

Pose-Normalized Image Generation for Person Re-identification

Xuelin Qian¹, Yanwei Fu^{1,*}, Wenzuan Wang¹, Tao Xiang², Yang Wu³, Yu-Gang Jiang¹, Xiangyang Xue¹
¹Fudan University; ²Queen Mary University of London;
³Nara Institute of Science and Technology;
*corresponding author, email: yanweifu@fudan.edu.cn

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 pose. The model is based on a generative adversarial network (GAN) and used 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 feature free of the influence of pose variations. We show that this feature is strong on its own and highly complementary to features learned with the original images. Importantly, we now have a model that generalizes to any new re-id dataset without the need for collecting any training data for model fine-tuning, thus making a deep re-id model truly scalable. Extensive experiments on five benchmarks show that our model outperforms the state-of-the-art models, often significantly. In particular, the features learned on Market-1501 can achieve a Rank-1 accuracy of 68.67% on VIPeR without any model fine-tuning, beating almost all existing models fine-tuned on the dataset.

1. Introduction

Person Re-identification (re-id) aims to match a person across multiple non-overlapping camera views [16]. It is a very challenging problem because a person's appearance can change drastically across views, due to the changes in 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 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. For instance, in the bottom row examples in Fig. 1, the big backpacks carried by the three persons are in



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

full display from the back, but reduced to mostly the straps from the front.

Most existing re-id approaches [29, 1, 11, 67, 45, 53, 81, 61] 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 so that the model has 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 human. 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 thus 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; 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 [80] and CUHK03 [29], they still struggle to beat hand-crafted feature based models on smaller datasets such as VIPeR [18], even when they are pretrained on the large 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 attributes such as gender, carrying, clothing style, color, and texture. The latter are the covariates mentioned earlier including pose. 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) and used specifically for pose normalization in re-id, 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 the canonical poses, and synthesize the new images for any given image, resulting in an increase in the training data size. The pose-normalized images are used to train a pose-normalized re-id model which produced a set of features that are complementary to the feature learned with the original images, and thus fused as the final feature representation. Critically, once trained, the model can be applied to a new dataset without any model fine-tuning as long as the test image’s pose is also normal-

ized.

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. Extensively experiments on five benchmarks including Market-1501 [80], CUHK03 [29], DukeMTMC-reID [47], VIPeR [18] and CUHK01 [30] show that our model outperforms the state-of-the-art models, often by big margins. In particular, the features learned on Market-1501 can achieve Rank-1 accuracies of 68.67% and 60.22% on the two small-scale datasets VIPeR and CUHK01 (486 testing ids) respectively without any model fine-tuning, beating almost all existing models finetuned on the two datasets.

2. Related Work

Deep re-id models Most recently proposed re-id models employ a DNN to learn discriminative view-invariant features [29, 1, 11, 67, 45, 53, 81, 61]. They differ in the DNN architectures – some adopt a standard DNN developed for other tasks, whilst others have architectures tailor-made. They also differ 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 [15]. 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 [19] and the standard identity classification loss. We show that once the pose variation problem is solved, such a general-purpose model can achieve the best re-id performance, beating many existing models with more elaborate architectures and losses.

Pose-guided deep re-id The negative effects of pose variation on deep re-id models have been recognised recently. A number of models [51, 79, 75, 74, 28, 64, 60] are proposed to address this problem. Most of them are pose-guided based on body part detection. For example, [51, 74] utilized detect normalized part regions from a person image, and then fused 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, [75] combined 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 defined on different body parts. A number of attributes based re-id models [59, 49, 68, 14] have been proposed. They use attributes to provide additional supervision for learning identity-sensitive features. In this work, attributes are used together with pose as conditions to learn our PN-GAN model as a conditional image generation model, rather than being directly used for learning the re-id model. Similar to pose, a pretrained attribute predictor is used for all re-id datasets without fine-tuning to normalize poses, whilst preserving person identity.

