Pose-Normalized Image Generation
for Person Re-identification

Xuelin Qian[1], Yanwei Fu[2,3], Tao Xiang[4], Wenxuan Wang[1], Jie Qiu[5],
Yang Wu[5], Yu-Gang Jiang[1(✉)], and Xiangyang Xue[1,2]

[1] Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University, Shanghai, China
{15110240002,17210240045,ygj,xyxue}@fudan.edu.cn
[2] School of Data Science, Fudan University, Shanghai, China
yanweifu@fudan.edu.cn
[3] Tencent AI Lab, Bellevue, USA
[4] Queen Mary University of London, London, UK
t.xiang@qmul.ac.uk
[5] Nara Institute of Science and Technology, Ikoma, Japan
qiu.jie.qf3@is.naist.jp, yangwu@rsc.naist.jp

Abstract. Person Re-identification (re-id) faces two major challenges: the lack of cross-view paired training data and learning discriminative identity-sensitive and view-invariant features in the presence of large pose variations. In this work, we address both problems by proposing a novel deep person image generation model for synthesizing realistic person images conditional on the pose. The model is based on a generative adversarial network (GAN) designed specifically for pose normalization in re-id, thus termed pose-normalization GAN (PN-GAN). With the synthesized images, we can learn a new type of deep re-id features free of the influence of pose variations. We show that these features are complementary to features learned with the original images. Importantly, a more realistic unsupervised learning setting is considered in this work, and our model is shown to have the potential to be generalizable to a new re-id dataset without any fine-tuning. The codes will be released at https://github.com/naiq/PN_GAN.

Keywords: Person re-id · GAN · Pose normalization

1 Introduction

Person Re-identification (re-id) aims to match a person across multiple non-overlapping camera views [14]. It is a very challenging problem because a person's appearance can change drastically across views, due to the changes in

---

Fig. 1. The same person's appearance can be very different across camera views, due to the presence of large pose variations.

various covariate factors independent of the person's identity. These factors include viewpoint, body configuration, lighting, and occlusion (see Fig. 1). Among these factors, pose plays an important role in causing a person's appearance changes. Here pose is defined as a combination of viewpoint and body configuration. It is thus also a cause of self-occlusion. For instance, in the bottom row examples in Fig. 1, the big backpacks carried by the three persons are in full display from the back, but reduced to mostly the straps from the front.

Most existing re-id approaches [2,9,25,34,40,47,51,63] are based on learning identity-sensitive and view-insensitive features using deep neural networks (DNNs). To learn the features, a large number of persons' images need to be collected in each camera view with variable poses. With the collected images, the model can have a chance to learn what features are discriminative and invariant to the camera view and pose changes. These approaches thus have a number of limitations. The first limitation is **lack of scalability** to large camera networks. Existing models require sufficient identities and sufficient images per identity to be collected from each camera view. However, manually annotating persons across views in the camera networks is tedious and difficult even for humans. Importantly, in a real-world application, a camera network can easily consist of hundreds of cameras (i.e. those in an airport or shopping mall); annotating enough training identities from all camera views are infeasible. The second limitation is **lack of generalizability** to new camera networks. Specifically, when an existing deep re-id model is deployed to a new camera network, view points and body poses are often different across the networks; additional data thus need to be collected for model fine-tuning, which severely limits its generalization ability. As a result of both limitations, although deep re-id models are far superior for large re-id benchmarks such as Market-1501 [61] and CUHK03 [25], they still struggle to beat hand-crafted feature based models on smaller datasets such as CUHK01 [24], even when they are pre-trained on the larger re-id datasets.

Even with sufficient labeled training data, existing deep re-id models face the challenge of learning identity-sensitive and view-insensitive features in the

presence of large pose variations. This is because a person's appearance is determined by a combination of identity-sensitive but view-insensitive factors and identity-insensitive but view-sensitive ones, which are inter-connected. The former correspond to semantic related identity properties, such as gender, carrying, color, and texture. The latter are the covariates mentioned earlier including poses. Existing models aim to keep the former and remove the latter in the learned feature representations. However, these two aspects of the appearance are not independent, *e.g.*, the appearance of the carrying depends on the pose. Making the learned features pose-insensitive means that the features supposed to represent the backpacks in the bottom row examples in Fig. 1 are reduced to those representing only the straps – a much harder type of features to learn.

In this paper, we argue that the key to learning an effective, scalable and generalizable re-id model is to remove the influence of pose on the person's appearance. Without the pose variation, we can learn a model with much less data thus making the model scalable to large camera networks. Furthermore, without the need to worry about the pose variation, the model can concentrate on learning identity-sensitive features and coping with other covariates such as different lighting conditions and backgrounds. The model is thus far more likely to generalize to a new dataset from a new camera network. Moreover, with the different focus, the features learned without the presence of pose variation would be different and complementary to those learned with pose variation.

To this end, a novel deep re-id framework is proposed. Key to the framework is a deep person image generation model. The model is based on a generative adversarial network (GAN) designed specifically for pose normalization in re-id. It is thus termed pose-normalization GAN (PN-GAN). Given any person's image and a desirable pose as input, the model will output a synthesized image of the same identity with the original pose replaced with the new one. In practice, we define a set of eight canonical poses, and synthesize eight new images for any given image, resulting in a 8-fold increase in the training data size. The pose-normalized images are used to train a pose-normalized re-id model which produces a set of features that are complementary to the feature learned with the original images. The two sets of feature are thus fused as the final feature.

