Pose Transferrable Person Re-Identification

Jinxian Liu[1*] Bingbing Ni[1*†] Yichao Yan[1*] Peng Zhou[1] Shuo Cheng[1] Jianguo Hu[2]
[1]Shanghai Key Laboratory of Digital Media Processing and Transmission,
Shanghai Institute for Advanced Communication and Data Science,
Shanghai Jiao Tong University, Shanghai 200240, China
[2]Minivision
{liujinxian, nibingbing, yanyichao, zhoupengcv}@sjtu.edu.cn
acccheng94@gmail.com hujianguo@minivision.cn

Abstract

*Person re-identification (ReID) is an important task in the field of intelligent security. A key challenge is how to capture human pose variations, while existing benchmarks (*i.e.*, Market1501, DukeMTMC-reID, CUHK03,* etc.*) do NOT provide sufficient pose coverage to train a robust ReID system. To address this issue, we propose a **pose-transferrable** person ReID framework which utilizes pose-transferred sample augmentations (*i.e.*, with ID supervision) to enhance ReID model training. On one hand, novel training samples with rich pose variations are generated via transferring pose instances from MARS dataset, and they are added into the target dataset to facilitate robust training. On the other hand, in addition to the conventional discriminator of GAN (*i.e.*, to distinguish between REAL/FAKE samples), we propose a novel guider sub-network which encourages the generated sample (*i.e.*, with novel pose) towards better satisfying the ReID loss (*i.e.*, cross-entropy ReID loss, triplet ReID loss). In the meantime, an alternative optimization procedure is proposed to train the proposed Generator-Guider-Discriminator network. Experimental results on Market-1501, DukeMTMC-reID and CUHK03 show that our method achieves great performance improvement, and outperforms most state-of-the-art methods without elaborate designing the ReID model.*

1. Introduction

Person re-identification (ReID) aims to match pedestrians across no-overlapping video cameras. It is a challenging task due to large variations in person pose, appearance, illumination, occlusion, *etc*. In recent years, this task has attracted a lot of attention for its great application potential in smart video surveillance.

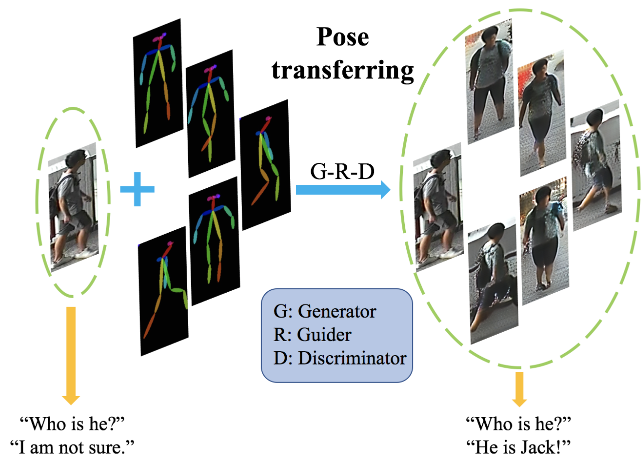


Figure 1. The motivation of proposed framework. Novel training samples with rich pose variations are generated via transferring pose instances from MARS dataset by the proposed pose-transfer module (*i.e.*, Generator-Guider-Discriminator) with ID information, which are utilized to enhance ReID model learning.

Pose variation is one of the key factors that prevent us from learning a robust ReID model. Existing benchmarks (*e.g.*, Market1501 [46], DukeMTMC-reID [51], CUHK03 [18]) only contain a limited number of pose changes, and therefore the learned ReID model is very easily over-fitted to certain poses. Many works have been proposed to address the discriminative and robust training issue for ReID. A large group of works focus on feature learning [19, 44, 42, 17] for more discriminative representation, and many other methods focus on metric learning [34, 19, 4, 2, 6].

Recently, deep convolutional networks (DCNN) based methods have been extensively studied, including learning deep feature [17, 42] and metric [2, 54] to cope with the "large-number-of-class and few-sample" problem in ReID

*Equal contribution
†Corresponding author

task, as well as design of novel discriminative loss functions suited for ReID task such as contrastive loss [49] and triplet loss [21]. While feature learning and discriminative metric could partially alleviate the sample insufficiency issue, as a rule of thumb, increasing the coverage of training samples (*i.e.*, to including more human samples with different poses) is the only way to essentially enhance the model. Unsupervised learning is explored in [43, 38] to enhance ReID model training. However, the promotion of performance achieved by unsupervised learning is quite limited since the unsupervised model is unable to learn discriminate features. Based on the recent success of GAN [26], Zheng *et al.* [51] propose to generate a large number of unlabeled human samples and assign a uniform label distribution to the generated unlabeled image for ReID feature learning. However, these unsupervised sample generation methods suffer from inherent issues: 1) unsupervised samples do not bring sufficient discriminative information for model training; 2) due to large complexity of human shape, directly applying traditional GAN (*e.g.*, DCGAN [26]) would only yield seriously distorted human samples; and 3) last but not least, previous GAN model only attempts to generate *visually* preferable samples, but it does not target at better discriminative power in ReID, which drastically limits the usage of the generated samples.

