Cross-Dataset Person Re-Identification via Unsupervised Pose Disentanglement and Adaptation

Yu-Jhe Li[1,2,3], Ci-Siang Lin[1,2], Yan-Bo Lin[1], Yu-Chiang Frank Wang[1,2,3]
[1] National Taiwan University, Taiwan
[2] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
[3] ASUS Intelligent Cloud Services, Taiwan
{d08942008, d08942011, r06942048, ycwang}@ntu.edu.tw

Abstract

Person re-identification (re-ID) aims at recognizing the same person from images taken across different cameras. On the other hand, cross-dataset/domain re-ID focuses on leveraging labeled image data from source to target domains, while target-domain training data are without label information. In order to introduce discriminative ability and to generalize the re-ID model to the unsupervised target domain, our proposed Pose Disentanglement and Adaptation Network (PDA-Net) learns deep image representation with pose and domain information properly disentangled. Our model allows pose-guided image recovery and translation by observing images from either domain, without predefined pose category nor identity supervision. Our qualitative and quantitative results on two benchmark datasets confirm the effectiveness of our approach and its superiority over state-of-the-art cross-dataset re-ID approaches.

1. Introduction

Given a query image containing a person (e.g., pedestrian, suspect, etc.), person re-identification (re-ID) [59] aims at matching images with the same identity across non-overlapping camera views. Person re-ID has been among active research topics in computer vision due to its practical applications to smart cities and large-scale surveillance systems. In order to tackle the challenges like visual appearance changes or occlusion in practical re-ID scenarios, several works have been proposed [4, 23, 36, 45, 46, 62]. However, such approaches require a large amount of labeled data for training, and this might not be applicable for real-work applications.

Since it might be computationally expensive to collect identity labels for the dataset of interest, one popular solution is to utilize an additional yet distinct source-domain dataset. This dataset contains fully labeled images (but with different identities) captured by a different set of cameras. Thus, the goal of cross-domain/dataset person re-ID is to extract and adapt useful information from source to the target-domain data of interest, so that re-ID at the target-domain can be addressed accordingly. Since no label is observed for the target-domain data during training, one typically views the aforementioned setting as a unsupervised learning task.

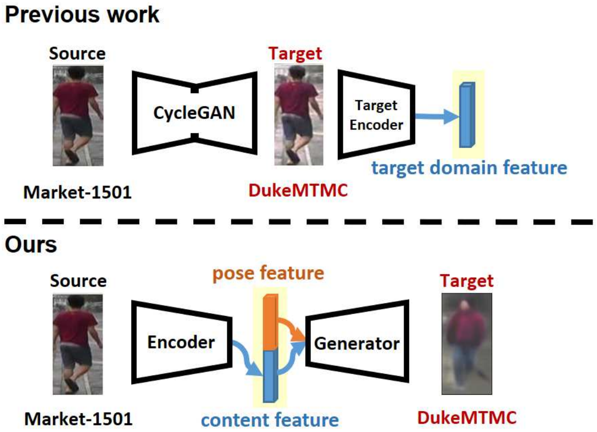

Figure 1: Existing cross-dataset re-ID methods like [12] perform style transfer followed by feature extraction for re-ID, which might limit image variants to be observed. We choose to perform pose disentanglement and adaption with domain-invariant features jointly learned, alleviating the above issue with improved image representation.

Several methods for cross-dataset re-ID have been proposed [13, 15, 42, 49, 54, 58, 61]. For example, Deng *et al.* [13] employ CycleGAN to covert labeled images from source to target domains, followed by performing re-ID at the target domain. Similarly, Zhong *et al.* [61] utilize StarGAN [11] to learn camera invariance and domain connectedness simultaneously. On the other hand, Lin *et al.* [35] employ Maximum Mean Discrepancy (MMD) for learning mid-level feature alignment across data domains for cross-dataset re-ID. However, as shown in Fig. 1, existing cross-domain re-ID approaches generally adapt style information across datasets, and thus pose information cannot be easily be described or preserved in such challenging scenarios.

To overcome the above limitations, we propose a novel deep learning framework for cross-dataset person re-ID. Without observing any ground truth label and pose information in the target domain, our proposed *Pose Disentanglement and Adaptation Network (PDA-Net)* learns domain-invariant features with the ability to disentangle pose information. This allows one to extract, adapt, and manipulate images across datasets without supervision in identity or label. More importantly, this allows us to learn domain and pose-invariant image representation using our proposed network (as depicted in Fig. 1). With label information observed from the source-domain images for enforcing the re-ID performance, our PDA-Net can be successfully applied to cross-dataset re-ID. Compare to prior unsupervised cross-dataset re-ID approaches which lack the ability to describe pose and content features, our experiments confirm that our model is able to achieve improved performances and thus is practically preferable.

We now highlight the contributions of our work below:

- To the best of our knowledge, we are among the first to perform pose-guided yet dataset-invariant deep learning models for cross-domain person re-ID.
- Without observing label information in the target domain, our proposed PDA-Net learns deep image representation with pose and domain information properly disentangled.
- The above disentanglement abilities are realized by adapting and recovering source and target-domain images in a unified framework, simply based on pose information observed from either domain image data.
- Experimental results on two challenging unsupervised cross-dataset re-ID tasks quantitatively and qualitatively confirm that our method performs favorably against state-of-the-art re-ID approaches.

2. Related Works

Supervised Person Re-ID. Person re-ID has been widely studied in the literature. Existing methods typically focus on tackling the challenges of matching images with viewpoint and pose variations, or those with background clutter or occlusion presented [2, 4, 7, 10, 27, 30, 31, 36, 37, 45, 46, 47, 50, 51]. For example, Liu *et al.* [37] develop a pose-transferable deep learning framework based on GAN [19] to handle image pose variants. Chen *et al.* [4] integrate conditional random fields (CRF) and deep neural networks with multi-scale similarity metrics. Several attention-based methods [5, 6, 9, 30, 34, 46, 47] are further proposed to focus on learning the discriminative image features to mitigate the effect of background clutter. While promising results have been observed, the above approaches cannot easily be applied for cross-dataset re-ID due to the lack of ability in suppressing the visual differences across datasets.