Contributions. Our contributions are as follows. (1) We identify pose as the chief culprit for preventing a deep re-id model from learning effective identity-sensitive and view-insensitive features, and propose a novel solution based on generating pose-normalized images. This also addresses the scalability and generalizability issues of existing models. (2) A novel person image generation model PN-GAN is proposed to generate pose-normalized images, which are realistic, identity-preserving and pose controllable. With the synthesized images of canonical poses, strong and complementary features are learned to be combined with features learned with the original images. Extensive experiments on several benchmarks show the efficacy of our proposed model. (3) A more realistic unsupervised transfer learning is considered in this paper. Under this setting, no data from the target dataset is used for model updating: the model trained from labeled source domain is applied to the target domain without any modification.

2   Related Work

Deep Re-id Models. Most recently proposed re-id models employ a DNN to learn discriminative view-invariant features [2,9,25,34,40,47,51,63]. They differ in the DNN architectures – some adopt a standard DNN developed for other tasks, whilst others have architectures tailor-made. They differ also in the training objectives. Different models use different training losses including identity classification, pairwise verification, and triplet ranking losses. A comprehensive study on the effectiveness of different losses and their combinations on re-id can be found in [12]. The focus of this paper is not on designing new re-id deep model architecture or loss – we use an off-the-shelf ResNet architecture [16] and the standard identity classification loss. We show that once the pose variation problem is solved, it could help to improve the performance of re-id.

Pose-Guided Deep Re-id. The negative effects of pose variation on deep re-id models have been recognised recently. A number of models [23,39,45,50, 57,58,62] are proposed to address this problem. Most of them are pose-guided based on body part detection. For example, [39,57] first detect normalized part regions from a person image, and then fuse the features extracted from the original images and the part region images. These body part regions are predefined and the region detectors are trained beforehand. Differently, [58] combine region selection and detection with deep re-id in one model. Our model differs significantly from these models in that we synthesize realistic whole-body images using the proposed PN-GAN, rather than only focusing on body parts for pose normalization. Note that body parts are related to semantic attributes which are often specific to different body parts. A number of attributes based re-id models [11,37,44,52] have been proposed. They use attributes to provide additional supervision for learning identity-sensitive features. In contrast, without using the additional attribute information, our PN-GAN is learned as a conditional image generation model for the re-id problem.

Deep Image Generation. Generating realistic images of objects using DNNs has received much interest recently, thanks largely to the development of GAN [15]. GAN is designed to find the optimal discriminator network D between training data and generated samples using a min-max game and simultaneously enhance the performance of an image generator network G. It is formulated to optimize the following objective functions:

$$\min_{G} \max_{D} \mathcal{L}_{GAN} = \mathbb{E}_{x \sim p_{data}(x)} \left[\log D\left(x\right) \right]$$
$$+ \mathbb{E}_{z \sim p_{prior}(z)} \left[\log \left(1 - D\left(G\left(z\right)\right)\right) \right] \quad (1)$$

where $p_{data}\left(x\right)$ and $p_{prior}\left(z\right)$ are the distributions of real data x and Gaussian prior $z \sim \mathcal{N}\left(\mathbf{0}, \mathbf{1}\right)$. The training process iteratively updates the parameters of G and D with the loss functions $\mathcal{L}_D = -\mathcal{L}_{GAN}$ and $\mathcal{L}_G = \mathcal{L}_{GAN}$ for the generator and discriminator respectively. The generator can draw a sample $z \sim p_{prior}\left(z\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{1}\right)$ and utilize the generator network G, i.e., $G(z)$ to generate an image.

Among all the variants of GAN, our pose normalization GAN is built upon deep convolutional generative adversarial networks (DCGANs) [35]. Based on a standard convolutional decoder, DCGAN scales up GAN using Convolutional Neural Networks (CNNs) and it results in stable training across various datasets. Many other variants of GAN, such as VAEGAN [21], Conditional GAN [18], stackGAN [53] also exist. However, most of them are designed for training with high-quality images of objects such as celebrity faces, instead of low-quality surveillance video frames of pedestrians. This problem is tackled in a very recent work [22,30]. Their objective is to synthesize person images in different poses, whilst our work aims to solve the re-id problem with the synthesized images. Besides, both of them utilized two generators/parts from coarse to fine to generate images. As a result, their models are more complicated and not easy to train.