Deep image generation Generating realistic images of objects using DNNs has received much interest recently, thanks largely to the development of GAN [17]. 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:

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

where $p_{data}(x)$ and $p_{prior}(z)$ are the distributions of real data x and Gaussian prior $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. 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 $z \sim p_{prior}(z) = \mathcal{N}(\mathbf{0}, \mathbf{1})$ 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) [46] and InfoGAN [8]. 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 [27], Conditional GAN [24], stackedGAN [69] 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 [38], which also aims to synthesize person images in different poses. Nonetheless, our model differs in several aspects: (1) Inspired by the recent Semi-Latent-GAN [65] and Attribute2Image [63], our image generation architecture learns to clip the visual attributes with the latent representation in the generator of GAN; thus our algorithm can synthesize the new “virtual” person image with not only different poses but also potentially different semantic visual attributes [34]. (2) The discriminator of our GAN architecture is trained to both discriminate real/fake

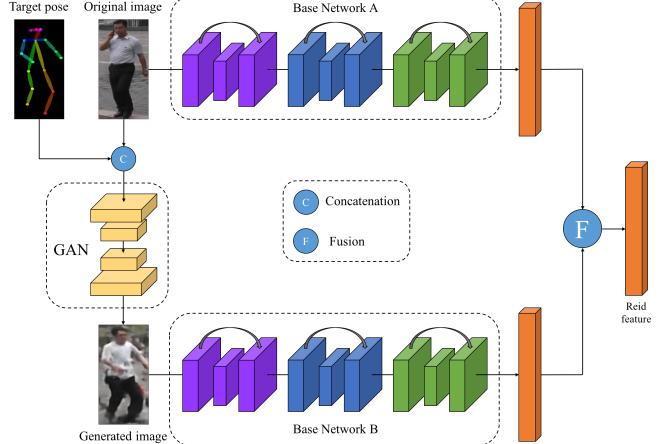


Figure 2. Overview of our framework.

images and predict the attribute vector. This design is inspired by InfoGAN [8], and leads to better visual quality. Note that the only work so far that uses deep image generator for re-id is [83]. Nevertheless, 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.

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 $\mathbf{f}_\mathbf{I} = \phi(\mathbf{I})$. 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 $\mathbf{f}_{\mathbf{I}_i}$ and $\mathbf{f}_{\mathbf{I}_j}$ 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 (Sec. 3.2) and a person re-id feature learning model (Sec. 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}_{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}_{P_j} . As in any GAN model, the image generator has two

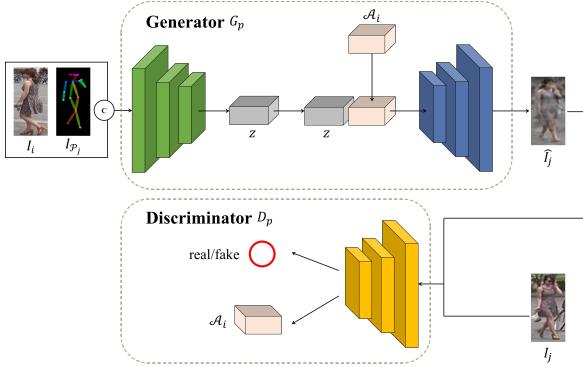


Figure 3. Schematic of our PN-GAN model.

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.

Pose and attribute estimation. The image generation process is conditional on two groups of factors: the identity-sensitive ones represented by a semantic attribute vector and the desired pose represented by a skeleton pose image. Both are obtained by pretrained off-the-shelf predictors. We denote the attribute vector of an image \mathbf{I}_i as \mathcal{A}_i [20]. The predictor is trained on the attribute labels of Market-1501 dataset training split, annotated by [34]. Once trained, it is applied to all re-id benchmarks without fine-tuning. For pose estimation, the off-the-shelf pose detection toolkit – OpenPose [3] 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}_{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.

