Leveraging Virtual and Real Person for Unsupervised Person Re-Identification

Fengxiang Yang, Zhun Zhong, Zhiming Luo, Sheng Lian, and Shaozi Li, *Senior Member, IEEE*

***Abstract*—Person re-identification (re-ID) is a challenging instance retrieval problem, especially when identity annotations are not available for training. Although modern deep re-ID approaches have achieved great improvement, it is still difficult to optimize the deep re-ID model and learn discriminative person representation without annotations in training data. To address this challenge, this study considers the problem of unsupervised person re-ID and introduces a novel approach to solve this problem by leveraging virtual and real data. Our approach includes two components: *virtual person generation* and *training of the deep re-ID model*. For virtual person generation, we learn a person generation model and a camera style transfer model using unlabeled real data to generate virtual persons with different poses and camera styles. The virtual data is formed as labeled training data, enabling subsequent training deep re-ID model in supervision. For training of the deep re-ID model, we divide it into three steps: 1) pre-training a coarse re-ID model by using virtual data; 2) collaborative filtering based positive pair mining from the real data; and 3) fine-tuning of the coarse re-ID model by leveraging the mined positive pairs and virtual data. The final re-ID model is achieved by iterating between step 2 and step 3 until convergence. Extensive experiments demonstrate the effectiveness of our method. Experimental results on two large-scale datasets, Market-1501 and DukeMTMC-reID, show the advantages of our method over state-of-the-art approaches in unsupervised person re-ID. Our code is now available online.[1]**

***Index Terms*—Person re-identification, generative adversarial network, collaborative filtering.**

I. Introduction

WITH the urgent demand for security and the rapid development of multimedia, surveillance camera systems have been deployed in a large number of public areas, such as airports, streets, malls, *et al.* This allows us to obtain massive image/video data for tracking [1] and retrieving [2] person of interest in a large-scale database, such as escaped criminals and missing children. Person re-identification (re-ID) is developed to find the same person from a gallery collected by different cameras. It is a challenging and attracting topic for computer vision and multimedia due to the significant image variations caused by changing of human poses and camera settings. During the past few years, person re-ID has achieved significant improvement [3]–[7], benefiting from the remarkable success of deep Convolutional Neural Nets (CNNs) [8]. Nevertheless, training deep re-ID model requires substantial annotated data, which is quite expenespecially when across a mass of cameras. Under such circumstances, there is an urgent demand for learning the discriminative deep re-ID model with large-scale unlabeled data. In this paper, we address the challenging unsupervised person re-ID problem, where large-scale training data is provided while no label information is available.

Unsupervised person re-ID has been studied in many previous works [9]–[11]. These works mainly focus on designing discriminative hand-crafted features and dealing with a small dataset but degenerate when applying on large-scale datasets. Deep CNNs have reached state-of-the-art performance on large-scale person re-ID datasets. Most of the existing deep CNNs based re-ID models were trained by using either ID-discriminative embedding (IDE) [5] or triplet (or pairwise) loss [6]. However, it is impossible to train these models without annotations on the training set, because both IDE and triplet loss require label information or the relationship (positive and negative) with other training data for the given image. There are limited works that make efforts on deep learning based unsupervised re-ID. Fan *et al.* [12] propose a framework called PUL, which progressively utilizes k-means clustering to find reliable positive pairs and fine-tunes the deep CNN model. The main drawbacks of PUL are that the initial re-ID model should be pre-trained on a labeled re-ID dataset and the rough number of unique identities in the target dataset should be given for clustering.

In this study, we consider the pure unsupervised setting of person re-ID, where no auxiliary labeled dataset is provided. We propose a novel deep CNN based approach, which consists of two components: 1) virtual person generation and 2) training of the deep re-ID model. For virtual person generation, we first employ DPG-GAN [13] and Star-GAN [14] to learn a person generation model and a camera style transfer model by using unlabeled real training data. As such, we can generate virtual persons with different poses and assign them with

Manuscript received November 3, 2018; revised June 28, 2019 and August 25, 2019; accepted November 25, 2019. Date of publication December 12, 2019; date of current version August 21, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61876159, 61806172, 61572409, U1705286, and 61571188, in part by the China Postdoctoral Science Foundation under Grant 2019M652257, and in part by the National Key Research and Development Program of China under Grant 2018YFC0831402. The associate editor coordinating the review of this manuscript and approving it for publication was Professor Mohammed Daoudi. *(Corresponding authors: Zhiming Luo; Shaozi Li.)*

F. Yang, Z. Zhong, S. Lian, and S. Li are with the Department of Artificial Intelligence, Xiamen University, Xiamen 361005, China (e-mail: yangfx@stu.xmu.edu.cn; zhunzhong007@gmail.com; lancerlian@stu.xmu.edu.cn; szlig@xmu.edu.cn).

Z. Luo is with the Post-doctoral Mobile Station of Information and Communication Engineering, Xiamen University, Xiamen 361005, China (e-mail: zhiming.luo@xmu.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2957928

