Pose Variation Adaptation for Person Re-identification

Lei Zhang[1†], Na Jiang[2†], Yue Xu[1], Qishuai Diao[1], Zhong Zhou[1*] and Wei Wu[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]Information Engineering College, Capital Normal University, Beijing, China
Email: {zhangleilei0125, xuyuevr, diaoqishuai, zz, wuwei}@buaa.edu.cn, jiangna@cnu.edu.cn

Abstract—Person re-identification (reid) aims at matching pedestrians observed from non-overlapping camera views. It has important applications in surveillance video analysis such as human retrieval, human tracking and activity analysis. Although a large number of effective feature learning and distance metric optimizing approaches have been proposed, it still suffers from pedestrian appearance variations caused by pose changing. Most of the previous methods address this problem by learning a pose-invariant descriptor subspace. In this paper, we propose a pose variation adaptation method for person reid. It can reduce the probability of deep learning network over-fitting. Specifically, we introduce a pose transfer generative adversarial network with a similarity measurement module. With the learned pose transfer model, training images can be transferred to any given poses, and with the original images, forming an augmented training dataset. It increases data diversity against over-fitting. In contrast to previous GAN-based methods, we consider the influence of pose variations on similarity measures to generate shaper and more realistic samples for person reid. Besides, we optimize hard example mining to introduce a novel manner of samples used with the learned pose transfer model. It focuses on the inferior samples which are caused by pose variations to increase the number of effective hard examples for learning discriminative features and improving the generalization ability. We extensively conduct comparative evaluations to demonstrate the advantages and superiorities of the proposed method over the state-of-the-art person reid approaches on Market-1501 and DukeMTMC-reID.

I. INTRODUCTION

Person re-identification (reid) is a retrieval task that aims to recognize and identify a pedestrian across multiple camera views at different times [1]. In recent years, this task has attracted increasing attention for its broad application potential in surveillance video analysis. In this task, pedestrian images often undergo intensive changes in appearance. One primary cause of such variations is that pose of the same pedestrian can differ greatly. Example images from Market-1501 [2] are shown in Figure 1, person's appearance can be very different due to large pose variations.

In addressing the challenge of pose variations, a previous strategy of the literature is implicit. That is, to learn stable feature representations that have invariance property under different poses. Examples in traditional methods include [3] [4] [5], etc. Examples in deep learning approaches include [6] [7], etc., which suffer from the number of images and pose variations in the closed datasets (Market-1501 [2], DukeMTMC-reID [8], CUHK03 [9] etc.). Therefore, with the success of generative adversarial network (GAN) [10], some works [11], [12] based on data augmentation have been proposed to improve the generalization ability. However, generative adversarial network aims at generating images that look realistic. It is easy to mix various appearance features and noise into the generated samples, which is harmful to the performance of person reid.

Fig. 1. Motivation of the proposed method. Appearance of the same person can be different due to the large pose variations.

In this paper, we resort to a strategy from the view of pose variation data augmentation. We are mainly motivated by the requirement for large data volume in deep learning-based person reid. To learn rich features that are robust to pose variations, annotating large-scale datasets are effective but prohibitively expensive. Nevertheless, if we can add more effective samples to the training set that is aware of the pose variations, we are able to 1) address the data scarcity problem in person reid, 2) learn invariant features under different poses.

Based on the above discussions, this paper proposes a pose variation adaptation method to regularize model training for person reid. More concretely, to maintain better appearance feature which is effective for reid task, we learn a pose transfer generative adversarial network for synthesizing realistic person images conditional on

* Corresponding Author.
† Authors contributed equally.

pose with appending a similarity measurement module. It considers the distance changing on pose variations which is shown in Figure 2, two images from the same person with similar pose appear similar (Figure 2(a)), while the same person with different poses appear quite different (Figure 2(b)). With the learned model, we can generate new training samples in any given poses which are labeled. During reid training, the dataset is a combination of the original images and the pose-transferred images. It is beneficial in reducing over-fitting and achieving pose-invariant property. Besides, we optimize the manner of samples used to extract discriminative features. Analyzing the methods which have achieved promising performance in the view of data augmentation, we note that most of them just combine the pose-transferred samples with the original images to obtain an extended dataset for training. Therefore, based on hard example mining, we introduce a novel strategy of data augmentation to make full use of the pose-transferred samples. Specifically, for hard example mining, several chosen triplets which we call them inferior examples, do not lie inside the given margin of triplet loss. They are invalid for the process of training. Therefore, seeking to increase effective hard examples, we propose a hard example mining strategy with replaceable sample, which replaces the inferior examples via pose transfer. It is beneficial to extract more discriminative features and improve the generalization.