To explicitly address these issues, this work proposes a novel pose-transferrable framework for supervised sample augmentation/generation for discriminative ReID model training in the case of single shot. Our motivation is shown in Figure 1. First, observing that MARS [45] dataset has rich variations of human poses, our proposed generative scheme extracts poses (*i.e.*, skeletons) from it, and pairs them with human appearance instances from existing ReID datasets (*i.e.*, with identity labels), and then generates supervised sample augmentation in new poses based on a novel variant of GAN models. Second, in contrast to previous GAN models which only consider whether the generated samples look realistic or not, our proposed generative network employs a *guider* module which is endowed with a ReID cross-entropy loss or triplet loss. The guider subnetwork works collaboratively with conventional discriminator to jointly pursuit good visual quality as well as great discriminative power for ReID. In addition, to balance the contribution between real samples and generated samples during training, we use a label smoothness scheme for cross entropy training and adjust triplet sampling strategy and margin for model training. Extensive experimental results show that our method can enhance the representation capability and discrimination of the learned ReID model. In the meantime, the proposed idea is general, which can be easily extended to other tasks which require human sample augmentation with pose variation, such as pedestrian detection and pedestrian tracking, *etc*.

2. Related Work

Generative Adversarial Networks [10] have been widely studied recently. On one hand, many works explore to improve model structure and optimization form [26, 1, 25, 28]. Their works make GAN model to generate more realistic samples or be more easily optimized. On the other hand, a lot of works focus on interesting applications of GAN. In this area, many similar methods propose to achieve conditional image generation. Conditional GAN (cGAN) [24] introduce a conditional version of GAN. Image-to-Image [15, 56] is one of the most interesting and meaningful research directions based on cGAN, which focus on image style transfer. Some derivative works like skeleton-to-image [33, 37] are proposed. The skeleton-to-image architecture takes an image and a skeleton as input and outputs a sample with the pose that is the same as input skeleton. Motivated by these works, we propose to generate labeled samples with multiple poses to improve the performance of ReID model. And a pre-trained ReID model is used to guide the training of GAN and make the generated samples more adapted to person ReID task.

Person re-identification attracts great attention due to its important application values. Most of the existing works focus on two ways, which are robust feature learning and distance metric learning. Before deep learning gets popular, there are many works explore to design handcrafted features [11, 23, 43] that are robust to changes in person pose and image condition. And there are also many works make efforts to utilize robust distance metric like Mahalanobis distance function, KISSME metric learning [16], *etc*. Recently, DCNN is widely used in the filed of ReID [44, 19, 36, 4, 6, 55, 48]. A large number of researchers design various DCNN structures to learn powerful features. Zhao *et al* [42] propose a novel DCNN named as Spindle Net to fuse whole body feature and body region feature, and Li *et al* [17] design a Multi-Scale Context-Aware Network to extract small visual cues that may be very useful to distinguish the pedestrian pairs. Some researchers combine DCNN with metric learning, and they propose various forms of metrics to guide the training of DCNN [2, 54].

However, designing deep network structure and distance metric easily result in overfitting. GAN is used to generate samples with the varied background to enhance ReID model in [5], however, various human poses are not considered. Zheng *et al* [51] propose to use generated unlabeled samples to improve performance, however, the serious distortion and unlabeled samples limit its improvement. In our work, we propose a pose-transferrable architecture to generate labeled samples with rich pose variations. Without designing complex models or distance metrics, our method can achieve great performance improvement. As a kind of data augmentation method, our work can combine with most methods and further enhance its performance.

3. Methodology

3.1. Motivation and Overview

Key to enhance ReID model learning is to provide sufficient training data which can cover a wide range of human pose variations. Very recently, few works [51] have attempted to utilize generative adversarial training schemes for data augmentation. However, as mentioned above, these methods have inherent limitations: 1) the generated samples carry no identity information (*i.e.*, unsupervised), which leads to ONLY marginal improvements for discriminative ReID model training; and 2) directly applying GAN ONLY considers whether the generated samples look *realistic* or NOT, which does not have any link to ReID performance.

To tackle these issues, we propose a pose-transferrable ReID architecture, as shown in Figure 2. Our framework contains two components. First, inspired by recent skeleton-to-image method [33, 37], we transfer a large variation of poses (*i.e.*, skeletons) from *pose-rich* datasets such as MARS onto the labeled human instances in ReID benchmarks, therefore numerous **labeled** data are generated . Second, we propose a *guider* sub-network, which is paired with conventional discriminator in GAN, so as to directly encourage discriminative power boosting (*i.e.*, cross-entropy loss or triplet ReID loss). The details of our proposed method are given as follows.

