Crossing Generative Adversarial Networks for Cross-View Person Re-identification

Chengyuan Zhang[♮], **Lin Wu**[‡], **Yang Wang**[†]
[♮]School of Information Science and Engineering, Central South University, Changsha 410083, China
[‡]ISSR, ITEE, The University of Queensland, Brisbane, QLD, 4072, Australia
[†]The University of New South Wales, Kensington, Sydney, Australia
Correspondence to lin.wu@uq.edu.au

Abstract

Person re-identification (*re-id*) refers to matching pedestrians across disjoint yet non-overlapping camera views. The most effective way to match these pedestrians undertaking significant visual variations is to seek reliably invariant features that can describe the person of interest faithfully. Most of existing methods are presented in a supervised manner to produce discriminative features by relying on labeled paired images in correspondence. However, annotating pair-wise images is prohibitively expensive in labors, and thus not practical in large-scale networked cameras. Moreover, seeking comparable representations across camera views demands a flexible model to address the complex distributions of images. In this work, we study the co-occurrence statistic patterns between pairs of images, and propose to crossing Generative Adversarial Network (Cross-GAN) for learning a joint distribution for cross-image representations in a unsupervised manner. Given a pair of person images, the proposed model consists of the variational auto-encoder to encode the pair into respective latent variables, a proposed cross-view alignment to reduce the view disparity, and an adversarial layer to seek the joint distribution of latent representations. The learned latent representations are well-aligned to reflect the co-occurrence patterns of paired images. We empirically evaluate the proposed model against challenging datasets, and our results show the importance of joint invariant features in improving matching rates of person re-id with comparison to semi/unsupervised state-of-the-arts.

1 Introduction

Nowadays person re-identification (re-id) is emerging as a key problem in intelligent surveillance system, which deals with maintaining identities of individuals at physically different locations through non-overlapping camera views. Cross-view person re-id enables automated discovery and analysis of person specific long-term structural activities over wide areas, and is fundamental to many surveillance applications such as multi-camera people tracking and forensic search.

More recently, deep learning methods gradually gain the popularity in person re-id, which are developed to incorporate two aspects of feature extraction and metric learning into an integrated framework [Li *et al.*, 2014; Ahmed *et al.*, 2015; Wang *et al.*, 2016a; Xiao *et al.*, 2016; Chen *et al.*, 2016b; Wu *et al.*, 2016; Yi *et al.*, 2014; Varior *et al.*, 2016a]. The basic idea is to feed-forward a pair of input images into two CNNs with shared weights to extract features, and a subsequent metric learning part compares the features to measure the similarity. This process is carried out essentially by a classification on *cross-image representation* whereby images are coupled to extract their features, after which a parameterized classifier based on some distance measure (*e.g.,* Euclidean distance) performs an ordinary binary classification task to predict whether the two pedestrian images are from the same person. The cross-image representation is effective in capturing the relationship across pairs of images, and several approaches have been suggested to address horizontal displacement by local patch matching. For instance, the FPNN [Li *et al.*, 2014] algorithm introduced a patch matching layer for the CNN part at early layers. An improved deep learning architecture is proposed in [Ahmed *et al.*, 2015] with cross-input neighborhood differences and patch summary features.These two methods are both dedicated to improve the CNN architecture with a purpose to evaluate the pair similarity early in the CNN stage, so that it could make use of spatial correspondence of feature maps. Adding on, in [Varior *et al.*, 2016a], a matching gate is embedded into CNN to extract more locally similar patterns in horizontal correspondence across viewpoints. As for the metric learning part, with the aim to reduce the distance of matched images while enlarging the distance of mismatched images, common choices are pairwise and/or triplet comparison constraints. For example, [Li *et al.*, 2014; Ahmed *et al.*, 2015; Wu *et al.*, 2016] use the logistic loss to directly form a binary classification problem of whether the input image pair belongs to the same identity. In some other works, [Varior *et al.*, 2016a] adopts the contrastive loss based on pairwise comparison. [Chen *et al.*, 2016b] uses Euclidean distance and triplet loss while [Wang *et al.*, 2016a] optimizes

Figure 1: Left: Pedestrian images selected from CUHK03 dataset. Each column indicates images in pairs regarding the same person observed by disjoint camera views. Right: Illustration of co-occurrence regions in positive image pairs.

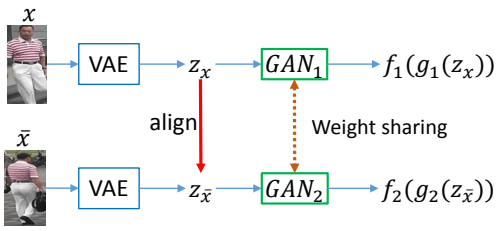


Figure 2: The schematic overview of the proposed crossing GAN for person re-id.

the combination loss function based on pairwise and triplet constraints.

However, these deep learning methods are inherently limited due to two presumable assumptions: the availability of large numbered labeled samples across views and the two fixed camera views are supposed to exhibit a unimodal inter-camera transform. In practice, building a training dataset with tuples of labeled corresponding images is impossible for every pair of camera views in the context of a large camera network in video surveillance. Thus, this correspondence dependency greatly limits the applicability of the existing approaches with training samples in correspondence. Secondly, the practical configurations (which are the combinations of view points, poses, lightings, and photometric settings) of pedestrian images are *multi-modal* and view-specific [Li and Wang, 2013] even if they are observed under the same camera. Therefore, the complex yet multi-modal inter-camera variations cannot be well learned with a generic metric which is incapable of handling multiple types of transforms across views. Last but not the least, existing deep learning methodologies directly compute the difference between intermediate CNN features and propagate only distance/similarity value to a ultimate scalar. This would lose important information since they did not consider feature alignment in cross-view.

1.1 Our Approach and Contributions

To overcome these limitations, we propose the crossing net based on a couple of generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] to seek effective cross-view representations for person re-id. To combat the first issue of relying on supervision, as shown in Fig.1, we observe some patterns that appear commonly across image pairs are distinct to discriminate positive pairs from negatives. Thus, these co-occurrence patterns should be mined out automatically to facilitate the task of re-id. Specifically, as shown in Fig.2, the proposed network starts from a tuple of variational auto-encoder (VAE) [Kingma and Welling, 2014], each for one image from a camera view, to encode the input images into their respective latent variables without any region-level annotations on person images. The technique of VAE has been established a viable solution for image distribution learning tasks while in this paper, we employ VAE to statistically generate latent variables for paired images without correspondence labeling. We remark that we don't use the Siamese Convolutional Neural Networks (CNNs) [Varior *et al.*, 2016a] to encode the input pair because CNNs are composed of fixed receptive fields which may not flexible to capture the varied local patterns. Also, the Siamese architecture enforces the weight sharing across CNN layers which are not suited for multi-modal view-specific variations.

To address the view disparity, we propose a cross-view alignment which is bridged over VAE outputs to allow the comparable matching. This alignment operation is to derive a shared latent space by modeling the statistical relationships between generative variables, and we empirically demonstrate this explicit alignment is crucial for cross-view representation learning (see Section 5.2). Then, the crossing net is coupled with adversarial networks to produce joint view-invariant distribution which gives a probability function to each joint occurrence of cross-view person images.

The major contributions of this paper can be summarized as follows:

- We extend the GAN to a dual setting, namely Cross-GAN, which is augmented with VAE to learn jointly invariant features for the task of person re-id in a unsupervised manner.
- The proposed Cross-GAN consists of a VAE layer to effectively encode image distributions w.r.t each camera view, a view-alignment layer to discover a shared latent space between cross-view images, and an adversarial network to produce the joint distribution of images.
- Extensive experiments are conducted to demonstrate our method outperforms semi/unsupervised state-of-the-art yet very comparable to supervised methods.