[1][Online]. Available: https://github.com/FlyingRoastDuck/PGPPM

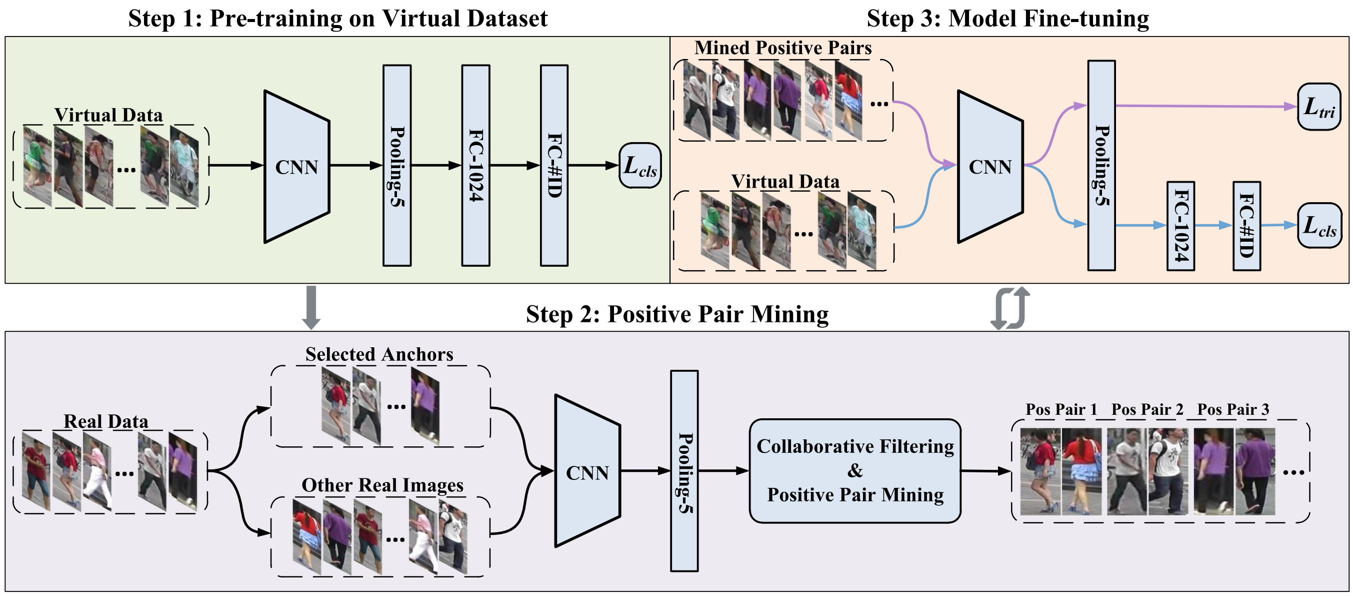


Fig. 1. The overall training procedures of the proposed unsupervised deep re-ID method contains three main steps. In step 1, we use virtual data generated by DPG-GAN and Star-GAN to train a coarse deep re-ID model. Then, a collaborative filtering based positive pair mining approach is utilized to find reliable positive pairs from the real data in step 2. In step 3, we refine the coarse re-ID model by leveraging the virtual data and mined positive pairs with a multi-task loss function. Finally, we alternate between step 2 and step 3 until the re-ID model converged.

corresponding pseudo labels. Then the same generated identity will be style transferred to different cameras. These virtual persons are formed as virtual training data and subsequently be utilized for training a coarse deep re-ID model in a supervised way.

Deep re-ID model training can be divided into three steps as shown in Fig. 1: 1) pre-training on virtual data, 2) positive pair mining, and 3) model fine-tuning.

For step 1, a coarse deep re-ID model is trained by using the generated virtual data. This coarse model can provide discriminative representation for similarity measuring of persons. However, the image quality of virtual data is lower than real data. Thus, the discriminative ability of the model trained on virtual data will be inferior to that trained on labeled real data. To address this problem, we further propose to mine reliable positive pairs from real data and jointly optimize the re-ID model with virtual and real images.

For step 2, we first use the previous pre-trained coarse re-ID model to extract features for each real image and compute its k-reciprocal nearest neighbors (k-RNNs) [15]. Although each image and one of its k-RNNs can be treated as a positive pair, there are large amount of false positive pairs which have negative effects for model refinement. To alleviate this issue, we leverage the relations of shared neighbors between samples and propose a novel collaborative filtering based positive pair mining approach to find the most reliable positive pairs in unlabeled data.

In step 3, the mined positive pairs and the virtual labeled training data are simultaneously leveraged for model refinement by using a multi-task loss function. At last, the final deep re-ID model is achieved by iterating between step 2 and step 3 until convergence.

To summarize, main contributions of this study are as follows:

- We propose a novel framework for unsupervised person re-ID by leveraging the generated pseudo labeled virtual data and the unlabeled real data for deep re-ID model training. Experiment shows the benefit of jointly training with the virtual and real data in unsupervised re-ID system.
- A collaborative filtering based positive pair mining approach is introduced to select reliable training pairs from unlabeled real data by leveraging person-to-person similarity relations. Experiment demonstrates the effectiveness of the proposed positive pair mining approach for model refinement.
- The proposed method achieves state-of-the-art performance in unsupervised person re-ID on two large-scale datasets, Market-1501 and DukeMTMC-reID.

II. Related Work

Unsupervised Person Re-identification: Unsupervised person re-ID attempts to learn discriminate features for pedestrians with unlabeled data. Hand-craft features can be directly employed for unsupervised person re-ID. Farenzena *et al.* [16] propose to use the weighted color histogram, maximally stable color regions, and recurrent high structured patches to separate the foreground of pedestrians from the background and compute appearance-based feature for re-ID. Gray and Tao [17] split input image into horizontal stripes and use eight color channels and 21 texture filters on the luminance channel to extract feature. Recently, Zhao *et al.* [18]–[20] propose to split images of pedestrians into 10×10 patches and combine LAB color histogram and SIFT feature as the final descriptor. Liao *et al.* [9] introduce local maximal occurrence descriptor (LOMO) by combining color feature and SILTP histogram. Zheng *et al.* [11] propose to extract global visual features by aggregating local color-name descriptors, and a bag-of-words model is then utilized for re-ID. Yang *et al.* [10] propose a weighted linear coding method for multi-level descriptor learning. These methods can be readily applied to unsupervised person re-ID but often fail to perform well on large-scale datasets.

Yu *et al.* [21] present an unsupervised metric learning approach for re-ID called CAMEL. It employs asymmetric metric learning to find the shared space where the data representations are less affected by view-specific bias. Liu *et al.* [22] propose a step-wise metric promotion model for unsupervised video person re-ID by iteratively estimating the annotations of training tracklets and optimizing the re-ID model.

Recently, many works [12], [23]–[25] try to transfer a pre-trained re-ID model to the unlabeled dataset (also called domain adaptation). Peng *et al.* [23] exploit a multi-task dictionary learning method to learn shared feature space between labeled dataset and unlabeled dataset. To take advantage of the strong discriminate ability of deep learning, Fan *et al.* [12] present a deep learning framework called PUL. They use a labeled dataset to initialize feature embedding and then fine-tune the network with positive sample pairs obtained through k-means clustering on the unlabeled dataset. TJ-AIDL [24] adopts a multi-branch network to establish an identity-discriminative and attribute-sensitive feature representation space most optimal for the target domain without any label information. Deng *et al.* [25] introduce SP-GAN by jointly preserving self-similarity and domain-dissimilarity in the process of image-to-image translation. The source set is transferred to the style of the target set and is then used to learn a re-ID model for the target set. Similarity, Wei *et al.* [26] present PT-GAN to reduce the domain gap by translating the given image to the style of the target dataset and train deep re-ID model in a supervised way. All the methods mentioned above require a labeled re-ID dataset to pre-train a re-ID model and then transfer it to the unlabeled target set. In this paper, we conduct unsupervised person re-ID under a more strict condition where there are only unlabeled target set.

Person Image Generation: Generating realistic person images is a challenging task because of the complexity of foreground, person pose, and background. The image generation models, *e.g.*, VAE [27] and GANs [28], have been demonstrated the effectiveness in person generation. Zhao *et al.* [29] combine variational inference into GAN to generate multi-view images of persons in a coarse-to-fine manner. Ma *et al.* [30] develop a framework to generate new person images in arbitrary poses given as input person images and a target pose. Despite the promising results, these two approaches require aligned person image pairs in the training stage. To solve this problem, Esser *et al.* [31] propose VAE-U-Net to train a person generation model by disentangling the shape and appearance of the input image. The new image is generated with U-Net for target shape, conditioned on the VAE output for appearance. Ma *et al.* [13] introduce DPG-GAN to generate virtual person images by simultaneously disentangling and encoding the foreground, background, and pose information into embedding features. The embedding features are then combined to reconstruct the input person image.

