A Unified Generative Adversarial Framework for Image Generation and Person Re-identification

Yaoyu Li[1,2], Tianzhu Zhang[1,2], Lingyu Duan[3], Changsheng Xu[1,2]
[1]National Lab of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences
[3]Institute of Digital Media, Peking University
{yaoyu.li,tzzhang,csxu}@nlpr.ia.ac.cn,lingyu@pku.edu.cn

ABSTRACT

Person re-identification (re-id) aims to match a certain person across multiple non-overlapping cameras. It is a challenging task because the same person's appearance can be very different across camera views due to the presence of large pose variations. To overcome this issue, in this paper, we propose a novel unified person re-id framework by exploiting person poses and identities jointly for simultaneous person image synthesis under arbitrary poses and pose-invariant person re-identification. The framework is composed of a GAN based network and two Feature Extraction Networks (FEN), and enjoys following merits. First, it is a unified generative adversarial model for person image generation and person re-identification. Second, a pose estimator is utilized into the generator as a supervisor in the training process, which can effectively help pose transfer and guide the image generation with any desired pose. As a result, the proposed model can automatically generate a person image under an arbitrary pose. Third, the identity-sensitive representation is explicitly disentangled from pose variations through the person identity and pose embedding. Fourth, the learned re-id model can have better generalizability on a new person re-id dataset by using the synthesized images as auxiliary samples. Extensive experimental results on four standard benchmarks including Market-1501 [69], DukeMTMC-reID [40], CUHK03 [23], and CUHK01 [22] demonstrate that the proposed model can perform favorably against state-of-the-art methods.

KEYWORDS

Person Re-identification; Multimedia System; GAN

ACM Reference Format:
Yaoyu Li, Tianzhu Zhang, Lingyu Duan, Changsheng Xu. 2018. A Unified Generative Adversarial Framework for Image, Generation and Person Re-identification. In *2018 ACM Multimedia Conference (MM 18), October 22C26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10. 1145/3240508.3240573

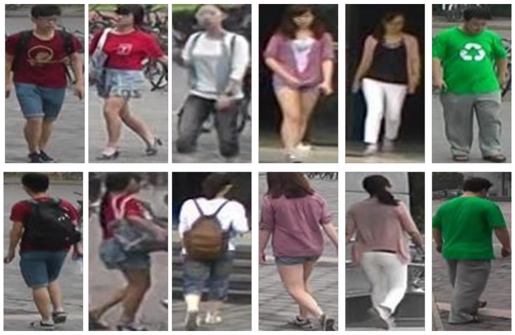

Figure 1: Due to the presence of large pose variations, the same person's appearance can be very different across multiple camera views.

1 INTRODUCTION

Person re-identification (re-id) is to match a certain person across multiple non-overlapping camera views. Given one query image of a certain person, an ideal person re-id system is expected to retrieve all the images of the same person from a set of gallery images. In the last few years, the re-id has attracted more and more research interest, because of its wide range of applications such as activity analysis, searching people of interest (e.g. criminals or terrorists) and long-term tracking. Despite of significant progress in recent years, it remains a difficult task for developing robust algorithms to match persons in scenarios with challenging factors such as cluttered backgrounds, severe occlusions, illumination changes and pose variations.

A variety of approaches have been proposed to address the above problems [46, 49, 52, 55], by representation learning or building robust signature matching. For example, in [46], the whole body is divided into a few fixed parts for person appearance representation learning without considering the alignment between parts. In [49], a Global-Local-Alignment Descriptor (GLAD) is proposed to detect key pose points and extract local features from corresponding regions. Among the above methods, most of them still show low re-id accuracy in real scenarios. A critical influencing factor on re-id accuracy is the large appearance changes of human body, which can be attributed to the changes in various covariate factors independent of the person's identity. These factors

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10. . . $15.00
https://doi.org/10.1145/3240508.3240573

include viewpoint, body configuration, lighting, and occlusion. Among these factors, pose plays the most important role in causing a person's appearance changes. Here pose is defined as a combination of viewpoint and body configuration, and it also affects self-occlusion. As shown in Figure 1, the same person's appearance can be very different across camera views, due to the presence of large pose variations. For instance, the person in the bottom row carry a big backpack, which is in full display from the back, but disappears from the front in the top row. As a result, it is challenging to build a robust re-id system with pose variations, but it is more applicable in real scenarios. Therefore, in this paper, we focus on the pose-invariant person re-id, which is to perform re-id by identifying or authorizing individuals identity with images captured under arbitrary poses.

However, it is not easy to extract robust pose-invariant features from person images. The factors to determine a person's appearance can be categorized into either identity-sensitive ones or identity-insensitive ones. The former factors mainly include some static physical attributes such as gender, carrying, clothing color, and texture. The latter factors are the dynamic covariates mainly associated with pose variations. The traditional methods aim to keep as much the effects of the former as possible and as little the ones of the latter as possible, and usually have two distinct perspectives: (1) utilizing hand-craft local invariant features extracted from images to represent the visual appearance of a person image [48, 66, 70]. (2) learning a discriminative distance metric to reduce the distance among features of person images with the same identity [15, 37, 54]. However, hand-craft features may be not robust enough to handle pose variations since human body has a complex non-rigid structure with lots of degrees of freedom. In addition, the distance metric is computed for each pair of cameras, which makes distance metric learning based person re-id methods suffer from high computational complexity. Furthermore, the identity-sensitive and identity-insensitive information are complexly interactional, e.g., the appearance of the carrying depends on the pose. Therefore, it is challenging for the traditional methods to perform person re-id free of the influence of pose variations.