Cross-dataset Person Re-ID. To handle cross-dataset person re-ID, a range of hand-crafted features have been considered, so that re-ID at the target domain can be performed in an unsupervised manner [16, 20, 33, 38, 40, 58]. To better exploit and adapt visual information across data domains, methods based on domain adaptation [8, 24] have been utilized [12, 14, 29, 35, 49, 61]. However, since the identities, viewpoints, body poses and background clutter can be very different across datasets, plus no label supervision is available at the target domain, the performance gains might be limited. For example, Fan *et al.* [14] propose a progressive unsupervised learning method iterating between K-means clustering and CNN fine-tuning. Li *et al.* [29] consider spatial and temporal information to learn tracklet association for re-ID. Wang *et al.* [49] learn a discriminative feature representation space with auxiliary attribute annotations. Deng *et al.* [12] translate images from source domain to target domain based on CycleGAN [63] to generate labeled data across image domains. Zhong *et al.* [61] utilize StarGAN [11] to learn camera invariance features. And, Lin *et al.* [35] introduce the Maximum Mean Discrepancy (MMD) distance to minimize the distribution variations of two domains.

Pose-Guided Re-ID. While impressive performances are presented in existing cross-dataset re-ID works, they typically require prior knowledge like the pose of interest, or do not exhibit the ability in describing such information in the resulting features. Recently, a number of models are proposed to better represent pose features during re-ID [28, 48, 52, 53, 55, 56, 57]. Ma *et al.* [39] generate person images by disentangling the input into foreground, background and pose with a complex multi-branch model which is not end-to-end trainable. While Qian *et al.* [43] are able to generate pose-normalized images for person re-ID, only eight pre-defined poses can be manipulated. Although Ge *et al.* [18] learn pose-invariant features with guided image information, their model cannot be applied for cross-dataset re-ID, and thus cannot be applied if the dataset of interest is without any label information. Based on the above observations, we choose to learn dataset and pose-invariant features using a novel and unified model. By disentangling the above representation, re-ID of cross-dataset images can be successfully performed even if no label information is available for target-domain training data.

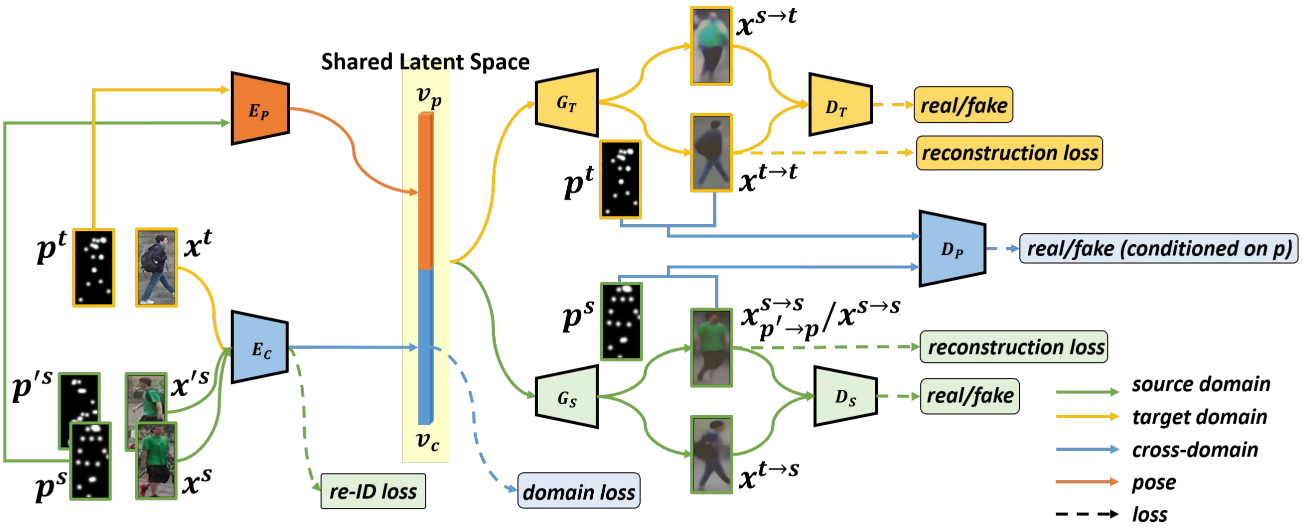

Figure 2: The overview of our Pose Disentanglement and Adaptation Network (PDA-Net). The content encoder E_C learns domain-invariant features $\boldsymbol{v}_c$ for input images from either domain. The pose encoder E_P transforms the pose maps (p^s and p^t) into the latent features $\boldsymbol{v}_p$ for pose guidance and disentanglement purposes. The generators G_T and G_S output domain-specific images via single-domain recovery or cross-domain translation ($x^{s\to s}_{p'\to p}$, $x^{s\to s}$, $x^{t\to s}$, $x^{t\to t}$ and $x^{s\to t}$), conditioned on the pose maps (p^s and p^t). The domain discriminators D_S and D_T preserve image perceptual quality, while the pose discriminator D_P is employed for pose disentanglement guarantees.

3. Proposed Method

3.1. Notations and Problem Formulation

For the sake of completeness, we first define the notations to be used in this paper. Assume that we have the access to a set of N_S images $X_S = \{x_i^s\}_{i=1}^{N_S}$ with the associated label set $Y_S = \{y_i^s\}_{i=1}^{N_S}$, where $x_i^s \in \mathbb{R}^{H\times W\times 3}$ and $y_i^s \in \mathbb{R}$ represent the i^{th} image in the source-domain dataset and its corresponding identity label, respectively. Another set of N_T target-domain dataset images $X_T = \{x_j^t\}_{j=1}^{N_T}$ without any label information are also available during training, where $x_j^t \in \mathbb{R}^{H\times W\times 3}$ represent the j^{th} image in the target-domain dataset. To extract the pose information from source and target-domain data, we apply the pose estimation model [1] on the above images to generate source/target-domain pose outputs $P_S = \{p_i^s\}_{i=1}^{N_S}$ and $P_T = \{p_j^t\}_{j=1}^{N_T}$, respectively. Note that $p_i^s \in \mathbb{R}^{H\times W\times N_L}$ and $p_j^t \in \mathbb{R}^{H\times W\times N_L}$ represent the i^{th} and j^{th} pose maps in the corresponding domains, respectively. Following [1], we set the number of pose landmarks $N_L = 18$ in our work.