2 Related Work

2.1 Person Re-identification

The task of person re-identification can be accomplished by two categories of methods: (i) learning distance or similarity measures to predict if two images describe the same person [Li *et al.*, 2013; Xiong *et al.*, 2014; Li and Wang, 2013; Zheng *et al.*, 2011; Zhang *et al.*, 2016; Wu *et al.*, 2013a; Wang *et al.*, 2014a; Wang *et al.*, 2013b; chen *et al.*, 2016a; Wang and Wu, 2017; Wang *et al.*, 2017b; Huang *et al.*, 2016; Shi *et al.*, 2016], and (ii) designing distinctive signature to represent a person under different cameras, which typically performs classification on cross-image representation [Li *et al.*, 2014; Ahmed *et al.*, 2015; Varior *et al.*, 2016a; Wang *et al.*, 2016a; Wu *et al.*, 2018; chen *et al.*, 2016a].

For the first category of methodologies, they usually use many kinds of hand-crafted features including local binary patterns [Xiong *et al.*, 2014; Kostinger *et al.*, 2012; Wang *et al.*, 2015a; Wu *et al.*, 2017b; Wang *et al.*, 2013a; Wu *et al.*, 2013b; Wang *et al.*, 2014b; Wang *et al.*, 2016b], color histogram [Kostinger *et al.*, 2012; Wu and Wang, 2017; Wu *et al.*, 2017a; Wu *et al.*, 2017d], local maximal occurrence (LOMO) [Liao *et al.*, 2015; Liao and Li, 2015], and focus on learning an effective distance/similarity metric to compare the features. For the second category, deep convolutional neural networks are very effective in localizing/extracting relevant features to form discriminative representations against view variations. However, all these re-id models are in a supervised manner and rely on substantial labeled training data, which are typically required to be in pair-wise for each pair of camera views. Their performance depends highly on the quantity and quality of labeled training data, which also limits their application to large-scale networked cameras. In contrast, our method is based on unsupervised generative modeling which does not require any labeled data, and thus is free from prohibitively high cost of manual labeling and the risk of incorrect labeling.

A body of unsupervised methods have been developed to address person re-id without dependency on labeling [Liao *et al.*, 2015; Zhao *et al.*, 2013b; Yu *et al.*, 2017; Farenzena *et al.*, 2010; Wang *et al.*, 2015b; Wang *et al.*, 2016c; Wang *et al.*, 2017c; Zhou *et al.*, 2017; Peng *et al.*, 2016; Wu *et al.*, 2017c; Wang *et al.*, 2017a; Bak and Carr, 2017; Wang *et al.*, 2015c]. These models differ from ours in two aspects. On the one hand, these models do not explicitly model the view-specific information, i.e., they treat feature transformation/optimization in every distinct camera view in the same manner. In contrast, our models is propertied to employ VAE to generate view-specific latent variables, and then aim to find a shared subspace through a view-alignment layer. Thus, view-specific interference can be alleviated and common patterns can be attained in the representation learning. On the other hand, our method is the first attempt to introduce the adversarial learning into cross-view representation learning which can automatically discover co-occurrence patterns across images. While co-occurrence based statistics has been studied in some work [Zhang *et al.*, 2014; Galleguillos *et al.*, 2008; Ladicky *et al.*, 2010; Liao *et al.*, 2015], our approach diverts from the literature by aiming to jointly optimized invariant feature distributions for cross-image representations.

2.2 Deep Generative Models

In recent years, generative models have received an increasing amount of attention. Several approaches including variational auto-encoders (VAE) [Kingma and Welling, 2014; Rezende *et al.*, 2014], generative adversarial networks (GAN) [Goodfellow *et al.*, 2014], and attention models [Gregor *et al.*, 2015] have shown that learned deep networks are capable of generating new data points after the completion of training to learn an image distribution from unlabeled samples. Typically, determining the underlying data distribution of unlabeled images can be highly challenging and inference on such distributions is highly computationally expensive and or intractable except in the simplest of cases. VAE and GAN are the most prominent ones which provide efficient approximations, making it possible to learn tractable generative models of unlabeled images.

Our proposed network is inspired by the coupled generative adversarial networks [Liu and Tuzel, 2016], which learn a joint distribution of images without any tuple of corresponding images. It is demonstrated to be applied into domain adaptation and image transformation. Whilst our method has the sharing of coupled GANs in terms of enforcing weight sharing across the streamed GANs, our model is different from [Liu and Tuzel, 2016] on two facets. First, the model of [Liu and Tuzel, 2016] is originated from the same source of random vector as the uniform distribution for the generator of GANs whereas our method uses two respective VAE to generate the random vectors for two GANs. Second, our model has a cross-view alignment layer to seek a shared latent space for two distributions which is not provided in [Liu and Tuzel, 2016].

3 Preliminaries

Let $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ represent a pair of observations (*e.g.,* two images of pedestrians). We aim to learn a set of latent random variables $\boldsymbol{z}$ and $\bar{\boldsymbol{z}}$ ($\boldsymbol{z}$ and $\bar{\boldsymbol{z}}$ are linked by an alignment mapping), designed to capture the variations in the observed inputs while maintaining co-occurrence therein. To this end, we wish to estimate a prior $p(\boldsymbol{x})$ ($p(\bar{\boldsymbol{x}})$) by modeling the generation process of $\boldsymbol{x}$ ($\bar{\boldsymbol{x}}$) by sampling some $\boldsymbol{z}$ ($\bar{\boldsymbol{z}}$) from an arbitrary distribution $p(\boldsymbol{z})$ ($p(\bar{\boldsymbol{z}})$) as $p(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ ($p(\bar{\boldsymbol{x}}) = \int_{\bar{\boldsymbol{z}}} p(\bar{\boldsymbol{x}}|\bar{\boldsymbol{z}})p(\bar{\boldsymbol{z}})d\bar{\boldsymbol{z}}$). Fitting $p(\boldsymbol{x})$ ($p(\bar{\boldsymbol{x}})$) directly is intractable which involves expensive inference. We therefore approximate $p(\boldsymbol{x})$ and $p(\bar{\boldsymbol{x}})$ using VAE on each, respectively, because VAE offers a combination of highly flexible nonlinear mapping between the latent states and the observed output and effective approximate inference. To further induce joint invariant distribution between $\boldsymbol{z}$ and $\bar{\boldsymbol{z}}$, two respective VAEs are connected with two GANs through which the shared latent representations to images in individual can be attained by an adversary acting on pairs of $(\boldsymbol{x}, \bar{\boldsymbol{x}})$ data points and their latent codes $(\boldsymbol{z}, \bar{\boldsymbol{z}})$. In the remainder of this section, we provide brief introduction of VAE and GAN which we use to model the prior of pedestrian images and joint invariant distributions.

3.1 Variational Autoencoder (VAE)

A VAE comprises an encoder which estimates the posterior of latent variable and a decoder generates sample from latent variable as follows,

$$\boldsymbol{z} \sim \text{encoder}(\boldsymbol{x}) = q(\boldsymbol{z}|\boldsymbol{x}), \hat{\boldsymbol{x}} \sim \text{decoder}(\boldsymbol{z}_{\boldsymbol{x}}) = p(\boldsymbol{x}|\boldsymbol{z}). \tag{1}$$

The VAE regularizes the encoder by imposing a prior over the latent distribution on $p(\boldsymbol{z})$ while at the same time reconstructing $\hat{\boldsymbol{x}}$ to be as close as possible to the original $\boldsymbol{x}$. Typically, $q(\boldsymbol{z}|\boldsymbol{x})$ is taken to be a Gaussian prior, *i.e.,* $\boldsymbol{z} \sim \mathcal{N}(0, 1)$, which can be incorporated into a loss in the form of Kullback-Leibler divergence D_{KL} between the encoded distribution $q(\boldsymbol{z}|\boldsymbol{x})$ and the prior $p(\boldsymbol{z})$. Thus, the VAE loss takes the form

of the sum of the reconstruction error and latent prior:

$$\mathcal{L}_{vae} = D_{KL}\left(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})]. \quad (2)$$

We use the VAE to be an effective modelling paradigm to recover the complex multi-modal distributions of images over the data space. A VAE introduces a set of latent random variables $\boldsymbol{z}$, designed to capture the variations in the observed variable $\boldsymbol{x}$.