3.2. Pose Transfer Module: Generator-Guider-Discriminator

3.2.1 Skeleton-to-Image Generation

In the following, we introduce in detail how to transfer poses (skeletons) from a source dataset (*i.e.*, which is considered to have large pose coverage) onto a target ReID dataset (*i.e.*, Market1501, DukeMTMC-reID, CUHK03). The source dataset we select in this work is the MARS dataset [45], which is a large video-based ReID dataset and has rich pose variations. For each human sample, we can collect corresponding skeleton representation by applying the pose detection algorithm proposed in [3]. Each pose is represented by an RGB image $\mathbf{s}$. Note that we directly use the pre-trained skeleton detector provided in [3] without any model re-training since it provides a robust and accurate general skeleton detection performance. To transfer a pose onto a static human image (*i.e.*, appearance) to form a new posed human sample, our method is mainly motivated by conditional GAN (cGAN) [24] as well as the previous skeleton-to-image works [33, 37]. In particular, training the skeleton-to-image network requires triplet data: an appearance image $\mathbf{x}$ of one person, a skeleton image $\mathbf{s}$, and the ground-truth image $\mathbf{y}$ that endows the human $\mathbf{x}$ with the corresponding pose $\mathbf{s}$. In testing, the generator G maps a paired input human image $\mathbf{x}$ and a new pose $\mathbf{s}$ to a new image $\hat{\mathbf{y}}$ corresponding to the new pose $\mathbf{s}$, via the mapping function $\hat{\mathbf{y}} = G(\mathbf{x}, \mathbf{s}, \mathbf{z})$. Here $\mathbf{z}$ denotes a random noise which we do not explicitly use in this work as in [15]. During training, the discriminator D try its best to classify the real triplet $(\mathbf{x}, \mathbf{s}, \mathbf{y})$ from generated triplet $(\mathbf{x}, \mathbf{s}, \hat{\mathbf{y}})$, while generator tries its best to fool the discriminator. In practice, the real triplet and generated triplet are stacked in the dimension of channel respectively and then sent into discriminator as in [15]. In the meantime, as in [22], we also include a ℓ_1 loss to enhance the quality of the generated image, *i.e.*, to minimize the reconstruction error in addition to the adversarial loss. Therefore, the combined value function of training skeleton-to-image generation network, denoted as $\mathcal{L}_c(G, D)$, could be expressed as:

$$\begin{aligned}\mathcal{L}_c(G, D) = \mathbb{E}[\log D(\mathbf{x}, \mathbf{s}, \mathbf{y})] \\ + \mathbb{E}[\log(1 - D(\mathbf{x}, \mathbf{s}, G(\mathbf{x}, \mathbf{s}, \mathbf{z})))],\end{aligned} \quad (1)$$

$$\mathcal{L}_{\ell_1}(G) = \mathbb{E}[\|\mathbf{y} - G(\mathbf{x}, \mathbf{s}, \mathbf{z})\|_1]. \quad (2)$$

For clarity, we use $\mathbb{E}$ to denote $\mathbb{E}_{\mathbf{x},\mathbf{s},\mathbf{y} \sim p_{data}(\mathbf{x},\mathbf{s},\mathbf{y})}$. Here G, D denote the generator and discriminator networks, respectively, which will be explained in detail later. And notice that the generated sample $\hat{\mathbf{y}}$ share the same appearance as $\mathbf{y}$, so the generated sample is **labeled**.

3.2.2 Guider Module: ReID Boosting

However, the above skeleton-to-image model ONLY considers that the generated human sample should visually look realistic, but how it helps enhance the ReID model training is NOT guaranteed. Namely, we DO NOT ask that the generated human sample should look like a human, but we DO ask whether training with these augmented/generated samples could boost the person ReID performance. To this end, we propose a novel *guider* module in addition to the conventional generator and discriminator, to guide the trained generative model more adapted to ReID problem, *i.e.*, to boost the discriminative power. In other words, the generated image $\hat{\mathbf{y}}$ goes through both discriminator D and guider R during model training. The guider R is a sub-network which distinguishes classes (*i.e.*, cross-entropy loss) or enforces intra-class samples closer and inter-classes farther (i.e, triplet loss).To this end, the guider R is pre-trained on the target ReID dataset with supervision and **fixed** (*i.e.*, to keep the discrimination ability) during joint G-R-D training. In particular, the guider could utilize supervision information to force the identity of $\hat{\mathbf{y}}$ to approach $\mathbf{y}$ via two types of loss function design: 1) cross-entropy loss and 2)triplet loss, which are elaborated as follows.

Cross-Entropy based Guider Loss. Denote by $\mathcal{L}_{R_{ce}}(G)$ a cross-entropy ReID loss. And the training objective of $\mathcal{L}_{R_{ce}}(G)$ can thus be expressed as:

$$\mathcal{L}_{R_{ce}}(G) = \mathbb{E}[-\sum q_a \log p_R(G(\mathbf{x_a}, \mathbf{s}, \mathbf{z}))], \quad (3)$$