To achieve cross-dataset person re-ID, we present an end-to-end trainable network, *Pose Disentanglement and Adaptation Network (PDA-Net)*. As illustrated in Figure 2, our PDA-Net aims at learning domain-invariant deep representation $\boldsymbol{v}_c \in \mathbb{R}^d$ (d denotes the dimension of the feature), while pose information is jointly disentangled from this feature space. To achieve this goal, a pair of encoders E_C and E_P for encoding the input images and pose maps into $\boldsymbol{v}_c$ and $\boldsymbol{v}_p \in \mathbb{R}^h$ (h denotes the dimension of the feature), respectively. Guided by the encoded pose features (from either domain), our domain specific generators (G_S and G_T for source and target-domain datasets, respectively) would recover/synthesize the desirable outputs in the associated data domain. We will detail the properties of each component in the following subsections.

To perform person re-ID of the target-domain dataset in the testing phase, our network encodes the query image by E_C for deriving the domain and pose-invariant representation $\boldsymbol{v}_c$, which is applied for matching the gallery ones via nearest neighbor search (in Euclidean distances).

3.2. Pose Disentanglement and Adaptation Network (PDA-Net)

As depicted in Figure 2, our proposed Pose Disentanglement and Adaptation Network consists of a number of network components. The content encoder E_C encodes input images across different domains/datasets and produces content feature $\boldsymbol{v}_c$ for person re-ID. The pose encoder E_P encodes the pose maps and produce pose feature $\boldsymbol{v}_p$ for pose disentanglement. The two domain-specific generators, G_S and G_T, output images in source and target domains respectively (by feeding both $\boldsymbol{v}_c$ and $\boldsymbol{v}_p$). The two domain specific discriminators, D_S and D_T, are designed to enforce the two domain-specific generators G_S and G_T produce perceptually realistic and domain-specific images. Finally, the pose discriminator D_P aims at enforcing the generators to output realistic images conditioned on the given pose.

3.2.1 Domain-invariant representation for re-ID

We encourage the content encoder E_C to generate similar feature distributions when observing both X_S and X_T. To accomplish this, we apply the Maximum Mean Discrepancy (MMD) measure [22] to calculate the difference between the associated feature distributions for the content feature $\boldsymbol{v}_c$ between the source and target domains. Given an source image $x^s \in X_S$ and an target image $x^t \in X_T$ [1], we first forward x^s and x^t to the content encoder E_C to obtain their content feature $\boldsymbol{v}_c^s$ and $\boldsymbol{v}_c^t$. Then we can formulate our MMD loss $\mathcal{L}_{\text{MMD}}$ as:

$$\mathcal{L}_{\text{MMD}} = \|\frac{1}{n_s}\sum_{g=1}^{n_s}\phi(\boldsymbol{v}^s_{c,g}) - \frac{1}{n_t}\sum_{l=1}^{n_t}\phi(\boldsymbol{v}^t_{c,l})\|^2_{\mathcal{H}}, \tag{1}$$

where ϕ is a map operation which project the distribution into a reproducing kernel Hilbert space $\mathcal{H}$ [21]. n_s and n_t are the batch sizes of the images in the associated domains. The arbitrary distribution of the features can be represented by using the kernel embedding technique. It has been proven that if the kernel is characteristic, then the mapping to the space $\mathcal{H}$ is injective while the injectivity indicates that the arbitrary probability distribution is uniquely represented by and element in the space $\mathcal{H}$.

It is also worth noting that, we do *not* consider the adversarial learning strategy for deriving domain-invariant features (e.g., [17]) in our work. This is because that this technique might produce pose-invariant features instead of domain-invariant ones for re-ID datasets, and thus the resulting features cannot perform well in cross-dataset re-ID.

Next, to utilize label information observed from source-domain training data, we impose a triplet loss $\mathcal{L}_{tri}$ on the derived feature vector $\boldsymbol{v}_c$. This would maximize the inter-class discrepancy while minimizing intra-class distinctness. To be more specific, for each input source image x^s, we sample a positive image x^s_{pos} with the same identity label and a negative image x^s_{neg} with different identity labels to form a triplet tuple. Then, the distance between x^s and x^s_{pos} (or x^s_{neg}) can be calculated as:

$$d_{\text{pos}} = \|\boldsymbol{v}^s_c - \boldsymbol{v}^s_{c,\text{pos}}\|_2, \tag{2}$$

$$d_{\text{neg}} = \|\boldsymbol{v}^s_c - \boldsymbol{v}^s_{c,\text{neg}}\|_2, \tag{3}$$

where $\boldsymbol{v}^s_c$, $\boldsymbol{v}^s_{c,\text{pos}}$, and $\boldsymbol{v}^s_{c,\text{neg}}$ represent the feature vectors of images x^s, x^s_{pos}, and x^s_{neg}, respectively.

With the above definitions, the triplet loss $\mathcal{L}_{tri}$ is

$$\mathcal{L}_{tri} = \mathbb{E}_{(x^s,y^s)\sim(X_S,Y_S)}\max(0, m + d_{\text{pos}} - d_{\text{neg}}), \tag{4}$$

where $m > 0$ is the margin enforcing the separation between positive and negative image pairs.

3.2.2 Pose-guided cross-domain image translation

To ensure our derived content feature is domain-invariant in cross-domain re-ID tasks, we need to perform additional image translation during the learning of our PDA-Net. That is, we have the pose encoder E_P in Fig. 2 encodes the inputs from source pose set inputs P_S and the target pose set P_T into pose features $\boldsymbol{v}^s_p$ and $\boldsymbol{v}^t_p$. As a result, both content and pose features would be produced in the latent space.