3.2 Generative Adversarial Networks (GAN)

A GAN consists of a generator and a discriminator. The objective of the generator is to synthesize images resembling real images, while the objective of the discriminator is to distinguish real images from synthesized ones. Let $\boldsymbol{x}$ be a natural image drawn from distribution p_X, and $\boldsymbol{z}$ be a random vector in $\mathbb{R}^d$. Let g and f be the generative and discriminative models, respectively. The generator synthesizes samples by mapping a random sample $\boldsymbol{z}$, from an arbitrary distribution, to a sample as output image $g(\boldsymbol{z})$, that has the same vector support as $\boldsymbol{x}$. Denote the distribution of $g(\boldsymbol{z})$ as p_G. The discriminator tries to distinguish between real data sample $\boldsymbol{x}$, and synthesized sample $g(\boldsymbol{z})$ by estimating the probability that an input image is drawn from p_X. The loss function for the GAN can be formulated as a binary entropy loss as follows:

$$\mathcal{L}_{gan}(f, g) = \log f(\boldsymbol{x}) + \log(1 - f(g(\boldsymbol{z}))). \quad (3)$$

Training on Eq.(3) alternatives between minimizing $\mathcal{L}_{gan}$ w.r.t. parameters of the generator while maximizing $\mathcal{L}_{gan}$ w.r.t. parameters of the discriminator. The generator tries to minimize the loss to generate more realistic samples to fool the discriminator while the discriminator tries to maximize the loss.

In practice, Eq.(3) is solved by alternating the following gradient update steps:

- $\boldsymbol{\theta}_f^{t+1} = \boldsymbol{\theta}_f^t - \lambda^t \nabla_{\boldsymbol{\theta}_f} \mathcal{L}_{gan}(f^t, g^t)$,
- $\boldsymbol{\theta}_g^{t+1} = \boldsymbol{\theta}_g^t - \lambda^t \nabla_{\boldsymbol{\theta}_g} \mathcal{L}_{gan}(f^{t+1}, g^t)$.

where $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$ are parameters of f and g, λ is the learning rate, and t is the iteration number. The GAN does not explicitly model reconstruction loss of the generator; instead, network parameters are updated by back-propagating gradients only from the discriminator. This strategy can effectively avoid pixel-wise loss functions that tend to produce overly smoothed results and enables realistic modeling of noise as present in the training set. Thus, GAN can be used to synthesize images, i.e., the distribution p_G converges to p_X, given enough capacity f and g and sufficient training iterations [Goodfellow *et al.*, 2014].

4 The Method

4.1 System Overview: Crossing GANs

The complete network is then trained end-to-end for learning a joint invariant distribution of images across camera views. Fig.3 illustrates the overview of our architecture. It consists of a pair of (VAE, GAN)s, that is, (VAE_1, GAN_1) and (VAE_2, GAN_2); each is responsible for synthesizing one image in one camera view. In Fig.3, the blue and green routes represent the forward paths of the VAE and GAN for images $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$, respectively. The blue route, *i.e.,* the VAE flow, is the use of expressive latent variables to model the variability observed in the data. It essentially captures the statistics of each individual image. The auto-encoding procedure is explained in Section 4.2. The red route denotes the cross-view alignment that links the latent variables $(\boldsymbol{z}_{\boldsymbol{x}}, \boldsymbol{z}_{\bar{\boldsymbol{x}}})$ to ensure the shared latent representations. The details of alignment is given in Section 4.3. The green routes represent the adversarial learning which works to optimize optimal latent features corresponding to the joint invariance across paired images. During training, the two GANs are enforced to share a subset of parameters (the brown routes), which results in synthesized pairs of corresponding images without correspondence supervision. The details are described in Section 4.4.

4.2 Auto-encoding

Given a pair of data points $(\boldsymbol{x}^{(i)}, \bar{\boldsymbol{x}}^{(i)})$ from a dataset $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}, \bar{\boldsymbol{x}}^{(i)}\}_{i=1}^{M}$ containing $N = 2M$ samples in M pairs. The auto-encoding algorithm uses unobserved random variable $\boldsymbol{z}^{(i)}$, to generate a data point $\boldsymbol{x}^{(i)}$. As the generating process can be repeated on either $\boldsymbol{x}^{(i)}$ or $\bar{\boldsymbol{x}}^{(i)}$, in the following, we describe $\boldsymbol{x}^{(i)}$ as illustration. The process is composed of two phases: (1) a value $\boldsymbol{z}^{(i)}$ is generated from some prior distribution $p(\boldsymbol{z}^{(i)})$; (2) a value $\boldsymbol{x}^{(i)}$ is generated from some conditional distribution $p(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i)})$. From a coding theory perspective, the unobserved variable $\boldsymbol{z}^{(i)}$ have an interpretation as a latent representation or *code*. Following VAE [Kingma and Welling, 2014] which introduces a recognition model $q(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$: an approximation to the intractable true posterior $p(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$, we will therefore refer to $q(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)})$ as a probabilistic *encoder*, since given a data point $\boldsymbol{x}^{(i)}$ its produces a distribution (*e.g.,* a Gaussian) over the possible values of the code $\boldsymbol{z}^{(i)}$ from which the data point $\boldsymbol{x}^{(i)}$ could be generated. In a similar vein, we refer to $p(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i)})$ as a probabilistic *decoder*, since given a code $\boldsymbol{z}^{(i)}$ it produces a distribution over the possible corresponding values of $\boldsymbol{x}^{(i)}$.

In this work, neural networks are used as probabilistic encoders and decoders, namely multi-layered perceptions (MLPs). Let the prior over the latent variables be the centered isotropic multivariate Gaussian $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$ whose distribution parameters are computed from $\boldsymbol{z}$ with a MLP. We assume the true posterior $p(\boldsymbol{z}|\boldsymbol{x})$ takes on an approximate Gaussian form with an approximately diagonal covariance. In this case, we can let the variational approximate posterior be a multivariate Gaussian with a diagonal covariance structure:

$$\log q(\boldsymbol{z}^{(i)}|\boldsymbol{x}^{(i)}) = \log \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\boldsymbol{I}) \quad (4)$$

where the mean and standard of the approximate posterior, $\boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{(i)}$ are outputs of the encoding MLP. *i.e.,* nonlinear functions of data point $\boldsymbol{x}^{(i)}$ and the variational parameters.

Specifically, we sample from the posterior $\boldsymbol{z}^{(i)} \sim q(\boldsymbol{z}|\boldsymbol{x}^{(i)})$ using $\boldsymbol{z}^{(i)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. With $\odot$ we signify an element-wise product. In this model, both $p(\boldsymbol{z})$ and $q(\boldsymbol{z}|\boldsymbol{x})$ are Gaussian. The resulting estimator

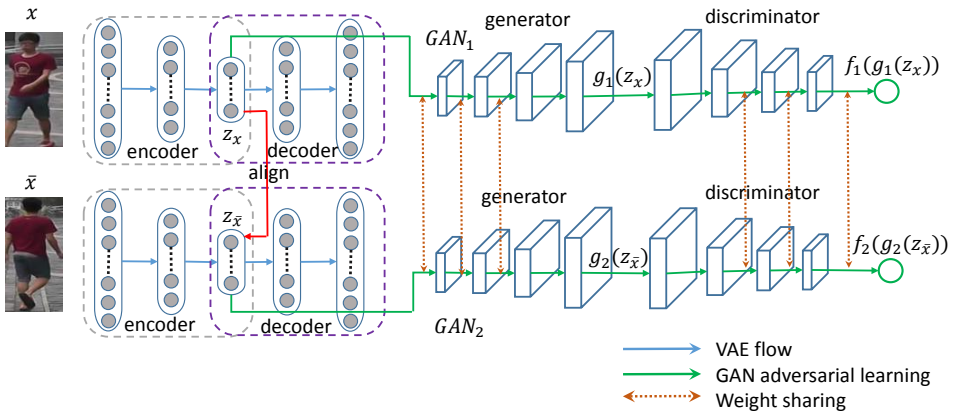


Figure 3: Architecture overview. Best view in color.

