# Unsupervised Person Re-identification via Cross-camera Similarity Exploration

Yutian Lin, Yu Wu, Chenggang Yan, Mingliang Xu, and Yi Yang

***Abstract*—Most person re-identification (re-ID) approaches are based on supervised learning, which requires manually annotated data. However, it is not only resource-intensive to acquire identity annotation but also impractical for large-scale data. To relieve this problem, we propose a cross-camera unsupervised approach that makes use of unsupervised style-transferred images to jointly optimize a convolutional neural network (CNN) and the relationship among the individual samples for person re-ID.**

**Our algorithm considers two fundamental facts in the re-ID task, *i.e.*, variance across diverse cameras and similarity within the same identity. In this paper, we propose an iterative framework which overcomes the camera variance and achieves across-camera similarity exploration. Specifically, we apply an unsupervised style transfer model to generate style-transferred training images with different camera styles. Then we iteratively exploit the similarity within the same identity from both the original and the style-transferred data. We start with considering each training image as a different class to initialize the Convolutional Neural Network (CNN) model. Then we measure the similarity and gradually group similar samples into one class, which increases similarity within each identity. We also introduce a diversity regularization term in the clustering to balance the cluster distribution. The experimental results demonstrate that our algorithm is not only superior to state-of-the-art unsupervised re-ID approaches, but also performs favorably compared with other competing unsupervised domain adaptation methods (UDA) and semi-supervised learning methods.**

## I. Introduction

Person re-identification (re-ID) aims at matching a target person in a set of gallery pedestrian images. In recent years, the widespread adoption of deep convolutional neural networks (CNN) has led to impressive progress in the field of re-ID [1], [2], [3], [4], where most of them are supervised approaches. However, supervised re-ID methods require intensive manual labeling, which is expensive and not applicable to real-world applications. To solve the scalability issue, we are motivated to study unsupervised approaches for the person re-ID task.

Traditional unsupervised methods focus on hand-crafted features [5], [6], [7], salience analysis [8], [9] and dictionary learning [10]. These methods produce much lower performance than supervised methods and are not applicable to large-scale real-world data. In recent years, some unsupervised domain adaptation methods (UDA) [11], [12], [13], [14] are proposed upon the success of deep learning [15], [16]. These methods usually learn an identity-discriminative feature embedding on the source dataset, and transfer the learned features to the unseen target domain. However, these methods require a large amount of annotated source data, which cannot be regarded as pure unsupervised approaches. Recently, in our conference version [17], a bottom-up clustering method is proposed to train the model and apply bottom-up clustering iteratively to achieve impressive performance. However, this method does not make use of the variance between diverse cameras. Compared with [17], we learned the image style of different cameras, which helps us to associate the images from different cameras of the same identity.


Fig. 1. Examples in Market-1501 with different cameras. The Market-1501 dataset contains 6 different cameras. Under different cameras, the images of the same pedestrian have different background, illumination, viewpoint, *etc*.

Without the manual annotation, there are two main challenges: (i) overcoming the variance of image style caused by different cameras, *e.g.* viewpoint, illumination. (ii) exploiting similarity within each identity. As shown in Fig. 1, images of the same person often undergo intensive appearance changes caused by variations of different camera views. To address the challenge of camera variations, we generate style-transferred images and use these images for similarity exploration to eliminate the effects of camera differences. Specifically, following [18], we learn the camera style transfer model with StarGAN [19]. With the StarGAN model, for a training image captured by a certain camera, we can generate several new training samples under the style of the other cameras. In this manner, the training set is a combination of the original training images and the style-transferred images. Then, to exploit the similarity within each identity, we start with viewing individual

(Corresponding author: Yu Wu.)