We enforce the two generators G_S and G_T for generating the person images conditioned on the encoded pose feature. For the source domain, we have the source generator G_S take the concatenated source-domain content and pose feature pair $(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c)$ and output the corresponding image $x^{s\to s}$. Similarly, we have G_T take $(\boldsymbol{v}^t_p, \boldsymbol{v}^t_c)$ for producing $x^{t\to t}$. Note that $x^{s\to s} = G_S((\boldsymbol{v}^s_p, \boldsymbol{v}^s_c))$, $x^{t\to t} = G_T(\boldsymbol{v}^t_p, \boldsymbol{v}^t_c)$ denote the reconstructed images in source and target domains, respectively. Since this can be viewed as image recovery in each domain, reconstruction loss can be applied as the objective during learning.

Since we have ground truth labels (i.e., image pair correspondences) for the source-domain data, we can further perform a unique image recovery task for the source-domain images. To be more precise, given two source-domain images x^s and x'^s of the same person but with different poses p^s and p'^s, we expect that they share the same content feature $\boldsymbol{v}^s_c$ but with pose features as $\boldsymbol{v}^s_p$ and $\boldsymbol{v}^s_{p'}$. Given the desirable pose $\boldsymbol{v}^s_p$, we then enforce G_S to output the source domain image x^s using the content feature $\boldsymbol{v}^s_c$ which is originally associated with $\boldsymbol{v}^s_{p'}$. This is referred to as *pose-guided* image recovery.

With the above discussion, image reconstruction loss for the source-domain data $\mathcal{L}^S_{\text{rec}}$ can be calculated as:

$$\begin{aligned}\mathcal{L}^S_{\text{rec}} &= \mathbb{E}_{x^s\sim X_S, p^s\sim P_S}[\|x^{s\to s} - x^s\|_1] \\ &+ \mathbb{E}_{\{x^s,x'^s\}\sim X_S, p^s\sim P_S}[\|x^{s\to s}_{p'\to p} - x^s\|_1],\end{aligned} \tag{5}$$

where $x^{s\to s}_{p'\to p} = G_S(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c|\boldsymbol{v}^s_{p'})$ denotes the generated image from the input x'^s and v^s_c describe the content feature of the same identity (i.e., x'^s, and x^s of the same person by with different poses p' and p).
As for the target-domain reconstruction loss, we have

$$\mathcal{L}^T_{\text{rec}} = \mathbb{E}_{x^t\sim X_T, p^t\sim P_T}[\|x^{t\to t} - x^t\|_1]. \tag{6}$$

Note that we adopt the L1 norm in the above reconstruction loss terms as it preserves image sharpness [25].

In addition to image recovery in either domain, our model also perform pose-guided image translation. That is, our decoders G_S and G_T allow input feature pairs whose content and pose representation are extracted from different domains. Thus, we would observe $x^{t\to s} = G_S(\boldsymbol{v}^t_p, \boldsymbol{v}^t_c)$ and $x^{s\to t} = G_T(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c)$ as the outputs, with the goal of having these translated images as realistic as possible.

[1] For simplicity, we would omit the subscript i and j, denote source and target images as x^s and x^t, and represent the corresponding labels for source images as y^s in this paper.

To ensure G_S and G_T produce perceptually realistic outputs in the associated domains, we have the image discriminator D_S discriminate between the real source-domain images x^s and the synthesized/translated ones (i.e., $x^{s\to s}$, $x^{t\to s}$). Thus, the source-domain discriminator loss $\mathcal{L}^S_{domain}$ as

$$\begin{aligned}\mathcal{L}^S_{domain} &= \mathbb{E}_{x^s\sim X_S}[\log(\mathcal{D}_S(x^s))] \\ &+ \mathbb{E}_{x^s\sim X_S, p^s\sim P_S}[\log(1-\mathcal{D}_S(x^{s\to s}))] \\ &+ \mathbb{E}_{x^t\sim X_T, p^t\sim P_T}[\log(1-\mathcal{D}_S(x^{t\to s}))].\end{aligned} \tag{7}$$

Similarly, the target domain discriminator loss $\mathcal{L}^T_{domain}$ is defined as

$$\begin{aligned}\mathcal{L}^T_{domain} &= \mathbb{E}_{x^t\sim X_T}[\log(\mathcal{D}_T(x^t))] \\ &+ \mathbb{E}_{x^t\sim X_T, p^t\sim P_T}[\log(1-\mathcal{D}_T(x^{t\to t}))] \\ &+ \mathbb{E}_{x^s\sim X_S, p^s\sim P_S}[\log(1-\mathcal{D}_T(x^{s\to t}))].\end{aligned} \tag{8}$$

3.2.3 Unsupervised pose disentanglement across data domains

With the above pose-guided image translation mechanism, we have our PDA-Net learn domain-invariant content features across data domains. However, to further ensure the pose encoder describes and disentangles the pose information observed from the input images, we need additional network modules for completing this goal.

To achieve this object, we introduce a pose discriminator D_P in Fig. 2, which focuses on distinguishing between real and recovered images, conditioned on the given pose inputs. Following previous FD-GAN [18], we adopt the PatchGAN [26] structure as our D_P. That is, the input to D_P is concatenation of the real/recovered image and the given pose map, which is processed by Gaussian-like heat-map transformation. Then, D_P produces a image-pose matching confidence map, each location of this output confidence map represents the matching degree between the input image and the associated pose map.

