-art method based on fully-supervised body region encoding, the rank-1 accuracy value goes from 80.3% up to 87.2%. This is particularly effective in Market-1501 dataset where each person has up to 10 samples, and our approach is able to address the view misalignment more carefully. Our approach with weight adaptation loss ($Ours^*$) can further improve the rank-1 accuracy and achieve 89.1%, which is better than the state-of-the-art SPReID* (rank-1=88.3%) [67] [2] and DPFL [69] (rank-1=88.6%). Experimental results on the DukeMTMC-reID dataset are reported in Table V. Our method outperforms the state-of-the-art DPFL [69] by 1% at rank-1 accuracy. It shows that the adaptive weighting scheme is very effective in training a balanced model on DukeMTMC-reID

[2]Please note that all results of SPReID [67] are reported by using reduced data augmentation backboned on ResNet-152 architecture.

	Method	R=1	R=5	R=20	mAP
	Ours	86.6	**98.6**	99.4	**91.4**
	*Ours**	**88.9**	**99.2**	**99.9**	**91.8**
Un/s-supervised	eSDC [12]	8.7	26.5	53.4	-
	OSML [10]	45.6	78.4	88.5	-
	LSRO [14]	84.6	97.6	99.8	87.4
	CAMEL [15]	31.9	54.6	80.6	-
	XQDA [70]	52.2	82.2	96.2	51.5
	UMDL [30]	1.6	5.4	10.2	-
	OL-MANS [16]	61.7	88.4	98.5	-
Supervised	DCSL [66]	80.2	97.7	99.8	-
	JSTL [1]	72.6	91.0	96.7	-
	Deep-Embed [2]	73.0	91.6	98.6	-
	SpindleNet [4]	88.5	97.8	99.2	-
	Part-Aligned [5]	85.4	97.6	99.9	90.9
	MSCAN [6]	74.2	94.3	99.3	-
	DNSL [29]	58.9	85.6	96.3	-
	SI-CI [8]	52.2	84.3	98.8	-
	PIE [7]	62.4	73.7	95.6	71.3
	SPReID [67]*	88.0	95.2	99.9	-
	SVDNet [68]	68.5	90.2	94.0	73.3
	DPFL [69]	86.7	97.0	98.2	83.8

TABLE III: Comparison results with state-of-the-arts on CUHK03. The best results are in bold.

	Method	R=1	R=5	R=20	mAP
	Ours	87.2	96.3	98.5	**74.7**
	*Ours**	**89.1**	**96.8**	**99.7**	**76.2**
Un/s-supervised	eSDC [12]	33.5	50.6	67.5	13.5
	LSRO [14]	83.9	93.6	97.5	66.1
	CAMEL [15]	54.5	74.6	87.0	-
	OL-MANS [16]	60.7	83.8	91.9	-
	PUL [31]	45.5	60.7	72.6	-
	UMDL [30]	34.5	52.6	68.0	-
	XQDA [70]	43.8	65.3	80.4	22.2
Supervised	JSTL [1]	44.7	67.2	82.0	-
	Deep-Embed [2]	68.3	87.2	96.7	40.2
	SpindleNet [4]	76.9	91.5	96.7	-
	Part-Aligned [5]	81.0	92.3	97.1	-
	MSCAN [6]	80.3	92.0	97.0	57.5
	DNSL [29]	55.4	75.0	87.3	35.7
	PIE [7]	79.3	90.7	96.5	55.9
	JLML [21]	85.1	97.9	99.5	65.5
	MTMCT [26]	82.1	93.5	98.1	68.0
	SPReID [67]*	88.3	93.6	98.5	72.9
	SVDNet [68]	80.5	91.7	93.7	62.1
	DPFL [69]	88.6	94.5	98.0	73.1

TABLE IV: Comparison results with state-of-the-arts on Market-1501. All results are evaluated on single-shot setting. The best results are in bold.

dataset which has severe imbalance classes in the probe (2,228 images) and gallery size (17,661 images).

c) Self-Ablation Studies: We first study the sensitivity of our model to the key parameter of γ in Eq.(2). The impact of γ is investigated and the results are shown in Fig.7. As γ is to balance the relative importance of the discriminative distance metric, it is proven to have higher rank-1 accuracy when $\gamma = 2.5$, while a larger γ does not bring more gains in accuracy. Thus, we empirically set $\gamma = 2.5$ in all experiments. We also study different network architectures to inspect the importance of backbone networks. In our experiment, we consider the VGG-16 and the ResNet [71]. Specifically, two variants of VGG: M-Net and D-Net are used to initialise (M_s, M_t), and two identical ResNet networks are employed to initialise (M_s, M_t) as a comparison. Experimental results are shown in Table VI. We can observe that performances of two identical ResNet networks are inferior to the asymmetric architectures with M-Net and D-Net on VIPeR and CUHK03 datasets. Thus, we use M-Net and D-Net as default backbone networks.