Y. Lin, and C. Yan are with the Institute of Information and Control, Hangzhou Dianzi University, Hangzhou, 310018, China. (e-mail: yutianlin477@gmail.com; cgyan@hdu.edu.cn) Y. Wu, and Y. Yang are with the Center for Artificial Intelligence, University of Technology Sydney, Ultimo 2007, NSW, Australia. (e-mail: yu.wu-3@student.uts.edu.au; yi.yang@uts.edu.au) M. Xu is with School of Information Engineering, Zhengzhou University, Zhengzhou, 450001, China. (e-mail: iexumingliang@zzu.edu.cn).

images as exemplars to initialize the network, *i.e.*, each image or transferred image belongs to a distinct cluster. We then gradually incorporate similarity within identities by bottom-up clustering, which is to merge similar images (clusters) into one cluster. Finally, during the iterative training and clustering procedure, our framework exploits the cross-camera similarity of the identities to learn discriminative features.

In order to better adapt to the unsupervised setting, we refine the framework with some specific design: (i) We adopt repelled loss to optimize the CNN model without labels. In the beginning, the repelled loss directly learns to discriminate between individual images that maximize the diversity among training images. As the images are merged into clusters, the repelled loss learns to minimize total intra-cluster variance and maximize the inter-cluster variance. (ii) In practice, different identities should have a similar probability to be captured by cameras, and thus the image number for different clusters should be balanced. To enforce this property, we incorporate a diversity regularization term in the clustering procedure.

The experimental results demonstrate that our approach is superior to the state-of-the-art methods on three large scale re-ID datasets. Moreover, the one-shot and UDA re-ID methods utilize more annotation than ours, whereas our approach also obtains a higher performance than them.

Our contributions are summarized in four-fold:

A camera style transfer model is adopted to generate images under different styles to decrease the camera variance. The transferred images allow us to easily divide images under different cameras into one class.

An iterative framework is proposed to solve the unsupervised re-ID problem. By gradually exploiting the similarity within each identity across cameras, our framework can learn robust and discriminative features.

The repelled loss is proposed to optimize the model without labels. It directly optimizes the cosine distance among each individual sample / cluster, which facilitates the model to exploit the similarity within each cluster and maximize the diversity among each identity.

A diversity regularization term is proposed to balance the image number in each cluster. It makes the clustering results align with the real world distribution.

## II. Related Work

Most re-ID methods are in a supervised manner, in which sufficient labeled person pairs across cameras are given. These methods mainly focus on designing feature representations [20] or learning robust distance metrics [21], [6]. Recently, deep learning methods achieve great success [2], [3], [22], [23], [4] by simultaneously learning the image representations and similarities. In this paper, we focus on the unsupervised person re-identification, and we do not discuss more supervised methods here.

### A. Unsupervised Person Re-identification

The existing fully unsupervised methods usually fall into three categories, designing hand-craft features [5], [6], [7], exploiting localized salience statistics [8], [9] or dictionary learning based methods [10], [24]. However, it is a challenging task to design suitable features for images captured by different cameras, under different illumination and view condition. In [25], camera information is used to learn view-specific projection for each camera view by jointly learning the asymmetric metric and seeking optimal cluster separations. These methods are unable to explicitly exploit the cross-view discriminative information without pairwise identity labels. Thus the performance of these methods is much weaker than supervised methods. Recently, Lin *et al.* [17] propose a bottom-up clustering framework that jointly optimize a convolutional neural network (CNN) and the relationship among the individual samples. However, [17] neglect the style variance caused by different cameras. In this paper, we introduce a style transfer model to generate images of other camera styles and use these generated images to exploit cross-camera similarity within each identity. Our framework is beneficial in achieving camera-invariant embeddings.

There are also some recent works [26], [27], [28], [29] focusing on the unsupervised video-based re-ID. However, these methods require some very useful annotations of the dataset, *i.e.*, the total number of identities and their appearance. To conduct experiments, they annotate each identity with a labeled video tracklet, which only reduces part of the annotation workload. As discussed in [30], these approaches are actually the one-example methods. In [31], an unsupervised graph association method is proposed to learn the view-invariant representations from the video pedestrian tracklets. This method suspects images under the same camera to be tracklets and thus have the same label. Different from these methods, our work focuses on the fully unsupervised setting in which there is ***no identity annotation*** on the dataset.