Experimental results show that our pose transfer generative adversarial network generates more realistic images in various poses which can be used to reduce over-fitting and achieves pose-invariant property, and the improved hard example mining strategy can make full use of the generated samples to improve the generalization of person reid.

In summary, our contributions can be summarized into three aspects as follows:

1) we propose a pose transfer generative adversarial network to synthesize images for data augmentation with a similarity measurement module which considers the distance changes on pose variations. It pushes pedestrians in similar poses closer than those in quite different poses to maintain better appearance and pose;
2) we optimize the manner of the pose-transferred sample usage, which replaces the inferior examples caused by pose variations with pose transfer model to learn discriminative representations;
3) we evaluate the performance on two popular datasets, our approach achieves rank-1 of 96.1% for Market-1501 [2] and 92.0% for DukeMTMC-reID [8]. Experimental results show that our approach outperforms or shows comparable results to the existing best perform methods on both datasets.

II. RELATED WORK

In this section, we briefly review two research topics which are related to our approach, including person re-identification and generative adversarial network.

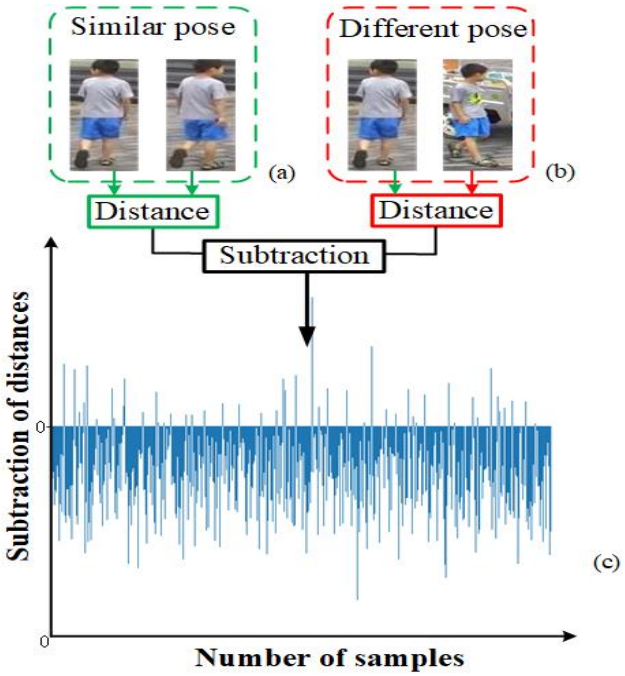

Fig. 2. Illustration of distances change with pose variations. Two images from the same person with similar poses appear more similar than the same person in quite different poses. (a), (b) denote person in similar and quite different poses respectively. (c) indicates subtraction of distances from a baseline method, the distances of person images in similar poses are generally larger than images from different poses.

A. Person Re-identification

Reid is an extremely challenging task because of various poses, illumination, domain differences, occlusions, etc., it attracts great attention due to its potential application in surveillance video analytics. Existing methods mainly focus on the following two categories: 1) extracting discriminative features to handle the variations in person's appearance, and 2) designing distance metrics to measure the similarity between images. We briefly discuss some of these works below.

Traditional person reid approaches usually explore to design handcrafted features and color features as representation descriptors [13], [14]. And many works utilize distance metric like cross-view quadratic discriminant analysis (XQDA) [3], KISSME metric learning [4], DNS [5], etc. Recently, deep learning is widely used in person reid. Yi et al. [15] first apply deep convolutional neural network to determine if a pair of input images belong to the same identity. Since then, deep learning-based methods such as Gated Network [16], SVDNet [17] and Deepreid [9] have been put forward to further improve the discrimination of features. Person's appearance can be very different due to the large pose variations which limits the performance of distance metric. Therefore, pose information is introduced to person reid. Sun et al. [18] introduce a Part-based Convolutional Baseline (PCB) with pose estimation which