Method	R=1	R=5	R=20	mAP
Ours	47.2	61.0	84.7	31.4
*Ours**	**80.1**	**89.5**	**96.9**	**67.2**
PUL [31]	30.0	43.4	57.6	16.4
UMDL [30]	18.5	31.4	55.2	7.3
SPGAN [22]	41.1	56.6	77.3	22.3
MTMCT [26]	74.2	81.9	93.2	54.9
SPReID [67]*	79.6	86.8	95.7	62.4
SVDNet [68]	67.6	80.5	83.7	45.8
DPFL [69]	79.2	85.7	94.6	60.6

TABLE V: Comparison results with state-of-the-arts on DukeMTMC-reID. All results are evaluated on single-query setting. The best results are in bold.

Architectures	VIPeR (R=1)	CUHK03 (R=1)
M-Net, D-Net = M_s, M_t	51.3	86.6
D-Net, M-Net = M_s, M_t	49.7	83.9
ResNet, ResNet = M_s, M_t	51.3	86.4

TABLE VI: The study on different architectures.

D. Comparison to Other Few-Shot Methods

We also compared to two recently proposed few-shot learning methods: matching networks [72] and model regression [73]. The matching networks propose a nearest neighbour approach that trains an embedding end-to-end for the task of few-shot learning. Model regression trains a small MLP to regress from the classifier trained on a small dataset to the classifier trained on the full dataset. Both of the two techniques are high-capacity in learning from few examples and facilitates the recognition in the small sample size regime on a broad range of tasks, including domain adaptation and fine-grained recognition. Comparison results are shown in Fig. 6. In terms of the overall performance, our method outperforms the two competitors constantly over the two datasets. Matching networks exhibit similar performance to our method, however, matching networks are based on nearest neighbours and use the entire training set in memory, and thus they are more expensive in testing time compared with our method and model regressors.

E. Comparison with View Adaptation Methods

In this experiment, we validate our approach in view adaptation by comparing to recent domain adaptation methods not limited from person re-ID: SPGAN [22], Deep Adaptation Networks (DAN) [43], Adversarial Discriminative Domain Adaptation (ADDA) [54], and CoGANs [56]. Experimental results are provided in Table VII. For these domain adaptation methods including DAN [43], ADDA [54], and CoGANs [56], their training are set and modified to adapt the gallery view (target) to match the probe view (source). For instance, CoGANs [56] can learn a joint distribution of multiple-domain data, the learning can be conducted by using two generative models with an identical architecture corresponding to the probe and the gallery images of a person. Then, through weight sharing, CcGANs are able to encode high-level semantics regarding identities into the low-level feature extraction. Our approach achieves the highest rank-1 value on the three datasets, despite being trained without a deep generator yet being a considerably simpler model. This also provides compelling evidence that generating images is not necessarily relevant to effective view adaptation. This discovery is consistent with ADDA [54] which does not
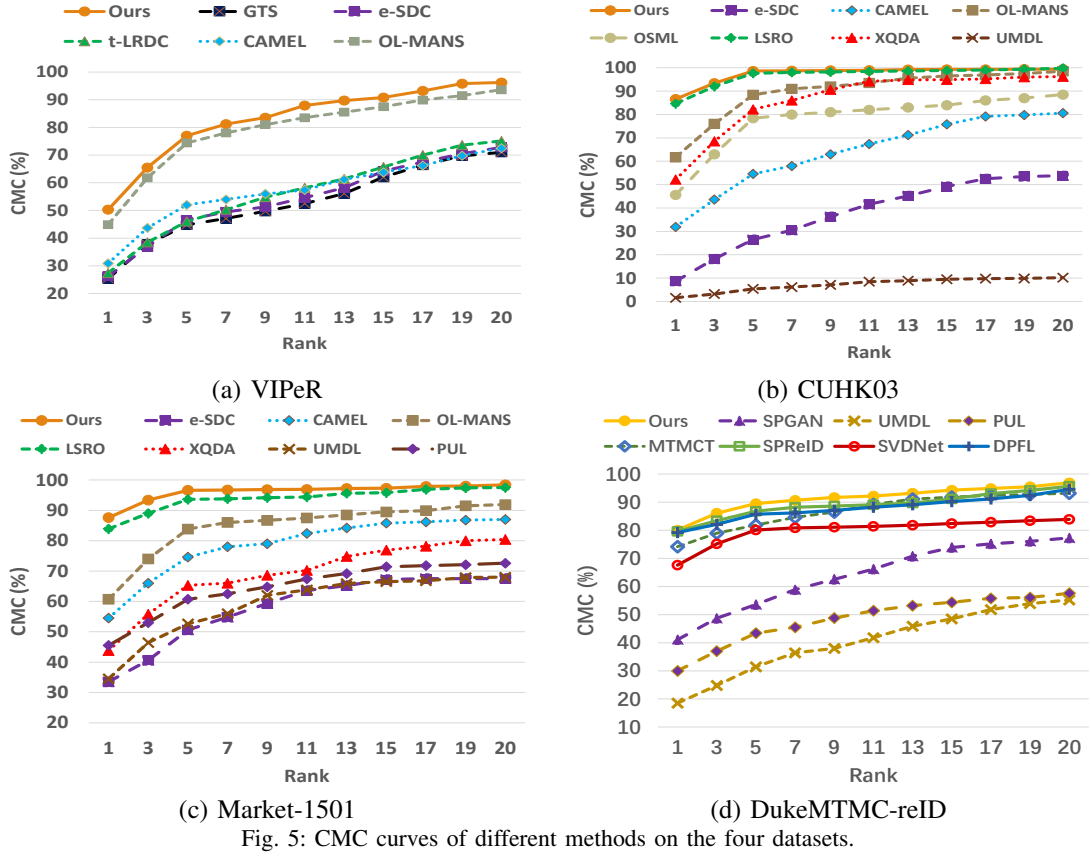
(a) VIPeR