### B. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [32] have achieved impressive success in many tasks, including image-to-image translation [33], [34], [19], style transfer [35], [36] and cross domain image generation [37]. In [33], a conditional GAN is proposed to learn a mapping from input to output images for image-to-image translation. However, this method requires pairs of corresponding images for training. To overcome this problem, Liu *et al.* [38] propose a coupled generative adversarial network (CoGAN) by employing weight-sharing networks to learn a joint distribution across domains. In [34], CycleGAN is proposed based on [33] to learn the image translation between two different domains without paired samples. Later in [19], StarGAN is proposed, which allows image translations among multiple domains with a single model. In this paper, we take the model of StarGan [19] to generate images in different camera styles.

Gan is also widely applied in re-ID field. In [39], a baseline DCGAN model [40] is adopted to generate unlabeled data, which are further mixed with the labeled real training images for simultaneous semi-supervised learning. In [41], [18], [42] the source data are transferred to styles of the target cameras and directly learns the deep re-ID model with labeled transferred samples in a supervised way. Different from these

Fig. 2. Examples of style-transferred images on the Market-1501 dataset and the DukeMTMC-reID dataset. The two datasets are captured by 6 and 8 cameras respectively. With the camera style transfer model, we can obtain additional images under a different style of the illumination, texture, and background.

works, we generate style-transferred images for unsupervised re-ID task, where the supervision is the exploited similarity between original and transferred training images.

### C. Unsupervised Domain Adaptation

Recently, unsupervised domain adaptation (UDA) is adopted in the unsupervised re-ID task [11], [43], [44], [45], where information from an external source dataset is utilized. In [46], a PatchNet pre-trained on the source dataset is used to generate pedestrian patches. A network is then designed to pull similar patches together and push the dissimilar patches. In [47], a soft multilabel is learned for each unlabeled person by comparing the unlabeled person with a set of known reference persons from the source domain. In [48], the model is first pre-trained on MSMT17 [49], then the cross-camera matching and intra-camera matching are applied to get the ranking result. In [14], the theoretical guarantees of unsupervised domain adaptive re-ID are introduced. This method first trains an encoder on the source dataset, and then a self-training scheme is adopted to iteratively employs clustering for unlabeled target data and trains the encoder with the triplet loss. In each iteration, [14] calculates the distance metric between the images in the target domain and images in the target and source domain with the time complexity of $O(N^2)$ and $O(N\ M)$, while our method calculates the distance metric between the real-real images and the real-fake images with the time complexity of $O(N^2)$ and $O(L\ N^2)$. Although our methods and [14] are both in an iterative clustering scheme, [14] is in a UDA manner, while our method does not utilize any annotated source domain for training or clustering measurement.

Some methods [13], [49] proposed to learn a similarity-preserving generative adversarial network based on CycleGAN [34] to translate images from the source domain to the target domain. In this way, high-quality person images are generated, and person identities are kept and the styles are effectively transformed. The translated images are utilized to train re-ID models in a supervised manner. These methods assume that the label of the source domain is available and apply the learned discriminative model to the target domain. In this work, we propose a fully unsupervised re-ID framework that gradually exploits the similarity within each identity.

## III. Proposed Method

### A. Problem definition

Given a training set $X^{train} = \{x_1, x_2, ..., x_N\}$ of $N$ images, our goal is to learn a feature embedding function $(\ ; x_i)$ from $X^{train}$ without any manual annotation, where parameters of are collectively denoted as. This feature embedding function can be applied to the testing set, $X^t = \{x_1^t, x_2^t, ..., x_{N_t}^t\}$ of $N_t$ images, and the query set $X^q = \{x_1^q, x_2^q, ..., x_{N_q}^q\}$ of $N_q$ images. During the evaluation, we use the feature of a query image $(\ ; x_i^q)$ to search the similar image features from the testing set. The query result is a ranking list of all testing images according to the Euclidean distance between the feature embedding of the query and testing data,i.e., $d(x_i^q, x_i^t) = \| (\ ; x_i^q) \quad (\ ; x_i^t)\|$. The feature embeddings are supposed to assign a higher rank to similar images and keep the images of a different person a low rank.