Overall, our model differs from the existing variants of GAN. In particular, built upon the residual blocks, our PN-GAN is learned to change the poses and yet keeps the identity of input person. Note that the only work so far that uses deep image generator for re-id is [65]. However, their model is not a conditional GAN and thus cannot control either identity or pose in the generated person images. As a result, the generated images can only be used as unlabeled or weakly labeled data. In contrast, our model generate strongly labeled data with its ability to preserve the identity and remove the influence of pose variation.
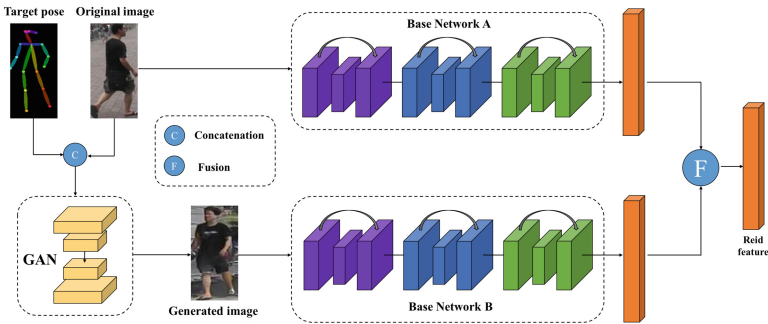

Fig. 2. Overview of our framework. Given an person image, we utilize PN-GAN to synthesize auxiliary images with different poses. Base Networks A and B are then deployed to extract features of original image and synthesized images, respectively. Finally, two types of features are merged for final re-identification task. (Color figure online)

3 Methodology

3.1 Problem Definition and Overview

Problem Definition. Assume we have a training dataset of N persons $\mathcal{D}_{Tr} = \{\mathbf{I}_k, y_k\}_{k=1}^{N}$, where $\mathbf{I}_k$ and y_k are the person image and person id of the k-th

person. In the training stage we learn a feature extraction function ϕ so that a given image $\mathbf{I}$ can be represented by a feature vector $\mathbf{f_I} = \phi(\mathbf{I})$. In the testing stage, given a pair of person images $\{\mathbf{I}_i, \mathbf{I}_j\}$ in the testing dataset $\mathcal{D}_{Te}$, we need to judge whether $y_i = y_j$ or $y_i \neq y_j$. This is done by simply computing the Euclidean distance between $\mathbf{f_{I_i}}$ and $\mathbf{f_{I_j}}$ as the identity-similarity measure.

Framework Overview. As shown in Fig. 2, our framework has two key components, *i.e.*, a GAN based person image generation model (Sect. 3.2) and a person re-id feature learning model (Sect. 3.3).

3.2   Deep Image Generator

Our image generator aims at producing the same person's images under different poses. Particularly, given an input person image $\mathbf{I}_i$ and a desired pose image $\mathbf{I}_{\mathcal{P}_j}$, our image generator aims to synthesize a new person image $\hat{\mathbf{I}}_j$, which contains the same person but with a different pose defined by $\mathbf{I}_{\mathcal{P}_j}$. As in any GAN model, the image generator has two components, a Generator G_P and a Discriminator D_P. The generator is learned to edit the person image conditional on a given pose; the discriminator discriminates real data samples from the generated samples and help to improve the quality of generated images.
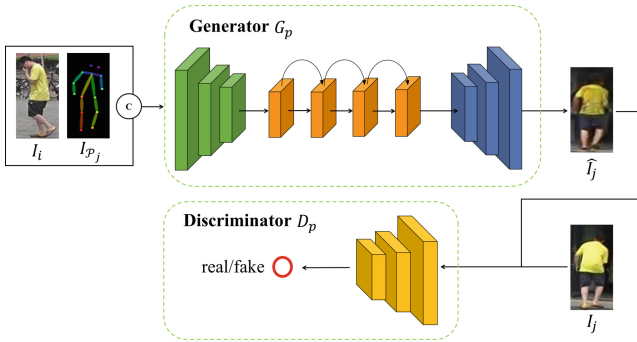

Fig. 3. Schematic of our PN-GAN model

Pose Estimation. The image generation process is conditional on the input image and one factor: the desired pose represented by a skeleton pose image. Pose estimation is obtained by a pretrained off-the-shelf model. More concretely, the off-the-shelf pose detection toolkit – OpenPose [4] is deployed, which is trained without using any re-id benchmark data. Given an input person image $\mathbf{I}_i$, the pose estimator can produce a pose image $\mathbf{I}_{\mathcal{P}_i}$, which localizes and detects 18 anatomical key-points as well as their connections. In the pose images, the orientation of limbs is encoded by color (see Fig. 2, target pose). In theory, any pose from any person image can be used as a condition to control the pose of

another person's generated image. In this work, we focus on pose normalization so we stick to eight canonical poses as shown in Fig. 4(a), to be detailed later.

Generator. As shown in Fig. 3, given an input person image $\mathbf{I}_i$, and a target person image $\mathbf{I}_j$ which contains the same person as $\mathbf{I}_i$ but a different pose $\mathbf{I}_{\mathcal{P}_j}$, our generator will learn to replace pose information in $\mathbf{I}_i$ with the target pose $\mathbf{I}_{\mathcal{P}_j}$ and generate the new pose $\hat{\mathbf{I}}_j$. The input to the generator is the concatenation of the input person image $\mathbf{I}_i$ and target pose image $\mathbf{I}_{\mathcal{P}_j}$. Specifically, we treat the target body pose image $\mathbf{I}_{\mathcal{P}_j}$ as a three-channel image and directly concatenate it with the three-channel source person image as the input of the generator. The generator G_P is designed based on the "ResNet" architecture and is an encoder-decoder network [17]. The encoder-decoder network progressively down-samples $\mathbf{I}_i$ to a bottleneck layer, and then reverse the process to generate $\hat{\mathbf{I}}_j$. The encoder contains 9 ResNet basic blocks[1].