Recently, inspired by the success of deep networks on a wide range of visual tasks [9, 10, 60, 72], a number of methods have been proposed to employ deep models to learn discriminative pose-invariant features. For example, Su et al.[43] utilize normalized part regions detected from a person image to learn pose-invariant feature representations. Differently, Zhao et al. [65] combine region selection and detection with deep re-id in a unified model. However, it is well known that deep models need to be trained with sufficient labeled samples, while the data annotation is expensive and time-consuming for the person re-id task. A camera network can easily consist of hundreds of cameras in a real-world scenario. It is difficult and tedious to manually collect and annotate sufficient identities and sufficient images per identity across views in the camera networks. Moreover, the generalizability of existing deep models to new camera networks is unsatisfactory. Generally, additional annotated data is needed for model fine-tuning when a trained deep re-id model is applied to a new camera network, or the performance would drastically declined. To deal with this issue, a better way is to generate training data automatically by exploiting human pose information. In recent times, the generative adversarial network (GAN) based approaches have been successfully used to generate impressively realistic faces [17, 30], house-numbers [64], bedrooms [39] and a variety of other image categories [13, 75] through a two-player game between a generator G and discriminator D. This inspires us to resort to the GAN to enlarge and enrich the training set. Despite many promising developments [17, 33, 39, 62, 74], image synthesis remains the main objective of GAN, which cannot be straightforwardly applied to the person re-identification task.

Inspired by the above discussions, a novel unified deep person re-id framework is proposed in this paper. Our full pipeline proceeds in two stages. At stage-I, we design a GAN-based structure to generate person images under arbitrary poses, which results in an increase in the training data. To disentangle the pose information from the identity-sensitive representation, we construct the generator G with an encoder-decoder structure, which serves as a person image changer. The input to the encoder G_{enc} is a combination of the condition person image and its pose embedding, and the output of the encoder G_{dec} is a synthesized realistic-looking person image under the desired pose, and the learnt identity-sensitive representation bridges the encoder G_{enc} and the decoder G_{dec}. Besides, a pose estimator is embedded into the GAN-based model as a supervisor to guide the desired pose structure generation. At stage-II, we utilize the synthesized images with different poses to train the re-id model, which can produce a set of pose-free features. Then, we combine the learned pose-free features with the features extracted from the original images, and thus obtain the final robust pose-invariant feature representations. Specifically, with the synthesized images as auxiliary samples, the learned re-id model is far more likely to generalize to a new person re-id dataset.

The major contributions of this work can be summarized as follows. (1) We propose a novel generative adversarial model by exploiting person poses and identities jointly for simultaneous person image synthesis under arbitrary poses and pose-invariant person re-identification. (2) In order to guide the image generation with any desired pose, we exploit a pre-trained human body pose estimator into the generator as a supervisor in the training process to help pose transfer. As a result, the proposed model can automatically generate a person image under an arbitrary pose. (3) The identity-sensitive representation is explicitly disentangled from pose variations through the person identity and pose embedding in G and D. (4) By using the synthesized images as auxiliary samples, the learned re-id model has better generalizability on a new person re-id dataset. Experimental results on four benchmarks including Market-1501 [69], DukeMTMC-reID [40], CUHK03 [23], and CUHK01 [22] demonstrate that the proposed model can achieve a competitive performance by comparing with state-of-the-art methods.

2 RELATED WORK

Image Generation by Deep Generative Models. With the emergence of deep generative models, many methods have been proposed to generate realistic images of objects [18]. Generally, the most commonly used generative models can be categorized into two groups. The first line of works is based on Variational Autoencoder (VAE) [18] framework. The VAE based models are trained by applying a re-parameterization method to maximize the lower bound of the data likelihood. The second group of works derives from the most popular generative model - Generative Adversarial Network (GAN) [12]. The GAN models are designed to simultaneously learn a discriminator D to distinguish generated samples from real ones and a image generator G to generate samples that can fool the discriminator by playing a min-max game.

With the rapidly development of the GAN models, amazing effects have been achieved in image generation. Isola et al. [16] propose a conditional GAN framework to transfer the low-level information of the condition image to the output one, which achieves incredible performance in image-to-image translation. Many other variants of GAN , such as VAEGAN [19], stacked-GAN [57] are also proposed in succession. However, most of them are designed for simple-texture high-quality sample generation, instead of person images which have complex background and non-rigid human body structure. There are some very recent works aiming at generating person images in surveillance scenario. Zheng et al. [74], firstly propose to use deep image generator for person re-id. Nevertheless, they directly adopt the DCGAN architecture [39] to generate person images from noise, and thus cannot control either identity or pose in the generated person images, which leads to unrealistic results. Ma et al. [32] propose a pose guided person generation network which allows to synthesize person images conditioned on a reference image and an intended pose. In [20], the generation process is divided into two stages: pose generation and appearance refinement, and a mask ℓ_1 loss is employed to pay more attention to transferring the human body appearance instead of background information. Qian et al. [38] propose a PN-GAN model which also allows to generate realistic, identity-preserving and pose controllable person images. The synthesized images of eight canonical poses are utilized to enhance the scalability and generalizability of the model. Compared with these approaches, the proposed model has the following differences. (1) It can generate person images with arbitrary poses. (2) A pose estimator is embedded in the proposed model to guide the image generation with any desired pose. (3) The proposed model can explicitly disentangle identity-sensitive information from different poses.

Person re-id by Deep Learning. Recently, Employing Deep Neural Network (DNN) has become main trend for person re-id task. The related work can be summarized into three categories based on their motivation and network structures, *i.e.,* classification-based network, metric learning-based network, and part-based network. *Classification-based Network:* In [51, 53, 71], the pre-trained classification network is fine-tuned on target person re-id datasets to generate the discriminative representation. For example, Zheng et al. [71] simply employ the DNN to extract features used for person Re-id. To overcome the distribution bias between different person re-id datasets, Xiao et al. [53] propose a novel dropout strategy to train a classification model. *Metric Learning based model:* Different from the classification-based methods, the metric learning-based model is trained to verify the similarity between images among each set. For person re-id, there are two popular metric learning-based models: siamese network and triplet network. Several works [41, 42, 46, 50, 56, 59, 73] employ the siamese network to verify whether the two input images contain the same person. The siamese network is trained with known pair-wise simialrity, which could be too strict and hard to collect. Therefore, Some works [7, 28, 29] study to train the network with relative similarity among three images, named as triplet, to learn the discriminative description for person re-id. *Part-based method:* One of the main challenges for person re-id is the diversity of human pose. Therefore, it is hard to generate a robust and aligned global representation. Targeting to generate a robust representation for person re-id, several works [21, 43, 43, 55, 63, 65] focus on how to generate aligned local person parts used for generating aligned person representation. Different from the existing methods, the proposed model is designed to exploit different poses and person identities jointly for simultaneous person image synthesis under arbitrary poses and pose-invariant person re-identification.