To learn the feature embedding, traditional methods usually learn the parameters with manual annotations. That is, each image$x_i$ is associated with a label$y_i$, where$1 \quad y_i \quad k$ and $k$ is the number of identities. A classifier$(w; (\ ; x_i)) \in R^k$ parameterized by$w$ is used to predict the identity of the image $x_i$. The classifier parameter$w$ and the embedding parameter are jointly optimized by the following objective function:

$$\min_{;w} \sum_{i=1}^{N} `(f(w; (\ ; x_i)); y_i); \quad (1)$$

where$`$ is the softmax cross entropy loss. However,$y_i$ is not available in the unsupervised setting, and it is challenging to find another objective function that can learn a robust embedding function .

### B. Camera Style Transfer Model

Without the manual annotation, we aim to exploit the cross-camera similarity from the training data as the supervision information. However, the same identity could look totally different under different cameras. To tackle the camera invariance, we propose to generate style-transferred images that preserve the person identity and reflect the style of another

Fig. 3. The pipeline of the proposed approach. It consists of two main components: 1) camera style transfer model, which generates style-transferred images together with the original training images for training; 2) iterative re-ID framework, which gradually exploits the similarity between the original and transferred images and gathers the images into a larger cluster to decrease the number of classes.

camera. In this paper, we employ StarGAN [19] to learn a camera style transfer model in the unlabelled training set. Different from [42] that adopts CycleGAN [34] for image translation, StarGAN allows us to train one model for the translations between multiple camera styles.

1) StarGAN review: The goal of StarGAN [19] is to learn a mapping function $G(x; c) \rightarrow y$ that translates an input image $x$ into an output image $y$ conditioned on the target domain label $c$. The image style of image $x$ is learned to be indistinguishable from the random generated domain $c$ using an adversarial loss. Meanwhile, the discriminator $D$ is adopted to distinguish if an image is translated fake image, and classify the real image to its corresponding domain, $D : x \rightarrow \{D_{src}(x); D_{cls}(x)\}$. The overall loss function is calculated as:

$$V(G; D; c; c') = V_{GAN}(D; G; c; c') + V_{cyc}(G; c; c'); \quad (2)$$

where $c'$ is the original domain label of the image, $V_{GAN}(G; D; c; c')$ is the loss functions for the mapping functions $G$ and for the discriminators $D$, $V_{cyc}(G; c; c')$ is the cycle consistency loss that forces $G(G(x; c); c') \approx x$, in which each image can be reconstructed after a cycle mapping, $\lambda$ balances the importance between $V_{GAN}$ and $V_{cyc}$.

2) Our practice: In our work, the images captured by different cameras are considered as different domains. Given a re-ID dataset containing images captured by $L$ different cameras, we aim to learn style transfer models for each camera pair with StarGAN. In conclusion, for a dataset captured by $L$ cameras, we generate $L - 1$ images under the style of the corresponding cameras for each training data. For a training data $x_i$, we define a support set $X_i^{cam} = \{x_i^1; x_i^2; :::; x_i^L\}$, which is the combination of the original training image and the style-transferred fake images. The examples of generated style-transferred images are shown in Fig. 2.

## C. Iterative Re-ID Framework

We propose an iterative re-ID framework to exploit the cross-camera unlabeled data gradually and steadily. As shown in Fig. 3, after generating the style-transferred images as external training data, we apply two components iteratively: (i) A network trained with a repelled loss to let the cluster centers repelled by each other. (ii) A clustering procedure in the feature embeddings space to merge existing clusters.