Style Transfer: Style transfer is a sub-domain of image-to-image translation. Recent works conducted on GANs [28] have achieved impressive results on image-to-image translation. Pix2pix achieves this goal by optimizing both adversarial and L1 loss of cGAN [32]. However, paired-samples are required in the training process, and this limits the application of pix2pix in practice. To alleviate this problem, Cycle-GAN [33] introduces cycle-consistent loss to preserve key attributes for both the source domain and the target domain. These two models can only transfer images from one domain to another and may not be flexible enough when dealing with multi-domain translation. To overcome this problem, Star-GAN [14] is proposed to combine classification loss and adversarial loss into the training process to translate an image into different styles with only one model.

III. The Proposed Method

In this section, we first describe the pipeline of virtual person generation in Section III-A. Then, the implementation of coarse Deep Re-ID model training is introduced in Section III-B. We present the details of collaborative filtering based positive pair mining in Section III-C and the final model fine-tuning in Section III-D.

A. Virtual Person Generation

In unsupervised person re-ID, identity annotations are not available in training set, which makes it challenging to train deep re-ID model in traditional way like IDE [5] and triplet loss [6]. In order to solve this problem, this paper considers learning the potential distribution of the unlabeled person data and generating labeled virtual person images for deep re-ID model training. For achieving this goal, this work employs DPG-GAN [13] to generate virtual person samples with different poses. In addition, the generated samples are transferred to styles of different cameras by Star-GAN [14] for overcoming the camera variations [34], [35]. Note that the training procedures of DPG-GAN and Star-GAN do not need any labeled identity information.

DPG-GAN: DPG-GAN is an unsupervised person generation method that can obtain the novel person image from Gaussian noise. A generator is proposed to disentangle pose information, foreground and background masks of unlabeled real data and encode them into embedding representations. These embeddings are decoded to reconstruct the input image with L1 loss. Besides, three generators are introduced to generate virtual embeddings from Gaussian noise, and corresponding discriminators try to distinguish the embeddings of real data from the virtual embeddings. In this way, DPG-GAN learns to synthesis virtual person samples with different appearances, backgrounds, and poses.

Star-GAN: Star-GAN contains a style transfer model $G(x, c)$ and a discriminator $D(x)$, where x and c represent input image and target domain label, respectively. In this paper, we regard each camera as an independent domain. During training, G is designed to generate virtual image in the style of target domain c. D learns to distinguish between real image and style transferred image, as well as to classify the real image to its corresponding camera domain. We alternatively optimize G and D as the training strategy in [14].

Virtual Dataset Generation: Given unlabeled real training data, we first learn a person generation model and a camera style transfer model with DPG-GAN and Star-GAN, respectively. Then, we use DPG-GAN to randomly generate person images with different poses and transfer them in the styles of

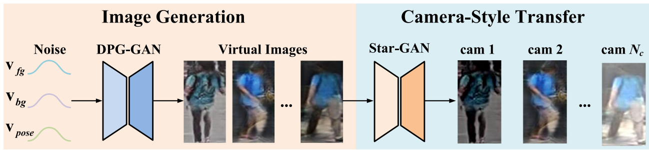


Fig. 2. The pipeline of virtual image generation. We first use DPG-GAN to generate virtual images from Gaussian noise. Then, we assign annotations to the virtual samples where person samples with the same foreground contain the same identity. Finally, we transfer the virtual person samples to the styles of different cameras with Star-GAN on average.

different cameras by Star-GAN. In Fig. 2, we show the pipeline of virtual person generation, which can be summarized into four steps:

1) Define the number of identities (classes) N_p and number of samples N_e for each person. In this way, the number of images in the virtual dataset will be $N_p \times N_e$.
2) Sample real-like foreground v_{fg}, background v_{bg} and pose v_{pose} embeddings from Gaussian noise and feed them into pre-trained DPG-GAN for composing virtual person image. For each identity of person, we fix v_{fg} and randomly sample v_{bg} and v_{pose} N_e times to generate person images with different poses and backgrounds.
3) Repeat step 2 N_p times to generate the whole virtual person dataset. Person images with the same foreground are assigned to the same identity.
4) Transfer virtual person images into styles of different cameras using pre-trained Star-GAN. For virtual person samples of each identity, we transfer them to N_c camera styles on average.

To this end, we generate virtual person data with different poses and camera styles. Examples of virtual person images are shown in Fig. 3.

B. Training Coarse Deep Re-ID Model

Given the labeled virtual person data with N_p identities, we are able to train a deep re-ID model in supervised way. In this work, we regard the re-ID model training as a classification problem and train a coarse re-ID model based on IDE [5]. We adopt ResNet-50 [8] as the backbone network and add two fully convolutional (FC) layers after the Pooling-5 layer. The first FC layer has 1024-dim named as "FC-1024". The second FC layer named as "FC-#ID" which has N_p-dim. N_p is the number of identities in the virtual person dataset. The cross-entropy loss is used to train the coarse re-ID model.

C. Collaborative Filtering Based Positive Pair Mining

Although person generation algorithm can produce high-quality samples, it still generates a certain proportion of poor instances *(e.g., broken limbs or blur background)* as shown in Fig. 3(c) and (f). These poor instances will degenerate the performance of the re-ID model. Coarse deep re-Id model trained on virtual data is insufficient to discriminate the real data in the testing set. To address this problem, we attempt to mine positive pairs from unlabeled data for model refinement.

Definition: We denote the unlabeled real data as $\mathcal{U}$. Given a query image $p \in \mathcal{U}$, our goal is to find the positive sample sharing the same identity with p from $\mathcal{U}$ (except p). Based on the pre-trained coarse re-ID model, we extract the output of pooling-5 as the feature for each real image and compute the pair-wise similarity matrix $\mathbf{S}$ between all real images as

$$\mathbf{S}_{p,q} = \exp(-||v_p - v_q||_2), \tag{1}$$

where v_p and v_q are normalized pooling-5 features of image p and q.

k-reciprocal nearest neighbors: Given the computed pair-wise similarity matrix, we could obtain the k-nearest neighbors (i.e., the top-k samples in the similarity ranking list) for each real image. We define the k-nearest neighbors of p as $N(p,k)$. In this paper, we adopt k-reciprocal nearest neighbors (k-RNNs) [15] instead of k-nearest neighbors as candidates that may contain positive samples of p. The k-RNNs for image p is defined as

$$R_k(p) = \{q_i | (q_i \in N(p,k)) \wedge (p \in N(q_i,k))\}, \tag{2}$$

where q_i is among the top-k similar samples of p, and p is also among the top-k of q_i. Intuitively, images in $R_k(p)$ are of high similarity with p and can be utilized to form positive pairs. We named this approach as k-reciprocal nearest neighbor based positive pair mining. However, it will be prone to form false positive pairs due to illumination, pose variation, and other uncontrollable factors. To filter false samples from the candidates of $R_k(p)$, we then propose a collaborative filtering based positive pair mining approach to find more reliable samples that share the same identity with p.