3 OUR APPROACH

In this section, we describe the design of our image generation model and our strategy for effectively using the synthesized images to train the Feature Extraction Network (FEN) for person re-id. Our model consists of a GAN based person image generation model and two FENs as shown in Figure 2. The details are elaborated as follows.

3.1 Person Image Generation

Our image generation model aims to simultaneously transfer the person on the referenced image from a given pose to a target pose and preserve important appearance information of the identity. To achieve the above goal, a conditional GAN architecture is employed to tackle such a challenging task. As in all GAN based models, the image generation is composed of a Generator G and a Discriminator D. The generator is trained to synthesize realistic identity-preserving person image under a desired pose conditioned on a given sample, and the discriminator is to distinguish generated samples from real data and help to improve the performance of the generator. The details of the proposed image generation model are shown in Figure 3.

Generator and Pose Embedding. To obtain human body poses, we employ an off-the-shelf state-of-the-art pose estimator [2] to avoid expensive annotation of poses, which is not fine-tuned on any re-id benchmark dataset. The pose estimator takes a person image of size $h \times w \times 3$ as input

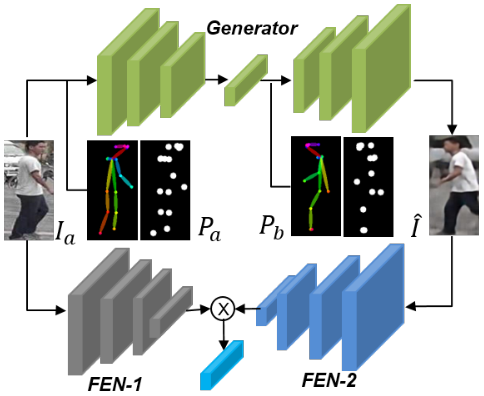

Figure 2: Overview of the proposed framework. It consists of the GAN-based model to generate identity-preserving person images under arbitrary poses and the two Feature Extraction Networks (FENs) to learn robust pose-invariant features from the original images and synthesized images for the re-id task.

and outputs 19 confidence maps of body part locations and corresponding Part Affinity Fields (PAFs) which encode the degree of association between body parts. Then, the coordinates of 18 keypoints and 3-channel skeleton map P^m can be inferred based on the confidence maps and PAFs. By using these information directly as input, the proposed model can map each keypoint to a position on the human body. Therefore, we encode the coordinates of the 18 keypoints as 18 heatmaps P^h. Each heatmap is filled with 1 in a radius of 4 pixels around the corresponding keypoints and 0 elsewhere (see Figure 3). We finally concatenate P^h and P^m into a 21-channel tensor P as pose embedding.

Given an input person image I_a and a target person image I_b in which the person has the same id y_a and a different pose P_b with I_a, we need to learn a generator to synthesize an image $\hat{I}$ that is as similar as the I_b. As shown in Figure 3, our generator G consists of an encoder G_{enc} and a decoder G_{dec}. The G_{enc} aims to learn an identity representation from input which is the concatenation of the condition person image I_a and its pose information P_a: $f_a = G_{enc}(I_a, P_a)$. The G_{dec} aims to synthesize a natural target person image $\hat{I} = G_{dec}(f_a, P_b, z)$ with the identity representation f_a, a target pose specified by P_b, and a random noise $z \in \mathrm{R}^d$ for modeling other variance besides identity or pose (e.g. background). Specifically, the G_{enc} integrates I_a and P_a from small local neighborhoods to global areas to encode as much identity-sensitive information as possible in f_a. Then, with a separate target pose embedding input to the G_{dec} and the person identity considered in D, the G_{enc} is trained to disentangle the pose variations from f_a in the adversarial learning process. Specifically, to avoid that a large amount of low-level information associated with appearance is lost in the bottleneck layer, skip connections are introduced to propagate these information from the bottom convolution layers of the encoder to the corresponding layers in decoder.

The employed network architecture of the generator is summarized as follows. The encoder of generator consists of N residual blocks and one fully-connected layer, where N depends on the size of input. Each residual block consists of two convolution layers with $stride = 1$ followed by one sub-sampling convolution layer with $stride = 2$ except the last block. The decoder is symmetric to the encoder. As shown in Figure 3, there are skip connections between the corresponding layers of the encoder and decoder which are built by concatenating the corresponding feature maps together. Note that rectified linear unit (ReLU) is applied to each convolution layer except the fully connected layer and the output layer, and no batch normalization or dropout are applied.

Discriminator. The discriminator is usually designed to differentiate whether the input images are real groundtruth images or fake generated images. To avoid that the generator G is misled to directly output the condition image I_a instead of synthesizing natural image of target pose I_b, we pair the G output $\hat{I}$ with I_a, and train the discriminator D to distinguish the fake pair $\left(\hat{I}, I_a\right)$ from the real pair of target image and condition image (I_b, I_a) (as shown in Figure 3).

Architecture Training. The G and D are trained using a combination of a standard conditional adversarial loss L_{GAN}, a masked ℓ_1 loss and a pose regression loss. In our conditional GAN, the objective function is formulated as follows:

$$\begin{aligned} L_{GAN}(G, D) = &\mathrm{E}_{I_a, I_b \sim p_{data}(x)} \left[\log D(I_a, I_b)\right] + \\ &\mathrm{E}_{z \sim p_z(z)} \left[\log\left(1 - D\left(I_a, \hat{I}\right)\right)\right], \end{aligned} \tag{1}$$

where $\hat{I} = G(I_a, P_a, P_b, z)$ is the synthesized target image.