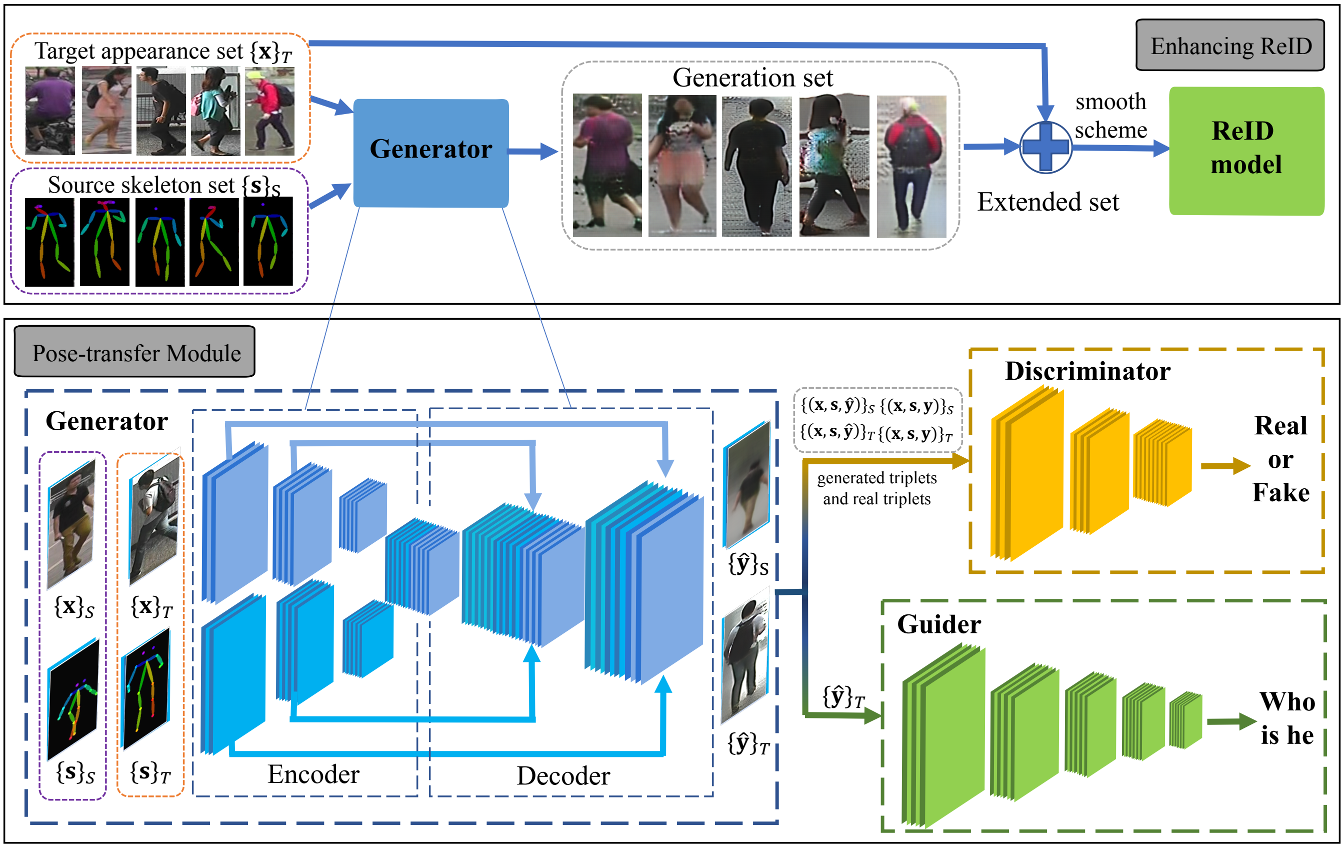


Figure 2. Framework of our pose-transferrable person ReID. The poses from source set (*i.e.*, MARS) are transferred to target set and form the generation set, which is combined with the target set to enhance the ReID model. The pose-transfer module (*i.e.*, G-R-D) is utilized to boost skeleton-to-image generation. The $\{\mathbf{x}\}_T$ denote appearances from target set, the $\{\mathbf{s}\}_S$ denote skeletons from source set, and so on.

where q_a denotes the label of class a and $\mathbf{x_a}$ denotes the appearance of class a. And p_R denotes the output probability distribution of the guider R.

Triplet based Guider Loss. In the case of triplet loss for the guider, triplets are chosen following the strategy in [29]. For every image $\mathbf{x_a}$ of class a, we denote by $\hat{\mathbf{y}}_\mathbf{a} = G(\mathbf{x_a}, \mathbf{s}, \mathbf{z})$ an anchor. Positive anchor $\mathbf{r_a}$ is chosen from the real images of class a (*i.e.*, realistic examples to force the generator to form images closer to real examples in class a). Negative anchor $\hat{\mathbf{y}}_\mathbf{b} = G(\mathbf{x_b}, \mathbf{s}, \mathbf{z})$ is sampled from generated examples of other class b, which encourages diversity. Given a constructed triplet set $\mathcal{T} = \{\hat{\mathbf{y}}_\mathbf{a}, \mathbf{r_a}, \hat{\mathbf{y}}_\mathbf{b}\}_{i=1}^m$ in this way, the guider training loss could be thus expressed as:

$$\mathcal{L}_{R_{tri}}(G) = \mathbb{E}[\sum_{\hat{\mathbf{y}}_\mathbf{a}, \mathbf{r_a}, \hat{\mathbf{y}}_\mathbf{b}, a \neq b} [\alpha + d_{\hat{\mathbf{y}}_\mathbf{a}, \mathbf{r_a}} - d_{\hat{\mathbf{y}}_\mathbf{a}, \hat{\mathbf{y}}_\mathbf{b}}]_+], \quad (4)$$

where $d_{i,j}$ is ℓ_2 distance between the output feature $R(i)$ and $R(j)$ of ReID model, and α denotes the margin.