Collaborative filtering based positive pair mining: Collaborative filtering (CF) is a technique utilized by recommender systems for preference prediction [36]. The underlying assumption of the user-based CF is that if two persons have a large overlap in opinions with items, they are very likely to have a similar taste. Inspired by the user-based CF, we argue that if an image p shares the same k-RNNs as an image q, they are more likely to be a positive pair. Based on the shared neighbors between p and q, we are able to leverage their potential relations and re-calculate their similarity. As shown in Fig. 4, our approach includes four steps:

1) *Obtaining k-reciprocal nearest neighbors:* Given the computed pair-wise similarity matrix, we first calculate the k-RNNs for each real image according to Eq. (2). For a query image p, we represent the k-RNNs of p as $R_k(p)$ and try to find the reliable positive sample from $R_k(p)$.
2) *Collaborator mining:* We denote collaborators as the shared k-RNNs of two images. Thus, given a query image p and a candidate image q in $R_k(p)$, the collaborator set C of p and q is defined as:

$$C(p,q) = \{c_i | (c_i \in R_k(p)) \wedge (c_i \in R_k(q))\}. \tag{3}$$

3) *Collaborative filtering similarity:* Based on the collaborator set of p and q, we calculate the filtered similarity

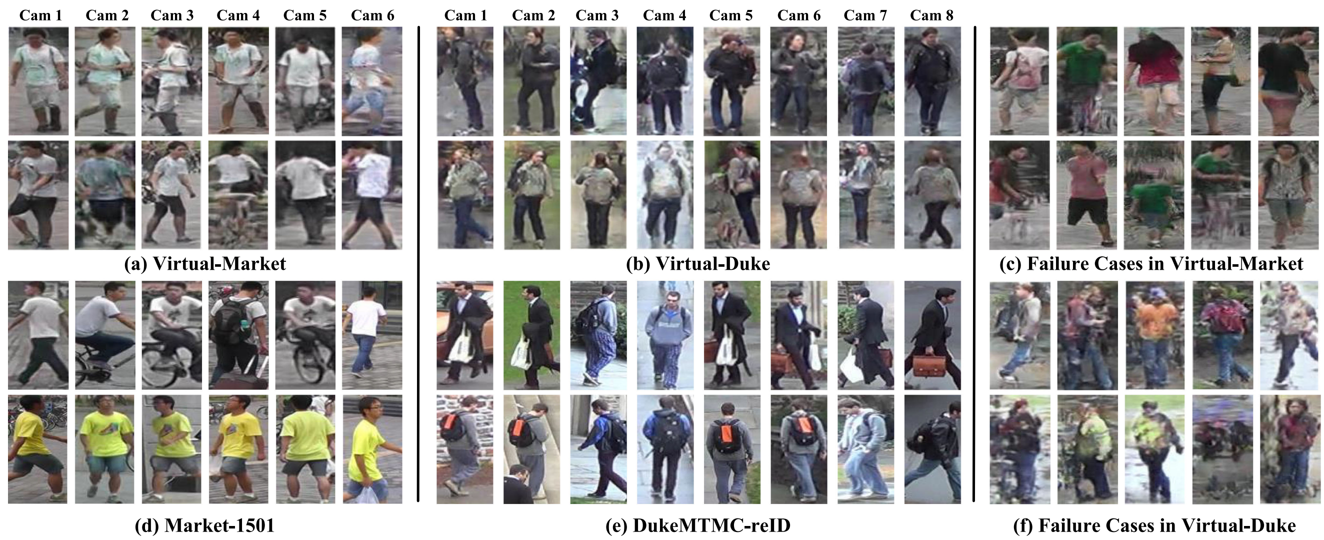


Fig. 3. Examples of virtual person images on Market-1501 and DukeMTMC-reID. Despite the successful virtual images, failure instances (*e.g.* incomplete body parts and blurred backgrounds) may influence the performance of deep re-ID model.

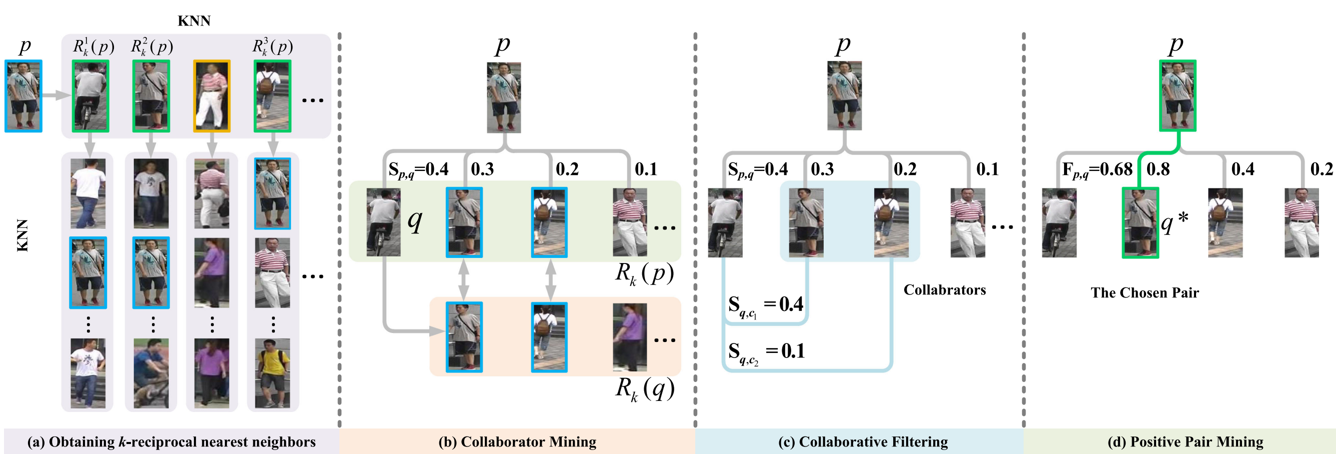


Fig. 4. Collaborative filtering based positive pair mining. Given a query image p (blue) of real data, we first compute the k-reciprocal nearest neighbors $R_k(p)$ of p (green). Then, the collaborator set (blue) of p, and each candidate q in $R_k(p)$ is mined in step (b). The collaborative filtering similarity of p and each candidate q in $R_k(p)$ is calculated by Eq. (4) in step (c). Finally, image pair with the highest re-calculated similarity is selected as the positive pair (green) in step (d).

as:

$$\mathbf{F}_{p,q} = \mathbf{S}_{p,q} + \sum_{i=1}^{|C|} \mathbf{w}_{q,c_i} \mathbf{S}_{p,c_i}, \tag{4}$$

where $|\cdot|$ denotes number of candidates in a set, and $\mathbf{w}_{q,c_i}$ is the normalized weight to measure the significance of collaborator c_i, defined as:

$$\mathbf{w}_{q,c_i} = \frac{\mathbf{S}_{q,c_i}}{\sum_{i=1}^{|C|} \mathbf{S}_{q,c_i}}. \tag{5}$$

The filtered similarity not only considers the original pair-wise distance of p and q, but also takes the similarities between p, q and the collaborator set into consideration.