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}_{P_j} , the generator is trained to generate an image $\hat{\mathbf{I}}_j$ that is as similar as possible to \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}_{P_j} . Specifically, we treat the target body pose image \mathbf{I}_{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 “U-net” architecture [48] and is an encoder-decoder network [21]. 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 bottleneck layer is directly clipped with the attribute vector \mathcal{A}_i which is shared with \mathbf{I}_j . However, the attribute vector only captures identity-sensitive information;

but in our person image generation task, in order to loyally generate the target image, there is a significant amount of identity-insensitive low-level information (*e.g.*, anything that is not covered by the attribute vector, and the background and lighting condition) that needs to be passed to the decoder to be used to generate $\hat{\mathbf{I}}_j$, which could otherwise be lost in the bottleneck layer. Thus the general shape of “U-net” with skip-connection¹ is introduced to pass low-level information from the bottom layers of encoder to the corresponding layers in the decoder. With this architecture (see Fig. 3), we have the best of both world: the encoder-decoder network can help learn to extract the middle-level semantic information, stored in the bottleneck layer, while the skip-connection in U-net can pass rich low-level information to help synthesize more realistic images.

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

$$\mathcal{L}_{G_P} = \mathcal{L}_{GAN} + \lambda_1 \cdot \mathcal{L}_{L_1} + \lambda_2 \cdot \mathcal{L}_{G-attrb}, \quad (2)$$

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.

Discriminator. The discriminator $D_P(\cdot)$ aims at learning to differentiate the input images is real or fake (*i.e.* a binary classification task), as well as predicting the attribute vector (*i.e.*, a multi-label classification task). To efficiently learn the latter task, we add a fully connected layer before the attribute prediction layer. Given the input image \mathbf{I}_i and target output image \mathbf{I}_j , the objective of the discriminator network can be formulated as

$$\mathcal{L}_{D_P} = -\mathcal{L}_{GAN} + \lambda_2 \cdot \mathcal{L}_{D-attrb}, \quad (3)$$

where $\mathcal{L}_{D-attrb}$ is the expected attribute loss; and $\hat{\mathcal{A}}_j = D_P(\tilde{\mathbf{I}}_j)$, $\tilde{\mathbf{I}}_j \in \{\hat{\mathbf{I}}_j, \mathbf{I}_j\}$. 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

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.

¹such a skip-connection is also used and shown good performance in [24].

Pose Normalization. First, 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 clusters with the mean pose images as the canonical poses. With these poses, given each image, our generator will synthesize the images by replacing the original pose with these poses.

Re-id Features We train one re-id model with the original training images. We name this network ResNet-50-A [19]. Given an input image \mathbf{I}_i , ResNet-50-A produce a feature set. The second model ResNet-50-B has the same structure as ResNet-50-A, but performs feature learning using the pose-normalized synthetic images. We thus obtain the sets of features for the poses.

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; and synthesize the images of the poses, feed them into ResNet-50-B to obtain the pose-free features. This can be done offline. Then given a query image \mathbf{I}_q , we do the same to obtain the feature vectors. We thus obtain one final feature vector by fusing the 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.

Dataset. Six benchmark datasets are used, including three large-scale datasets Market-1501 [80], CUHK03 [29] and DukeMTMC-reID [47], and three small-scale datasets VIPeR [18], CUHK01 [30] and GRID [37]. **Market-1501.** It is collected from 6 different view cameras. It has 32,668 bounding boxes of 1,501 identities obtained using a Deformable Part Model (DPM) person detector. Following the standards split [80], 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.** It 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, validation and testing sets consist of 1,367 identities, 100 identities and 100 identities respectively. The testing process is repeated with 20 random splits following [29]. **DukeMTMC-reID.** Constructed from the multi-camera tracking dataset DukeMTMC, DukeMTMC-reID contains 1,812 identities. Following the evaluation protocol [83], 702 identities are used as the training set and the remaining 1,110 identities as the testing set. During test-

ing, one query image for each identity in each camera is used for query and the remaining as the gallery set. **VIPeR.** As a widely used small dataset for re-id, the VIPeR dataset contains 632 pedestrian image pairs taken from two camera views and each person has only one image per camera view. Following the standard test protocol, we split the dataset into half, 316 identities for training and the rest for testing. **CUHK01.** It has 971 identities with 2 images captured in two disjoint camera views per person. As in [30], 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. **GRID.** The QMUL underGround Re-IDentification (GRID) dataset contains 250 pedestrian image pairs. Each pair contains two images of the same individual seen from different camera views. As in [31], the training/test was split to 125/125 with 775 distractor people included in the gallery.