The motivation of designing such a generator is to take advantage of learning residual information in generating new images. The general shape of "ResNet" is learning $y = f(x) + x$ which can be used to pass invariable information from the bottom layers of the encoder to the decoder, and change the variable information of pose. To this end, the other features (e.g., clothing, and the background) will also be reserved and passed to the decoder in order to generate $\hat{\mathbf{I}}_j$. With this architecture (see Fig. 3), we have the best of both worlds: the encoder-decoder network can help learn to extract the semantic information, stored in the bottleneck layer, while the ResNet blocks can pass rich invariable information of person identity to help synthesize more realistic images, and change variable information of poses to realize pose normalization at the same time.

Formally, let $G_P(\cdot)$ be the generator network which is composed of an encoder subnet $G_{Enc}(\cdot)$ and a decoder subnet $G_{Dec}(\cdot)$, the objective of the generator network can be expressed as

$$\mathcal{L}_{G_P} = \mathcal{L}_{GAN} + \lambda_1 \cdot \mathcal{L}_{L_1}, \tag{2}$$

where $\mathcal{L}_{GAN}$ is the loss of the generator in Eq. (1) with the generator $G_P(\cdot)$ and discriminator $D_P(\cdot)$ respectively,

$$\mathcal{L}_{GAN} = \mathbb{E}_{\mathbf{I}_j \sim p_{data}(\mathbf{I}_j)} \{\log D_P(\mathbf{I}_j) \\ + \log(1 - D_P(G_P(\mathbf{I}_i, \mathbf{I}_{\mathcal{P}_j})))\} \tag{3}$$

and $\mathcal{L}_{L_1} = \mathbb{E}_{\mathbf{I}_j \sim p_{data}(\mathbf{I}_j)} \left[\left\| \mathbf{I}_j - \hat{\mathbf{I}}_j \right\|_1 \right]$, and $\hat{\mathbf{I}}_j = G_{Dec}(G_{Enc}(\mathbf{I}_i, \mathbf{I}_{\mathcal{P}_j}))$ is the reconstructed image for $\mathbf{I}_j$ from the input image $\mathbf{I}_i$ with the body pose $\mathbf{I}_{\mathcal{P}_j}$. Here the L_1-norm is used to yield sharper and cleaner images. λ_1 is the weighting coefficient to balance the importance of each term.

Discriminator. The discriminator $D_P(\cdot)$ aims at learning to differentiate the input images as real or fake (i.e., a binary classification task). Given the input

---

[1] Details of structure are in the Supplementary.

image $\mathbf{I}_i$ and target output image $\mathbf{I}_j$, the objective of the discriminator network can be formulated as

$$\mathcal{L}_{D_P} = -\mathcal{L}_{GAN}, \qquad (4)$$

Since our final goal is to obtain the best generator G_P, the optimization step would be to iteratively minimize the loss function $\mathcal{L}_{G_P}$ and $\mathcal{L}_{D_P}$ until convergence. Please refer to the Supplementary Material for the detailed structures and parameters of the generator and discriminator.

3.3 Person Re-id with Pose Normalization

As shown in Fig. 2, we train two re-id models. One model is trained using the original images in a training set to extract identity-invariant features in the presence of pose variation. The other is trained using the synthesized images with normalized poses using our PN-GAN to compute re-id features free of pose variation. They are then fused as the final feature representation.
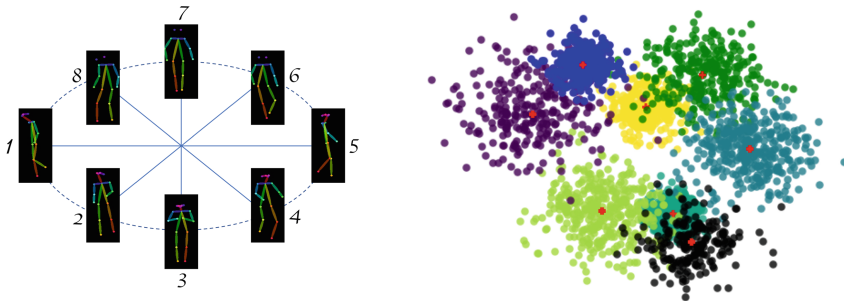
(a) Eight canonical poses on Market-1501 (b) t-SNE visualization of different poses.

Fig. 4. Visualization of canonical poses. Note that red crosses in (b) indicates the canonical pose obtained as the cluster means. (Color figure online)

Pose Normalization. We need to obtain a set of canonical poses, which are representative of the typical viewpoint and body-configurations exhibited by people in public captured by surveillance cameras. To this end, we predict the poses of all training images in a dataset and then group the poses into eight clusters $\{\mathbf{I}_{\mathcal{P}_C}\}_{c=1}^{8}$. We use VGG-19 [5] pre-trained on the ImageNet ILSVRC-2012 dataset to extract the features of each pose images, and K-means algorithm is used to cluster the training pose images into canonical poses. The mean pose images of these clusters are then used as the canonical poses. The eight poses obtained on Market-1501 [61] is shown in Fig. 4(a). With these poses, given each image $\mathbf{I}_i$, our generator will synthesize eight images $\left\{\hat{\mathbf{I}}_{i,\mathcal{P}_C}\right\}_{C=1}^{8}$ by replacing the original pose with these poses.