4) *Positive pair mining:* With the calculated collaborative filtering similarities between query image p and images in $R_k(p)$, image q^* with the highest similarity F is selected to construct a positive pair (p, q^*) for re-ID model fine-tuning:

$$q^* = \arg\max_{q \in R_k(p)} \mathbf{F}_{p,q}. \tag{6}$$

5) *Camera constraint:* In practice, we find that positive pairs obtained by our algorithm are always in the same camera. This phenomenon may make the re-ID model sensitive to camera variations, while the primary goal of re-ID is to retrieval a person across different cameras. To alleviate this problem, we attempt to add the constraint of removing image sharing the same camera during the computation of k-RNNs for p and q. We evaluate three types of constraint:

- **Free**: there are no constraints for p and q;
- **Single**: we add the constraint for p, while not for q;

- **Double**: we add the constraint for both p and q, and this is the default setting.

D. Model Fine-Tuning

After mining the positive pairs of real data, we combine them together with the generated virtual data to refine the previous coarse deep Re-ID model. Triplet loss project similar pairs into a feature space with a smaller distance than dissimilar pairs, which can be adopted for the selected positive training pairs. Another reason to use triplet loss on positive pairs is that we do not have the real label for selected real images, cross-entropy loss can not be obtained.

During training, we randomly select N anchor images from real data and their corresponding mined positive samples to form the training batch. For each anchor p_i, we directly assign the same pseudo label of p_i to its mined positive sample q_i^*, and select the hardest (closest) sample z_i as the negative sample within the other $N-1$ anchor images and their corresponding positive samples. The final triple loss function is as following,

$$L_{tri} = \sum_{i=1}^{N_r} \left[||f(p_i) - f(q_i^*)||_2 - ||f(p_i) - f(z_i)||_2 + m \right]_+, \tag{7}$$

where m is a margin that is enforced between positive and negative pairs, and $f(\cdot)$ is the pooling-5 feature of the deep re-ID model. N_r is the number of anchors in the training batch.

As we already have the pseudo labels of the generate virtual data, we directly use the IDE cross-entropy loss function L_{cls}. By merging these two losses into a multi-task training framework, we then have the final loss as:

$$L_{loss} = L_{cls} + \lambda L_{tri}, \tag{8}$$

where λ is a hyper-parameter controlling the influence of L_{cls} and L_{tri}.

When finished training the re-ID model for each epoch, the parameters of the deep re-ID model will be updated and the adjacent matrix **S** of the real data will also be updated. As a result, we need to proceed a positive pair mining step for each epoch. The final model can be trained by using loss function (8). By doing so, the real data can help increase the final re-ID accuracy by eliminating negative effects of distorted virtual images while virtual data stabilizes the training process and the keep basic performance of re-ID model.

IV. Experiments

To evaluate the performance of our proposed method, we conduct experiments on two large-scale benchmark datasets: Market-1501 [11] and DukeMTMC-reID [37], [38]. The mAP and rank-1 accuracy are adopted as evaluation metrics.

Market-1501 dataset contains 32,668 bounding boxes of 1,501 identities obtained from six cameras. 751 persons are used for training while the rest for testing (750 identities, 19,732 images). The probe set contains 3,338 images for querying true person images from gallery set.

DukeMTMC-reID dataset is a subset of DukeMTMC [38] which consists of 36,411 labeled bounding boxes of 1,404 identities pictured by 8 different cameras. Similar to the protocol of Market-1501, this dataset split 16,522 images of 702 identities for training, 2,228 probe images and 17,661 gallery images from the rest for testing.

A. Experiment Settings

DPG-GAN: We train the DPG-GAN by 120,000 epochs with a batch size of 16. The learning rates of all networks are set to 0.00008 and divided by 10 in every 10,000 epochs. All input images are resized to 128 $\times$ 64. We use the same network architectures following [13].

In virtual person generation stage, we use N_p to represent the number of individuals/identities included in virtual dataset while N_e denotes the number of images generated for each person. Unless otherwise specified, we generate virtual datasets with $N_p = 600$ and $N_e = 36$ for Market-1501, and with $N_p = 600$ and $N_e = 48$ for DukeMTMC-reID.

Star-GAN: The Adam solver is employed to train G and D of Star-GAN for a total 200 epochs with a batch-size of 40. Input images are resized to 128 $\times$ 128. The learning rates for D and G are initialized to 0.0001 and linearly reduced to 0 for the last 100 epochs. We employ the network structures following [14].

During camera style translation, one-hot label of target camera is tiled and concatenated with input images to form a $128 \times 128 \times (N_c + 3)$ tensor, the tensor is then sent to U-Net-like generator for style translation. N_c is the total number of cameras for corresponding real dataset. We convert images from virtual data to different camera styles on average. In other words, each image is transferred to one style of cameras.

Re-ID Model Training: We resize input image to 256 $\times$ 128, and employ random horizontal flipping and random cropping for data argumentation. The SGD solver is used for optimization with a learning rate initialized as 0.1 and divided by 10 after 100 epochs. We train the re-ID model with 150 epochs in total. For positive pair mining, we first train the re-ID model by only using the virtual data for 100 epochs. After that, we add the mined positive pairs from real data for fine-tuning with another 50 epochs. Other parameters are set as follows: the triplet loss with anchor batch size $N_r = 50$ and a margin $m = 0.3$, k-reciprocal nearest neighbors with $k = 50$, and $\lambda = 1$ in Eq. (8).