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.

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

4.2. Transfer Learning Results

We report our results obtained under the TL settings on the VIPeR, CUHK01, GRID, CUHK03, and DukeMTMC-reID dataset in Table 4, Table 5, Table 6, Table 3, and Table 2 respectively. In this setting, we only use the training data from Market-1501 dataset to train the models, and testing on the test split of each individual dataset.

On VIPeR, Tab. 4 shows that our model (the row “Our (TL)”) achieves 68.67% Rank-1 accuracy which is still 12% higher than the second best results DeepTransfer [15] which are fine-tuned on the training set of VIPeR. On CUHK01

Methods	Single-Query		Multi-Query	
	R-1	mAP	R-1	mAP
TMA [40]	47.90	22.3	—	—
SCSP [4]	51.90	26.40	—	—
DNS [70]	61.02	35.68	71.56	46.03
LSTM Siamese [57]	—	—	61.60	35.31
Gated_Sia [56]	65.88	39.55	76.50	48.50
HP-net [36]	76.90	—	—	—
Spindle [74]	76.90	—	—	—
Basel. + LSRO [83]*	78.06	56.23	85.12	68.52
PIE [79]	79.33	55.95	—	—
Verif.-Identif. [82]	79.51	59.87	85.84	70.33
DLPAR[75]	81.00	63.40	—	—
DeepTransfer [15]	83.70	65.50	89.60	73.80
Verif-Identif. + LSRO[83]*	83.97	66.07	88.42	76.10
PDC [51]	84.14	63.41	—	—
JLML [31]	85.10	65.50	89.70	74.50
Ours	95.52	89.94	95.90	91.37

Table 1. Results on Market-1501. ‘-’ indicates not reported. Note that *: on [83], 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	R-1	R-5	R-10	mAP
LOMO+XQDA[33]	30.80	—	—	17.00
ResNet50 [19]	65.20	—	—	45.00
Basel. +LSRO [83]	67.70	—	—	47.10
AttIDNet [34]	70.69	—	—	51.88
SVDNet [53]	76.70	86.40	89.90	56.80
DPFL [9]	79.20	—	—	60.60
Ours(SL)	91.47	96.68	98.07	81.39
Ours(TL)	61.72	78.46	85.37	53.02

Table 2. Results on DukeMTMC-reID.

dataset, we can achieve 60.22% Rank-1 accuracy in Table 5 which is comparable to many models trained under the supervised learning setting. These results thus show that our model has the potential to be truly generalizable to a new reid data from new camera networks – when operating in a ‘plug-and-play’ mode, it already beats most existing models trained on the target dataset.

4.3. Supervised Learning Results

Results on large-scale datasets. Tables 1, 2 and 3 compare our model with the best performing alternative models. We can make the following observations: (1) On all three datasets, our model achieves the highest results. The improvement on Market-1501 is particularly striking - Under single-query setting, our model obtains 95.52% Rank-1 accuracy which is 10% higher than the second best result [31]; our mAP is 89.94% which is around 23% higher than the second best reported result of [83]. (2) Compared with

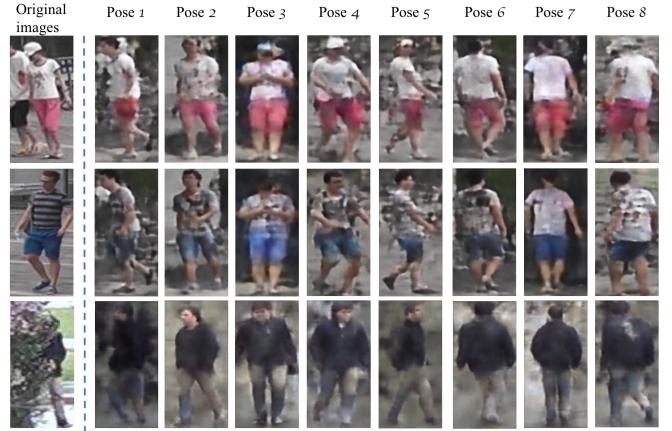


Figure 4. The visualization of the generated pose images.