Because the condition person image and the target person image are captured under disjoint camera views respectively, it is difficult for the model to imagine what the target image background would look like. Therefore, we adopt a masked ℓ_1 loss proposed by [32] for the G, which encourages the model to focus on transferring the human body appearance instead of background information. The masked ℓ_1 loss is defined by:

$$L_{l1-mask} = \left\| \left(\hat{I} - I_b\right) \odot (1 + M_b) \right\|_1, \tag{2}$$

where M_b is the target pose mask that is set to 0 for background and 1 for foreground. However, a well-known problem caused by the use of ℓ_1 loss is that the synthesized image is blurry to a certain degree. To further improve the generated human pose structure, we attach a pre-trained human body pose estimator [2] to the generator as a supervisor in the training process. The pose regression loss is formulated as:

$$L_{pose} = \sum_{t=1}^{T} \left(L_H^t + L_V^t\right), \tag{3}$$

where L_H^t and L_V^t are defined as follows:

$$L_H^t = \sum_{j=1}^{J} \sum_{p} \left\| H_j^t(p) - H_j^*(p) \right\|_2^2, \tag{4}$$

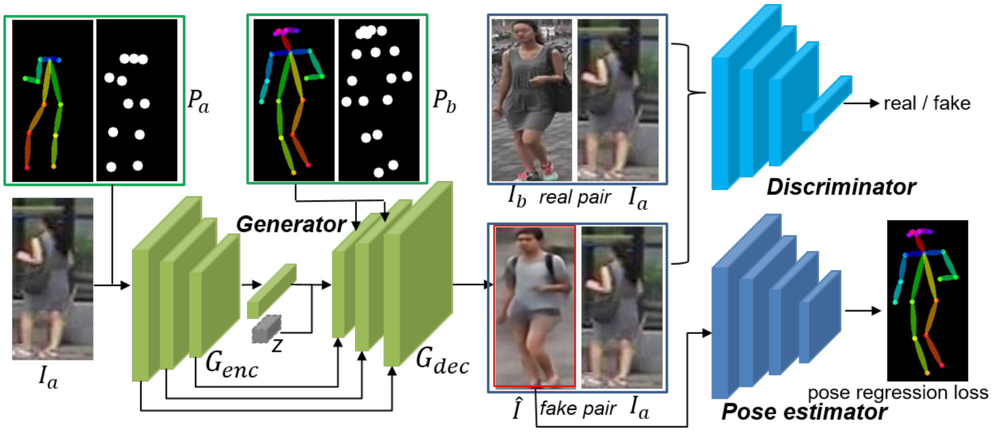

Figure 3: The proposed GAN-based image generation model embedded with an effective pose estimator. The generator G is constructed with an encoder-decoder structure to synthesize identity-preserving person images and the discrminator D is a binary classifier to distinguish fake image pairs from real image pairs.

$$L_V^t = \sum_{c=1}^{C}\sum_{p}\left\|V_c^t(p) - V_c^*(p)\right\|_2^2, \qquad (5)$$

where H_j^* is a part confidence map of the j-th channel and V_c^* is a part affinity vector field of the c-th channel, which are the final output of the employed pre-trained pose estimator ϕ with the target person image I_b input to it, i.e., $(H^*, V^*) = \phi(I_b)$. Similarly, $\left(H^t, V^t\right) = \phi\left(\hat{I}\right)$ are generated from the synthesized image $\hat{I}$. Specifically, the pose estimator refines the predictions over successive stages, $t \in \{1, \ldots, T\}$, with intermediate supervision at each stage, which can address the vanishing gradient problem by replenishing the gradient periodically. Note that we do not tune the parameters of the pose estimator ϕ and back propagate the gradient only for G in the training process.

Finally, we have the following loss function for the generator G and the discriminator D, respectively.

$$L_G = L_{GAN}^G + \lambda_1 \cdot L_{l1-mask} + \lambda_2 \cdot L_{pose}, \qquad (6)$$

$$L_D = -L_{GAN}^D, \qquad (7)$$

where λ_1 is the weight of $L_{l1-mask}$. It controls low-level information transferring in the image generation process. λ_2 is the weight of L_{pose} to control the refinement of the generated pose structure. λ_1 , and λ_2 of the generated images In the training process of the person image generation model, the optimization step would be to iteratively minimize the loss function L_G and L_D until convergence.

3.2 Person re-id with synthesized images

Person re-id aims to retrieve the images that are about the same identity with the query image from a large-scale gallery dataset. Assume that there is a training dataset of N persons $S_{train} = \left\{\left\{I_i^j\right\}_{i=1}^{C_j}, y_j\right\}_{j=1}^{N}$, where I_i^j is the i-th image (out of C_j images) of the j-th person whose person id is y_j. In the training process we learn a feature extractor φ which maps person images into a feature space, so that a given image I can be represented by a feature vector $\varphi_I = \varphi(I)$. In the testing process, given a query image I_q and a gallery set $\{I_k\}_{k=1}^{G}$, we need to find all the images of the same id with I_q from gallery set. This is done by ranking the identity-similarity which measured by the Euclidean distance between φ_{I_q} and $\{\varphi_{I_k}\}_{k=1}^{G}$.

Training Stage. To learn more robust features, we train two Feature Extraction Networks (FENs) as shown in Figure 2. Here, the FEN-1 and FEN-2 are both built as the architecture of ResNet-50 [14] which has proven to be effective for deep feature learning. The FEN-1 is trained using the original images to extract identity-sensitive features in the presence of pose variation. The FEN-2 is finetuned on FEN-1 using the synthesized images with 20 arbitrary poses to compute re-id features which are free of pose variation. Specifically, we only adopt classification loss for training our person re-id model. Since the image generation model inevitably loses or distorts some information associated with identity in the process of pose transferring, we introduce a label smoothing regularization [74] trick to the classification loss function when fine-tuning FEN-2 using the synthesized images.