As a summary, the whole human sample generation network contains three component, *i.e.*, a generator, a discriminator and a guider and the integrated loss function to train this generation network is a weighted sum of all losses mentioned above:

$$\mathcal{L}(G, D) = \mathcal{L}_c(G, D) + \lambda \mathcal{L}_{\ell_1}(G) + \beta \mathcal{L}_R(G), \quad (5)$$

where λ and β are the weighting factors for the two loss terms. The $\mathcal{L}_R(G)$ is a cross-entropy loss $\mathcal{L}_{R_{ce}}(G)$ or a triplet loss $\mathcal{L}_{R_{tri}}(G)$. During training of skeleton-to-image model, the guider conveys discriminative identity information and propagates this supervision signal from the guider to the generator, thus to form a human sample which are more readily to be classified into the correct person class. We therefore regard our ReID model objective as a **Identity Oriented Generation Model**, in contrast into previous **Appearance Oriented Generation Model**. The optimization target is thus defined as:

$$G^* = \arg \min_G \max_D \mathcal{L}(G, D). \quad (6)$$

To transfer poses from source dataset onto target ReID dataset, during training we sample as many paired skeleton and the appearance instances as the target dataset in MARS to form auxiliary training data, and these data only pass through the discriminator D, but are NOT sent into the guider R as they do not possess label information. These auxiliary training data is combined with labeled appearance and skeleton instances in the target ReID dataset for joint model training. During testing, we pair the skeleton from the source dataset (e.g, MARS) and appearance from the target dataset, so the generated samples share the same identity as samples from target dataset and with a pose from source dataset. Through this way, variations of poses in source dataset (*i.e.*, MARS) could be FULLY explored and transferred towards the target datasets to form pose-rich data augmentations as well as discriminative model training. We present some generated examples with and without the guider respectively in Figure 3 based on ResNet-50 and cross-entropy loss. We note that the generated images are significantly shaper and realistic if there is a guider.

3.3. Training with Balanced Data

We add generated samples into the target train set and train the ReID model with this extended train set. For every person instance, we now have samples with a larger variation of poses, therefore the trained ReID model will gain more representation capability in terms of human pose variation. However, even with such carefully designed generator, to generate human samples as realistic as true images in the target dataset is still difficult. Therefore, we cannot regard a generated example of class k equally *trustable* as a real example of class k during the process of ReID model training. In face, our experiments show that the performance of the trained ReID model will be negatively affected if too many generated samples are utilized. To alleviate this problem, we use a soft labeling scheme for the generated samples, instead of assigning each generated sample a definite (hard) class label of person.

In particular, in the case of cross-entropy loss, we use the label smoothing regularization (LSR) that is rediscovered by Szegedy *et al* [31] to re-weight the generated samples. The LSR label distribution can be formulated as:

$$q_{LSR}(k) = \begin{cases} \frac{\varepsilon}{K} & k \neq y \\ 1 - \varepsilon + \frac{\varepsilon}{K} & k = y \end{cases}, \tag{7}$$

where $k \in \{1, 2, ..., K\}$ denotes the pre-defined classes of the training data, and K is the number of classes, y is the ground truth. $\varepsilon \in [0, 1]$ denotes how confident we are with the ground truth, which is a hyper-parameter that can be adjusted according to the quality of generated images. ε is set to 0 for real images. The cross-entropy loss with $q_{LSR}(k)$ is easily derived as:

$$\mathcal{L}_{LSR} = -(1 - \varepsilon) \log (p(y)) - \frac{\varepsilon}{K} \sum_{k=1}^{K} \log (p(k)). \tag{8}$$

In the case of triplet loss, a similar idea is to adjust the margin α when the triplet contains generated images. Also, the strategy of choosing triplet could be improved as follows. After adding some generated images into the target dataset, we conduct the triplet $\{\mathbf{x_a}, \mathbf{x_p}, \mathbf{x_n}\}$ construction based on the strategy used in [29]. In a batch, every sample will be an anchor. Positive anchor and negative anchor are chosen from real images when the anchor is a real image. When an anchor is a generated image, positive anchor and negative anchor are chosen from the whole augmented train set, and we then reduce the margin α for these triplets. This sampling method can prevent us from giving the generated images too much weight and affecting the performance of the trained ReID model.

Figure 3. Examples of generated images. This three lines of samples are from Market1501, DukeMTMC-reID and CUHK03 (labeled) respectively.

3.4. Network Architecture

The generator and discriminator of our method are the same as [37], where the generator has a siamese and U-net structure and discriminator has a simple stacked structure. The input of the generator is an image and a skeleton, and the skeleton is extracted by the real-time human pose estimator [3]. The output of the generator is not only sent into a discriminator but also sent into the guider except when the inputs are from MARS. In our work, we adopt ResNet-50 [12] and Densenet-169 [14] as the backbone of the ReID model respectively.

4. Experiments

To verify the effectiveness of our method, we carry out experiments on three public person ReID datasets, including Market1501 [46], DukeMTMC-reID [51], and CUHK03 [18]. We transfer pose instances to generate samples from MARS dataset, because it is a video-based

Methods	Market-1501		DukeMTMC-reID		CUHK03 (labeled)		CUHK03 (detected)	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
Basel (R) [47]	73.90	47.78	65.22	44.99	22.2	21.0	21.3	19.7
Basel (R)+LSRO [51]	78.06	56.23	67.68	47.13	-	-	-	-
Pose-transfer (R)	**79.75**	**57.98**	**68.64**	**48.06**	**33.8**	**30.5**	**30.1**	**28.2**

Table 1. Comparison of the proposed method with baseline and the unsupervised sample generation method on the three benchmarks. Rank-1(%) and mAP(%) are shown. R: ResNet-50.

dataset, and contains rich pose variations. The experimental results suggest that the proposed architecture can significantly improve the performance of both feature learning and metric learning. In addition, based on DenseNet-169 which is in the same order of magnitude as the parameter of ResNet-50 and triplet loss [29], our method outperforms most existing methods.