Method	R-1	R-5	R-10
DeepReid [29]	19.89	50.00	64.00
Imp-Deep [1]	44.96	76.01	83.47
EMD [50]	52.09	82.87	91.78
SI-CI [58]	52.17	84.30	92.30
LSTM Siamese [57]	57.30	80.10	88.30
PIE [79]	67.10	92.20	96.60
Gated_Sia [56]	68.10	88.10	94.60
Basel. + LSRO [83]	73.10	92.70	96.70
PDC [51]	78.92	94.83	97.15
DLPAR[75]	81.60	97.30	98.40
Verif-Identif. + LSRO [83]	84.60	97.60	98.90
Ours(SL)	92.66	99.84	100
Ours(TL)	73.54	98.70	100

Table 3. Results on CUHK03 (Detected). Note that both Spindle [74] and HP-net [36] 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.

the existing pose-guided re-id models [75, 74, 79, 51], 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 only other re-id model that uses synthesized images for re-id model training [83], our model yields better performance for all datasets, the gap on Market-1501 and DukeMTMC-reID being particularly big (over 20% on mAP). This is because our model can synthesize images with controllable identity and pose, which can thus be used for supervised training. In contrast, the synthesized images in [83] do not correspond to any particular person identities or poses, so can only be used as unlabeled or weakly-labeled data. In adjunct to the work [72] and on Market-1501 dataset, our results (R-1: 95.52%) are comparing against the human performance evaluation (see Tab. 7 in [72]) which is 93.5%. In

Method	R-1	R-5	R-10
RPLM [22]	27.00	55.30	69.00
MtMCML [39]	28.83	59.34	75.82
Mid-Filter [77]	29.11	52.34	65.95
kCCA[35]	30.16	62.69	76.04
LADF [32]	30.22	64.70	78.92
WARCA [25]	37.50	70.80	82.00
XQDA [33]	40.00	68.13	80.51
DNS [70]	42.30	71.50	82.0
TMA [40]	43.80	—	83.90
SCSP [4]	53.50	82.60	91.50
Imp-Deep [1]	34.81	63.61	75.63
SI-CI [58]	35.76	67.40	83.50
G-Dropout [54]	37.70	—	—
Gated_Sia [56]	37.90	66.90	76.30
Deep Ranking [6]	38.40	69.20	81.30
EMD [50]	40.91	67.41	79.11
DLPAR[75]	48.70	74.70	85.10
PDC [51]	51.27	74.05	84.18
SSM [80]	53.73	—	91.49
Spindle [74]	53.80	74.10	83.20
PIE [79]	54.49	84.43	92.18
DeepTransfer [15]	56.30	—	—
Ours(SL)	78.17	96.20	99.05
Ours(TL)	68.67	93.99	99.05

Table 4. Results on the VIPeR dataset. SL and TL indicates the supervised learning and transfer learning settings respectively.

Method	R-1	R-5	R-10
ITML[13]	15.98	35.22	45.60
eSDC [76]	19.76	32.72	40.29
kLFDA [62]	32.76	59.01	69.63
mFilter [78]	34.30	55.00	65.30
Imp-Deep [1]	47.53	71.50	80.00
DeepRanking [6]	50.41	75.93	84.07
Ensembles [44]	53.40	76.30	84.40
ImpTrpLoss [12]	53.70	84.30	91.00
GOG [42]	57.80	79.10	86.20
Quadruplet [7]	62.55	83.44	89.71
NullReid [71]	64.98	84.96	89.92
G-Dropout [54]	71.70	88.60	92.60
Ours(SL)	86.22	96.60	98.31
Ours(TL)	60.22	86.93	95.06

Table 5. Results of CUHK01 dataset.

other words, our results on Market-1501 are 2% higher than human performance.