(b) CUHK03

(c) Market-1501

(d) DukeMTMC-reID

Fig. 5: CMC curves of different methods on the four datasets.

Fig. 6: Comparison to recent few-shot learning methods.

Fig. 7: γ in Eq.(2) w.r.t re-ID accuracy. A larger γ indicates a larger weight of similarity discrimination constraint.

Method	VIPeR	CUHK03	Market-1501
SPGAN [22]	-	-	58.1
DAN [43]	39.6	71.0	53.8
CoGAN [56]	41.7	-	48.1
ADDA [54]	39.4	73.1	56.4
Ours	**51.3**	**86.6**	**87.2**

TABLE VII: The comparison results with view adaptation methods.

the view changes are very disparate, and it is unable to train coupled generators for them simultaneously.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduce an effective view adaptation model to person re-identification to produce asymmetric transformations that can fully characterise view specificity. The

use a generative model while also shows convincing results in comparison with CoGANs [56]. For CoGANs [56], it is sometimes hard to get convergence, e.g. on CUHK03 when

approach is based on adversarial learning to minimise view-discrepancy through view confusion objective with an entropy regularisation to align and form view-invariant feature space. The network is trained with a cross-entropy loss to optimise view confusion objective and jointly with a discriminative distance metric through a margin-based separability criterial. Also, training imbalance is explicitly described as weight distribution on hard samples, and the proposed adaptive weighting loss can address it more effectively. Experimental results show that the adversarial neural networks are able to produce feature space with cross-view variations being reduced. The proposed approach works effectively for labeled training samples with large visual divergence, and our method shows clear promise as it sets new state-of-the-art performance in experiments.

In future work, we would explore the direction of view adaptation in the case when such training pairs are not given. One possibility is to learn a probe to gallery encoder-decoder under a generative adversarial objective with some reconstruction term which can be applied to predict the clothing people are wearing. The other interesting direction is towards intriguing few-shot learning principles to learn to match persons with more powerful augmented memory networks.

REFERENCES

[1] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representation with domain guided dropout for person re-identification," in *CVPR*, 2016.

[2] L. Wu, Y. Wang, J. Gao, and X. Li, "Deep adaptive feature embedding with local sample distributions for person re-identification," *Pattern Recognition*, vol. 73, pp. 275–288, 2018.

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

[4] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017.

[5] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017.

[6] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017.

[7] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," in *arXiv:1701.07732*, 2017.

[8] 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 *CVPR*, 2016, pp. 1288–1296.

[9] L. Wu, C. Shen, and A. van den Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, vol. 65, p. 238250, 2017.

[10] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *CVPR*, 2017.

[11] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *BMVC*, 2014.

[12] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[13] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *TPAMI*, vol. 38, no. 3, pp. 591–606, March 2016.

[14] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[15] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *ICCV*, 2017.

[16] J. Zhou, P. Yu, W. Tang, and Y. Wu, "Efficient online local metric adaptation via negative samples for person re-identification," in *ICCV*, 2017.