B. Comparison With State-of-The-Art

In order to compare with other competing unsupervised re-ID methods, we train two models with generated virtual datasets for the Market-1501 and DukeMTMC-reID dataset, respectively. All the experimental results of our method and other methods are reported in Table I. As can be seen, our method outperforms all previous unsupervised re-ID methods. On Market-1501, we can get a rank-1 accuracy of 63.9%, which is 9.4% higher than the previous state-of-the-art method CAMEL [21]. On the DukeMTMC-reID, our method can also beat PUL [12] with a 5.9% higher rank-1 accuracy. Additionally, we compare our

TABLE I
UNSUPERVISED PERSON RE-ID PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID

Methods	DukeMTMC-reID → Market					Market → DukeMTMC-reID				
(Domain Adaptation)	mAP	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20
UMDL [23]	12.4	34.5	52.6	59.6	-	7.3	18.5	31.4	37.6	-
PT-GAN [26]	-	38.6	-	66.1	-	-	27.4	-	50.7	-
SP-GAN [25]	22.8	51.5	70.1	76.8	-	22.3	41.1	56.6	63.0	-
Methods	Market-1501					DukeMTMC-reID				
(Unsupervised)	mAP	rank-1	rank-5	rank-10	rank-20	mAP	rank-1	rank-5	rank-10	rank-20
LOMO [9]	8.0	27.2	41.6	49.1	-	4.8	12.3	21.3	26.6	-
Bow [11]	14.8	35.8	52.4	60.3	-	8.3	17.1	28.8	34.9	-
DPG-GAN [13]	13.8	33.8	-	-	-	9.0	19.5	33.3	39.9	47.9
PUL [12]	20.1	44.7	59.1	65.6	71.7	16.4	30.4	44.5	50.7	56.0
CAMEL [21]	26.3	54.5	-	-	-	-	-	-	-	-
Our Method	33.9	63.9	81.1	86.4	90.8	17.9	36.3	54.0	61.6	67.8

TABLE II
ABLATION STUDY OF OUR APPROACH. BASED ON THE RE-ID MODEL TRAINED ON DPG-GAN, WE ADD STAR-GAN, POSITIVE PAIR MINING GRADUALLY INTO IT TO EVALUATE THE RE-ID ACCURACY

Method	Market-1501		DukeMTMC-reID	
	mAP	rank-1	mAP	rank-1
DPG-GAN	13.4	33.8	9.0	19.5
+ Star-GAN	25.1	51.7	13.9	30.3
+ Star-GAN+Mining	33.9	63.9	17.9	36.3

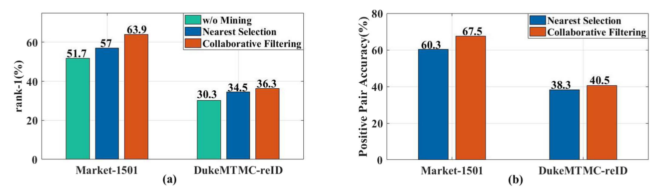


Fig. 5. The effectiveness of collaborative filtering based positive pair mining. We compare with the re-ID models trained without mining and with nearest selection mining. (a) The rank-1 accuracies of the model trained with different strategies, w/o mining, nearest selection, and collaborate filtering. (b) The true positive rate (TPR) during the whole train phase of nearest selection and collaborate filtering. The proposed mining strategy can increase TPR by a large margin.

method with three domain adaptation methods. Domain adaptation methods train a model with a labeled dataset and then transfer it to another dataset. As can be seen from Table I, our method outperform all three methods on Market-1501 dataset, with a 12.4% higher rank-1 accuracy compared with the best SP-GAN. On DukeMTMC-reID, the accuracy of our method is higher than UMDL and PT-GAN, but lower than SP-GAN. The main reason is that the generated virtual images still contain lots of low-quality samples which directly affect the accuracy of our method.

C. Ablation Study

The method discussed in Section III contains three main components: DPG-GAN, Star-GAN and positive pair mining. In order to figure out which component contributes most for the accuracy, we evaluate the performance by gradually adding Star-GAN and positive pair mining into re-ID model training. As can be seen in Table II, after adding the Star-GAN, the rank-1 on Market-1501 dataset can boost from 33.8% to 51.7%, which demonstrates that camera-style transfer plays a significant role in re-ID model initialization. Then including the positive pair mining step, we observe a further 12.2% rank-1 accuracy improvement on Market-1501. Adding real data into training can help reducing the gap between the generated virtual data and unlabeled real data. On DukeMTMC-reID dataset, we have similar findings.

To assess the effectiveness of the proposed collaborative filtering based mining procedure, we perform another comparison between *method without mining*, *nearest selection* and *collaborative filtering*. During the mining process, nearest selection takes the most similar image from k-reciprocal neighbors of the anchor image as positive sample under the default ”double constraint” setting. As shown in Fig. 5(a), the nearest selection based mining step and our proposed collaborative filtering based mining step can improve the re-ID result compared with the one without mining. The proposed collaborative filtering outperforms the nearest selection on rank-1 accuracy by 6.9% and 1.8% on the two datasets respectively. We also validate the accuracy of the mined pairs belonging to the same identity during the whole fine-tuning step by using the ground-truth information of two datasets in Fig. 5(b). The accuracy of nearest selection is 60.3% and 38.3% for Market-1501 and DukeMTMC-reID, respectively. After employing the collaborative filtering, the accuracies increase to 67.5% and 40.5%, which means the quality of mined positive pairs is improved by using our proposed method.

In triplet training phase, we randomly select N anchor images and their corresponding mined positive samples to form the training batch. For each anchor, we randomly select hardest sample as the negative within the other $N-1$ images and their corresponding positive samples. In this way, the probability of selecting the real positive sample as the negative is very low when sampling a few images from a dataset containing a large number of images and identities. We evaluate the overall false negative rate throughout the training process, which is 2.6% and 3.1% for Market-1501 and DukeMTMC-reID, respectively. These low rates might likely have a slightly negative effect on performance. Due to the competitive performance, we would rather consider the effect of selecting the real positive sample as the negative to be very small.

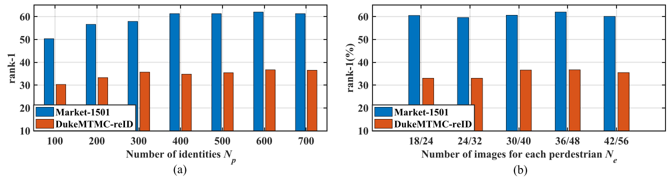


Fig. 6. Sensitive analysis for N_e and N_p. Increasing the size of virtual dataset may help improve re-ID accuracy in a certain degree.

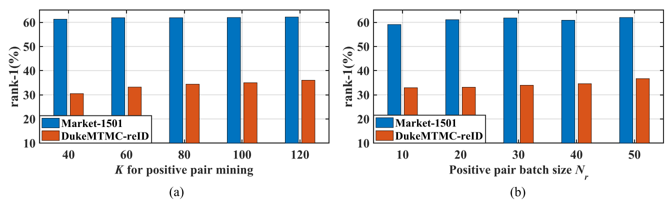


Fig. 7. Sensitive analysis for k and N_r. Our approach is robust to the changes of k and N_r.

D. Sensitive Analysis

To check the sensitive of method with different hype-parameters, we do a thorough evaluation of: (1) the scale of generated virtual dataset (N_p and N_e), (2) the batch-size of real positive pair N_r and the value of k, (3) the influence of the λ for the L_{cls} and L_{tri}.