loss for data point $\boldsymbol{x}^{(i)}$ is:

$$\begin{aligned}
&\tilde{L}_{vae}(\boldsymbol{x}^{(i)}) = D_{KL}(q(\boldsymbol{z}|\boldsymbol{x}^{(i)}||p(\boldsymbol{z}))) - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p(\boldsymbol{x}^{(i)}|\boldsymbol{z})] \\
&= D_{KL}(q(\boldsymbol{z}|\boldsymbol{x}^{(i)}||p(\boldsymbol{z}))) - \log p(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i)}) \\
&\simeq \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2\right) - \log p(\boldsymbol{x}^{(i)}|\boldsymbol{z}^{(i)}) \\
&\boldsymbol{z}^{(i)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})
\end{aligned} \tag{5}$$

where the KL-divergence $D_{KL}(q(\boldsymbol{z}|\boldsymbol{x}^{(i)}||p(\boldsymbol{z})))$ can be integrated analytically, such that only the expected reconstruction error $\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p(\boldsymbol{x}^{(i)}|\boldsymbol{z})]$ requires estimation by sampling. Given multiple data points from a dataset $\boldsymbol{X}$ with M pairs of data points, we can construct an estimator loss of as follows:

$$\tilde{L}_{vae}(\boldsymbol{X}) = \frac{1}{M}\sum_{i=1}^{M}\left(\tilde{L}_{vae}(\boldsymbol{x}^{(i)}) + \tilde{L}_{vae}(\bar{\boldsymbol{x}}^{(i)})\right). \tag{6}$$

4.3 Learning Cross-View Alignment on Latent Codes

In this section, we introduce cross-view alignment over latent representations provided by VAE, which is capable of modeling complex multi-modal distributions over data space. Note that for notation convenience, we use $\boldsymbol{z}_{\boldsymbol{x}}$ and $\boldsymbol{z}_{\bar{\boldsymbol{x}}}$ to distinguish the latent representation for $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$.

$$\mathcal{L}_{align} = \max(||\boldsymbol{z}_{\boldsymbol{x}} - \text{ Align}(\boldsymbol{z}_{\bar{\boldsymbol{x}}})||^2, \tau), \tag{7}$$

where we model Align$(\cdot)$ as a single fully connected neuron with a $tanh$ activation function. The threshold τ is $\tau = 1$. In essence, Align$(\cdot)$ is implicitly learning a mapping across two normal distributions $(\boldsymbol{z}_{\boldsymbol{x}}, \boldsymbol{z}_{\bar{\boldsymbol{x}}})$. The parameters of the mapping $\boldsymbol{\theta}_{Align}$ are optimized through back-propagation. Since both the VAE and the GAN are able to learn low-dimensional representations (in our case, both $\boldsymbol{z}_{\boldsymbol{x}}, \boldsymbol{z}_{\bar{\boldsymbol{x}}}$ are set to be 100 dimensions.), we are able to fit the cross-view alignment with moderate pairs.

The strategy of alignment is designed to align the transformation across cameras by revealing underlying invariant properties among different views. As a result, unsupervised matching pedestrian images can be statistically inferred through aligned latent representations. This is motivated by the observation that some regions are distributed similarly in images across views and robustly maintain their appearance in the presence of large cross-view variations.

4.4 Adversarial Learning

Generator

Let g_1 and g_2 be the generators of GAN_1 and GAN_2, which map corresponding inputs $\boldsymbol{z}_{\boldsymbol{x}}$ and $\boldsymbol{z}_{\bar{\boldsymbol{x}}}$ to images that have the same support as $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$, respectively. Both g_1 and g_2 are realized as convolutions [Radford *et al.*, 2015]:

$$\begin{aligned}
g_1(\boldsymbol{z}_{\boldsymbol{x}}) &= g_1^{(m)}(g_1^{(m-1)}(\ldots g_1^{(2)}(g_1^{(1)}(\boldsymbol{z}_{\boldsymbol{x}})))), \\
g_2(\boldsymbol{z}_{\bar{\boldsymbol{x}}}) &= g_2^{(m)}(g_2^{(m-1)}(\ldots g_2^{(2)}(g_2^{(1)}(\boldsymbol{z}_{\bar{\boldsymbol{x}}})))),
\end{aligned} \tag{8}$$

where $g_1^{(i)}$ and $g_2^{(i)}$ are the i-th layer of g_1 and g_2 and m is the number of layers in generators. Through layers of convolution operations, the generator gradually decode information from more abstract concept to more material details. The first layer decode high-level semantics while the last layer decode low-level details. Note this information flow is opposite to that in a standard deep neural network [Krizhevsky *et al.*, 2012] where the first layers extract low-level features while the last layers extract high-level features. Based on the observation that a pair of person images from two camera views share the same high-level concept (i.e., they belong to the same identity but with different visual appearance), we enforce the first layers of g_1 and g_2 to have identical structures and share the weights, which means $\boldsymbol{\theta}_{g_1^{(i)}} = \boldsymbol{\theta}_{g_2^{(i)}}$, for $i = 1, 2, \ldots, k$ where k is the number of shared layers, and $\theta_{g_1^{(i)}}$ and $\theta_{g_2^{(i)}}$ are the parameters of $g_1^{(i)}$ and $g_2^{(i)}$, respectively. This constraint can force the high-level semantics to be decoded in the same way in g_1 and g_2, which can also be propagated into the VAE to update the parameters simultaneously. Thus, the generator can gradually decode the information from more abstract concepts to more finer details, and the view-alignment is embedded to ensure the common finer regions can be preserved with high correlations.

Discriminator

Let f_1 and f_2 be the discriminators of GAN_1 and GAN_2 given by

$$\begin{aligned} f_1(\boldsymbol{x}) &= f_1^{(n)}(f_1^{(n-1)}(\dots f_1^{(2)}(f_1^{(1)}(\boldsymbol{x})))), \\ f_2(\bar{\boldsymbol{x}}) &= f_2^{(n)}(f_2^{(n-1)}(\dots f_2^{(2)}(f_2^{(1)}(\bar{\boldsymbol{x}})))); \end{aligned} \tag{9}$$

where $f_1^{(i)}$ and $f_2^{(i)}$ are the i-th layer of f_1 and f_2, and n is the number of layers. Note that GAN_1 and GAN_2 have the identical network structure. The discriminator maps an input image to a probability score, estimating the likelihood that the input is drawn from a true data distribution. The first layers of the discriminator extract low-level features while the last layers of layers extract high-level features. Considering that input image pair are realizations of the same person in two camera views, we force f_1 and f_2 to have the same last layers, which is achieved by sharing the weights of the last layers via $\boldsymbol{\theta}_{f_1^{(n-i)}} = \boldsymbol{\theta}_{f_2^{(n-i)}}$, for $i = 0, 1, \dots, l-1$ where l is the number of weight-sharing layers in the discriminator, and $\boldsymbol{\theta}_{f_1^{(i)}}$ and $\boldsymbol{\theta}_{f_2^{(i)}}$ are the network parameters of $f_1^{(i)}$ and $f_2^{(i)}$, respectively. The weight-sharing constraints herein helps reduce the number of trainable parameters of the network, and also effective in deriving view-invariant features in joint distribution across $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$.

Therefore, we cast the problem of learning jointly invariant feature distribution as a constrained objective function with the training loss given by

$$\begin{aligned} \mathcal{L}_{gan}(f_1, f_2, g_1, g_2) &= \log f_1(\boldsymbol{x}) + \log(1 - f_1(g_1(\boldsymbol{z}))) \\ &\quad + \log f_2(\bar{\boldsymbol{x}}) + \log(1 - f_2(g_2(\boldsymbol{z}_{\bar{\boldsymbol{x}}}))) \\ &\text{subject to } \boldsymbol{\theta}_{g_1^{(j)}} = \boldsymbol{\theta}_{g_2^{(j)}},\ j = 1, 2, \dots, k \\ &\quad \boldsymbol{\theta}_{f_1^{(n-i)}} = \boldsymbol{\theta}_{f_2^{(n-i)}},\ i = 0, 1, \dots, l-1 \end{aligned} \tag{10}$$

The crossing GAN can be interpreted as minimax game with two teams and each team has two players.