$$L_{cls} = -\sum_{k=1}^{K}\log\left(p(k)\right)q(k) \qquad (8)$$

Assume that $k \in \{1, \ldots, K\}$ are the pre-defined classes of the training data, where K is the number of classes. The cross-entropy loss is formulated as in Eq.(8), where $p(k) \in [0, 1]$ is the predicted probability of the input belonging to class k, $q(k)$ is defined as:

$$q(k) = \begin{cases} 0 & k \neq y \\ 1 & k = y, \end{cases} \qquad (9)$$

where y is the groundtruth class label. The label smoothing regularization takes the distribution of the non-groundtruth classes into account and encourages the network not to be too confident towards the groundtruth. The regularized label distribution $q_{LSR}(k)$ is formulated as:

$$q_{LSR}(k) = \begin{cases} \frac{\rho}{K} & k \neq y \\ 1 - \frac{(K-1)\rho}{K} & k = y \end{cases}, \tag{10}$$

where $\rho \in [0, 1]$ is a hyperparameter. If ρ is zero, Eq. (10) reduces to Eq. (9). If ρ is too large, the model may fail to predict the ground truth label. In our case, ρ is set to 0.1. According to Eq. (10), the cross-entropy is re-written as:

$$L_{LSR} = -(1-\rho)\log(p(y)) - \frac{\rho}{K}\sum_{k=1}^{K}\log(p(k)). \tag{11}$$

Testing Stage. Once the training of our person re-id model is finished, during testing, given a query image I_q, we feed it into FEN-1 to output a feature vector. Then we synthesize 20 images of I_q with arbitrary poses and feed them into FEN-2 to obtain the pose-free features. A final feature vector is obtained by fusing the above two feature vectors by element-wise maximum operation. For each gallery image, we do the same process to obtain the gallery feature vectors in an off-line manner. We then rank the identity-similarity of gallery images by measuring the Euclidean distance between the final feature vectors of the query and gallery images.

4 EXPERIMENTS

In this section, we show experimental results of the proposed model for person images synthesis and person re-identification. For the former task, we show qualitative results of the generated images under different poses. For the latter one, we quantitatively evaluate the re-id performance.

4.1 Datasets

To demonstrate the effectiveness of the proposed model, we conduct extensive experiments on four widely-used datasets including Market-1501 [69], DukeMTMC-reID [40], CUHK03 [23], and CUHK01 [22]. The details are as follows.

Market-1501: This dataset is collected from six different view cameras. It has 32,668 bounding boxes of 1,501 identities obtained with a Deformable Part Model person detector. Following the standard split, we use 751 identities with 12,936 images as training set and the rest 750 identities with 19,732 images for testing. **DukeMTMC-reID**: It is constructed from the multi-camera tracking dataset DukeMTMC and contains 1,812 identities. Following the evaluation protocol [74], 702 identities are adopted as the training set and the remaining 1,110 identities as the testing set. During testing, one query image for each identity in each camera is used for query and the remaining as the gallery set. **CUHK03**: It includes 14,096 images of 1,467 identities, captured by six camera views with 4.8 images for each identity in each camera on average. We adopt the more realistic yet harder detected person images setting. The training, validation and testing sets consist of 1,367 identities, 100 identities and 100 identities, respectively. The testing process is repeated with 20 random splits as in [24]. **CUHK01**: It contains 971 identities with 2 images captured in two disjoint camera views per person. As in [22], we utilize as probe the images of camera A and adopt those from camera B as gallery. 486 identities are randomly selected for testing and the remaining are used for training. The experiments are repeated for 10 times with the average results reported.

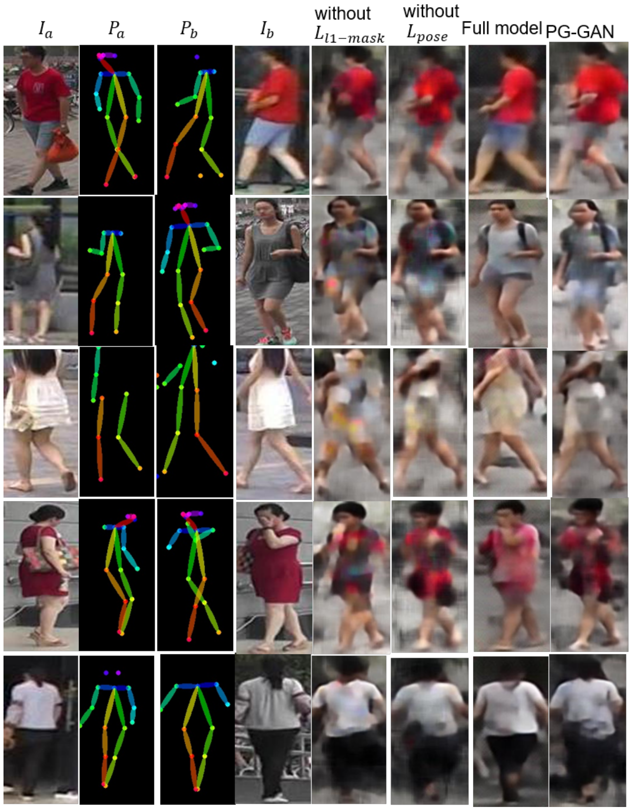

Figure 4: Qualitative results on the Market-1501 dataset. The columns 1, 2 and 3 represent the inputs of our model. The column 4 corresponds to target images. The last four columns show the outputs of our three baselines and PG-GAN [32], respectively.

4.2 Image Generation Results

Experimental Settings. We evaluate the proposed image generation model on the Market-1501 dataset. Persons' appearance in this dataset vary significantly due to pose variation, illumination change, different viewpoints and occlusion, which make the person generation task more challenging. We need pairs of images of the same person in two different poses. Following [32], we adopt 439,420 pairs person images in the training set to train our image generation model.

Qualitative Analysis. We provide qualitative results on the Market-1501 dataset to compare the proposed model with recently published deep person image generation model [32]. Moreover, we present an ablation study to clarify the impact of two supervised losses introduced to the generator G on the final performance, namely the masked ℓ_1 loss $\ell_{l1-mask}$ and the pose regression loss ℓ_{pose}. Specifically, the following models are compared for ablation study, which are obtained

Table 1: Results on the Market-1501 dataset.