Large-scale virtual dataset has positive impact: Intuitively, we conduct a series experiments to evaluate the influence of the scale of the generated virtual datasets. Fig. 6(a) presents the relationship between N_p and rank-1 accuracy by changing N_p from 100 to 700 with a fixed N_e. We set N_e to 36 and 48 for Market-1501 and DukeMTMC-reID, respectively In general, re-ID model performs better with larger N_p, but the accuracy will begin to saturate when N_p is large enough. The same is true for N_e as shown in Fig. 6(b). We fix N_p to 600 and vary N_e. The rank-1 accuracy will saturate when $N_e = 36$ for Market-1501 and 48 for DukeMTMC-reID, respectively. The results demonstrate that enlarging the scale of virtual dataset can improve performance of model in a certain degree.

Various N_r and k have less effect: We perform another experiment to check how many positive pairs are needed for fine-tuning the final re-ID model. As shown in Fig. 7(a) and (b), the rank-1 accuracy are fluctuated in a very small range around 63.9% and 36.3% for Market-1501 and DukeMTMC-reID by using various N_r and k. But in practice, we still suggest using a large k to ensure that k-reciprocal nearest neighbors can always be found.

Both L_{cls} and L_{tri} are important for model optimization: We evaluate our model with different λ values to find out which part contributes most for the accuracy of model. Fig. 8 shows that rank-1 accuracy of our model improves with the increase of λ when λ is in the range of [0, 1]. However, when λ exceeds 1, the rank-1 score begins to decrease. The best result is achieved when λ is around 1. The results prove our claims that both L_{cls} and L_{tri} are important for our model. L_{cls} helps model to learn robust features while L_{tri} eliminates the negative effects brought by virtual images.

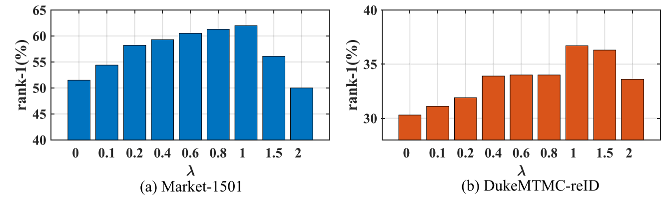


Fig. 8. Sensitive analysis for λ shows that L_{tri} and L_{cls} contribute equally to the accuracy of our method. Best result is achieved when λ is around 1.

TABLE III
COMPARISON BETWEEN DIFFERENT TYPES OF CAMERA CONSTRAINTS IN THE K-RNN COMPUTATION STEP

Method	Market-1501		DukeMTMC-reID	
	mAP	rank-1	mAP	rank-1
Free	31.6	60.8	14.2	30.3
Single	34.4	63.1	17.4	35.7
Double	33.9	63.9	17.9	36.3

TABLE IV
RESULTS ON THE TWO-CAMERA SUBSET OF THE MARKET-1501 DATASET

Method	Two-camera subset		Whole dataset (6 cameras)	
	mAP	rank-1	mAP	rank-1
Ours (w/o Mining)	9.4	9.7	18.4	42.5
Ours (Free)	8.9	8.3	16.8	40.9
Ours (Single)	27.8	34.4	24.7	51.3
Ours (6 Cameras)	28.1	35.2	33.9	63.9
SP-GAN (6 Cameras)	8.8	9.5	22.8	51.8
Supervised ResNet50	41.6	47.0	38.7	66.1

Removing positive pairs from the same camera is necessary for improving the accuracy: We also compare the influence of three camera constraint settings discussed in Section III during the k-RNN computation procedure on the whole Market-1501 and DukeMTMC-reID dataset. The results are reported in Table III. As can be seen, the "Double" constraint obtains slightly better performance than the "Single" constraint, but both of them clearly outperform the "Free" constraint. This is because that positive pairs from the same camera are usually the same person with an extremely similar appearance, these easy pairs are not very helpful for cross-camera retrieval. Therefore, it is preferred to remove these pairs during the step of positive pair mining.

E. Proposed Framework in Two-Camera System

In this section, we conduct experiments in a two-camera re-ID system by using a two-camera subset of the Market-1501 dataset *(camera 1 and camera 6)*. The illumination of camera 1 and 6 of Market-1501 are quite different. Then we use the two-camera subset for model training, and testing on both the two-camera subset and the whole six-camera test set. Since the proposed framework will not work under the default "Double constraint, we only test the "Free" and "Single" constraints.

We first train the DPG-GAN and Star-GAN by using the two-camera subset and generate a virtual dataset with 600 IDs and 21600 images ($N_p = 600$, $N_e = 36$) for re-ID model initialization. Then we fine-tune the re-ID model by mining with "Free" and "Single" constraints. We report these results in Table IV. The model without the mining step only gives 9.4% mAP and 18.4% mAP in two test settings. However, when

TABLE V
EXPERIMENTAL RESULTS OF THE MODEL TRAINED WITH THE VIRTUAL DATASET AFTER CLEANING THOSE DISTORTED SAMPLES. WE GENERATE SEVERAL DIFFERENT VIRTUAL DATASETS AND REMOVE THE BOTTOM 10% TO 50% OF IMAGES WITH A LOWER CONFIDENCE SCORE BY USING A DISCRIMINATOR D TRAINED FROM BOTH REAL AND VIRTUAL IMAGES. IS IS THE INCEPTION SCORE (IS) [39] THAT MEASURES THE QUALITY AND DIVERSITY OF GENERATED IMAGES

Method	Market-1501			DukeMTMC-reID			Experiment Settings
	mAP	rank-1	IS	mAP	rank-1	IS	
Ours w/o cleaning	33.9	63.9	3.83	17.9	36.3	3.46	No virtual images removed
Ours w/ cleaning	33.3	64.1	3.94	17.3	35.6	3.47	10% of virtual images removed
	34.8	64.5	4.08	17.8	34.8	3.51	20% of virtual images removed
	32.6	63.4	3.42	15.6	32.9	3.35	30% of virtual images removed
	32.0	60.2	3.40	14.7	32.4	3.10	40% of virtual images removed
	31.5	61.0	3.41	14.2	32.1	3.02	50% of virtual images removed

adding the "Free" constraint, the mAP will decrease to 8.9% and 16.8%. This is because most positive pairs mined under the "Free" constraint are from the same camera. Using such easy positive pairs will result in overfitting of the model and thus has negative effects during the triplet loss fine-tuning step. When training the model with "Single" constraint, we can see a significant boost, improving the mAP to 27.8% and 24.7% in two test settings, respectively. These results suggest that it is essential to remove positive pairs from the same camera.

We also compare our model with SP-GAN, which is trained on images of all six cameras. Even trained with samples of two cameras, "ours (Single)" outperforms SP-GAN on the two-camera subset and achieves competitive results on the whole six-camera test set.