1) Network with Repelled Loss: Since we do not have ground truth labels, we assign each image to a different cluster initially, i.e., $\{\hat{y}_i = i \mid 1 \le i \le N\}$, where $\hat{y}_i$ is the cluster index for $x_i$ and is dynamically changed. Note that, despite we have more transferred images for augmentation, the number of images for training in each epoch is not increased. Instead, for each original unlabeled training image, we use the pedestrian image under a randomly selected camera, which could be either the original image or the stale transferred image. In this way, in the initialization, the network learns to recognize each training sample across cameras instead of the identities and thus obtain an initial discriminative ability. In the later iterations, we gradually incorporate similarity within identities by grouping similar images into clusters. The cluster ID is then used as the training label, and the network is trained to minimize total intra-cluster variance and maximize the inter-cluster variance. We define the probability that image $x$ belongs to the $c$-th cluster as,

$$p(c|x; V) = \frac{\exp(V_c^T v/\tau)}{\sum_{j=1}^{C} \exp(V_j^T v/\tau)}; \quad (3)$$

where $v = \frac{f(x)}{\|f(x)\|}$, $V \in R^{C \times n}$ is a lookup table that stores the feature of each cluster, $V_j$ is the $j$-th colum of $V$, and $C$ is the number of clusters at the current stage. At the first training stage, $C = N$. At the following stages, our approach will merge similar images into one class, and $C$ will gradually decrease. $\tau$ is a temperature parameter [50] that controls the softness of probability distribution over classes. Following [51], we set $\tau = 0.1$ in this paper. In the forward operation, we compute cosine similarities between data $x_i^{cam}$ and all the other data by $V^T v_i^{cam}$. During backward, we update the $\hat{y}_i$-th column of the table $V$ by $V_{\hat{y}_i} \leftarrow \frac{1}{2}(V_{\hat{y}_i} + v_i^{cam})$. Finally, we minimize the repelled loss, which is formulated as,

$$L = -\log(p(\hat{y}_i|x_i; V)). \quad (4)$$

During the optimization, $V_j$ will contain the information of all images within the $j$-th cluster. It can be considered as a kind of "centroid" of this cluster. We do not directly

Fig. 4. The cluster merging procedure. Each circle denotes an individual image for training. $K$ denotes the number of clusters in each iteration. After the current training iteration, we apply clustering based on the feature similarity of the current stage. By applying bottom-up clustering, individual pedestrian samples are gathered to represent an identity.

calculate the centroid feature in each training stage due to the high time complexity. The lookup table $V$ can avoid exhaustive computation of extracting features from all data at each training step. The proposed objective has two advantages. First, it can maximize the cosine distance between each image feature $v_i$ and each centroid feature $\sum_{j \neq y_i} v_j$. Second, it can minimize the cosine distance between each image feature $v_i$ and the corresponding centroid feature $\sum_{j=y_i} v_j$. With these two advantages, our network achieves the discriminative property without any annotation.

2) Cluster Merging: After the first training stage, the training samples are prone to be away from each other in the learned feature space. However, the style-transferred images and images of the same identity are usually visually similar and should be close, which we call similarity. As shown in 4, we apply bottom-up clustering on the CNN features to gradually exploit the similarity between clusters and merge the closest clusters into a larger one. In this way, the images of the same identity under the same camera style are likely to be merged into one cluster. The transferred images are also easy to be merged with the original image. Then we can achieve the goal of grouping the images across cameras into one class.

As shown in 4, in the start, each image is treated as an individual cluster. Then pairs of clusters are merged into one by measuring their similarity. In order to decide which clusters should be merged, we calculate the shortest distance between images in two clusters as the dissimilarity value $D(A, B)$ between cluster A and cluster B. The advantage is that the transferred images and images of the same identity under the same camera style are visually alike and tend to be merged into one cluster under this criterion, which guarantees the accuracy of merged images. $D(A, B)$ is formulated as:

$$D_{distance}(A, B) = \min_{x_a \in A, x_b \in B} d(x_a, x_b); \quad (5)$$

where $d(x_a, x_b)$ is defined as the Euclidean distance between the feature embeddings of two images, i.e., $v_a$ and $v_b$. Specifically, $d(x_a, x_b) = \|v_a - v_b\|$.

At each iteration, we aim to reduce $m$ clusters. We define $m = N \cdot mp$, where $mp \in (0, 1)$. Here $mp$ is the merge percent, that denotes the speed of cluster merging. Each time, the clusters with the shortest distance are merged. The number of clusters is initialed as $C = N$, i.e., the number of training samples. After $t$ iterations, the number of clusters is dynamically decreased to $C = N - t \cdot m$.