4.5 Implementation Details

Given a dataset $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}, \bar{\boldsymbol{x}}^{(i)}\}_{i=1}^{M}$ where $N = 2M$ is the total number of data points.

$$\mathcal{L}_{align} = \frac{1}{M} \sum_{i=1}^{M} \max(||\boldsymbol{z}_{\boldsymbol{x}^{(i)}} - \text{Align}(\boldsymbol{z}_{\bar{\boldsymbol{x}}^{(i)}})||^2, \tau), \tag{11}$$

$$L = \mathcal{L}_{vae} + \mathcal{L}_{align} - \mathcal{L}_{gan} \tag{12}$$

In this work, we adopt a deep convolutional GAN framework architecture [Radford *et al.*, 2015] and feature matching strategy [Salimans *et al.*, 2016] for stable and fast-converging training. The visualization of model is shown in Table 1. Specifically, we use all convolutional nets to replace deterministic spatial pooling functions (such as max pooling) with strided convolutions. This allows the network to learn its own spatial down-sampling. We use this approach in our generator, allowing it to learn its own spatial up-sampling, and discriminator. The overview architecture of Cross-GAN is shown in Table 1, and the training procedure is summarized in Algorithm 1.

Algorithm 1 Mini-batch stochastic gradient descent for training crossing generative adversarial nets.

input : Mini-batch of training samples in pairs $\boldsymbol{X} = \{\boldsymbol{x}^{(j)}, \bar{\boldsymbol{x}}^{(j)}\}_{j=1}^{M}$

output: Parameters of VAE, alignment, and two GANs

Initialize parameters for VAE and alignment: $\boldsymbol{\theta}_{vae}$, $\boldsymbol{\theta}_{align}$

Initialize parameters $\boldsymbol{\theta}_{f_1^{(i)}}, \boldsymbol{\theta}_{f_1^{(i)}}, \boldsymbol{\theta}_{g_1^{(j)}}, \boldsymbol{\theta}_{g_2^{(j)}}$ with the shared network connection weights set to the same values. **for** $t = 0, 1, 2, \dots, T$ **do**

`/* update parameters of VAE */`

repeat

Draw M samples from camera view A, $\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\}$ Draw M samples from camera view B, $\{\bar{\boldsymbol{x}}^{(1)}, \dots, \bar{\boldsymbol{x}}^{(M)}\}$ $\boldsymbol{\epsilon} \leftarrow$ random samples from noise distribution $p(\boldsymbol{\epsilon})$ Compute gradients of the estimator of Eq.(5): $\boldsymbol{e} \leftarrow \bigtriangledown_{\boldsymbol{\theta}_{vae}} \frac{1}{M} \sum_{j=1}^{M} \left(\tilde{L}_{vae}(\boldsymbol{x}^{(j)}) + \tilde{L}_{vae}(\bar{\boldsymbol{x}}^{(j)}) \right)$ Update parameters of $\boldsymbol{\theta}_{vae}$ using gradients $\boldsymbol{e}$ (*e.g.*, SGD or Adagrad [Duchi *et al.*, 2010])

until *convergence of parameters* $\boldsymbol{\theta}_{vae}$;

`/* update parameters of` $\boldsymbol{\theta}_{align}$ `*/`

Compute the gradients of the parameters of the alignment $\bigtriangledown \mathcal{L}_{align}$ (eq.(7)) `/* update parameters of two GANs */`

Draw M samples from $p(\boldsymbol{z})$, $\{\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(M)}\}$ Compute the gradients of the parameters of the discriminator, f_1^t, $\Delta\boldsymbol{\theta}_{f_1^{(i)}}$ $\bigtriangledown_{\boldsymbol{\theta}_{f_1^{(i)}}} \frac{1}{M} \sum_{j=1}^{M} \log f_1^t(\boldsymbol{x}^{(j)}) + \log(1 - f_1^t(g_1^t(\boldsymbol{z}_{\boldsymbol{x}^{(j)}}^{(j)})))$ Compute the gradients of the parameters of the discriminator, f_2^t, $\Delta\boldsymbol{\theta}_{f_2^{(i)}}$ $\bigtriangledown_{\boldsymbol{\theta}_{f_2^{(i)}}} \frac{1}{M} \sum_{j=1}^{M} \log f_2^t(\bar{\boldsymbol{x}}^{(j)}) + \log(1 - f_2^t(g_2^t(\boldsymbol{z}_{\bar{\boldsymbol{x}}^{(j)}}^{(j)})))$

end

Table 1: The network architecture of Cross-GANs.

Layer	Generator		
	View 1	View 2	Shared?
1	Conv (N=20, K=5 × 5, S=1), BN, ReLU	Conv (N=20, K=5 × 5, S=1), BN, ReLU	Yes
2	Conv (N=20, K=5 × 5, S=1), BN, ReLU	Conv (N=20, K=5 × 5, S=1), BN, ReLU	Yes
3	Conv (N=20, K=5 × 5, S=1), BN, ReLU	Conv (N=20, K=5 × 5, S=1), BN, ReLU	Yes
4	Conv (N=20, K=3 × 3, S=1), BN, ReLU	Conv (N=20, K=3 × 3, S=1), BN, ReLU	Yes
5	Conv (N=20, K=3 × 3, S=1), BN	Conv (N=20, K=3 × 3, S=1), BN	No
	Discriminator		
	View 1	View 2	Shared?
1	Conv (N=20, K=5 × 5, S=1), MAX-POOL (S=2), LeakyReLU	Conv (N=20, K=5 × 5, S=1), MAX-POOL (S=2), LeakyReLU	No
2	Conv (N=20, K=5 × 5, S=1), MAX-POOL (S=2), LeakyReLU	Conv (N=20, K=5 × 5, S=1), MAX-POOL (S=2), LeakyReLU	No
3	Conv (N=20, K=5 × 5, S=1), MAX-POOL (S=2), LeakyReLU	Conv (N=20, K=5 × 5, S=1), MAX-POOL (S=2), LeakyReLU	No
4	FC (N=1024), ReLU	FC (N=1024), ReLU	No
5	FC (N=1024), Sigmoid	FC (N=1024), Sigmoid	Yes

Figure 4: Examples from person re-identification datasets: VIPeR (left), CUHK03 (middle), and Market-1501 (right). Columns indicate the same identities.

5 Experiments

5.1 Datasets and Settings

We perform experiments on three benchmarks: VIPeR [Gray *et al.*, 2007], CUHK03 [Li *et al.*, 2014], and Market-1501 data set [Zheng *et al.*, 2015].

- The **VIPeR** data set [Gray *et al.*, 2007] contains 632 individuals taken from two cameras with arbitrary viewpoints and varying illumination conditions. The 632 person's images are randomly divided into two equal halves, one for training and the other for testing.
- The **CUHK03** data set [Li *et al.*, 2014] 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 pedestrian bounding boxes and bounding boxes automatically obtained by running a pedestrian detector [Felzenszwalb *et al.*, 2010]. In our experiment, we report results on labeled data set. The dataset is randomly partitioned into training, validation, and test with 1160, 100, and 100 identities, respectively.
- The **Market-1501** data set [Zheng *et al.*, 2015] contains 32,643 fully annotated boxes of 1501 pedestrians, making it the largest person re-id dataset to date. Each identity is captured by at most six cameras and boxes of person are obtained by running a detector of Deformable Part Model (DPM) [Huang *et al.*, 2015]. The dataset is randomly divided into training and testing sets, containing 750 and 751 identities, respectively.