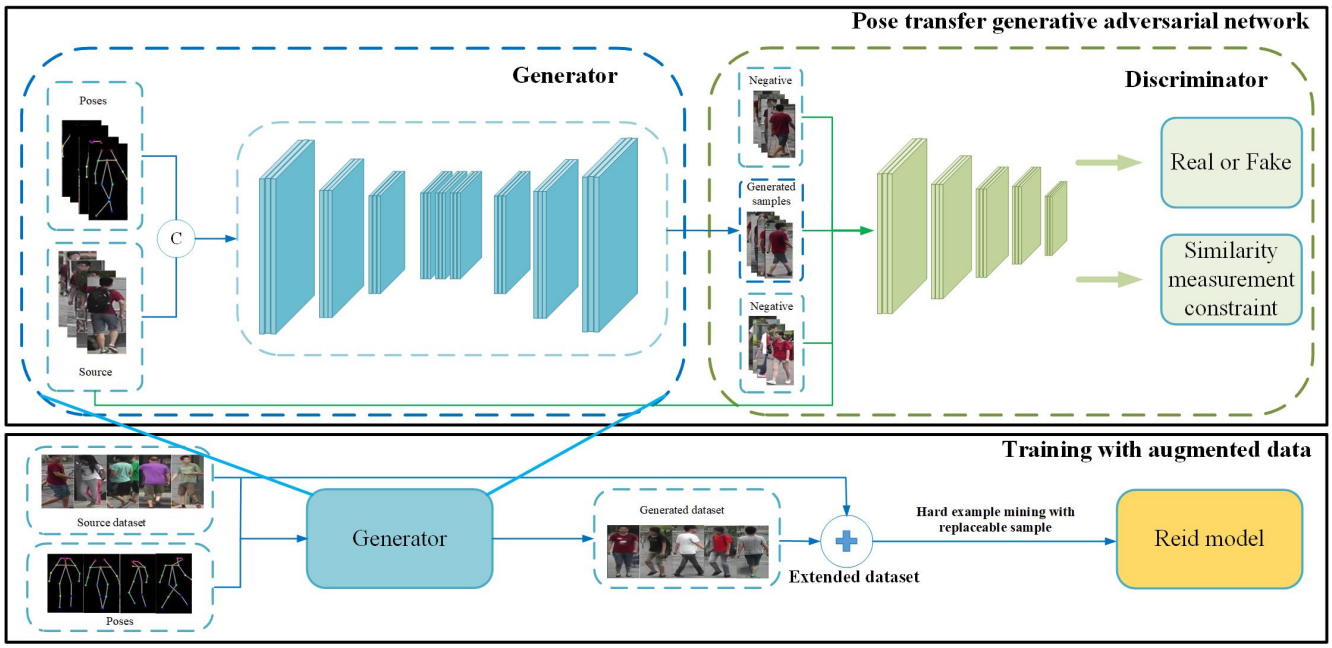

Fig. 3. Outline of our proposed method. Our methods are learned to synthesize auxiliary samples in various poses for data augmentation. We generate training data which is combined with the real image to train jointly for alleviating the influence of pose variations in reid. And improving hard example mining with replaceable sample to learn discriminative features and enhance the generalization ability.

conducts uniform partition on the conv-layer for learning part-level features. In [19], a Visibility-aware Part Model (VPM) is proposed, which learns to perceive the visibility of regions through self-supervision.

The above-mentioned approaches have proposed multifarious contributions, most of them just design complicated frameworks and distance metrics for reid training with increasing computational cost which cannot tackle the influence of pose variations on the similarity measures. The public datasets also contain a limited number of samples, making the training hard to converge and easily overfitting.

B. Generative Adversarial Network

Generative adversarial networks (GANs) have shown promising performance for generating realistic and shaper images in recent years. Goodfellow et al. [20] first propose the adversarial process to learn generative models. It generally consists of a generator and a discriminator, where the generator attempts to generate realistic images to cheat the discriminator and the discriminator learns to distinguish the generated images from real distribution. It can be regarded as a minimax two-player game optimized simultaneously. After that, DCGAN [10] scales up GAN using CNNs and Mirza et al. [21] proposed a conditional version of GAN (cGAN) where the generator and discriminator are conditioned on some auxiliary information.

Since GAN proposed, many variants of GAN have been proposed to tackle various problems, e.g., image-to-image translation, style transfer, pose-to-image generation, etc. Among them, image-to-image translation has received lots of attention. Isola et al. [22] utilize cGAN learning a mapping from input to output images for image-to-image translation application (pix2pix). The major defect is that it requires paired training data, which is difficult to acquire in many works. Therefore, based on "pix2pix", a cycle consistency loss is introduced to train with unpaired images. Neural style transfer and pose-to-image can also be regarded as a strategy of image-to-image translation. Style transfer aims at transferring the style of an input image to another, and the pose-to-image framework takes an image and a pose as inputs and generates a sample with the pose of inputs.