4.1. Datasets and Evaluation Protocols

Market1501 is an image-based ReID dataset. It consists of 12,936 images for training, and each person has 17.2 images on average in the train set. **DukeMTMC-reID** is a subset of the DukeMTMC [27] for image-based ReID. Its train set contains 16522 images of 702 identities. The evaluation protocol is the same as that of [51]. **CUHK03** contains 14,096 images of 1,467 identities which are captured from two cameras in the CUHK campus. We use the new training and testing protocol for CUHK03 [52]. The train set of these three datasets only have limited pose variations, which limits its performance, therefore, experiments on these three datasets can effectively verify the effectiveness of our method. **MARS** is a large video-based person ReID dataset. It consists of 20,478 tracklets and 1,191,003 bounding boxes of 1,261 identities and contains rich pose variations. We transfer various poses from MARS to the three datasets mentioned above to enhance the ReID model.

We use two evaluation metrics to evaluate the performance of our ReID algorithm, *i.e*., Rank-1 identification rate and mean Average Precision (mAP). In all our experiments, we use the single query mode.

4.2. Implementation Details

Our method is divided into three steps. First, we train a ReID model on the target dataset (*i.e*., Market1501, DukeMTMC-reID, CUHK03). Second, we use the pretrained model, which we call it a guider, to guide the training of the skeleton-to-image generator. The input pairs (skeletons and appearances) are from the training set of target dataset and MARS. Third, we combine the appearances in target training set and skeleton sampled from MARS, and use them to generate a large number of labeled pose-varied samples to enhance the ReID model.

Pre-training a ReID model to be the guider. For Market1501 and DukeMTMC-reID, we train the ResNet-50 with cross-entropy loss that is the same as [47]. The learning rate is set to 0.001 and decay to 0.0001 after 30 epochs. We train the DenseNet-169 with cross-entropy loss and triplet loss respectively. In the case of cross-entropy loss, the learning rate is set to 0.01 and decay to 0.001 after 25 epochs. Dropout is not used in this case. In the case of triplet loss, we sample triplet $\{\mathbf{x_a}, \mathbf{x_p}, \mathbf{x_n}\}$ following the strategy proposed in [29]. And we take the online sampling methods. Due to memory limitations, the batch size is set to 128. The learning rate is set to 0.001, and the margin α is set to 1. We use stochastic gradient descent with momentum 0.9 to optimize the ReID model mentioned above. For CUHK03, we train the ResNet-50 with cross-entropy loss and triplet loss respectively, *i.e*., the DenseNet-169 model is not used because its performance is similar to the ResNet-50 model on this dataset. The settings are the same as those in the experiment of training R-CE and D-Tri on Market1501 and DukeMTMC-reID. We use **R-CE** and **D-Tri** to denote ResNet-50 with cross-entropy loss and DenseNet-169 with triplet loss respectively, and so on.

In all above experiments, input images are resized to 256×256 and randomly cropped to 224×224 with random horizontal flipping during training. When testing, the input images are resized to 256×256 and center cropped to 224×224. No other data augmentation methods are taken.

Boosting training of skeleton-to-image model. The inputs (*i.e*., appearances and skeletons) of the generator are resized to 256×256 and rescaled into $[-1, 1]$, which are from the training set of target dataset and MARS. The outputs of the generator are sent into the discriminator and guider. The exception to this is that the outputs of the generator are only sent into the discriminator when inputs come from MARS. The parameters of the guider are fixed when training G-R-D module. We train the generator and discriminator with Adam optimizer, and the learning rate is set to 0.0002. In all our experiments, the λ and β are set to 10.0 and 1.0 respectively.

Improving ReID model with generated samples. For every image in target dataset, we sample two skeletons in MARS, then we use the generator to generate two samples with poses transferred from MARS. We analyze the effects of the number of added samples on performance in Section 4.5. When we train the ReID model with cross-entropy loss, ε is set to 0.4 for generated samples in all our experiments. The margin α is set to 0.5 for triplets that contain generated samples. We also analyze the influence of the

Methods	Market-1501		Duke-R	
	rank-1	mAP	rank-1	mAP
BoW+kissme [46]	44.42	20.76	25.13	12.17
LOMO+XQDA [19]	-	-	30.75	17.04
FisherNet [35]	48.15	29.94	-	-
Null Space [40]	55.43	29.87	-	-
Gated SCNN [32]	65.88	39.55	-	-
Basel (R)* [47]	73.90	47.78	65.22	44.99
ReRank [52]	77.11	63.63	-	-
Basel (R)+LSRO [51]	78.06	56.23	67.68	47.13
Verif + Identif* [49]	79.51	59.87	68.9	49.3
PAN* [50]	82.81	63.35	71.59	51,55
Transfer* [9]	83.7	65.5	-	-
APR [20]	84.29	64.67	70.69	51.88
SVDNet [30]	82.3	62.1	76.7	56.8
DPFL [7]	-	-	79.2	60.6
TriNet* [13]	84.92	69.14	-	-
DML* [41]	87.73	68.83	-	-
SVDNet+REDA* [53]	87.08	71.13	79.31	62.44
Pose-transfer (R)	79.75	57.98	68.64	48.06
Basel (D)	84.47	64.17	73.92	50.79
Pose-transfer (D)	85.52	65.33	75.17	52.25
Basel (D, Tri)	86.73	67.78	77.03	55.34
Pose-transfer (D, Tri)	87.65	68.92	78.52	56.91