It can be seen that, the two generators G_S and G_T in PDA-Net tend to fool the pose discriminator D_P to obtain high matching confidences for the generated images. Intuitively, since only source-domain data are with ground truth labels, our D_P is designed to authenticate the recovered images in each corresponding domain but not the translated ones across domains. In other words, the adversarial loss of D_P is formulated as:

$$\mathcal{L}_{pose} = \mathcal{L}^S_{pose} + \mathcal{L}^T_{pose}, \tag{9}$$

where

$$\begin{aligned}\mathcal{L}^S_{pose} &= \mathbb{E}_{x^s\sim X_S, p^s\sim P_S}[\log(\mathcal{D}_P(p^s, x^s))] \\ &+ \mathbb{E}_{x^s\sim X_S, p^s\sim P_S}[\log(1-\mathcal{D}_P(p^s, x^{s\to s}))] \\ &+ \mathbb{E}_{x^s\sim X_S, p'^s\sim P_S}[\log(1-\mathcal{D}_P(p'^s, x^s))] \\ &+ \mathbb{E}_{\{x^s, x'^s\}\sim X_S, p^s\sim P_S}[\log(1-\mathcal{D}_P(p^s, x^{s\to s}_{p'\to p}))]\end{aligned} \tag{10}$$

Algorithm 1: Learning of PDA-Net

```
Data: Source domain: $X_S$, $P_S$, and $Y_S$; Target domain: $X_T$
      and $P_T$
Result: Configurations of PDA-Net
$\theta_{E_C}, \theta_{E_P}, \theta_{G_S}, \theta_{G_T}, \theta_{D_S}, \theta_{D_T}, \theta_{D_P} \leftarrow$ initialize
for Num. of training Iters. do
    $x^s, p^s, y^s, x^t, p^t, x'^s, p'^s \leftarrow$ sample from $X_S, P_S, Y_S,$
      $X_T, P_T$
    $v^s_c, v^t_c \leftarrow$ obtain by $E_C(x^s/x'^s), E_C(x^t)$
    $v^s_p, v^t_p \leftarrow$ obtain by $E_P(p^s), E_P(p^t)$
    $\mathcal{L}_{\mathrm{MMD}}, \mathcal{L}_{tri} \leftarrow$ calculate by (1), (4)
    $\theta_{E_C} \xleftarrow{+} -\nabla_{\theta_{E_C}}(\mathcal{L}_{\mathrm{MMD}} + \lambda_{tri}\mathcal{L}_{tri})$
    $x^{s\to s}, x^{t\to s} \leftarrow$ obtain by $G_S(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c), G_S(\boldsymbol{v}^t_p, \boldsymbol{v}^t_c)$
    $x^{s\to t}, x^{t\to t} \leftarrow$ obtain by $G_T(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c), G_T(\boldsymbol{v}^t_p, \boldsymbol{v}^t_c)$
    $x^{s\to s}_{p'\to p} \leftarrow$ obtain by $G_S(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c | \boldsymbol{v}^s_{p'})$
    $\mathcal{L}^S_{rec}, \mathcal{L}^T_{rec}, \mathcal{L}^S_{domain}, \mathcal{L}^T_{domain}, \mathcal{L}_{pose} \leftarrow$ calculate by
      (5), (6), (7), (8), (9)
    for Iters. of updating generator do
        $\theta_{E_C,E_P,G_S} \xleftarrow{+} -\nabla_{\theta_{E_C,E_P,G_S}}(\lambda_{rec}\mathcal{L}^S_{rec} -$
          $\mathcal{L}^S_{domain} - \lambda_{pose}\mathcal{L}_{pose})$
        $\theta_{E_C,E_P,G_T} \xleftarrow{+} -\nabla_{\theta_{E_C,E_P,G_T}}(\lambda_{rec}\mathcal{L}^T_{rec} -$
          $\mathcal{L}^T_{domain} - \lambda_{pose}\mathcal{L}_{pose})$
    for Iters. of updating discriminator do
        $\theta_{D_S} \xleftarrow{+} -\nabla_{\theta_{D_S}}\mathcal{L}^S_{domain}$
        $\theta_{D_T} \xleftarrow{+} -\nabla_{\theta_{D_T}}\mathcal{L}^T_{domain}$
        $\theta_{D_P} \xleftarrow{+} -\nabla_{\theta_{D_P}}\mathcal{L}_{pose}$
```

and

$$\begin{aligned}\mathcal{L}^T_{pose} &= \mathbb{E}_{x^t\sim X_T, p^t\sim P_T}[\log(\mathcal{D}_P(p^t, x^t))] \\ &+ \mathbb{E}_{x^t\sim X_T, p^t\sim P_T}[\log(1-\mathcal{D}_P(p^t, x^{t\to t}))].\end{aligned} \tag{11}$$

Note that $x^{s\to s}_{p'\to p} = G_S(\boldsymbol{v}^s_p, \boldsymbol{v}^s_c | \boldsymbol{v}^s_{p'})$ represents the synthesized image from the input x'^s (with the same content feature v^s_c with x^s but with a different pose feature v'^s_p).

From (9), we see that while our pose disentanglement loss enforces the matching between the output image and its conditioned pose in each domain, additional guidance is available in the source domain to update our D_P. That is, as shown in (7), we are able to verify the authenticity of the source-domain output image which is given by the input image of the same person but with a different pose (i.e., p' instead of p). While our decoder is able to output such a image with its ground truth source-domain image observed (as noted in (5), the introduced D_P would further improve our capability of pose disentanglement and pose-guided image recovery.

It is worth repeating that the goal of PDA-Net is to perform cross-dataset re-ID without observing label information in the target domain. By introducing the aforementioned network module, our PDA-Net would be capable

Table 1: Performance comparisons on Market-1501 with cross-dataset/unsupervised Re-ID methods. The number in bold indicates the best result.

Method	Source: DukeMTMC, Target: Market			
	Rank-1	Rank-5	Rank-10	mAP
BOW [58]	35.8	52.4	60.3	14.8
UMDL [42]	34.5	52.6	59.6	12.4
PTGAN [51]	38.6	-	66.1	-
PUL [15]	45.5	60.7	66.7	20.5
CAMEL [54]	54.5	-	-	26.3
SPGAN [13]	57.7	75.8	82.4	26.7
TJ-AIDL [49]	58.2	74.8	81.1	26.5
MMFA [35]	56.7	75.0	81.8	27.4
HHL [61]	62.2	78.8	84.0	31.4
CFSM [3]	61.2	-	-	28.3
ARN [32]	70.3	80.4	86.3	39.4
TAUDL [29]	63.7	-	-	41.2
PDA-Net (Ours)	**75.2**	**86.3**	**90.2**	**47.6**

of performing cross-dataset re-ID via pose-guided cross-domain image translation. More precisely, with the joint training of cross-domain encoders/decoders and the pose disentanglement discriminators, our model allows learning of domain-invariant and pose-disentangled feature representation. The pseudo code for training our PDA-Net is summarized in Algorithm 1.