We use the deep convolutional networks to instantiate the GANs in Cross-GAN. The two generative models have an identical structure with 5 convolutional layers. The generator is realized using the convolutions of ResNet-50 [He *et al.*, 2016] with fine-tuned parameters on re-id [Zheng *et al.*, 2017a]. Following [Liu and Tuzel, 2016], we use the batch normalization processing and the parameter sharing is applied on all convolutional layers except the last convolution. For the discriminative models, we use three fully connected layers with hidden units of 1,024 on each layer. The inputs to the discriminative models are batches containing the output images from the generators and images from each training subsets. Also, each training set is equally divided into two non-overlapping subsets, which are used to train two GANs respectively. The Adam algorithm [Kingma and Ba, 2015] is used for training, the learning rate is set to be 0.002, the momentum parameter is 0.5, and the mini-batch size is 128. The training is performed 30,000 iterations.

The evaluation protocol we adopt is the widely used single-shot modality to allow extensive comparison. Each probe image is matched against the gallery set, and the rank of the true match is obtained. The rank-k recognition rate is the expectation of the matches at rank k, and the cumulative values of the recognition rate at all ranks are recorded as the one-trial Cumulative Matching Characteristic (CMC) results. This evaluation is performed ten times, and the average CMC results are reported.

5.2 Ablation Studies

The Impact of Cross-View Alignment

In this experiment, we study the impact of cross-view alignment which is demonstrated to be essential to the person matching. To quantifying the performance with/without cross-view alignment, we transform the query images generated by g_1 to the gallery view by using the same method employed for generating the training image in the gallery camera view. Then we compare the transformed images with the images generated by the g_2. The performance is measured by the average of the ratios of agreed pixels between the transformed image and the corresponding image in the gallery view. The pixel agreement ratio is the number of corresponding pixels that have the same value in the two images divided by the total image size. The experimental results are shown in Fig.5, and it can be observed that with the cross-view alignment strategy, the rendered pairs of images (positive or negative) resembled true pairs drawn from the joint distributions.

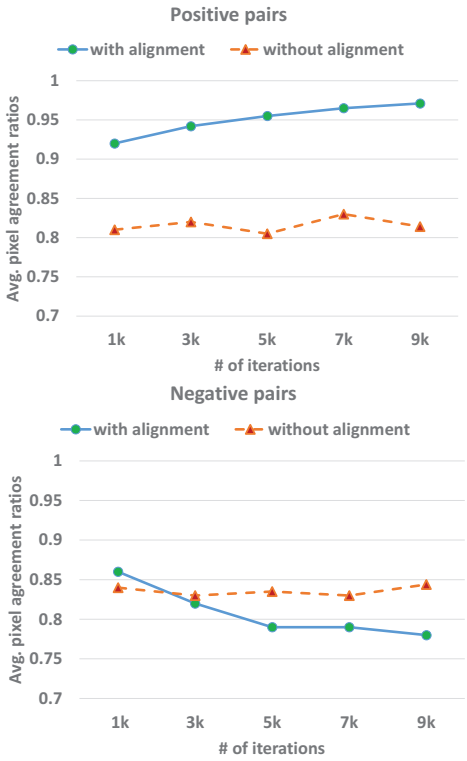


Figure 5: The average agreement ratios of the Cross-GAN with/without cross-view alignment on VIPeR dataset.

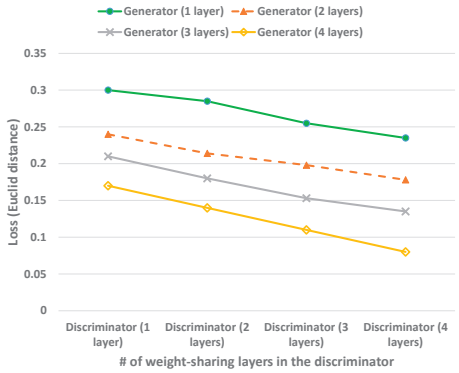


Figure 6: The loss function (Euclidean distance) measuring the difference between two images from VIPeR with respect to different weight-sharing configurations in the coupled generators and discriminators. It can be seen that the performance is positively correlated with the number of weight-sharing layers in the generative models but less correlated with the number of weight-sharing layers in the discriminators.

The Impact of Weight-Sharing in GAN

The weight-sharing constraint and adversarial learning are crucial for co-occurrence pattern encoding/generation across images without requirement on the labeled pair in correspondence. In our model, each sample can be separately drawn from the marginal distribution $p_{\boldsymbol{x}_1}$ and $p_{\boldsymbol{x}_2}$, and not rely on samples in correspondence with joint distribution of $p_{\boldsymbol{x}_1,\boldsymbol{x}_2}$. The adversarial learning encourages the generators to produce realistic images individually resembling to respective view domains, while the weight-sharing can capture the correspondence between two views automatically.

In this experiment, we study the weight-sharing effect for the adversarial training by varying the number of weight-sharing layers in both generative and discriminative models. If image x_1 is from the probe view, and the Cross-GAN is trained to find the correct matching image $\bar{x}_1$ from the gallery view such that the joint probability density $p(x_1, \bar{x}_1)$ is maximized. Let L be the loss function measuring the difference between two images, e.g., L is implemented to be the Euclidean distance in this experiment. Given g_1 and g_2, we aim to seek the transformation by finding the random vector that generates the query image via $z^* = \arg\min_z L(g_1(z), x_1)$. With z^* found, one can apply g_2 to produce the transformed image $\bar{x}_1 = g_2(z^*)$. In Fig.6, we show the loss computed on cross-image transformation matching by using Euclidean distance on VIPeR with varied weight-sharing configurations. It can be seen that the matching performance is positively correlated with the number of weight-sharing layers in the generative models, while less correlated with the number of weight-sharing layers in the discriminative models.

Table 2: Comparison results with state-of-the-arts on the VIPeR dataset (test person =316).

	Method	R=1	R=10	R=20
Semi/un-supervised	Cross-GAN	**49.28**	**91.66**	**93.47**
	LADF [Li *et al.*, 2013]	29.34	75.98	88.10
	SDALF [Farenzena *et al.*, 2010]	19.87	49.37	65.73
	eSDC [Zhao *et al.*, 2013b]	26.31	58.86	72.77
	t-LRDC [Zheng *et al.*, 2016]	27.40	46.00	75.10
	OSML [Bak and Carr, 2017]	34.30	-	-
	CAMEL [Yu *et al.*, 2017]	30.90	52.00	72.50
	OL-MANS [Zhou *et al.*, 2017]	44.90	74.40	93.60
	SalMatch [Zhao *et al.*, 2013a]	30.16	62.50	75.60
Supervised	MLF [Zhao *et al.*, 2014]	29.11	65.20	79.90
	LocallyAligned [Li and Wang, 2013]	29.60	69.30	86.70
	JointRe-id [Ahmed *et al.*, 2015]	34.80	74.79	82.45
	SCSP [chen *et al.*, 2016a]	53.54	91.49	**96.65**
	Multi-channel [Cheng *et al.*, 2016]	47.80	84.80	91.10
	DNSL [Zhang *et al.*, 2016]	42.28	82.94	92.06
	JSTL [Xiao *et al.*, 2016]	38.40	-	-
	SI-CI [Wang *et al.*, 2016a]	35.80	83.50	-
	S-LSTM [Varior *et al.*, 2016b]	42.40	79.40	-
	S-CNN [Varior *et al.*, 2016a]	37.80	77.40	-
	SpindleNet [Zhao *et al.*, 2017a]	**53.80**	90.10	96.10
	Part-Aligned [Zhao *et al.*, 2017b]	48.70	87.70	93.00
	Deep-Embed [Wu *et al.*, 2018]	49.00	91.10	96.20

5.3 Comparison with State-of-the-arts

In this subsection, we extensively compare the proposed Cross-GAN with a number of state-of-the-art semi/unsupervised and supervised methods on three datasets. Semi/unsupervised methods include LADF [Li *et al.*, 2013], SDALF [Farenzena *et al.*, 2010], eSDC [Zhao *et al.*, 2013b], t-LRDC [Zheng *et al.*, 2016], OSML [Bak and Carr, 2017],

Table 3: Rank-1, -10, -20 recognition rate of various methods on the CUHK03 data set (test person =100).