Methods	Single-Query		Multi-Query	
	R-1	mAP	R-1	mAP
TMA [34]	47.90	22.3	-	-
SCSP [3]	51.90	26.40	-	-
DNS [58]	61.02	35.68	71.56	46.03
LSTM Siamese [46]	-	-	61.60	35.31
Gated_Sia [45]	65.88	39.55	76.50	48.50
HP-net [31]	76.90	-	-	-
Spindle [63]	76.90	-	-	-
Basel. + LSRO [74]	78.06	56.23	85.12	68.52
PIE [68]	79.33	55.95	-	-
Verif.-Identif. [73]	79.51	59.87	85.84	70.33
DLPAR [65]	81.00	63.40	-	-
DeepTransfer [11]	83.70	65.50	89.60	73.80
PDC [43]	84.14	63.41	-	-
JLML [25]	85.10	65.50	89.70	74.50
Zhang et al. [61]	92.60	82.30	-	-
Ours	**92.81**	**82.67**	**93.62**	**84.50**

by “amputating one of the two supervised losses. Note that the model architecture is the same for these models.

- *Our model without* ℓ_{pose}: This model is trained with the adversarial loss and the masked $\ell 1$ loss $\ell_{l1-mask}$.
- *Our model without* $\ell_{l1-mask}$: This model training is performed using the adversarial loss together with the pose regression loss ℓ_{pose}.
- *Our full model*: This model is trained using the combination of the adversarial loss, the masked $\ell 1$ loss $\ell_{l1-mask}$, and the pose regression loss ℓ_{pose}.

In Figure 4, we show the qualitative results of [32] and our three baseline methods. These images show the progressive improvement through the three baselines. From Figure 4 we observe that: (1) Comparing to [32], the images synthesized by our full model are more realistic, sharper and with local details more similar to the details of the conditioning image, e.g., the carrying condition and other details such as hair style and shoe-wear are better preserved in each synthesized image. we deem that it is because the identity-sensitive information is disentangled from different poses in the image generation process and thus the negative effects of pose variations are overcome to a large extent. (2) When the ℓ_{pose} is not used for model training, the synthesised poses are still similar to the target poses silhouette but some local details are lost. (3) When the $\ell_{l1-mask}$ is removed, we can see a clear degradation of the persons’ appearance quality. This indicates that it is important to focus on transferring the human body appearance instead of background information.

4.3 Person Re-identification Results

Evaluation metrics. Two evaluation metrics are used to quantitatively measure the re-id performance. The first one is Rank-1, Rank-5 and Rank-10 accuracy. For Market-1501 and DukeMTMC-reID datasets, the mean Average Precision (mAP) is also used.

Table 2: Results on the DukeMTMC-reID dataset. The SL and TL indicate the supervised learning and transfer learning settings respectively.

Methods	R-1	R-5	R-10	mAP
LOMO+XQDA [26]	30.80	-	-	17.00
ResNet50 [14]	65.20	-	-	45.00
Basel.+LSRO [74]	67.70	-	-	47.10
AttIDNet [27]	70.69	-	-	51.88
SVDNet [44]	76.70	86.40	89.90	56.80
DPFL [6]	79.20	-	-	60.60
Ours-SL	**88.67**	**93.06**	**97.11**	**79.32**
Ours-TL	61.05	76.61	83.74	54.13

Experimental Settings. In the following experiments, we have two different settings. One is the standard Supervised Learning (SL) setting on all datasets: the models are trained on the training set of the dataset, and evaluated on the corresponding testing set. The other one is the Transfer Learning (TL) setting on the DukeMTMC-reID, CUHK01, and CUHK03 datasets. Specifically, the re-id model is trained on Market- 1501. We then directly adopt the trained single model to run the testing on the test set of DukeMTMC-reID, CUHK01, and CUHK03 datasets. Here, no model updating is done by using any data from these datasets. The TL setting is especially useful in real-world scenarios, where a pre-trained model needs to be deployed to a new camera network without any model fine-tuning. This setting can also evaluate how generalizable a re-id model is.

Supervised Learning Results. We report our results obtained under the supervised learning settings on the Market-1501, DukeMTMC-reID, CUHK03, and CUHK01 datasets in Table 1, Table 2, Table 3, and Table 4, respectively. Here, we compare the proposed model with the best performing re-id models on the four re-id datasets. Based on the results, it is clear that our model outperforms most of existing methods or achieves comparable results, and we can make the following observations: (1) The proposed model achieves the best results on the Market-1501, DukeMTCM-reID and CUHK01 datasets. Particularly on the DukeMTCM-reID dataset, our model obtains 88.67% Rank-1 accuracy and 79.32% mAP which are around 9% and 19% higher than the second best reported results in [6], respectively. (2) Compared with the method [74] that utilizes generated images for the re-id model training, our model achieves notable improvement on all four datasets (e.g, over 20% at mAP on the Market-1501 and DukeMTCM-reID datasets). This is because the proposed model can synthesize identity-preserving images under desired poses, which can thus be used for supervised training to learn pose-invariant feature representations. In contrast, the synthesize images in [74] can only be used as unlabeled or weakly-labeled data because of no identity and pose information modeled in it. Note that, the baseline performance of the ResNet-50 is reported in [74]. The rank-1 accuracies are 73.69%, 71.5% and 60.28% on the Market-1501, CUHK03, and DukeMTMC-reID datasets, respectively. Therefore, it is helpful to utilize the synthesized person images for the re-id

Table 3: Results on the CUHK03 dataset. The SL and TL indicate the supervised learning and transfer learning settings respectively.