Table 2. Comparison of the proposed method with the state-of-the-art on Market1501 and DukeMTMC-reID. Rank-1 (%) and mAP (%) are shown. Duke-R denotes DukeMTMC-reID. R: ResNet-50. D: DenseNet-169. * denotes unpublished paper. Tri denotes that the model is trained with triplet loss [29]. The best, second and third results are highlighted in green, red, blue respectively. Best viewed in colors.

two parameters in Section 4.5. Other settings are the same as those in the experiments of pre-training ReID model.

4.3. Comparison Results and Discussions

Comparison with baseline. We compare our method based on R-CE with the ResNet-50 baseline [47] on all three benchmarks mentioned above. The results are shown in Table 1. We can observe significant performance improvements over the baseline. Especially, our method greatly improves the performance on CUHK due to fact that the pose variations are extremely limited on it. It verifies that our pose-transferrable framework greatly alleviates the sample insufficiency issue.

Comparison with unsupervised sample generation method. To verify the pose-transferred samples that are more effective than unlabeled generated samples, we compare our methods with the work [51] on the Market1501 and DukeMTMC-reID, and the results are shown in Table 1. The Rank-1 rises by 1.69% and 0.96% correspondingly, and the mAP rises by 1.75% and 0.93% correspondingly. Notice that the number of added samples in our work is almost the same as [51]. So the labeled generations with various poses enhance the ReID system more compared with unlabeled generations.

Methods	Labeled		Detected	
	rank-1	mAP	rank-1	mAP
BoW+XQDA [46]	7.9	7.3	6.4	6.4
PUL* [8]	-	-	9.1	9.2
LOMO+XQDA [19]	14.8	13.6	12.8	11.5
Basel (R)* [47]	22.2	21.0	21.3	19.7
Basel (R)+DaF* [39]	27.5	31.5.	26.4	30.0
Basel (R)+XQ+Re [52]	38.1	40.3	34.7	37.4
PAN* [50]	36.9	35.0	36.3	34.0
DPFL [7]	43.0	40.5	40.7	37.0
SVDNet [30]	40.9	37.8	41.5	37.3
TriNet+REDA* [53]	58.1	53.8	55.5	50.7
Pose-Transfer (R)	33.8	30.5	30.1	28.2
Basel (R, Tri)	42.8	39.2	39.1	36.6
Pose-Transfer (R, Tri)	45.1	42.0	41.6	38.7

Table 3. Comparison of the proposed method with the state-of-the-art on CUHK03. Rank-1 (%) and mAP (%) are shown. R: ResNet-50. * denotes unpublished paper. Tri denotes that the model is trained with triplet loss. The best, second and third results are highlighted in green, red, blue respectively. Best viewed in colors.

Comparison with state-of-the-arts. We compare our work with state-of-the-art methods on the three benchmarks. Based on the DenseNet-169 and triplet loss, our method outperforms most previous methods. The results on Market1501 and DukeMTMC-reID are shown in Table 2, the results on CUHK03 are shown in Table 3. Some methods [7, 13, 41] achieve great performance by elaborately designing network structures or loss functions. An effective augmentation method is proposed in [53], which greatly improves the performance. Above all, our methods can combine with all these methods and further improve their performance.

4.4. Component Analysis

Effectiveness of the guider. In this part, we analyze the effectiveness of the guider. We present some generated samples of the three datasets in Figure 3, which are trained with or without the guider of R-CE respectively. We observe that the generated samples with the guider are more realistic and shaper than the ones without the guider. To further verify this point, we add the two kinds of samples into target dataset and train ReID model with them respectively. The ε for generated samples without guider decreases to 0.2 for getting best results in this case. The experimental results are summarized in Table 4. With the guider, the generated samples achieve better performance.

Analysis of the different form of the guider. We compare the guider of R-CE and D-Tri on Market1501, which have different network structures and loss functions. Some

Methods	Market-1501		DukeMTMC-reID		CUHK03 (labeled)		CUHK03 (detected)	
	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP
No Guider	76.93	54.22	66.70	45.98	28.1	26.0	25.2	23.9
With Guider	**79.75**	**57.98**	**68.64**	**48.06**	**33.8**	**30.5**	**30.1**	**28.2**

Table 4. Quantifying the effectiveness of guider with the ReID evaluation protocols.

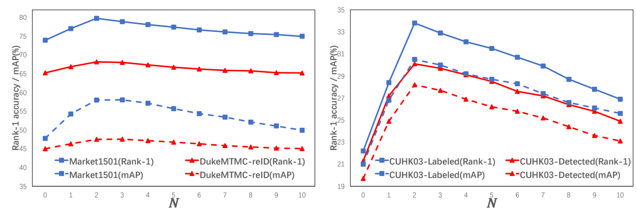
Figure 4. Examples of generated images from Market1501 with two different forms of guider.

generated samples with the two form guiders respectively are present in Figure 4. Intuitively, the visual quality of generated samples has no difference with the two forms of the guider. In addition, we conduct a cross experiment. We use the generated samples with one form of the guider to improve a ReID model in another form. The Rank-1 and the mAP are both down by about 0.5% on Market1501. It shows that the performance will be a little lower if the forms of the guider and the ReID model to be improved are different, but the decrease is not significant. Therefore, we can use a guider of simple form, and then utilize the generated samples to enhance ReID model of complex form. But to get better performance, it is best to keep the form of the guider and the ReID model to be consistent.