3) Dynamic Network Updating: The framework iteratively trains the network and merges the clusters. The clustering results are then fed to the network for further updating. The whole updating process is described in Algorithm 1. In this way, the cross-camera similarity is exploited gradually by clustering, and the network is gradually trained with more supervision to be more discriminative. The number of clusters is initialized as the number of training images. After each cluster merging, the labels of the training images are re-assigned as the new cluster ID. The memory layer of the optimizer is re-initialized to zero vector to avoid getting stuck in local optima. We constantly train the network until we observe a performance drop on the validation set. The model that produces the best result on the validation set is adopted as the final model.

**Algorithm 1** The Unsupervised Framework

**Require:** Unlabeled data $X = \{x_1, x_2, \ldots, x_N\}$;
Support set for each data $x_i$: $X_i^{cam} = \{x_i^1, \ldots, x_i^L\}$;
Merge percent $mp \in (0, 1)$; CNN model: $(\cdot\,; \theta_0)$.
**Ensure:** Best CNN model $(\cdot\,; \theta)$.

```
Initialize: Cluster label Y = {y_i | 1 ≤ i ≤ N}
            Number of cluster C = N
            Number of merging image m = ⌊mp · C⌋
while C > m do
    Train CNN model (x; θ) with X, each X_i^cam and Y
    Clustering with m:
        C ← C − m
        Update Y with the new cluster labels
    Initialize the lookup table V with new dimensions
    Evaluate on the validation set, get performance P_t
    if P_t > P then
        P ← P_t
        Best model = (x; θ)
    end if
end while
```

D. Clustering Strategy

1) Clustering constraint: To merge the clusters, distances between each image are calculated as illustrated in Eq. (5). However, with the style transfer model, the training set is enlarged $L$ times with generated style-transferred fake images. The distances between images are then can be divided into three types: "real-real", "real-fake", and "fake-fake". The camera variations could be reduced by integrating the "real-fake" link during clustering. However, merging two fake images will introduce noise into the framework, and calculating the distance between fake images is also time-consuming. In this work, we adopt a clustering constraint that we only consider the "real-real" and "real-fake" relations during clustering, while neglecting the "fake-fake" relations. As shown in Fig.

Fig. 5. The different clustering strategies. The solid circles represent the original training data, while the hollow circles represent the style-transferred images. Circles of the same color represent images from the same cluster. Three clustering strategies are shown: (a) Without clustering constraint. (b) Without diversity regularization. (c) Our strategy.

5 (a) and (b), with calculating the distance between "fake-fake" images, the two green circles are merged into one cluster because they are the closest. However, with the clustering constraint we merges the blue and yellow circles. The dissimilarity between two clusters $D(A;B)$ can be formulated as:

$$D_{distance}(A;B) = \min_{x_a \in A; x_b \in B; x_a \in X^{cam} \text{ or } x_b \in X^{cam}} d(x_a;x_b); \quad (6)$$

2) Diversity Regularization: With the clusters being merged, the number of classes is decreasing, and the number of images in the clusters is increasing. Although we do not know the exact number of images in each identity, we can assume that the images are evenly distributed to the identities, and different identities should be scattered in different clusters. This implies that one cluster should not contain much more images compared to other clusters.

To avoid one cluster being redundant and boost the small clusters to merge together, we incorporate a diversity regularization term into the distance criterion.

$$D_{diversity}(A;B) = |A| + |B|; \quad (7)$$

where $|A|$ denotes the number of samples belonging to the cluster $A$. Then, the final dissimilarity is calculated as:

$$D(A;B) = D_{distance}(A;B) + D_{diversity}(A;B); \quad (8)$$

where is a parameter that balances the impact of distance and regularization. The reason for adding a diversity regularization term is that, there exist some visually similar identities wearing almost the same clothes. Without the regularization term, the algorithm might merge these similar but different identities into one tremendous cluster by mistake. We tend to merge small clusters, unless the distance $d(x_a;x_b)$ is small enough. This procedure is illustrated in Fig. 5 (b) and (c). (b) shows the cluster merging result without diversity regularization: the yellow and blue images have the shortest distances (neglecting the distance between two fake images), and are then merged into one cluster. However, the yellow and blue clusters are too large and should not be merged. In our strategy, the blue and green clusters are merged instead.