	Method	R=1	R=10	R=20
Semi/un-supervised	Cross-GAN	**83.23**	**96.73**	**99.47**
	OSML [Bak and Carr, 2017]	45.61	85.43	88.50
	LSRO [Zheng *et al.*, 2017b]	84.62	97.64	99.80
	CAMEL [Yu *et al.*, 2017]	31.90	76.62	80.63
	eSDC [Zhao *et al.*, 2013b]	8.76	38.28	53.44
	UMDL [Peng *et al.*, 2016]	1.64	8.43	10.24
	OL-MANS [Zhou *et al.*, 2017]	61.70	92.40	98.52
	XQDA [Liao *et al.*, 2015]	52.20	92.14	96.25
Supervised	FPNN [Li *et al.*, 2014]	20.65	51.32	83.06
	kLFDA [Xiong *et al.*, 2014]	48.20	66.38	76.59
	DNSL [Zhang *et al.*, 2016]	58.90	92.45	96.30
	JointRe-id [Ahmed *et al.*, 2015]	54.74	91.50	97.31
	E-Metric [Shi *et al.*, 2016]	61.32	96.50	97.50
	S-LSTM [Varior *et al.*, 2016b]	57.30	88.30	-
	S-CNN [Varior *et al.*, 2016a]	61.80	88.30	-
	Deep-Embed [Wu *et al.*, 2018]	73.00	94.60	98.60
	SpindleNet [Zhao *et al.*, 2017a]	**88.50**	98.80	99.20
	Part-Aligned [Zhao *et al.*, 2017b]	85.40	98.60	99.90

Table 4: Rank-1, -10, -20 recognition rate and mAP of various methods on the Market-1501 data set (test person =751). All results are evaluated on single-shot setting.

	Method	R=1	R=10	R=20	mAP
Semi/un-supervised	Cross-GAN	**72.15**	**94.3**	**97.5**	**48.24**
	eSDC [Zhao *et al.*, 2013b]	33.54	60.61	67.53	13.54
	SDALF [Farenzena *et al.*, 2010]	20.53	-	-	8.20
	LSRO [Zheng *et al.*, 2017b]	**83.97**	95.64	97.56	66.07
	CAMEL [Yu *et al.*, 2017]	54.56	84.67	87.03	-
	OL-MANS [Zhou *et al.*, 2017]	60.72	89.80	91.87	-
	PUL [Fan *et al.*, 2017]	45.53	72.75	72.65	-
	UMDL [Peng *et al.*, 2016]	34.54	62.60	68.03	-
	XQDA [Liao *et al.*, 2015]	43.79	75.32	80.41	22.22
	BoW [Zheng *et al.*, 2015]	34.40	-	-	14.09
Supervised	JSTL [Xiao *et al.*, 2016]	44.72	77.24	82.00	-
	KISSME [Kostinger *et al.*, 2012]	39.35	-	-	19.12
	kLFDA [Xiong *et al.*, 2014]	44.37	-	-	23.14
	SCSP [chen *et al.*, 2016a]	51.90	-	-	26.35
	DNSL [Zhang *et al.*, 2016]	61.02	-	-	35.68
	S-CNN [Varior *et al.*, 2016a]	65.88	-	-	39.55
	Deep-Embed [Wu *et al.*, 2018]	68.32	94.59	96.71	40.24
	SpindleNet [Zhao *et al.*, 2017a]	76.90	-	-	-
	Part-Aligned [Zhao *et al.*, 2017b]	81.00	-	-	-

CAMEL [Yu *et al.*, 2017], OL-MANS [Zhou *et al.*, 2017], SalMatch [Zhao *et al.*, 2013a], UMDL [Peng *et al.*, 2016], XQDA [Liao *et al.*, 2015], and PUL [Fan *et al.*, 2017]. Supervised methods include MLF [Zhao *et al.*, 2014], LocallyAligned [Li and Wang, 2013], JointRe-id [Ahmed *et al.*, 2015], SCSP [chen *et al.*, 2016a], Multi-channel [Cheng *et al.*, 2016], DNSL [Zhang *et al.*, 2016], JSTL [Xiao *et al.*, 2016], SI-CI [Wang *et al.*, 2016a], S-CNN [Varior *et al.*, 2016a], SpindleNet [Zhao *et al.*, 2017a], Part-Aligned [Zhao *et al.*, 2017b], FPNN [Li *et al.*, 2014], S-LSTM [Varior *et al.*, 2016b], kLFDA [Xiong *et al.*, 2014], KISSME [Kostinger *et al.*, 2012], E-Metric [Shi *et al.*, 2016] and Deep-Embed [Wu *et al.*, 2018]. Please note that not all methods report their matching results on three datasets and the CMC values are quoted from their papers.

The comparison results are reported in Table 2, Table 3, and Table 4 for VIPeR, CUHK03, and Market-1501 respectively. The CMC curves of unsupervised/semi-supervised methods on three datasets are shown in Fig.7.

In the VIPeR dataset, Cross-GAN notably outperforms all semi/unsupervised competitors by achieving rank-1=49.28. Compared with unsupervised feature encoding methods such as SDALF [Farenzena *et al.*, 2010], eSDC [Zhao *et al.*, 2013b], and SalMatch [Zhao *et al.*, 2013a], the proposed method of Cross-GAN is able to learn deep local features with joint distribution and thus robust against visual variations. Also, our method is very comparable to the state-of-the-art supervised method of SpindleNet [Zhao *et al.*, 2017a] which obtains rank-1=53.80.

In the CUHK03 dataset, the proposed Cross-GAN outperforms all state-of-the-art unsupervised method except LSRO [Zheng *et al.*, 2017b] whereby Cross-GAN achieves rank-1=83.23 versus LSRO [Zheng *et al.*, 2017b] achieves rank-1=84.62. The main reason is that LSRO [Zheng *et al.*, 2017b] is a semi-supervised approach which uses GANs to generate complex realistic images to augment the number of training data, and a uniform labeling on the generated samples and semantic labeling on existing training samples are performed respectively. However, the proposed method doesn't require any labeling in training.

In the Market-1501 dataset, the matching rate of Cross-GAN is only secondary to LSRO [Zheng *et al.*, 2017b]. The primary reason is that on Market-1501, many persons exhibit similar visual appearance and it is more difficult to distinguish people without any supervision aid. In this aspect, LSRO [Zheng *et al.*, 2017b] generates more realistic images regarding each person to enable discriminative feature learning. However, generating sophisticated images in large numbers is very computationally expensive, which is not feasible in practice. In contrast, the proposed Cross-GAN can still achieve very comparable performance to LSRO [Zheng *et al.*, 2017b] without any labeling.

6 Conclusions and Future Work

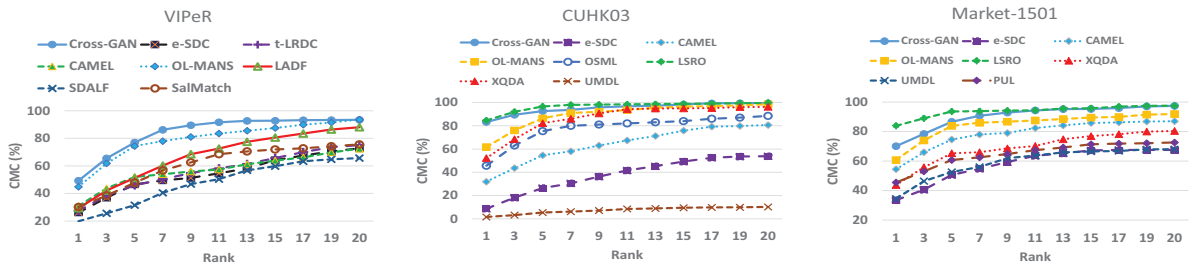
This paper presents a unsupervised generative model to learn jointly invariant features for person re-id without relying on labeled image pairs in correspondence. The proposed method is built atop variational auto-encoders, a cross-view alignment, and dual GANs to seek a series of non-linear transformations into a shared latent space which allows comparable matching across camera views. The learned joint feature distribution effectively captures the co-occurrence patterns in person image against dramatic visual variations. Extensive experiments are conducted to demonstrate the effectiveness of our method in person re-id by setting the state-of-the-art performance.