4.5. Parameter Analysis

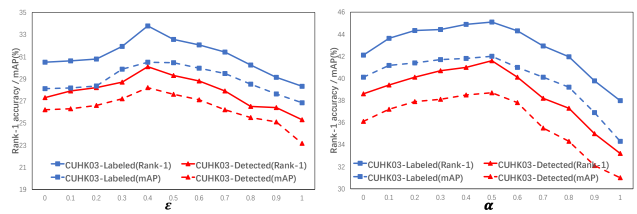
Analysis of No. of pose-transferred samples N. We analyze how the numbers of generated samples for every image in target dataset affects the performance of ReID model. We use the ResNet-50 trained with cross-entropy loss as the guider and ReID model to be improved. For every image, we test the impact of 1 to 10 pose-transferred samples on performance separately. ε is constantly adjusted to get best results with the change of N. We observe that 2 samples for every image get the best performance. And as the number of extended samples increases further, the performance decreases slightly. Experimental results on the three datasets are shown in Figure 5.

Analysis of the ε **and** α **hyper-parameters**. ε and α are two other parameters that affect the ReID performance, which are used to smooth cross-entropy training and triplet training with generated samples respectively. We analyze



(a) (b)

Figure 5. The impact of the N on the ReID performance. (a): Market1501 and DukeMTMC-reID. (b): CUHK03 labeled and detected.



(a) (b)

Figure 6. The impact of the ε and α on the ReID performance. (a): The impact of ε. (b): The impact of α.

the impact of the ε and α on CUHK03, and the results are shown in Figure 6(a) and Figure 6(b) respectively. We observe that our method achieves the best performance when ε is set to 0.4 and α is set to 0.5 for generated samples. Notice that setting ε to 0 and α to 1 for generated samples (i.e, equate the generated sample with the real one) limits the performance improvement.

5. Conclusion

In this work, we proposed a pose-transferrable person re-identification framework which transfers various pose instances from one dataset to another. The generated samples with transferred poses increase the richness of pose variations in target dataset and greatly enhance the ReID model. As a special kind of data augmentation method, our work can be utilized to enhance both feature learning and metric learning based methods.

6. Acknowledgements

This work was supported by National Science Foundation of China (U1611461, 61502301, 61671298, 61521062). The work was partially supported by China's Thousand Youth Talents Plan, State Key Research and Development Program (2016YFB1001003) and 18DZ2270700.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[2] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *CVPR*, pages 1571–1580, 2017.

[3] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 1302–1310, 2017.

[4] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268–1277, 2016.

[5] L. Chen, H. Yang, S. Wu, and Z. Gao. Data generation for improving person re-identification. In *ACM MM*, pages 609–617, 2017.

[6] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, pages 1320–1329, 2017.

[7] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCV workshop*, Oct 2017.

[8] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *CoRR*, abs/1705.10444, 2017.

[9] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *CoRR*, abs/1611.05244, 2016.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[11] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *CoRR*, abs/1703.07737, 2017.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017.

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017.

[16] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[17] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 7398–7407, 2017.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[20] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *CoRR*, abs/1703.07220, 2017.

[21] J. Liu, Z. Zha, Q. I. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei. Multi-scale triplet CNN for person re-identification. In *ACM MM*, pages 192–196, 2016.

[22] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015.

[23] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012.

[24] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[25] G. Qi. Loss-sensitive generative adversarial networks on lipschitz densities. *CoRR*, abs/1701.06264, 2017.

[26] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[27] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV workshop*, pages 17–35, 2016.

[28] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *NIPS*, pages 2226–2234, 2016.

[29] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[30] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, Oct 2017.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[32] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016.

[33] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, pages 3560–3569, 2017.

[34] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009.

[35] L. Wu, C. Shen, and A. van den Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65:238–250, 2017.

[36] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016.

[37] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang. Skeleton-aided articulated motion generation. In *ACM MM*, pages 199–207, 2017.

[38] Y. Yang, L. Wen, S. Lyu, and S. Z. Li. Unsupervised learning of multi-level descriptors for person re-identification. In *AAAI*, pages 4306–4312, 2017.

[39] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. *CoRR*, abs/1708.04169, 2017.

[40] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016.

[41] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *CoRR*, abs/1706.00384, 2017.

[42] 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 *CVPR*, pages 907–915, 2017.

[43] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[44] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151, 2014.

[45] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[46] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[47] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *CoRR*, abs/1610.02984, 2016.

[48] W. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *TPAMI*, 38(3):591–606, 2016.

[49] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned CNN embedding for person re-identification. *CoRR*, abs/1611.05666, 2016.

[50] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. *CoRR*, abs/1707.00408, 2017.

[51] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, Oct 2017.

[52] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 3652–3661, 2017.

[53] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *CoRR*, abs/1708.04896, 2017.

[54] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, pages 5028–5037, 2017.

[55] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785, 2017.

[56] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, Oct 2017.