In addition, some recent approaches make use of style transfer and pose-to-image to generate auxiliary data for person reid. Deng et al. [23] introduce a similarity preserving generative adversarial network (SPGAN) to preserve self-similarity and domain-dissimilarity, which consisting of a Siamese network and a CycleGAN. In [24], a person transfer GAN is proposed to bridge the domain gap between two different datasets. Except to be used in cross-domain task, GAN also can be utilized for data augmentation. In [8], DCGAN is adopted to generate more training data and propose the label smoothing regularization for outliers. Zheng et al. [25] introduce a joint learning architecture, it couples data generation and reid training end-to-end.

The work most relevant to ours is PN-GAN [26] which proposes a novel deep person image generation model for synthesizing realistic person images. It is based on a generative adversarial network designed specifically for pose normalization in reid. Different from PN-GAN, this paper introduces a pose variation adaptation method for person

re-identification. We propose a pose transfer generative adversarial network with a similarity measurement module to generate pose-rich images for data augmentation. It pulls persons in similar poses closer than those in quite different poses to maintain better appearance and pose, which is useful for reid task. Simultaneously, we optimize the manner of samples using based on hard example mining, which focuses on the hard examples that are caused by pose variations. Experimental results show that our methods can learn discriminative features and improve the generalization ability.

III. OUR PROPOSED METHOD

Our goal is to learn a pose-transferred model for generating auxiliary samples which contains rich pose information. Then combining the pose-transferred samples with the source images for training to improve the generalization ability and reduce over-fitting in reid. To this end, we propose a pose variations adaptation method for person reid, which consists of a pose transfer generative adversarial network and an improved hard example mining strategy. The overall framework is shown in Figure 3, pose transfer generative adversarial network is utilized to generate samples for data augmentation (Sec. 3.1), and the improved hard example mining strategy replaces the inferior examples which are caused by pose variations (Sec. 3.2) to learn discriminative features and enhance the generalization ability. The details of our introduced methods are given as follows.

A. Pose Transfer Generative Adversarial Network

In this work, we employ GAN to generate new training samples: the poses between images from the same pedestrian are considered as different domains. We propose a pose transfer generative adversarial network, which consists of an image generator G and a discriminator D. The generator aims at producing images of the same person in different poses. The Discriminator focuses on learning to differentiate whether the input image is real or fake.

Generative adversarial network aims at generating images which look realistic. It is easy to mix various appearance features and noise into the generated samples, which influences the performance of distance metric and limiting the generalization ability in reid. To tackle this problem, we propose a pose transfer generative adversarial network constraining the similarity of pedestrians in different and similar poses. Our generator takes a source image x_s, a pose image p_t and the target image (ground-truth) x_t which is in pose p_t as input, it aims at learning to replace pose information in x_s with pose p_t to generate a new image x_f via the mapping function $x_f = G(x_s, p_t)$. Pose representation is obtained by a pre-trained model. More concretely, OpenPose [27] is deployed, which is trained without using any reid dataset. The generator is an encoder-decoder network based on the ResNet architecture. The encoder-decoder network learns to extract semantic information and by progressively down-sampling x_s to a bottleneck layer, and then reverse the process to generate x_f. For the discriminator, as is shown in Figure 2, we consider that two images from the same person with similar pose appear more alike than the same person with a greatly different pose. Inspired by it, we introduce a novel similarity measurement module to optimize the quality of the generated images further. We use the source image x_s, the target image x_t, the generated image x_f and a random negative image x_n to constitute a quadruplet as input of discriminator, where x_s, x_t and x_f are in the same identity and x_n is from a different person. Therefore, different from the traditional discriminator, except to distinguish whether the input image is real or fake, our method learns to maintain the appearance features which is effective for reid simultaneously. The introduced constrain can be formulated as follows:

$$\begin{aligned}\mathcal{L}_s = \mathbb{E}_{x,p\sim p_{data}}\left[\left\|f\left(x_f\right)-f\left(x_s\right)\right\|_2^2-\left\|f\left(x_f\right)-f\left(x_n\right)\right\|_2^2+\alpha_1\right]_+ \\ +\mathbb{E}_{x,p\sim p_{data}}\left[\left\|f\left(x_f\right)-f\left(x_t\right)\right\|_2^2-\left\|f\left(x_f\right)-f\left(x_s\right)\right\|_2^2+\alpha_2\right]_+\end{aligned} \tag{1}$$

where $f(x)$ denotes the feature of image x, the threshold α_1 and α_2 are margins, and $[\mathcal{L}]_+ = \max(\mathcal{L}, 0)$.