Results on small-scale datasets. On the three smaller datasets, Tables 4, 5, and 6 show that, again our model produces the best performance and the gap between ours and the second best becomes even bigger. For example, on VIPeR, our model achieves a 78.17% Rank-1 accuracy which is around 21% higher than the second best result re-

Method	R-1	R-5	R-10
KEPLER[41]	18.40	39.12	50.24
LSSCDL [73]	20.40	—	51.28
DR-KISS [55]	20.60	59.01	51.40
SSDAL+XQDA [52]	22.40	39.20	48.00
NLML [23]	24.54	35.86	43.53
SCSP [5]	24.24	44.56	54.08
GOG+XQDA [43]	24.80	—	58.40
SSM [2]	27.20	—	61.12
HIPHOP+CRAFT [10]	22.40	49.90	61.80
HIPHOP+LOMO+CRAFT [10]	26.00	50.60	62.50
JLML [31]	37.50	61.40	69.40
Ours(TL)	59.14	95.91	99.14

Table 6. Results of GRID dataset.

ported in [15]. Note that on the smallest dataset VIPeR, the handcrafted feature + metric learning based models (*e.g.*, SCSP [4]) 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, most deep re-id models such as [74, 79] are pre-trained on some large-scale training datasets, such as CUHK03 and Market-1501. But the models still struggle to fine-tune on the small datasets such as VIPeR due to the covariate condition differences between them. With the pose normalization, our model is more adaptive to the small datasets and the model pretrained only on Market-1501 can be easily fine-tuned on the small datasets, achieving much better result than existing models.

4.4. Further Evaluations

Examples of the synthesized images. Figure 4, 5, and 6 give some examples of the synthesized image poses. Given one input image, our image generator can produce realistic images under different poses, while keeping the same attributes as the input person image. Interestingly, our model can help alleviate the problems caused by occlusion as shown in the third row of Fig. 4: a man with black jacket and grey trousers is blocked by a tree, while our image generator can generate synthesized images to keep his key attributes (*i.e.*, black jacket and grey trousers) without occlusion. More examples can be found in the Supplementary Material, where we show that our PN-GAN generates magical image quality.

Qualitative comparisons of GANs. We also provide qualitative results to compare our PN-GAN with a number of alternative deep image generation models. We also evaluate the contributions of two important design choices in the architecture of PN-GAN, namely the use of the skip-connection and the concatenation with an attribute vector in the bottleneck layer condition. Specifically, the following models are compared:

VAE [26]. A coding scheme based on Variational Auto-Encoder (VAE) is applied to the output of the encoder Z . Under this encoding scheme, Z is not a deterministic output for a given input image, but a random vector conditioned on the input image. All other layers are the same as in our PN-GAN model for a fair comparison.

SL-GAN [66]. SL-GAN is a recently proposed variant of conditional GAN. It was originally designed to automatically generate and modify hairstyles in face images. Similar to our PN-GAN, SL-GAN uses the attribute vector of the input image as a condition and supervision signal for training the discriminator to predict attributes for the generated image. By using the target pose image as the input to the encoder, SL-GAN is re-purposed here to perform pose-normalized person image synthesis.

PN-GAN without skip-connections. This is a variant of our PN-GAN. The only modification we make is to remove the skip-connections between the encoder and the decoder.

PN-GAN without attributes. In this variant of PN-GAN, we remove the concatenation operation with the input image attribute vector in the bottleneck layer. Without this operation, the PN-GAN becomes similar to the model proposed in [24].