4. Experiments

4.1. Datasets and Experimental Settings

To evaluate our proposed method, we conduct experiments on Market-1501 [58] and DukeMTMC-reID [44, 60], both are commonly considerd in recent re-ID tasks.

Market-1501. The Market-1501 [58] is composed of 32,668 labeled images of 1,501 identities collected from 6 camera views. The dataset is split into two non-overlapping fixed parts: 12,936 images from 751 identities for training and 19,732 images from 750 identities for testing. In testing, 3368 query images from 750 identities are used to retrieve the matching persons in the gallery.

DukeMTMC-reID. The DukeMTMC-reID [44, 60] is also a large-scale Re-ID dataset. It is collected from 8 cameras and contains 36,411 labeled images belonging to 1,404 identities. It also consists of 16,522 training images from 702 identities, 2,228 query images from the other 702 identities, and 17,661 gallery images.

Evaluation Protocol. We employ the standard metrics as in most person Re-ID literature, namely the cumulative matching curve (CMC) used for generating ranking accuracy, and the mean Average Precision (mAP). We report rank-1 accuracy and mean average precision (mAP) for evaluation on both datasets.

Table 2: Performance comparisons on DukeMTMC-reID with cross-dataset/unsupervised Re-ID methods. The number in bold indicates the best result.

Method	Source: Market, Target: DukeMTMC			
	Rank-1	Rank-5	Rank-10	mAP
BOW [58]	17.1	28.8	34.9	8.3
UMDL [42]	18.5	31.4	37.6	7.3
PTGAN [51]	27.4	-	50.7	-
PUL [15]	30.0	43.4	48.5	16.4
SPGAN [13]	46.4	62.3	68.0	26.2
TJ-AIDL [49]	44.3	59.6	65.0	23.0
MMFA [35]	45.3	59.8	66.3	24.7
HHL [61]	46.9	61.0	66.7	27.2
CFSM [3]	49.8	-	-	27.3
ARN [32]	60.2	73.9	79.5	33.4
TAUDL [29]	61.7	-	-	43.5
PDA-Net (Ours)	**63.2**	**77.0**	**82.5**	**45.1**

4.2. Implementation Details

Configuration of PDA-Net. We implement our model using PyTorch. Following Section 3, we use ResNet-50 pre-trained on ImageNet as our backbone of cross-domain encoder E_C. Given an input image x (all images are resized to size $256 \times 128 \times 3$, denoting width, height, and channel respectively.), E_C encodes the input into 2048-dimension content feature $\boldsymbol{v}_c$. As mentioned in the Section. 3.1, the pose-map is represented by an 18-channel map, where each channel represents the location of one pose landmark. Such landmark location is converted to a Gaussian heat map. The pose encoder E_P then employs 4 convolution blocks to produce the 256-dimension pose feature vector $\boldsymbol{v}_p$ from these pose-maps. The structure of the both the domain generators (G_S, G_T) are 6 convolution-residual blocks similar to that proposed by Miyato *et al.* [41]. The structure of the both the domain discriminator (D_S, D_T) employ the ResNet-18 as backbone while the architecture of shared pose disciminator D_P adopts PatchGAN structure following FD-GAN [18] and is composed of 5 convolution blocks in our PDA-Net. Domain generators (G_S, G_T), domain discriminator (D_S, D_T), shared pose discriminator D_P are all randomly initialized. The margin for the $\mathcal{L}_{tri}$ is set as 0.5, and we fix λ_{tri}, λ_{rec}, and λ_{pose} as 1.0, 10.0, 0.1, respectively.

4.3. Quantitative Comparisons

Market-1501. In Table 1, we compare our proposed model with the use of Bag-of-Words (BoW) [58] for matching (i.e., no transfer), four unsupervised re-ID approaches, including UMDL [42], PUL [15], CAMEL [54] and TAUDL [29], and seven cross-dataset re-ID methods, including PTGAN [51], SPGAN [12], TJ-AIDL [49], MMFA [35], HHL [61], CFSM [3] and ARN [32]. From this table, we see that our model achieved very promising

Table 3: Ablation studies of the proposed PDA-Net under two experimental settings. "Share D_P" incidates whether to build separate pose discriminators, i.e. D_P^S and D_P^T, instead of one shared D_P.

Experimental setting	Loss functions and component						Source: DukeMTMC-reID Target: Market-1501		Source: Market-1501 Target: DukeMTMC-reID	
	$\mathcal{L}_{tri}$	$\mathcal{L}_{\text{MMD}}$	$\mathcal{L}_{rec}^{S/T}$	$\mathcal{L}_{domain}^{S/T}$	$\mathcal{L}_{pose}$	Share D_P	Rank-1	mAP	Rank-1	mAP
Baseline (ResNet-50)	✓	✗	✗	✗	✗	✗	44.2	18.1	33.5	16.3
Baseline (ResNet-50 w/ MMD)	✓	✓	✗	✗	✗	✗	50.4	22.6	39.5	23.1
PDA-Net (w/o $\mathcal{L}_{rec}^S$,$\mathcal{L}_{rec}^T$)	✓	✓	✗	✓	✓	✓	52.3	24.7	42.5	24.0
PDA-Net (w/o $\mathcal{L}_{pose}$)	✓	✓	✓	✓	✗	✓	55.1	25.2	45.5	26.1
PDA-Net (w/o share D_P)	✓	✓	✓	✓	✓	✗	59.4	27.8	50.9	29.7
PDA-Net (w/o $\mathcal{L}_{domain}^S$, $\mathcal{L}_{domain}^T$)	✓	✓	✓	✗	✓	✓	65.3	30.7	56.5	31.2
PDA-Net (w/o MMD)	✓	✗	✓	✓	✓	✓	71.2	39.8	60.1	35.8
PDA-Net (Ours)	✓	✓	✓	✓	✓	✓	**75.2**	**47.6**	**63.2**	**45.1**

results in Rank-1, Rank-5, Rank-10, and mAP, and observed performance margins over recent approaches. For example, in the single query setting, we achieved **Rank-1 accuracy=75.2%** and **mAP=52.6%**.