In summary, the integral objective loss function of our method is as below:

$$\mathcal{L}(G,D) = \mathcal{L}_{cGAN}(G,D) + \lambda_1\mathcal{L}_{L_1} + \lambda_2\mathcal{L}_{\mathrm{s}} \tag{2}$$

where $\mathcal{L}_{cGAN}(G,D)$ is the loss function of the conditional GAN, which is expressed as:

$$\begin{aligned}\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,p\sim p_{\mathrm{data}}}\left[\log D(x_s,p_t)\right] \\ +\mathbb{E}_{x,\ \mathrm{p}\ \sim p_{\mathrm{data}}}\left[\log\left(1-D(x_s,G(x_s,p_t))\right)\right]\end{aligned} \tag{3}$$

$\mathcal{L}_{L_1}$ is L1 loss which is denoted as:

$$\mathcal{L}_{L_1} = \mathrm{E}_{x,p\sim p_{\mathrm{data}}}\left[\left\|x_f - G(x_s,p_t)\right\|_1\right] \tag{4}$$

and λ_1, λ_2 are the weighting factors.

During training, we sample as many paired poses and the appearance instances as possible which are constituted quadruplets for learning to transfer source dataset in various poses, and these images are not only passed through the generator, but are also sent into the discriminator. During testing, we pair images from reid dataset and random poses to generate auxiliary training samples. Experimental results show that our results are significantly realistic and shaper. And the training set is augmented to a combination of the original images and the pose-transferred samples. Since each pose-transferred sample preserves the content of its original image and introduces new pose variations, the new image is considered to be of the same pedestrian as the original image. This allows our method to utilize pose-transferred samples as well as their associated labels to train reid model together with the original images.

B. Hard Example Mining with Replaceable Sample

Analyzing the methods which have achieved promising performance in the view of data augmentation, we note that most of them just combine the generated samples

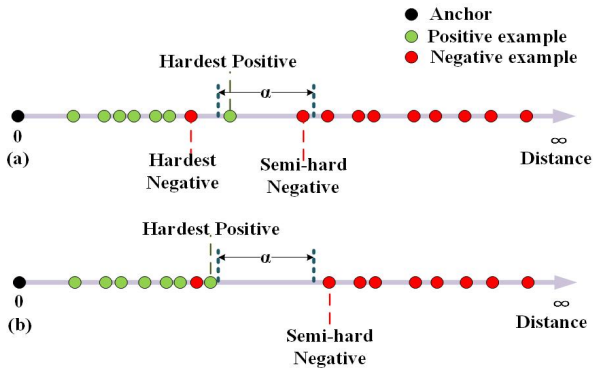

Fig. 4. Examples of hard example mining. Given a margin α, (a) is an example of semi-hard negative which is inside the margin. And (b) indicates an ineffective example which lies outside of the margin. And it is which our method focuses on.

with the original images directly to obtain an extended dataset for training. In this work, based on hard example mining, we optimize the manner of samples (the pose-transferred images) used to extract discriminative features. The hard example mining scheme which focuses on the hard samples in the dataset has been proved effective. Considering the influence of pose variations on distance metric, we improve hard example mining with replaceable sample, which replaces the inferior example to increase the number of effective triplets. We divide hard examples into two types: 1) person image degradation due to camera view changes; 2) person appearance changes dramatically because of pose variations. Note that our method focuses on the latter type. In this section, we first introduce the triplet selection method in FaceNet [28] to analyze the influence of pose on distance metric and propose hard example mining with replaceable sample.

FaceNet first introduce triplet selection (triplet loss) method which having proved that it is crucial for fast convergence in training stage. This means that, as shown in Figure 4, the green and red dots indicate positive and negative examples respectively, and α is a margin that is enforced between positive and negative pairs. Given an image x_i^a (anchor), selecting the hardest positive x_i^p such that argmax $\left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2$ and similarly hardest negative x_i^n such that argmin $\left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2$. In order to avoid model collapse, FaceNet replaces the hardest negatives with semi-hard exemplars such that $\left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2 < \left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2$. We divide the selected hard examples into two types: effective and ineffective for training. As shown in Figure 4(a), the positive is hardest, as it is further away from the anchor than other positive exemplars, and the negative (semi-hard) lies inside the margin α, it is effective. In Figure 4(b), the positive and the negative are selected with the same method, they are not inside the margin α. This kind of triplets are invalid for training. To tackle this problem, utilizing the generator G (Sec. 3.1), we introduce our hard example mining strategy with replaceable sample to increase effective hard examples for improving the generalization ability of reid.