The results in Figure 7 show that: (1) Comparing to the two alternative image generators, the images synthesized by our PN-GAN are sharper and preserve some fine-grained visual attributes more loyally, *e.g.*, the carrying condition and other details such as hair style and shoe-wear are better preserved in each synthesized image. Importantly, comparing with the canonical poses, the synthesized poses are more accurate. (2) When the attribute vector is not used as an condition for the decoder, the synthesised poses are still accurate but some fine-grained characteristics are lost. (3) When the skip-connections are removed, we can see a clear degradation of the image quality. This indicates that it is critical to pass some the low-level information, such as colour and texture, directly from the encoder to the decoder; this information would otherwise be lost in the bottleneck layer.

5. Conclusion

We 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 discriminator 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 five benchmarks showed that our model achieves the best performance. More importantly, we demonstrated that our model can be generalized to new re-id datasets collected from new camera networks without any additional data collection and

model fine-tuning.

References

- [1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2, 4.3, 4.3, 4.3
- [2] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. *arXiv preprint arXiv:1703.08359*, 2017. 4.3
- [3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3.2
- [4] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 4.3, 4.3, 4.3
- [5] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1268–1277, 2016. 4.3
- [6] S.-Z. Chen, C.-C. Guo, and J.-H. Lai. Deep ranking for person re-identification via joint representation learning. 2016. 4.3, 4.3
- [7] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. 2017. 4.3
- [8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016. 2
- [9] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2590–2600, 2017. 4.3
- [10] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 4.3
- [11] D. Cheng, Y. Gong, S. Zhou, JinjunWang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, 2016. 1, 2
- [12] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 4.3
- [13] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007. 4.3
- [14] Y. Deng, P. Luo, C. C. Loy, and X. Tang. Learning to recognize pedestrian attribute. *arXiv preprint arXiv:1501.00901*, 2015. 2
- [15] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. In *arXiv:1611.0524*, 2016. 2, 4.2, 4.3, 4.3, 4.3



Figure 5. The visualization of generated images on Market-1501 and DukeMTMC-reID.



Figure 6. The visualization of generated images on VIPeR and CUHK01. Note that the generator model didn't fine-tune on these small datasets.

- [16] S. Gong and T. Xiang. Person re-identification. In *Visual Analysis of Behaviour*, pages 301–313. Springer, 2011. [1](#)
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. [2](#)
- [18] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *IEEE PETS Workshop*, 2007. [1](#), [4.1](#)
- [19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2015. [2](#), [3.3](#), [4.3](#)
- [20] K. He, Zhanxiong Wang, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue. Adaptively weighted multi-task deep network for person attribute classification. In *ACM MM*, 2017. [3.2](#)
- [21] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. 2006. [3.2](#)
- [22] M. Hirzer, P. M. Roth, M. Kostinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, 2012. [4.3](#)
- [23] S. Huang, J. Lu, J. Zhou, and A. K. Jain. Nonlinear local metric learning for person re-identification. *arXiv preprint arXiv:1511.05169*, 2015. [4.3](#)
- [24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. 2017. [2](#), [1](#), [4.4](#)
- [25] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. In *ECCV*, 2016. [4.3](#)
- [26] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [4.4](#)
- [27] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. [2](#)
- [28] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on*



Figure 7. Comparing images with several canonical poses synthesised using our PN-GAN and a number of alternative models.

Computer Vision and Pattern Recognition, pages 384–393, 2017. 2

[29] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 1, 2, 4.1, 4.3

[30] W. Li, R. Zhao, and X. Wang. Human re-identification with transferred metric learning. In *ACCV*, 2012. 1, 4.1

[31] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *IJCAI*, 2017. 4.1, 4.3, 4.3, 4.3

[32] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *ECCV*, 2014. 4.3

[33] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 4.3, 4.3

[34] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017. 2, 3.2, 4.3

[35] G. Lisanti, I. Masi, and A. D. Bimbo. Matching people across camera views using kernel canonical correlation analysis. In *ICDSC*, 2014. 4.3

[36] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. 2017. 4.3, 3

[37] C. C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1988–1995. IEEE, 2009. 4.1

[38] L. Ma, Q. Sun, X. Jia, B. Schiele, T. Tuytelaars, and L. V. Gool. Pose guided person image generation. In *NIPS*, 2017. 2