Methods	R-1	R-5	R-10
DeepReid [24]	19.89	50.00	64.00
Imp-Deep [1]	44.96	76.01	83.47
EMD [42]	52.09	82.87	91.78
SI-CI [47]	52.17	84.30	92.30
LSTM Siamese [46]	57.30	80.10	88.30
PIE [68]	67.10	92.20	96.60
Gated_Sia [45]	68.10	88.10	94.60
Basel.+LSRO [74]	73.10	92.70	96.70
PDC [43]	78.92	94.83	97.15
DLPAR [65]	81.60	97.30	98.40
Verif-Identif. + LSRO [74]	84.60	97.60	98.90
Zhang et al. [61]	**91.9**	**98.7**	99.4
Ours-SL	89.53	97.65,	**99.87**
Ours-TL	71.24	95.62	99.48

Table 4: Results on the CUHK01 dataset. The SL and TL indicate the supervised learning and transfer learning settings respectively.

Methods	R-1	R-5	R-10
ITML [8]	15.98	35.22	45.60
eSDC [66]	19.76	32.72	40.29
kLFDA [54]	32.76	59.01	69.63
mFilter [67]	34.30	55.00	65.30
Imp-Deep [1]	47.53	71.50	80.00
DeepRanking [4]	50.41	75.93	84.07
Ensembles [36]	53.40	76.30	84.40
ImpTrpLoss [7]	53.70	84.30	91.00
GOG [35]	57.80	79.10	86.20
Quadruplet [5]	62.55	83.44	89.71
NullReid [58]	64.98	84.96	89.92
G-Dropout [53]	71.70	88.60	92.60
Ours-SL	**84.79**	**94.88**	**98.14**
Ours-TL	59.31	82.42	92.72

model training. (3) Compared with the existing pose-guided re-id models [3, 63, 65, 68], the improvements of the proposed model are quite striking. The proposed model achieves 92.81% and 82.67% in terms of Rank-1 and mAP, respectively, on the Market-1501 dataset. In [65], the method achieves the best performance with the results of 81.00% and 63.4% in terms of Rank-1 and mAP, respectively. Compared with the best method in [65], the proposed model outperforms around 11% and 19% higher accuracy in terms of Rank-1 and mAP, respectively. The results imply that it is quite effective to train the re-id model by utilizing the synthesized person images under different poses for removing the negative effects of pose variations on the final feature representations. (4) On the small-scale dataset CUHK01, the proposed model achieves the highest performance 84.79% in term of Rank-1 as shown in Table 4. The gap between ours and the second best is much bigger (over 13% on Rank-1 accuracy). Note that the traditional methods based on handcrafted features and metric learning (e.g., GOG [35]) are still quite competitive, even beating some deep models (around 4% Rank-1 accuracy higher than [7]), which shows that limitations of the existing deep models on generalizability. With the synthesized person images, the proposed model is more adaptive to the small dataset and can achieve much better performance than existing models. (5) On the CUHK03 dataset, the proposed model achieves comparable performance by comparing with the best existing method [61]. It is because the strong generalizability of our model may hurt the performance on one specific dataset. In order to further demonstrate the generalization ability, we evaluate our model on transfer learning setting.

Transfer Learning Results. We show our results obtained under the transfer settings on the DukeMTMC-reID, CUHK03, and CUHK01 datasets in Table 2, Table 3, and Table 4, respectively. In this experimental setting, we only adopt the training data from the Market-1501 dataset to train the models, and evaluate the trained model on the test set of each individual dataset. On the CUHK03 dataset, Table 3 shows that our model achieves 71.24% Rank-1 accuracy, which beats some deep re-id models [45, 46, 68] which are fine-tuned on the training set of the CUHK03 dataset. On the CUHK01 dataset, we can achieve 59.31% Rank-1 accuracy in Table 4 which is comparable to many existing models trained under the supervised learning setting. These results thus demonstrate that the proposed model has the potential to be truly generalizable to a new re-id data from new camera networks.

5 CONCLUSIONS

We proposed a novel deep person re-id model in this work. The proposed method contains a deep person image generation model which can synthesize person images for the re-id model training. Moreover, the identity-sensitive information can be disentangled from pose variations in the image generation process. As a result, the proposed image generation model can automatically generate identity-preserving person images under arbitrary poses. In contrast to previous re-id approaches extracting discriminative features which are identity-sensitive but view-insensitive, the proposed re-id model can learn robust pose-invariant features from both original person images and synthesized images. Besides, by using the synthesized images as auxiliary samples to perform the model training, the learned re-id model can have better generalizability on a new person re-id dataset.

6 ACKNOWLEDGEMENT

This work is supported by National Nature Science Foundation of China under Grant 61751211, U1705262, 61432019, 61572498, 61532009, 61728210, 61721004, and 61572296, the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation, the Key Research Program of Frontier Sciences, CAS, Grant NO. QYZDJ-SSW-JSC039, and the Beijing Natural Science Foundation 4172062. the National Key Research and Development Program of China, No. 2017YFB1002804.

REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks. 2015. An improved deep learning architecture for person re-identification. In *CVPR*.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Real-time multi-person 2d pose estimation using part affinity fields. In *CVPR*.

[3] Dapeng Chen, Zejian Yuan, Badong Chen, and Nanning Zheng. 2016. Similarity learning with spatial constraints for person re-identification. In *CVPR*. 1268–1277.

[4] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. 2016. Deep ranking for person re-identification via joint representation learning. *TIP* 25, 5 (2016), 2353–2367.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, Vol. 2.

[6] Y. Chen, X. Zhu, and S. Gong. 2017. Person re-identification by deep learning multi-scale representations. In *ICCVW*. 2590–2680.

[7] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*. 1335–1344.

[8] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *ICML*. ACM, 209–216.

[9] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2017. A unified personalized video recommendation via dynamic recurrent neural networks. In *ACM MM*. 127–135.

[10] Junyu Gao, Tianzhu Zhang, Xiaoshan Yang, and Changsheng Xu. 2018. P2t: Part-to-target tracking via deep regression learning. *IEEE Transactions on Image Processing* 27, 6 (2018), 3074–3086.

[11] Mengyue Geng, Yaowei Wang, Tao Xiang, and Yonghong Tian. 2016. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*. 2672–2680.

[13] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and Venkatesh Babu Radhakrishnan. 2017. DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data. In *CVPR*.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[15] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. 2012. Relaxed pairwise learned metric for person re-identification. In *ECCV*. 780–793.