In detail, we divide the dataset into three subsets of the front, back and side as calculated in [7]. As shown in Figure 5, given the generator G, for the triplet $\left\{x_i^a, x_i^p, x_i^n\right\}$ which is ineffective (image (a), (p) and (n)), and their poses $\{p^a, p^p, p^n\}$. x_i^a, x_i^p, x_i^n denote the anchor, the positive and the negative, respectively. For the negative (image (n)), we transfer it to the pose p^a directly and express as:

$$x_i^N = G\left(x_i^n, p^a\right) \tag{5}$$

we know that $\left\|f\left(x_i^a\right)-f\left(x_i^N\right)\right\|_2^2 < \left\|f\left(x_i^a\right)-f\left(x_i^n\right)\right\|_2^2$. Therefore, we obtain a harder negative example (image (N)). And for the positive (image (p)), first, we compare the orientation of the anchor and the positive example, if they do not belong to the same subset, we do nothing. Otherwise, we transfer the positive example to one of the other two subsets expressed as x_i^P using the generator G. We know that $\left\|f\left(x_i^a\right)-f\left(x_i^P\right)\right\|_2^2 > \left\|f\left(x_i^a\right)-f\left(x_i^p\right)\right\|_2^2$. Similar to the discussion above, we obtain a harder positive example(image (P)). Finally, we replace the original positive and negative with x_i^P and x_i^N to obtain an effective triplet $\left\{x_i^a, x_i^P, x_i^N\right\}$ for training.

In the experiments, our proposed method can extract more discriminative features and improve the generalization ability significantly.

IV. EXPERIMENTAL RESULTS

In this section, we conduct the following experiments to verify the effectiveness of our proposed methods on two publicly datasets from real-world surveillance video. We describe the experimental settings and provide experimental results along with discussions.

A. Datasets

Market-1501 [2] is an image-based reid dataset which contains 32668 labeled images of 1501 identities captured from 6 camera views. The detector used is deformable part model (DPM). The dataset is divided into two parts: 12936 images for training and 19732 images for testing.

DukeMTMC-reID [8] is a large-scale person reid dataset which is a subset of DukeMTMC. It consists of 36411 labeled images captured from 8 camera views and 1404 identities in which 702 identities for training and the remaining 702 identities for testing.

B. Implementation Details

Our method consists of three steps. First, we train the proposed pose transfer generative adversarial network to generate realistic and shaper images in different poses. Second, we choose four classic poses to generate labeled pose-varied augmented data for training to improve the generalization ability and reduce over-fitting in reid. Third, we optimize the manner of generated samples using to extract discriminative features.

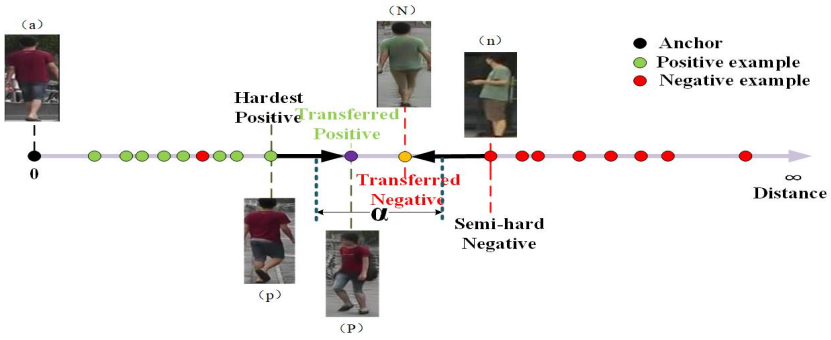

Fig. 5. An example of the proposed hard example mining with replaceable sample. We transfer the negative example to the pose of anchor directly. For the positive example, if the orientation is same as the anchor's, we transfer it to a different pose.

Training of pose transfer generative adversarial network. Pose information is obtained by the pre-trained model of OpenPose [27]. The input pairs (i.e., images and poses) of generator are resized into 256×128 and the outputs are sent into discriminator. Our proposed pose transfer generative adversarial network is trained with adam optimizer, and the learning rate is set to 0.0002. The dropout ratio is set as 0.5. In our experiments, λ_1 and λ_2 are set to 10.0 and 1.0 respectively. Simultaneously, the margins α_1, α_2 and α_3 are set to 1.0, 0.7 and 0.5 respectively.