Compared to SPGAN [12] and HHL [61], we note that our model is able to generate cross-domain images conditioned on various poses rather than few camera styles. Compared to MMFA [35], our model further disentangles the pose information and learns a pose invariant cross-domain latent space. Compared to the second best method, *i.e.*, TAUDL [29], our results were higher by **11.5%** in Rank-1 accuracy and by **11.4%** in mAP, while no additional spatial and temporal information is utilized (but TAUDL did).

DukeMTMC-reID. We now consider the DukeMTMC-reID as the target-domain dataset of interest, and list the comparisons in Table 2. From this table, we also see that our model performed favorably against baseline and state-of-art unsupervised/cross-domain re-ID methods. Take the single query setting for example, we achieved **Rank-1 accuracy=63.2%** and **mAP=45.1%**. Compared to the second best method, our results were higher by **1.5%** in Rank-1 accuracy and by **1.6%** in mAP. From the experiments on the above two datasets, the effectiveness of our model for cross-domain re-ID can be successfully verified.

4.4. Ablation Studies and Visualization

Analyzing the network modules in PDA-Net. As shown in Table 3, we start from two baseline methods, i.e., naive Resnet-50 (w/o $\mathcal{L}_{\text{MMD}}$) and advanced Resnet-50 (w/ $\mathcal{L}_{\text{MMD}}$), showing the standard re-ID performances. We then utilize ResNet-50 as the backbone CNN model to derive representations for re-ID with only triplet loss $\mathcal{L}_{tri}$, while the advanced one includes the MMD loss $\mathcal{L}_{\text{MMD}}$. We observe that our full model (the last row) improved the performance by a large margin (roughly $20 \sim 25\%$) at Rank-1 on both two benchmark datasets. The performance gain can be ascribed to the unique design of our model for deriving both domain-invariant and pose-invariant representation.

Loss functions To further analyze the importance of each introduced loss function, we conduct an ablation study from third row to seventh rows shown in Table 3. Firstly, the reconstruction loss $\mathcal{L}_{rec}$ is shown to be vital to our PDA-Net, since we observe 23% and 20% drops on Market-1501 and DukeMTMC-reID, respectively when the loss was excluded. This is caused by no explicit supervision to guide our PDA-Net to generate human-perceivable images, and thus the resulting model would suffer from image-level information loss.

Secondly, without the pose loss $\mathcal{L}_{pose}$ on both domains, our model would not be able to perform pose matching based on each generated image, causing failure on the pose disentanglement process and resulting in the re-ID performance drop (about 20% on both settings). Thirdly, when $\mathcal{L}_{domain}^{S/T}$ is turned off, our model is not able to preserve the domain information, indicating that only pose information would be observed. We credited such a 10% performance drop to the negative effect in learning pose-invariant feature, which resulted in unsatisfactory pose disentanglement. Lastly, the MMD loss $\mathcal{L}_{\text{MMD}}$ is introduced to our PDA-NET to mitigate the domain shift due to dataset differences. Its effectiveness is also confirmed by our studies.

Shared pose discriminator D_P. To demonstrate the effectiveness and necessity of the pose discriminator D_P introduced to our PDA-Net, we first consider replacing D_P by two separate pose discriminators D_P^S and D_P^T, and report the re-ID performance in the fifth row of Table 3. With a clear performance drop observed, we see that the resulting PDA-Net would not be able to transfer the substantiated pose-matching knowledge from source to target domains. In other words, a shared pose discriminator would be preferable since pose guidance can be provided by both domains.

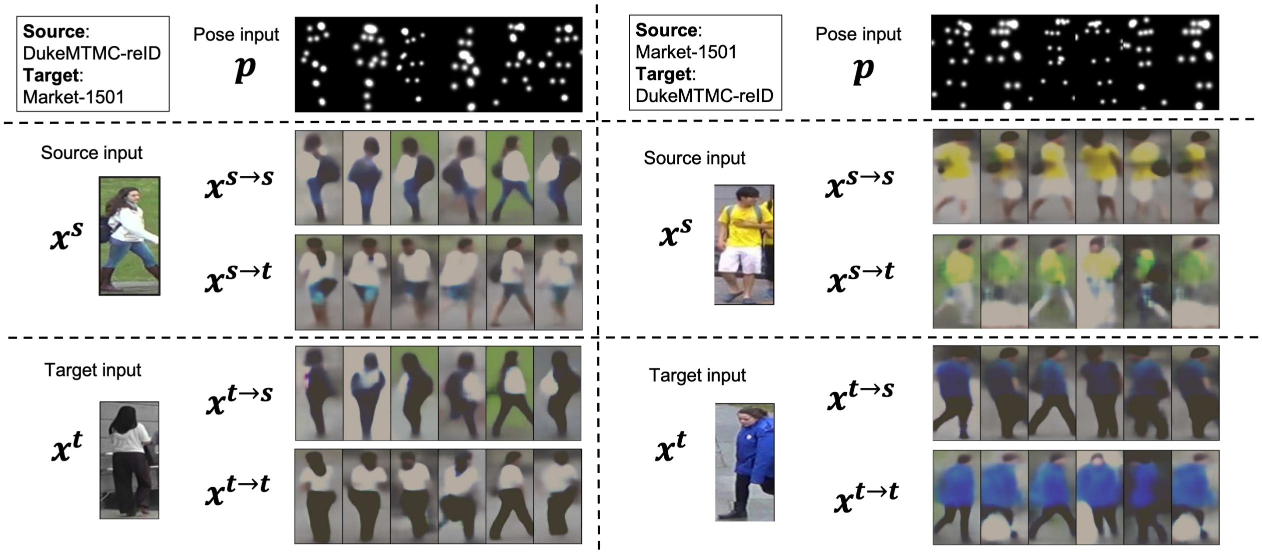

Figure 3: Visualization examples of our PDA-Net for pose-guided image translation across datasets. Given six pose conditions (the first row) and the input image (x^s or x^t), we present the six generated images for each dataset pair: $x^{s\to s}$ (the second row), $x^{t\to s}$ (the third row), $x^{t\to t}$ (the fourth row) and $x^{s\to t}$ (the fifth row).