References

[Ahmed *et al.*, 2015] Ejaz Ahmed, Michael Jones, and Tim K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[Bak and Carr, 2017] Sawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *CVPR*, 2017.

[chen *et al.*, 2016a] Dapeng chen, Zejian Yuan, Badong Chen, and Nanning Zhang. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016.



(a) VIPeR (b) CUHK03 (c) Market-1501

Figure 7: CMC curves of unsupervised/semi-supervised methods on three datasets.

[Chen *et al.*, 2016b] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016.

[Cheng *et al.*, 2016] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zhang. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016.

[Duchi *et al.*, 2010] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2010.

[Fan *et al.*, 2017] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. In *Arxiv*, 2017.

[Farenzena *et al.*, 2010] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[Felzenszwalb *et al.*, 2010] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[Galleguillos *et al.*, 2008] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Gray *et al.*, 2007] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. Int'l. Workshop on Perf. Eval. of Track. and Surv'l.*, 2007.

[Gregor *et al.*, 2015] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *ICML*, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Huang *et al.*, 2015] 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.

[Huang *et al.*, 2016] Siyuan Huang, Jinwen Lu, Jie Zhou, and Anil K. Jain. Nonlinear local metric learning for person re-identification. In *CVPR*, 2016.

[Kingma and Ba, 2015] Diederik Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *ICLR*, 2015.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Kostinger *et al.*, 2012] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Ladicky *et al.*, 2010] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and Philip H.S. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[Li and Wang, 2013] Wei Li and Xiaogang Wang. Locally alligned feature transforms across views. In *CVPR*, 2013.

[Li *et al.*, 2013] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J.R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Xiaoou Tang, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[Liao and Li, 2015] Shengcai Liao and Stan Z. Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCR*, 2015.

[Liao *et al.*, 2015] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[Liu and Tuzel, 2016] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NIPS*, 2016.

[Peng *et al.*, 2016] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *CVPR*, 2016.

[Radford *et al.*, 2015] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv:1511.06434*, 2015.

[Rezende *et al.*, 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *arXiv:1606.03498*, 2016.

[Shi *et al.*, 2016] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016.

[Varior *et al.*, 2016a] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016.

[Varior *et al.*, 2016b] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016.

[Wang and Wu, 2017] Yang Wang and Lin Wu. Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. In *arXiv:1708.02288*, 2017.

[Wang *et al.*, 2013a] Yang Wang, Xiaodi Huang, and Lin Wu. Clustering via geometric median shift over riemannian manifolds. *Information Sciences*, 220:292–305, 2013.

[Wang *et al.*, 2013b] Yang Wang, Xuemin Lin, and Qing Zhang. Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In *ACM CIKM*, 2013.

[Wang *et al.*, 2014a] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, and Qing Zhang. Exploiting correlation consensus: Towards subspace clustering for multi-modal data. In *ACM Multimedia*, 2014.

[Wang *et al.*, 2014b] Yang Wang, Xuemin Lin, Qing Zhang, and Lin Wu. Shifting hypergraphs by probabilistic voting. In *PAKDD*, pages 234–246, 2014.

[Wang *et al.*, 2015a] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. Effective multi-query expansions: Robust landmark retrieval. In *ACM Multimedia*, 2015.

[Wang *et al.*, 2015b] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, and Qing Zhang. Lbmch: Learning bridging mapping for cross-modal hashing. In *ACM SIGIR*, 2015.

[Wang *et al.*, 2015c] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, Qing Zhang, and Xiaodi Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 24(11):3939–3949, 2015.

[Wang *et al.*, 2016a] Faqiang Wang, Wangmeng Zuo andLiang Lin, David Zhang, and Lei Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016.

[Wang *et al.*, 2016b] Yang Wang, Xuemin Lin, Lin Wu, Qing Zhang, and Wenjie Zhang. Shifting multi-hypergraphs via collaborative probabilistic voting. *Knowledge and Information Systems*, 46:515–536, 2016.

[Wang *et al.*, 2016c] Yang Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, Meng Fang, and Shirui Pan. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In *International Joint Conference on Artificial Intelligence*, 2016.

[Wang *et al.*, 2017a] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Transactions on Image Processing*, 26(3):1393–1404, 2017.

[Wang *et al.*, 2017b] Yang Wang, Lin Wu, Xuemin Lin, and Junbin Gao. Multi-view spectral clustering via structured low-rank matrix factorization. In *arXiv:1709.01212*, 2017.

[Wang *et al.*, 2017c] Yang Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, and Xiang Zhao. Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. *IEEE Transactions on Neural Networks and Learning Systems*, 28(1):57–70, 2017.

[Wu and Wang, 2017] Lin Wu and Yang Wang. Robust hashing for multi-view data: Jointly learning low-rank kernelized similarity consensus and hash functions. *Image and Vision Computing*, 57:58–66, 2017.

[Wu *et al.*, 2013a] Lin Wu, Yang Wang, and John Shepherd. Efficient image and tag co-ranking: a bregman divergence optimization method. In *ACM Multimedia*, 2013.

[Wu *et al.*, 2013b] Lin Wu, Yang Wang, John Shepherd, and Xiang Zhao. Max-sum diversification on image ranking with non-uniform matroid constraints. *Neurocomputing*, 118:10–20, 2013.

[Wu *et al.*, 2016] Lin Wu, Chunhua Shen, and Anton van den Hengel. Personnet: Person re-identification with deep convolutional neural networks. In *CoRR abs/1601.07255*, 2016.

[Wu *et al.*, 2017a] Lin Wu, Chunhua Shen, and Anton van den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017.

[Wu *et al.*, 2017b] Lin Wu, Yang Wang, Zongyuan Ge, Qichang Hu, and Xue Li. Structured deep hashing

with convolutional neural networks for fast person re-identification. *Computer Vision and Image Understanding*, 2017.

[Wu *et al.*, 2017c] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. *Pattern Recognition*, 2017.

[Wu *et al.*, 2017d] Lin Wu, Yang Wang, and Shirui Pan. Exploiting attribute correlations: A novel trace lasso based weakly supervised dictionary learning method. *IEEE Transactions on Cybernetics*, 47(12):4479–4508, 2017.

[Wu *et al.*, 2018] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognition*, 73:275–288, 2018.

[Xiao *et al.*, 2016] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representation with domain guided dropout for person re-identification. In *CVPR*, 2016.

[Xiong *et al.*, 2014] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014.

[Yi *et al.*, 2014] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *ICPR*, 2014.

[Yu *et al.*, 2017] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

[Zhang *et al.*, 2014] Ziming Zhang, Yuting Chen, and Venkatesh Saligrama. A novel visual word co-occurrence model for person re-identification. In *ECCV workshop on visual surveillance and re-identification*, 2014.

[Zhang *et al.*, 2016] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016.

[Zhao *et al.*, 2013a] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *ICCV*, 2013.

[Zhao *et al.*, 2013b] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013.

[Zhao *et al.*, 2014] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identfiation. In *CVPR*, 2014.

[Zhao *et al.*, 2017a] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[Zhao *et al.*, 2017b] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017.

[Zheng *et al.*, 2011] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[Zheng *et al.*, 2016] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Towards open-world person re-identification by one-shot group-based verification. *TPAMI*, 38(3):591–606, March 2016.

[Zheng *et al.*, 2017a] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. In *arXiv:1701.07732*, 2017.

[Zheng *et al.*, 2017b] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zhou *et al.*, 2017] Jiahuan Zhou, Pei Yu, Wei Tang, and Ying Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *ICCV*, 2017.