Re-id Feature with Pose Variation. We train one re-id model with the original training images to extract re-id features with pose variation. The ResNet-50 model [16] is used as the base network. It is pre-trained on the ILSVRC-2012 dataset, and fine-tuned on the training set of a given re-id dataset to classify the training identities. We name this network ResNet-50-A (Base Network A), as shown in Fig. 2. Given an input image $\mathbf{I}_i$, ResNet-50-A produces a feature set $\{\mathbf{f}_{\mathbf{I}_i, layer}\}$, where $layer$ indicates from which layer of the network, the re-id features are extracted. Note that, in most existing deep re-id models, features are computed from the final convolutional layer. Inspired by [29] which shows that layers before the final layer in a DNN often contain useful mid-level identity-sensitive information. We thus merge the $5a$, $5b$ and $5c$ convolutional layers of ResNet-50 structures into a 1024–d feature vector after an FC layer.

Re-id Feature Without Pose Variation. The second model called ResNet-50-B has the same architecture as ResNet-50-A, but performs feature learning using the pose-normalized synthetic images. We thus obtain eight sets of features for the eight poses $\mathbf{f}_{\hat{\mathbf{I}}_i, \mathcal{P}_C} = \left\{ \mathbf{f}_{\hat{\mathbf{I}}_i, \mathcal{P}_C} \right\}_{C=1}^8$.

Testing Stage. Once ResNet-50-A and ResNet-50-B are trained, during testing, for each gallery image, we feed it into ResNet-50-A to obtain one feature vector; as for synthesize eight images of the canonical poses, in consideration of confidence, we feed them into ResNet-50-B to obtain 8 pose-free features and one extra FC layer for the fusion of original feature and each pose feature. This can be done offline. Then given a query image $\mathbf{I}_q$, we do the same to obtain nine feature vectors $\left\{ \mathbf{f}_{\mathbf{I}_q}, \mathbf{f}_{\hat{\mathbf{I}}_{q, \mathcal{P}_C}} \right\}$. Since Maxout and Max-pooling have been widely used in multi-query video re-id, we thus obtain one final feature vector by fusing the nine feature vectors by element-wise maximum operation. We then calculate the Euclidean distance between the final feature vectors of the query and gallery images and use the distance to rank the gallery images.

4   Experiments

4.1   Datasets and Settings

Experiments are carried out on four benchmark datasets:

Market-1501. [61] is collected from 6 different camera views. It has 32,668 bounding boxes of 1,501 identities obtained using a Deformable Part Model (DPM) person detector. Following the standard split [61], we use 751 identities with 12,936 images as training and the rest 750 identities with 19,732 images for testing. The training set is used to train our PN-GAN model.

CUHK03. [25] contains 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 utilize the more realistic yet harder detected person images setting. The training and testing sets consist of 1,367 identities and 100 identities respectively. The testing process is repeated with 20 random splits following [25].

Table 1. Results on Market-1501. '–' indicates not reported. Note that *: on [65], we report the results of using both Basel. + LSRO and Verif.-Identif. + LSRO. Our model only uses the identification loss, so should be compared with Basel. + LSRO which uses the same ResNet-50 base network and the same loss.

Methods	Single-query		Multi-query	
	R-1	mAP	R-1	mAP
TMA [31]	47.90	22.3	–	–
SCSP [6]	51.90	26.40	–	–
DNS [54]	61.02	35.68	71.56	46.03
LSTM Siamese [41]	–	–	61.60	35.31
Gated_Sia [42]	65.88	39.55	76.50	48.50
HP-net [29]	76.90	–	–	–
Spindle [57]	76.90	–	–	–
Basel. + LSRO [65]*	78.06	56.23	85.12	68.52
PIE [62]	79.33	55.95	–	–
Verif.-Identif. [64]	79.51	59.87	85.84	70.33
DLPAR [58]	81.00	63.40	–	–
DeepTransfer [12]	83.70	65.50	89.60	73.80
Verif-Identif. + LSRO [65]*	83.97	66.07	88.42	76.10
PDC [39]	84.14	63.41	–	–
DML [56]	87.7	68.8	–	–
SSM [3]	82.2	68.8	88.2	76.2
JLML [26]	85.10	65.50	89.70	74.50
ResNet-50-A	87.26	69.32	91.81	77.85
Ours (SL)	**89.43**	**72.58**	**92.93**	**80.19**

DukeMTMC-reID. [36] is constructed from the multi-camera tracking dataset – DukeMTMC. It contains 1,812 identities. Following the evaluation protocol [65], 702 identities are used 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.

CUHK01. [24] has 971 identities with 2 images per person captured in two disjoint camera views respectively. As in [24], we use as probe the images of camera A and utilize 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.

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 CUHK01 and CUHK03 datasets. Note that both Spindle [57] and HP-net [29] reported higher results on CUHK03. But their results are obtained using a very different setting: six auxiliary re-id datasets are used and both labeled and detected bounding boxes are used for both training and testing. So their results are not comparable to those in this table.