[16] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*. 5967–5976.

[17] Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. 2017. Generative Attribute Controller with Conditional Filtered Generative Adversarial Networks. In *CVPR*. 6089–6098.

[18] D. P. Kingma and M. Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

[19] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2015. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*.

[20] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. 2017. A Generative Model of People in Clothing. In *CVPR*. 853–862.

[21] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. 2017. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*. 384–393.

[22] Wei Li, Rui Zhao, and Xiaogang Wang. 2012. Human Reidentification with Transferred Metric Learning. In *ACCV*.

[23] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Reidentification. In *CVPR*.

[24] W. Li, R. Zhao, T. Xiao, and X. Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*.

[25] Wei Li, Xiatian Zhu, and Shaogang Gong. 2017. Person reidentification by deep joint learning of multi-loss classification. *IJCAI* (2017).

[26] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

[27] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. 2017. Improving person re-identification by attribute and identity learning. In *arXiv preprint arXiv:1703.07220*.

[28] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2016. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404* (2016).

[29] Jiawei Liu, Zheng-Jun Zha, QI Tian, Dong Liu, Ting Yao, Qiang Ling, and Tao Mei. 2016. Multi-scale triplet cnn for person re-identification. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 192–196.

[30] Ming Yu Liu and Oncel Tuzel. 2016. Coupled generative adversarial networks. In *NIPS*. 469–477.

[31] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. 2017. Hydraplus-net: Attentive deep features for pedestrian analysis. In *arXiv preprint arXiv:1709.09930*.

[32] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *NIPS*. 405–415.

[33] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2016. Adversarial autoencoders. In *ICLR*.

[34] Niki Martinel, Abir Das, Christian Micheloni, and Amit K Roy-Chowdhury. 2016. Temporal model adaptation for person reidentification. In *ECCV*. 858–877.

[35] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Y-oichi Sato. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*. 1363–1372.

[36] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. 2015. Learning to rank in person re-identification with metric ensembles. In *CVPR*. 1846–1855.

[37] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*. 3318–3325.

[38] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. 2017. Pose-Normalized Image Generation for Person Re-identification. *arXiv preprint arXiv:1712.02225*.

[39] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*.

[40] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*.

[41] Chen Shen, Zhongming Jin, Yiru Zhao, Zhihang Fu, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. 2017. Deep siamese network with multi-level similarity perception for person reidentification. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1942–1950.

[42] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. 2016. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*.

[43] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. 2017. Pose-driven Deep Convolutional Model for Person Re-identification. In *ICCV*. 3980–3989.

[44] Y. Sun, L. Zheng, D. Weijian, and W. Shengjin. 2017. SVDNet for pedestrian retrieval. In *ICCV*.

[45] R. R. Varior, M. Haloi, and G. Wang. 2016. Gated siamese convolutional neural network architecture for human reidentification. In *ECCV*.

[46] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. 2016. A siamese long short-term memory architecture for human reidentification. In *ECCV*.

[47] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*.

[48] Xiaogang Wang and Rui Zhao. 2014. Person re-identification: System design and evaluation overview. In *Person Re-Identification*. 351–370.

[49] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, and Qi Tian. 2017. Glad: Global-local-alignment descriptor for pedestrian retrieval. In *ACM Multimedia*. 420–428.

[50] L. Wu, C. Shen, and A.v.d. Hengel. 2016. PersonNet: Person Reidentification with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1601.07255* (2016).

[51] Shangxuan Wu, Ying-Cong Chen, Xiang Li, An-Cong Wu, Jin-Jie You, and Wei-Shi Zheng. 2016. An enhanced deep feature representation for person re-identification. In *WACV*.

[52] Qiqi Xiao, Kelei Cao, Haonan Chen, Fangyue Peng, and Chi Zhang. 2016. Cross domain knowledge transfer for person re-identification.

In *arXiv preprint arXiv:1611.06026*.

[53] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*. 1249–1258.

[54] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. 2014. Person re-identification using kernel-based metric learning methods. In *ECCV*. 1–16.

[55] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. 2017. Deep representation learning with part loss for person re-identification. In *arXiv preprint arXiv:1707.00798*.

[56] D. Yi, Z. Lei, and S.Z. Li. 2014. Deep metric learning for practical person re-identification. In *ICPR*.

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*. 5907–5915.

[58] Li Zhang, Tao Xiang, and Shaogang Gong. 2016. Learning a discriminative null space for person re-identification. In *CVPR*. 1239–1248.

[59] Tianzhu Zhang, Adel Bibi, and Bernard Ghanem. 2016. In Defense of Sparse Tracking: Circulant Sparse Tracker. In *CVPR*.

[60] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. 2018. Learning Multi-task Correlation Particle Filters for Visual Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2018), 1–1.

[61] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. 2017. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*.

[62] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *CVPR*.

[63] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*. 1077–1085.

[64] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2017. Energy-based generative adversarial network. In *ICLR*.

[65] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-Learned Part-Aligned Representations for Person Re-Identification. In *CVPR*. 3219–3228.

[66] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2013. Unsupervised salience learning for person re-identification. In *CVPR*. 3586–3593.

[67] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. 2014. Learning mid-level filters for person re-identification. In *CVPR*. 144–151.

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

[69] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable Person Re-identification: A Benchmark. In *ICCV*.

[70] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian. 2015. Query-adaptive late fusion for image search and person re-identification. In *CVPR*. 1741–1750.

[71] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. 2016. Person Re-identification in the Wild. *arXiv preprint arXiv:1604.02531* (2016).

[72] Yuhui Zheng, Le Sun, Shunfeng Wang, Jianwei Zhang, and Jifeng Ning. 2018. Spatially Regularized Structural Support Vector Machine for Robust Visual Tracking. *IEEE Transactions on Neural Networks and Learning System* (2018).

[73] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A Discriminatively Learned CNN Embedding for Person Reidentification. *TOMM* 14, 1 (2017).

[74] Z Zheng, L Zheng, and Y Yang. 2017. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. In *ICCV*.

[75] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *CVPR*.