## IV. Experimental Results

### A. Datasets

The Market1501 dataset [22] is a large-scale dataset captured by 6 cameras for person re-ID. It contains 12,936 images of 751 identities for training and 19,732 images of 750 identities for testing. With the style transfer model, 5 images in the corresponding camera style are generated for each training data. Finally, we get a training set of 77,616 images.

The DukeMTMC-reID dataset [39] is a subset of the DukeMTMC dataset [54]. It contains 1,812 identities captured by 8 cameras. A number of 1,404 identities appear in more than two cameras, and the rest 408 IDs are distractor images. Using the evaluation protocol specified in [39], the training and testing set both have 702 IDs. There are 2,228 query images, 16,522 training images and 17,661 gallery images. With the style transfer model, 7 images in the corresponding camera style are generated for each training data. Finally, we get a training set of 132,176 images.

The MSMT17 dataset [49] is the largest re-ID dataset, which contains 126,441 images of 4,101 identities captured by 15 cameras. In practice, 32,621 images of 1,041 identities are used for training and 93,820 images of 3,060 identities. With the style transfer model, 14 images in the corresponding camera style are generated for each training data.

### B. Experimental Settings

Evaluation Metrics. For person re-ID, we use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) to evaluate the performance of each method. For each query, its average precision (AP) is computed from its precision-recall curve. The mAP is calculated as the mean value of average precision across all queries. We report the Rank-1, Rank-5, Rank-10 scores to represent the CMC curve. These CMC scores reflect the retrieval precision, while mAP reflects the recall.

Implementation Details. For the style transfer model, following the training strategy in [18], we train StarGAN models for Market-1501, DukeMTMC-reID and MSMT17, respectively. During training, we adopt random flipping and random cropping. We train the generator and discriminator for 200 epochs, and the learning rate is 0.0001 at the first 100 epochs and linearly decays to zero in the remaining 100 epochs. Finally, for each training image, we generated 1 (5 for Market-1501, 7 for DukeMTMC-reID and 14 for MSMT17) style-transferred images with their original image for training. For the re-ID network, we adopt ResNet-50 as the CNN backbone. We initialize it by the ImageNet [55] pre-trained model with the last classification layer removed. For all the experiments if not specified, we set the number of training epochs in the first stage to be 20, and the training epochs in later iterations to be 2. We set the dropout rate to be 0.5, to be 0.05 and in Eq. (8) to be 0.03. For Market-1501 and DukeMTMC-reID, we set the batch size to be 16, and for MSMT17, we set the batch size to be 64. We use stochastic gradient descent with a momentum of 0.9 to optimize the model. The learning rate is initialized to 0.1 and changed to 0.01 after 15 epochs. For Market-1501, DukeMTMC-reID

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON MARKET-1501 AND DUKEMTMC-REID. "TRANSFER" DENOTES THE METHODS USE INFORMATION FROM ANOTHER REID DATASET WITH FULL ANNOTATIONS. "ONE-EXAMPLE" MEANS THE METHODS USE THE ONE-EXAMPLE ANNOTATION, IN WHICH EACH PERSON IN THE DATASET IS ANNOTATED WITH ONE LABELED EXAMPLE MEANS RESULTS REPRODUCED BY US