F. Re-ID Accuracy After Cleaning Distorted Images

Since the badly distorted virtual images may still be harmful for the accuracy, in this section, we test the performance of the model trained with cleaned virtual datasets.

Instead of removing those distorted images manually, we train a discriminator D with both real and virtual images. Then we estimate the confidence score for the generated training set by using the discriminator D and remove those images with lower scores.

In Table V, we report how the cleaning rate (CR) influences the re-ID accuracy and inception score (IS) [39] from 10% to 50%. IS measures the quality and diversity of generated images. In order to keep the number of training images to be the same for each CR, we generate virtual datasets with different sizes and remove images with lower scores at a certain rate. For instance, when the CR is 10% on Market-1501, we first generate virtual dataset with $N_e = 40$ and $N_p = 600$, then remove 4 ($40 \times 10\%$) images with lower confidence for each ID. Under this scenario, all virtual sets for experiment are roughly the same size.

As can be seen from Table V, when tested on Market-1501, our model achieves slight improvement after removing bottom 20% of images with low confidence, and the IS is increased from 3.83 to 4.08. This demonstrates that our cleaning scheme could discard badly distorted virtual images to some extent. On the other hand, we also notice a significant decline of the IS when CR is greater than 20%, which may be caused by the drop in diversity in the cleaned dataset. On DukeMTMC-reID, we have a similar observation that a large decrease in IS will result in a large drop in accuracy. However, we also notice that a slightly improvement of IS when removing the bottom 20% distorted images, while the final re-ID accuracy still decreases. The possible reason is that the overall quality of generated images is low due to the high complexity of DukeMTMC-reID dataset. Therefore, removing those low confident images can not effectively increase the overall quality of the virtual dataset, and does not help to improve the re-ID accuracy.

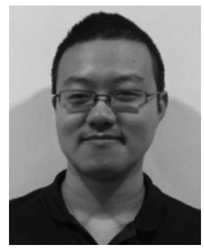
V. CONCLUSION

In this paper, we consider a challenging problem in person re-identification (re-ID), where labels are not provided in training data. To optimize deep re-ID model in supervised way, this work generates virtual dataset with a person generation model and a camera style model. Moreover, a collaborative filtering based positive pair mining approach is proposed to explore reliable positive samples from real data. This enables us to refine the re-ID model with virtual and real data, and thus improves the discriminative representation of the re-ID model. Experiments on two benchmark datasets show that our method outperforms current unsupervised re-ID algorithms. In the future work, we will focus on learning a person generation model that jointly considers the pose and camera variations and produces higher quality virtual images.

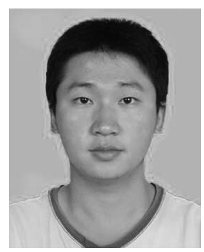
REFERENCES

[1] K.-W. Chen, C.-C. Lai, P.-J. Lee, C.-S. Chen, and Y.-P. Hung, "Adaptive learning for target tracking and true linking discovering across multiple non-overlapping cameras," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 625–638, Aug. 2011.
[2] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2035–2045, Dec. 2013.
[3] S. Zhou *et al.*, "Large margin learning in set-to-set similarity comparison for person re-identification," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 593–604, Mar. 2018.
[4] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global–local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.
[5] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*.
[6] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
[7] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
[9] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2197–2206.

[10] Y. Yang, L. Wen, S. Lyu, and S. Z. Li, "Unsupervised learning of multi-level descriptors for person re-identification," in *Proc. 31st Assoc. Advancement Artif. Intell.*, 2017, pp. 4306–4312.
[11] L. Zheng *et al.*, "Scalable person re-identification: A benchmark," in *Proc. Int. Conf. Comput. Vision*, 2015, pp. 1116–1124.
[12] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 14, no. 4, 2018, Art. no. 83.
[13] L. Ma *et al.*, "Disentangled person image generation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 99–108.
[14] Y. Choi *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8789–8797.
[15] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proc. IEEE, Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 777–784.
[16] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE,Conf. Comput. Soc. Comput. Vision Pattern Recognit.*, 2010, pp. 2360–2367.
[17] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. 10th Eur. Conf. Comput. Vision*, 2008, pp. 262–275.
[18] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3586–3593.
[19] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *Proc. Int. Conf. Comput. Vision*, 2013, pp. 2528–2535.
[20] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 144–151.
[21] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 994–1002.
[22] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2429–2438.
[23] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1306–1315.
[24] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2275–2284.
[25] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 994–1003.
[26] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 79–88.
[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Mach. Learn.*, 2013.
[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[29] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng, "Multi-view image generation from a single-view," 2017, *arXiv:1704.04886*.
[30] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. 31st Neural Inf. Process. Syst.*, 2017, pp. 406–416.
[31] P. Esser, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 8857–8866.
[32] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2223–2232.
[34] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5157–5166.
[35] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camstyle: A novel data augmentation method for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1176–1190, Mar. 2019.
[36] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 43–52.
[37] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 3754–3762.
[38] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vision*, Berlin, Germany: Springer, 2016, pp. 17–35.
[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.



Fengxiang Yang is currently working toward the master's degree at Xiamen University, Xiamen, China. His research interests include person re-identification and domain adaptation.



Zhun Zhong received the M.S. degree in computer science and technology from the China University of Petroleum, Qingdao, China, in 2015. He is currently working toward the Ph.D. degree at Xiamen University, Xiamen, China. He is also a Joint Ph.D. Student at the University of Technology, Sydney, Australia. His research interests include person re-identification and domain adaptation.

Zhiming Luo received the B.S. degree from the Cognitive Science Department, Xiamen University, Xiamen, China, in 2011, and the Ph.D. degree in computer science from Xiamen University and the University of Sherbrooke, Sherbrooke, QC, Canada, in 2017. His research interests include traffic surveillance video analytics, computer vision, and machine learning.

Sheng Lian received the bachelor's degree in computer science from the Huazhong University of Science and Technology, China. He is currently working toward the Ph.D. degree at Xiamen University, Xiamen, China. His research interests include medical image analysis, computer vision and machine learning.

Shaozi Li (SM'18) received the B.S. degree from Hunan University, Changsha, China, the M.S. degree from Xian Jiaotong University, Xi'an, China, and the Ph.D. degree from the National University of Defense Technology, Changsha, China. He currently serves as the Chair and a Professor with the Cognitive Science Department, Xiamen University, Xiamen, China. He has directed and completed more than 20 research projects, including several national 863 programs, National Nature Science Foundation of China, and Ph.D. Programs Foundation of Ministry of Education of China. His research interests cover artificial intelligence and its applications, moving objects detection and recognition, machine learning, computer vision, and multimedia information retrieval. He is a Senior Member of ACM and the China Computer Federation (CCF). He is a Vice Director of the Technical Committee on Collaborative Computing of CCF and the Fujian Association of Artificial Intelligence.