Enhancing performance of reid with proposed method. For every image in the dataset, we utilize the generator to synthesize four samples in four selected classic poses and add them into the dataset obtaining an extended set. As described in Sec. 2.1, there are variety of methods proposed to learn features for person reid task, among them, ResNet50 and Densenet121 are the most used two. Recently, a strong baseline has been proposed, it achieves state-of-the-art performance on several public datasets. Therefore, we adopt the strong baseline without center loss as our base network. In our experiments, we train with label smoothing regularization and triplet loss. The epsilon is set to 0.1 and 0.5 for real image and generated images respectively, the margin is set as 0.3 for triplets which contain augmented data. Each of input images is resized to 256×128 pixels and utilizing zero values to pad the resized image 10 pixels. Then crop it into 256×128 randomly. Simultaneously, horizontal flip each image with 0.5 probability. Adam used to optimize the training. The initial learning rate is set to be 0.00035 and we spend 10 epochs increasing it to 0.0035 linearly. Then, the learning rate is decreased by 0.1 at the 40th epoch and 70th epoch, respectively. There are 200 epochs to train totally.

C. Comparison Results and Discussions

On the Market-1501 [2] and DukeMTMC-reID [8] dataset, we evaluate the effectiveness of the introduced method comprehensively.

Effectiveness of the proposed pose transfer generative adversarial network. In this part, we analyze the effectiveness of the proposed pose transfer generative adversarial network.We compare our method with other generative approaches, including one style transfer GAN (Camstyle [30]) and two open-source conditional GAN (Deformable [29] and PN_GAN [26]). As shown in Figure 6, for every part, the first two columns indicate source image and target pose, respectively. The last two columns are samples generated with our proposed method and ground-truth. We can clearly see that the pedestrian images generated with our proposed method are more realistic and shaper. We argue that the key is that the proposed pose transfer generative adversarial network utilizes the similarity measurement constraint to make the encoder learn better appearance and pose information.

Performance comparison on public datasets with the state-of-the-arts. In this part, we report the performance of comparing our approach with several state-of-the-art methods in recent years on Market-1501 [2] and DukeMTMC-reID [8] in Table 1. We compare our approach with three types of person re-identification methods, which are feature learning methods, metric learning methods and data augmentation methods. On each dataset, our method outperforms the performance of methods in the table. As shown in Table 1, we use the terms w/ and w/o to stand for with and without re-ranking scheme respectively. PN_GAN [26] and DG-net [25] are similar to our method which utilize data augmentation for training. Comparing with PN_GAN [26], our method achieves clear gains of 6.3% and 16.3% for rank-1 on Market-1501 and DukeMTMC-reID. And our approach also performs significantly better than DG-net by 0.9% and 3.3% for rank-1. In addition, our method outperforms other non-generative approaches on the two datasets in the table. Specifically, we achieve rank-1 accuracy=95.7% for Market-1501 and rank-1 accuracy=89.9% for DukeMTMC-reID. With the re-ranking scheme, our proposed method achieves rank-1 accuracy=96.1% for Market-1501, 92.0% for DukeMTMC-reID and achieves mAP accuracy=94.5% for Market-1501, 89.3% for DukeMTMC-reID. Experimental results show that the generated images of our proposed pose transfer generative adversarial network maintain better appearance features, and our hard example mining with replaceable

Fig. 6. Comparison of the generated images and real images on Market-1501 across the different methods including Camstyle, DeformGAN, PN_GAN, and our approach.

sample can extract discriminative features further.

Effectiveness of the proposed methods on two baselines. In this part, we evaluate the effectiveness of our proposed methods for reid training, including the generated data and hard example mining with replaceable sample. We compare our methods with the ResNet50 baselines [30] and [11], which are trained with triplet loss. As is shown in Table 2, we use the terms w/ and w/o to stand for with and without similarity measurement constraint respectively. On Market-1501, comparing with baseline [30] [11], our method achieves a gain of 1.6% and 5.2% for rank-1 respectively. On DukeMTMC-reID datasets, the gains are 3.7% and 10.5%. The results show that our proposed pose transfer generative adversarial network generates pose-rich augmented data which can be effective for reid training. And our hard example mining with replaceable sample extracts discriminative features further to improve the generalization ability.

V. CONCLUSION

In this paper, we propose a pose variation adaptation method for person re-identification. The pose transfer models are learned for each pair of pedestrian images in different poses, which are used to generate new training samples. The real images and the pose-transferred samples form the new training dataset. Moreover, to extract discriminative features, we optimize the manner of the pose-transferred samples using in the view of data augmentation. It can make full use of the generated images to enhance the training of reid. Experimental results on two public datasets (Market-1501 [2] and DukeMTMC-reID [8]) demonstrate that our approach can effectively reduce the impact of over-fitting and achieve substantial improvements on both image generation and reid accuracy.