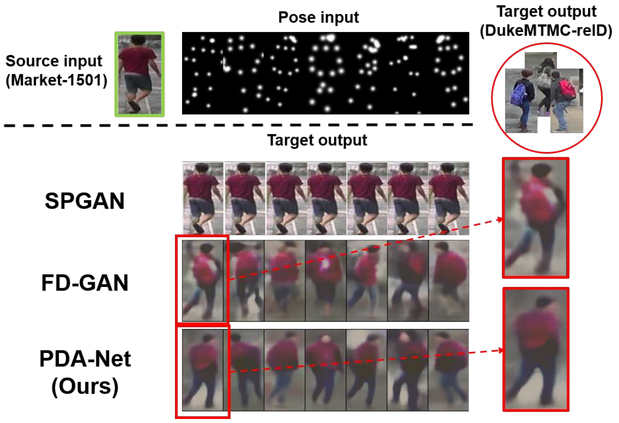

Figure 4: Visualization of cross-dataset or pose-guided re-ID. Note that SPGAN [13] performs style-transfer for converting images across datasets but lacks the ability to exhibit pose variants, while FD-GAN [18] disentangles pose information but cannot take cross-domain data.

Visualization comparisons of cross-dataset and pose-guided re-ID models. In Figure 3, we visualize the generated images: $x^{s\to s}$, $x^{s\to t}$, $x^{t\to s}$, and $x^{t\to t}$ in two cross-domain settings. Given an input from either domain with pose conditions, our model was able to produce satisfactory pose-guided image synthesis within or across data domains.

In Figure 4, we additionally consider the cross-dataset re-ID appoach of SPGAN [13] and the pose-disentanglement re-ID method of FD-GAN [18]. We see that, since SPGAN performed style transfer for synthesizing cross-domain images, pose variants cannot be exploited in the target domain. While FD-GAN was able to generate pose-guided image outputs with supervision on target target domain, their model is not designed to handle cross-domain data so that cannot produce images across datasets with satisfactory quality. From the above qualitative evaluation and comparison, we confirm that our PDA-Net is able to perform pose-guided single-domain image recovery and cross-domain image translation with satisfactory image quality, which would be beneficial to cross-domain re-ID tasks.

5. Conclusions

In this paper, we presented a novel Pose Disentanglement and Adaptation Network (PDA-Net) for cross-dataset re-ID. The main novelty lies in the unique design of our PDA-Net, which jointly learns domain-invariant and pose-disentangled visual representation with re-ID guarantees. By observing only image input (from either domain) and any desirable pose information, our model allows pose-guided singe-domain image recovery and cross-domain image translation. Note that only label information (image correspondence pairs) is available for the source-domain data, any no pre-defined pose category is utilized during training. Experimental results on the two benchmark datasets showed remarkable improvements over existing works, which support the use of our proposed approach for cross-dataset re-ID. Qualitative results also confirmed that our model is capable of performing cross-domain image translation with pose properly disentangled/manipulated.

Acknowledgements. This work is supported by the Ministry of Science and Technology of Taiwan under grant MOST 108-2634-F-002-018.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[3] Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M Hospedales. Disjoint label space transfer learning with common factorised space. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 6

[4] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[5] Yun-Chun Chen and Winston H Hsu. Saliency aware: Weakly supervised object localization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019. 2

[6] Yun-Chun Chen, Po-Hsiang Huang, Li-Yu Yu, Jia-Bin Huang, Ming-Hsuan Yang, and Yen-Yu Lin. Deep semantic matching with foreground detection and cycle-consistency. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018. 2

[7] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representation for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2

[8] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[9] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint learning of semantic matching and object co-segmentation. *arXiv*, 2019. 2

[10] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[12] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6, 7

[13] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 6, 8

[14] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 2

[15] Hehe Fan, Liang Zheng, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. In *arXiv preprint*, 2017. 1, 6

[16] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 2016. 4

[18] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 5, 6, 8

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[20] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008. 2

[21] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems (NIPS)*, 2007. 4

[22] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems (NIPS)*, 2009. 4

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. In *arXiv preprint*, 2017. 1

[24] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 2

[25] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 4

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[27] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[28] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[29] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 6, 7

[30] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[31] Yu-Jhe Li, Yun-Chun Chen, Yen-Yu Lin, Xiaofei Du, and Yu-Chiang Frank Wang. Recover and identify: A generative dual model for cross-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[32] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 6

[33] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[34] Jhih-Yuan Lin, Min-Sheng Wu, Yu-Cheng Chang, Yun-Chun Chen, Chao-Te Chou, Chun-Ting Wu, and Winston H Hsu. Learning volumetric segmentation for lung tumor. *IEEE ICIP VIP Cup Tech. Report*, 2018. 2

[35] Shan Lin, Haoliang Li, Chang-Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 1, 2, 6, 7

[36] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. In *arXiv preprint*, 2017. 1, 2

[37] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[38] Bingpeng Ma, Yu Su, and Frederic Jurie. Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 2014. 2

[39] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[40] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[41] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 6

[42] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 6

[43] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[44] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cuc-

chiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 6

[45] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[46] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[47] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[48] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[49] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6

[50] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[51] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[52] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *Proceedings of the ACM Conference on Multimedia (MM)*, 2017. 2

[53] Hantao Yao, Shiliang Zhang, Richang Hong, Yongdong Zhang, Changsheng Xu, and Qi Tian. Deep representation learning with part loss for person re-identification. *IEEE Transactions on Image Processing (TIP)*, 2019. 2

[54] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 6

[55] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[56] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[57] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. In *arXiv preprint*, 2017. 2

[58] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 6

[59] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. In *arXiv preprint*, 2016. 1

[60] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[61] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6, 7

[62] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2