Method	R-1	R-5	R-10
DeepReid [25]	19.89	50.00	64.00
Imp-Deep [2]	44.96	76.01	83.47
EMD [38]	52.09	82.87	91.78
SI-CI [43]	52.17	84.30	92.30
LSTM Siamese [41]	57.30	80.10	88.30
PIE [62]	67.10	92.20	96.60
Gated_Sia [42]	68.10	88.10	94.60
Basel. + LSRO [65]	73.10	92.70	96.70
DGD [46]	75.30	–	–
OIM [48]	77.50	–	–
PDC [39]	78.92	94.83	97.15
DLPAR[58]	**81.60**	**97.30**	98.40
ResNet-50-A (SL)	76.83	93.79	97.27
Ours (SL)	79.76	96.24	**98.56**
ResNet-50-A (TL)	16.50	38.60	52.84
Ours (TL)	16.85	39.05	53.32

(a) Results on CUHK03

Method	R-1	R-5	R-10
eSDC [59]	19.76	32.72	40.29
kLFDA [49]	32.76	59.01	69.63
mFilter [60]	34.30	55.00	65.30
Imp-Deep [2]	47.53	71.50	80.00
DeepRanking [7]	50.41	75.93	84.07
Ensembles [33]	53.40	76.30	84.40
ImpTrpLoss [10]	53.70	84.30	91.00
GOG [32]	57.80	79.10	86.20
Quadruplet [8]	62.55	83.44	89.71
NullReid [55]	64.98	84.96	89.92
ResNet-50-A (SL)	64.56	83.66	89.74
Ours (SL)	**67.65**	**86.64**	**91.82**
ResNet-50-A (TL)	27.20	48.60	59.20
Ours (TL)	27.58	49.17	59.57

(b) Results on CUHK01

Implementation Details. Our model is implemented on Tensorflow [1] (PN-GAN part) and Caffe [19] (re-id feature learning part) framework. The λ_1 in Eq. (2) is empirically set as 10 in all experiments. We utilize the two-stepped fine-tuning strategy in [13] to fine-tune re-id networks. The input images are resized into 256×128. Adam [20] is used to train both the PN-GAN model and re-id networks with a learning rate of 0.0002, $\beta_1 = 0.5$, a batch size of 32, and a learning rate of 0.00035, $\beta_1 = 0.9$, a batch size of 16, respectively. The dropout ratio is set as 0.5. PN-GAN models and re-id networks are converged in 19 h and 8 h individually on Market-1501 with one NVIDIA 1080Ti GPU card.

Experimental Settings. Experiments are conducted under two settings. The first 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 testing set. The other one is the **Transfer Learning** (TL) setting only for the datasets, CUHK03, CUHK01, and DukeMTMC-reID. Specifically, the re-id model is trained on Market-1501 dataset. We then directly utilize the trained single model to do the testing (*i.e.*, to synthesize images with canonical poses and to extract the nine feature vectors) on the test set of CUHK03, CUHK01, and DukeMTMC-reID. That is, no model updating is done using any data from these three 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 thus tests how generalizable a re-id model is.

Table 3. Results on DukeMTMC-reID.

Methods	R-1	R-10	mAP
LOMO + XQDA [27]	30.80	–	17.00
ResNet50 [16]	65.20	–	45.00
Basel. + LSRO [65]	67.70	–	47.10
AttIDNet [28]	70.69	–	51.88
ResNet-50-A (SL)	72.80	87.90	52.48
Ours (SL)	**73.58**	**88.75**	**53.20**
ResNet-50-A (TL)	27.87	51.12	13.94
Ours (TL)	29.94	51.62	15.77

4.2   Supervised Learning Results

Results on Large-Scale Datasets. Tables 1, 3 and 2(a) compare our model with the best performing alternative models. We can make the following observations:

(1) On all three datasets, the results clearly show that, in the supervised learning settings, our results are improved over those of ResNet-50-A baselines by a clear margin. This validates that the synthetic person images generated by PN-GAN can indeed help the person re-id tasks.
(2) Compared with the existing pose-guided re-id models [39,57,62], our model is clearly better, indicating that synthesizing multiple normalized poses is a more effective way to deal with the large pose variation problem.
(3) Compared with the other re-id model that uses synthesized images for re-id model training [65], our model yields better performance for all datasets, the gap on Market-1501 and DukeMTCM-reID being particularly clear. This is because our model can synthesize images with different poses, which can thus be used for supervised training. In contrast, the synthesized images in [65] do not correspond to any particular person identities or poses, so can only be used as unlabeled or weakly-labeled data.

Results on Small-Scale Dataset. On the smaller dataset – CUHK01, Table 2(b) shows that, again our ResNet-50-A is a pretty strong baseline which can beat almost all the other methods. And by using the normalized pose images generated by PN-GAN, our framework further boosts the performance of ResNet-50-A by more than 3% in the supervised setting. This demonstrates the efficacy of our framework. Note that on the small dataset CUHK01, the handcrafted feature + metric learning based models (e.g., NullReid [55]) are still quite competitive, often beating the more recent deep models. This reveals the limitations of the existing deep models on scalability and generalizability. In particular, previous deep re-id models are pre-trained on some large-scale training datasets, such as CUHK03 and Market-1501. But the models still struggle

Table 4. The Ablation study of Rank-1 and Rank-5 on benchmarks.