[39] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. In *IEEE TIP*, 2014. 4.3

[40] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person reidentification. In *ECCV*, 2016. 4.3, 4.3

[41] N. Martinel, C. Micheloni, and G. L. Foresti. Kernelized saliency-based person re-identification through multiple metric learning. *IEEE Transactions on Image Processing*, 24(12):5645–5658, 2015. 4.3

[42] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Pro-*

- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016. 4.3
- [43] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1363–1372, 2016. 4.3
- [44] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1855, 2015. 4.3
- [45] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-scale deep learning architecture for person re-identification. 2017. 1, 2
- [46] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. 2016. 2
- [47] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop on Benchmarking Multi-Target Tracking*, 2016. 1, 4.1
- [48] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 3.2
- [49] M. S. Sarfraz, A. Schumann, Y. Wang, and R. Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017. 2
- [50] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei1, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 4.3, 4.3
- [51] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. 2017. 2, 4.3, 4.3, 4.3, 4.3
- [52] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian. Deep attributes driven multi-camera person re-identification. In *European Conference on Computer Vision*, pages 475–491. Springer, 2016. 4.3
- [53] Y. Sun, L. Zheng, D. Weijian, and W. Shengjin. Svdnet for pedestrian retrieval. 2017. 1, 2, 4.3
- [54] X. T. W.Ouyang, H. Li, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 4.3, 4.3
- [55] D. Tao, Y. Guo, M. Song, Y. Li, Z. Yu, and Y. Y. Tang. Person re-identification by dual-regularized kiss metric learning. *IEEE Transactions on Image Processing*, 25(6):2726–2738, 2016. 4.3
- [56] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 4.3, 4.3, 4.3
- [57] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 4.3, 4.3
- [58] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 4.3, 4.3
- [59] J. Wang, X. Zhu, S. Gong, and W. Li. Attribute recognition by joint recurrent learning of context and correlation. 2017. 2
- [60] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. *arXiv preprint arXiv:1709.04329*, 2017. 2
- [61] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. 2017. 1, 2
- [62] F. Xiong, M. Gou, O. Camps, and M. Sznajer. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014. 4.3
- [63] X. Yan, J. Yang, K. Sohn, and H. Lee. attribute2image: conditional image generation from visual attributes. In *ECCV*, 2016. 2
- [64] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*, 2017. 2
- [65] W. Yin, Y. Fu, L. Sigal, and X. Xue. Semi-latent gan: Learning to generate and modify facial images from attributes. In *arxiv*, 2017. 2
- [66] W. Yin, Y. Fu, L. Sigal, and X. Xue. Semi-latent gan: Learning to generate and modify facial images from attributes. *arXiv preprint arXiv:1704.02166*, 2017. 4.4
- [67] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. 2017. 1, 2
- [68] K. Yu, B. Leng, Z. Zhang, D. Li, and K. Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *arXiv preprint arXiv:1611.05603*, 2016. 2
- [69] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. 2017. 2
- [70] L. Zhang and T. X. S. Gong. Learning a discriminative null space for person re-identificatio. In *CVPR*, 2016. 4.3, 4.3
- [71] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016. 4.3
- [72] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification. In *arXiv:1711.0818*, 2017. 4.3
- [73] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2016. 4.3
- [74] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 2, 4.3, 4.3, 3, 4.3
- [75] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. 2017. 2, 4.3, 4.3, 4.3, 4.3
- [76] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, 2013. 4.3

- [77] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, 2014. 4.3
- [78] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151, 2014. 4.3
- [79] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017. 2, 4.3, 4.3, 4.3, 4.3
- [80] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 4.1, 4.3
- [81] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. *arXiv preprint arXiv:1604.02531*, 2016. 1, 2
- [82] Z. Zheng, L. Zheng, , and Y. Yang. A discriminatively learned cnn embedding for person re-identification. In *arXiv:1611.05666*, 2016. 4.3
- [83] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *ICCV*, 2017. 2, 4.1, 4.3, 1, 4.3, 4.3, 4.3, 4.3