[17] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised l1 graph learning," in *ECCV*, 2016.

[18] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3939–3949, 2015.

[19] W. Li, R. Zhao, X. Tang, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[20] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Transactions on Multimedia*, 2018.

[21] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *IJCAI*, 2017.

[22] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *CVPR*, 2018.

[23] N. Martinel, M. Dunnhofer, G. L. Foresti, and C. Micheloni, "Person re-identification via unsupervised transfer of learned visual representations," in *International Conference on Distributed Smart Cameras*, 2017.

[24] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," in *arXiv:1611.05244*, 2016.

[25] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,

"Generative adversarial nets," in *NIPS*, 2014.

[26] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and person re-identification," in *CVPR*, 2018.

[27] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based person re-identification," in *ICCV*, 2015.

[28] L. Wu, Y. Wang, L. Shao, and M. Wang, "3d personvlad: Learning deep global representations for video-based person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2019.2891244, 2019.

[29] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *CVPR*, 2016.

[30] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016.

[31] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," in *Arxiv*, 2017.

[32] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[33] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, and T. Brox, "Learning to generate chairs, tables and cars with convolutional networks," in *arXiv:1411.5928*, 2017.

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[35] S. Motiian, Q. Jones, S. M. Iranmanesh, and G. Derotto, "Few-shot adversarial domain adaptation," in *NIPS*, 2017.

[36] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1602–1612, 2019.

[37] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: a deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

[38] C. Zhang, L. Wu, and Y. Wang, "Crossing generative adversarial networks for cross-view person re-identification," *Neurocomputing*, vol. 340, no. 7, pp. 259–269, 2019.

[39] B. Sun and K. Saenko, "Deep coral: correlation alignment for deep domain alignment," in *ECCV Workshop*, 2016.

[40] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *AAAI*, 2016.

[41] P. Morerio, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," in *ICLR*, 2018.

[42] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Deep domain confusion: maximizing for domain invariance," in *arXiv:1412.3474*, 2014.

[43] M. Long and J. Wang, "Learning transferable features with deep adaptation networks," in *ICML*, 2015.

[44] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Link prediction across networks by biased cross-network sampling," in *ICDE*, 2013, pp. 793–804.

[45] ——, "Online community detection in social sensing," in *ACM international conference on Web search and data mining*, 2013, pp. 617–626.

[46] X.-S. Hua and G.-J. Qi, "Online multi-label active annotation: towards large-scale content-based video search," in *ACM Multimedia*, 2008, pp. 141–150.

[47] G.-J. Qi, C. C. Aggarwal, and T. Huang, "On clustering heterogeneous social media objects with outlier links," in *ACM international conference on Web search and data mining*, 2012, pp. 553–562.

[48] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *ICCV*, 2017.

[49] S. Chang, G.-J. Qi, C. C. Aggarwal, J. Zhou, M. Wang, and T. S. Huang, "Factorized similarity learning in networks," in *ICDM*, 2014, pp. 60–69.

[50] X. Wang, T. Zhang, G.-J. Qi, J. Tang, and J. Wang, "Supervised quantization for similarity search," in *CVPR*, 2016, pp. 2018–2026.

[51] J. Tang, X.-S. Hua, G.-J. Qi, and X. Wu, "Typicality ranking via semi-supervised multiple-instance learning." in *ACM Multimedia*, 2007, pp. 297–300.

[52] J. Wang, Z. Zhao, J. Zhou, H. Wang, B. Cui, and G. Qi, "Recommending flickr groups with social topic model," *Inf. Retr.*, vol. 15, no. 3-4, pp. 278–295, 2012.

[53] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *ICCV*, 2015.

[54] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, 2017.

[55] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015.

[56] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016.

[57] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[59] L. Wu, Y. Wang, X. Li, and J. Gao, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 5, pp. 1791–1802, 2019.

[60] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," in *arXiv preprint, arXiv:1409.0575*, 2014.

[61] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *arXiv:1611.07004*, 2016.

[62] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2007.

[63] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[64] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCV Workshop on Benchmarking Multi-Target Tracking*, 2016.

[65] B. Huang, J. Chen, Y. Wang, C. Liang, Z. Wang, and K. Sun, "Sparsity-based occlusion handling method for person re-identification," in *Multimedia Modeling*, 2015.

[66] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification," in *IJCAI*, 2016.

[67] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *CVPR*, 2018.

[68] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017.

[69] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *ICCV workshops*, 2017.

[70] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[72] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *arXiv:1606.04080*, 2016.

[73] Y. Wang and M. Hebert, "Learning to learn: model regression networks for easy small sample learning," in *ECCV*, 2016.