Dataset	Market-1501		DukeMTMC-reID		CUHK03		CUHK01	
Methods	R-1	mAP	R-1	mAP	R-1	R-5	R-1	R-5
ResNet-50-A	87.26	69.32	72.80	52.48	76.83	93.79	64.56	83.66
ResNet-50-B	63.75	41.29	26.62	14.30	32.54	55.12	36.18	51.17
Ours	**89.43**	**72.58**	**73.58**	**53.20**	**79.76**	**96.24**	**67.65**	**86.64**

Table 5. The Ablation study of Market-1501 on 1 pose feature and 8 pose features.

Feature(s)	1 pose		8 poses	
Methods	R-1	mAP	R-1	mAP
ResNet-50-A	87.26	69.32	87.26	69.32
ResNet-50-B	58.70	36.69	63.75	41.29
Ours (SL)	87.65	69.60	89.43	72.58

Table 6. The Rank-1/mAP results of ensembling two networks and ours. 'A+B' means training one ResNet-50-A and one ResNet-50-B model.

Methods	ResNet-50-A	ResNet-50-A	ResNet-50-B	Fusion
Ensemble (A+A)	87.26/69.32	87.29/69.36	–	87.38/69.57
Ours (A+B)	87.26/69.32	–	63.75/41.29	89.43/72.58

to fine-tune on the small datasets such as CUHK01 due to the covariate condition differences between them. With the pose normalization, our model is more adaptive to the small datasets and the model pre-trained on only Market-1501 can be easily fine-tuned on the small datasets, achieving much better result than existing models.

4.3 Transfer Learning Results

We report our results obtained under the TL settings on the three datasets – CUHK03, CUHK01, and DukeMTMC-reID in Tables 2(b) and 3 respectively. On CUHK01 dataset, we can achieve 27.58% Rank-1 accuracy in Table 2(b) which is comparable to some models trained under the supervised learning setting, such as eSDC [59]. These results thus show that our model has the potential to be generalizable to a new re-id data from new camera networks – when operating in a 'plug-and-play' mode. Our results are also compared against those of ResNet-50-A (TL) baseline. On all three datasets, we can observe that our model gets improved over those of ResNet-50-A (TL) baseline. Again, this demonstrates that our pose normalized person images can also help the person re-id in the transfer learning settings. Note that due to the intrinsic difficulty of transfer setting, the results are still much lower than those in supervised setting.
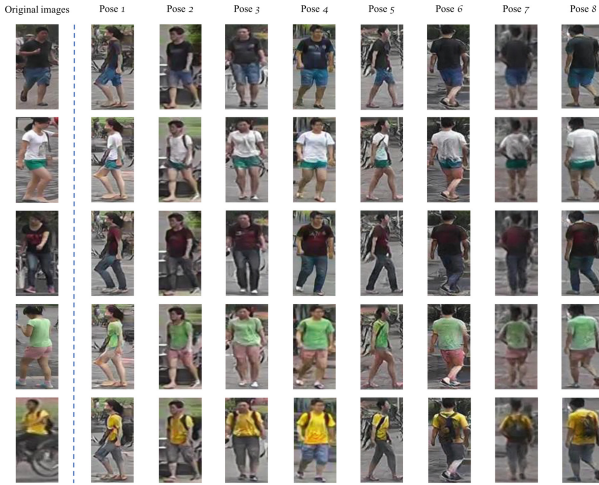
Fig. 5. Visualization of different poses generated by PN-GAN model. (Color figure online)

4.4   Further Evaluations

Ablation Studies. (1) We first evaluate the contributions from the two types of features computed using ResNet-50-A and ResNet-50-B respectively towards the final performance. Table 4 shows that although ResNet-50-B alone performs poorly compared to other methods, when the two types of features are combined, there is an improvement in the final results on all four datasets. This clearly indicates that the two types of features are complementary to each other. (2) In a second study, we compare the result obtained when features are merged with 8 poses and that obtained with only one pose, in Table 5. The result drops from 72.58 to 69.60 on Market-1501 on mAP. This suggests that having eight canonical poses is beneficial – the quality of generated image under one particular pose may be poor; using all eight poses thus reduces the sensitivity to the quality of the generated images for specific poses. (3) In order to prove that the performance gain comes from synthesized images instead of ensembling 2 networks, we conducted experiments on ensembling two ResNet-50-A models. As shown in Table 6, the gain from ensembling two ResNet-50-A is clearly less than that of ensembling one ResNet-50-A and one ResNet-50-B, despite the fact that the ResNet-50-B is much weaker than the second ResNet-50-A. These results thus suggest that our approaches performance gain is not due to ensembling but complementary features extracted from the ResNet-50-B model.

Examples of the Synthesized Images. Figure 5 gives some examples of the synthesized image poses. Given one input image, our image generator can produce realistic images under different poses, while keeping the similar visual appearance as the input person image. We find that, (1) Even though we did

not explicitly use the attributes to guide the PN-GAN, the generated images of different poses have roughly the same visual attributes as the original images. (2) Our model can help alleviate the problems caused by occlusion as shown in the last row of Fig. 5: a man with yellow shirt and grey trousers is blocked by a bicycle, while our image generator can generate synthesized images to keep his key attributes whilst removing the occlusion.