Acknowledgement

This work is supported by the National Key R&D Program of China under Grant No.2018YFB2100603 and the Natural Science Foundation of China under Grant No.61872024.

TABLE I
Comparison with state-of-the-art on Market-1501 and DukeMTMC-reID

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	**mAP**	**Rank-1**	**mAP**
BoW+kissme [2]	44.4	20.8	25.1	12.1
XQDA [3]	-	-	30.8	17.0
DNS [5]	55.4	29.9	-	-
Gated [16]	65.9	39.6	-	-
IDE [1]	72.5	46.0	65.2	45.0
SVDNet [17]	82.3	62.1	76.7	56.8
TriNet [31]	84.9	69.1	72.4	53.5
Part-aligned [32]	91.7	79.6	84.4	69.3
VPM [19]	93.0	80.8	83.6	72.6
Mance [33]	93.1	82.3	84.9	71.8
M^3 [34]	95.4	82.6	84.7	68.5
LSRO(w/o) [8]	84.0	66.1	67.7	47.1
PT(w/o) [12]	87.7	68.9	78.5	56.9
PN-GAN(w/o) [26]	89.4	72.6	73.6	53.2
Camstyle(w/o) [11]	89.5	71.6	78.3	57.6
FD-GAN(w/o) [35]	90.5	77.7	80.0	64.5
DG-net(w/o) [25]	94.8	86.0	86.6	74.8
Base1(w/o) [30]	94.1	85.7	86.2	75.9
Ours	**95.7**	**88.0**	**89.9**	**78.2**
DG-net(w/) [25]	95.4	92.5	90.3	88.3
Base1(w/) [30]	95.4	94.2	90.3	89.1
Auto-ReID(w/) [36]	95.4	94.2	91.4	89.2
Ours+re-ranking	**96.1**	**94.5**	**92.0**	**89.3**

TABLE II
Effectiveness of the proposed pose transfer person re-identification with improved hard example mining on two different baselines [11] [30]

Methods	Martket-1501		DukeMTMC-reID	
	Rank-1	**mAP**	**Rank-1**	**mAP**
Base1 [30]	94.1	85.7	86.2	75.9
+Generated data(w/o)	94.3	86.0	87.1	76.0
+Generated data (w/)	95.0	86.2	88.0	76.5
+Replaceable sample	95.7	88.0	89.9	78.2
Base1 [11]	85.6	65.8	72.3	51.8
+Generated data(w/o)	88.2	70.4	76.7	67.4
+Generated data (w/)	89.4	73.2	78.1	70.0
+Replaceable sample	90.8	85.9	82.8	75.2

References

[1] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984, 2016.

[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In Proceedings of the IEEE international conference on computer vision, pages 1116–1124, 2015.

[3] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2197–2206, 2015.

[4] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In European Conference on Computer Vision, pages 701–716. Springer, 2016.

[5] Li Zhang, Tao Xiang, and Shaogang 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.

[6] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 3960–3969, 2017.

[7] Na Jiang, Junqi Liu, Chenxin Sun, Yuehua Wang, Zhong Zhou, and Wei Wu. Orientation-guided similarity learning for person re-identification. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 2056–2061. IEEE, 2018.

[8] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, pages 3754–3762, 2017.

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 152–159, 2014.

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

[11] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5157–5166, 2018.

[12] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4099–4108, 2018.

[13] A Mignon and FJ Pcca. A new approach for distance learning from sparse pairwise constraints. In 2012 IEEE Conference on Computer Vision and Pattern Recognition.

[14] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3586–3593, 2013.

[15] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In 2014 22nd International Conference on Pattern Recognition, pages 34–39. IEEE, 2014.

[16] Rahul Rama Varior, Mrinal Haloi, and Gang Wang. Gated siamese convolutional neural network architecture for human re-identification. In European conference on computer vision, pages 791–808. Springer, 2016.

[17] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, pages 3800–3808, 2017.

[18] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), pages 480–496, 2018.

[19] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 393–402, 2019.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017.

[23] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 994–1003, 2018.

[24] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 79–88, 2018.

[25] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2138–2147, 2019.

[26] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 650–667, 2018.

[27] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7291–7299, 2017.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[29] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3408–3416, 2018.

[30] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019.

[31] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.

[32] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 402–419, 2018.

[33] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 365–381, 2018.

[34] Jiahuan Zhou, Bing Su, and Ying Wu. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2909–2918, 2020.

[35] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In Advances in Neural Information Processing Systems, pages 1222–1233, 2018.

[36] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 3750–3759, 2019.