5   Conclusion

We have proposed a novel deep person image generation model by synthesizing pose-normalized person images for re-id. In contrast to previous re-id approaches that try to extract discriminative features which are identity-sensitive but view-insensitive, the proposed method learns complementary features from both original images and pose-normalized synthetic images. Extensive experiments on four benchmarks showed that our model achieves state-of-the-art performance. More importantly, we demonstrated that our model has the potential to be generalized to new re-id datasets collected from new camera networks without any additional data collection and model fine-tuning.

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: OSDI, vol. 16, pp. 265–283 (2016)
2. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
3. Bai, S., Bai, X., Tian, Q.: Scalable person re-identification on supervised smoothed manifold. In: CVPR, vol. 6, p. 7 (2017)
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
5. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: BMVC (2014)
6. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: CVPR (2016)
7. Chen, S.Z., Guo, C.C., Lai, J.H.: Deep ranking for person re-identification via joint representation learning. IEEE TIP **25**, 2353–2367 (2016)
8. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR (2017)

9. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR (2016)
10. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: 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, pp. 1335–1344 (2016)
11. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Learning to recognize pedestrian attribute. arXiv preprint arXiv:1501.00901 (2015)
12. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv:1611.0524 (2016)
13. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244 (2016)
14. Gong, S., Xiang, T.: Person re-identification. In: Visual Analysis of Behaviour, pp. 301–313. Springer, London (2011). https://doi.org/10.1007/978-0-85729-670-2_14
15. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2015)
17. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**, 504–507 (2006)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
19. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. arXiv (2014)
20. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300 (2015)
22. Lassner, C., Pons-Moll, G., Gehler, P.V.: A generative model of people in clothing. In: Proceedings of the IEEE International Conference on Computer Vision, vol. 6 (2017)
23. Li, D., Chen, X., Zhang, Z., Huang, K.: 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, pp. 384–393 (2017)
24. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 31–44. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_3
25. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: CVPR (2014)
26. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In: IJCAI (2017)
27. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
28. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y.: Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220 (2017)
29. Liu, X., et al.: HydraPlus-Net: attentive deep features for pedestrian analysis. In: ICCV (2017)
30. Ma, L., Sun, Q., Jia, X., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NIPS (2017)

31. Martinel, N., Das, A., Micheloni, C., Roy-Chowdhury, A.K.: Temporal model adaptation for person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 858–877. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_52

32. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical Gaussian descriptor for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1363–1372 (2016)

33. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1855 (2015)

34. Qian, X., Fu, Y., Jiang, Y.G., Xiang, T., Xue, X.: Multi-scale deep learning architecture for person re-identification. In: ICCV (2017)

35. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)

36. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2

37. Sarfraz, M.S., Schumann, A., Wang, Y., Stiefelhagen, R.: Deep view-sensitive pedestrian attribute inference in an end-to-end model. arXiv preprint arXiv:1707.06089 (2017)

38. Shi, H., et al.: Embedding deep metric for person re-identification: a study against large variations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 732–748. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_44

39. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)

40. Sun, Y., Zheng, L., Weijian, D., Shengjin, W.: SVDNet for pedestrian retrieval. In: ICCV (2017)

41. Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A Siamese long short-term memory architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 135–153. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_9

42. Varior, R.R., Haloi, M., Wang, G.: Gated Siamese convolutional neural network architecture for human re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 791–808. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_48

43. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: CVPR (2016)

44. Wang, J., Zhu, X., Gong, S., Li, W.: Attribute recognition by joint recurrent learning of context and correlation. In: ICCV (2017)

45. Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q.: GLAD: global-local-alignment descriptor for pedestrian retrieval. arXiv preprint arXiv:1709.04329 (2017)

46. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1249–1258. IEEE (2016)

47. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR (2017)

48. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 3376–3385. IEEE (2017)

49. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_1
50. Yao, H., Zhang, S., Zhang, Y., Li, J., Tian, Q.: Deep representation learning with part loss for person re-identification. arXiv preprint arXiv:1707.00798 (2017)
51. Yu, H.X., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV (2017)
52. Yu, K., Leng, B., Zhang, Z., Li, D., Huang, K.: Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. arXiv preprint arXiv:1611.05603 (2016)
53. Zhang, H., et al.: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
54. Zhang, L., Gong, T.X.S.: Learning a discriminative null space for person re-identification. In: CVPR (2016)
55. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248 (2016)
56. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. arXiv preprint arXiv:1706.00384 (2017)
57. Zhao, H., et al.: 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, pp. 1077–1085 (2017)
58. Zhao, L., Li, X., Wang, J., Zhuang, Y.: Deeply-learned part-aligned representations for person re-identification. In: ICCV (2017)
59. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)
60. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 144–151 (2014)
61. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)
62. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose invariant embedding for deep person re-identification. arXiv preprint arXiv:1701.07732 (2017)
63. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Tian, Q.: Person re-identification in the wild. arXiv preprint arXiv:1604.02531 (2016)
64. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned CNN embedding for person re-identification. arXiv:1611.05666 (2016)
65. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV (2017)