| Methods | Setting | Market-1501 | | | | DukeMTMC-reID | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | rank-1 | rank-5 | rank-10 | mAP | rank-1 | rank-5 | rank-10 | mAP |
| BOW [22] | Unsupervised | 35.8 | 52.4 | 60.3 | 14.8 | 17.1 | 28.8 | 34.9 | 8.3 |
| OIM [51] | Unsupervised | 38.0 | 58.0 | 66.3 | 14.0 | 24.5 | 38.8 | 46.0 | 11.3 |
| UMDL [12] | Transfer | 34.5 | 52.6 | 59.6 | 12.4 | 18.5 | 31.4 | 37.6 | 7.3 |
| PUL [11] | Transfer | 44.7 | 59.1 | 65.6 | 20.1 | 30.4 | 46.4 | 50.7 | 16.4 |
| EUG [30] | One-Example | 49.8 | 66.4 | 72.7 | 22.5 | 45.2 | 59.2 | 63.4 | 24.5 |
| Progressive [52] | One-Example | 55.8 | 72.3 | 78.4 | 26.2 | 48.8 | 63.4 | 68.4 | 28.5 |
| SPGAN [13] | Transfer | 58.1 | 76.0 | 82.7 | 26.7 | 46.9 | 62.6 | 68.5 | 26.4 |
| TJ-AIDL [43] | Transfer | 58.2 | - | - | 26.5 | 44.3 | - | - | 23.0 |
| Ours (w/o clustering) | Unsupervised | 39.1 | 60.2 | 69.1 | 13.9 | 29.2 | 44.4 | 51.7 | 10.7 |
| Ours (w/o style transfer) | Unsupervised | 61.0 | 71.6 | 76.4 | 30.6 | 40.2 | 52.7 | 57.4 | 21.9 |
| Ours (w/o diversity regularization) | Unsupervised | 63.6 | 78.1 | 83.3 | 28.3 | 49.0 | 60.1 | 64.3 | 24.1 |
| Ours (w/o clustering constraint) | Unsupervised | 70.2 | 80.1 | 85.9 | 38.8 | 53.9 | 66.2 | 71.6 | 28.5 |
| Ours | Unsupervised | 73.7 | 84.0 | 87.9 | 38.0 | 56.1 | 66.7 | 71.5 | 30.6 |

TABLE II
PERSON REIDENTIFICATION PERFORMANCES ON THE MSMT17 DATASET.

| Methods | rank-1 | rank-5 | rank-10 | mAP |
|---|---|---|---|---|
| OIM [51] | 7.3 | 14.4 | 18.5 | 1.7 |
| PTGAN [49] | 11.8 | - | 27.4 | 3.3 |
| ENC (Market-1501) [53] | 25.3 | 36.3 | 42.1 | 8.5 |
| Ours (w/o clustering) | 15.7 | 24.5 | 29.0 | 3.8 |
| Ours (w/o StyleTrans) | 26.4 | 35.2 | 39.4 | 9.2 |
| Ours | 31.4 | 41.4 | 45.7 | 9.9 |

and MSMT17, we validate the network on DukeMTMC-reID, Market-1501 and Market-1501, respectively.

Fig. 6. Performance curve with different values of the diversity regularization parameter on Market-1501.

### C. Comparison with the State of the Art

The comparisons with the state-of-the-art algorithms on Market-1501 and DukeMTMC-reID are shown in Table I and the performances on MSMT17 are shown in Table II. Note that the performances in [12] are reproduced by [11] and we borrow the numbers to our table. On Market-1501, we obtain the best performance among the compared methods with rank-1 = 73.7%, mAP = 38%. Compared to the state-of-the-art unsupervised method [51], we achieve 35.7 points (absolute) and 24 points improvement in rank-1 accuracy and mAP, respectively. On DukeMTMC-reID, we achieve the best accuracy with rank-1 = 56.1%, mAP = 30.6%. Compared to [51], our method achieves 31.6% and 19.3% point of improvement of rank-1 accuracy and mAP, respectively. On MSMT17, we obtain the best performances of rank-1 = 31.4%, mAP = 9.9%, which beat [51] by 24.1 points and 8.2 points on rank-1 and mAP, respectively. The impressive performance indicates that the style-transferred images eliminate the variance of different cameras by merging with the origin images during clustering, and the cluster merging effectively exploits the similarity from the instances for better supervision.

We also compare our method to the state-of-the-art UDA methods in Table I and Table II. Although these methods utilize external images and human annotations, our method with zero annotation still surpasses them by a large margin. On Market-1501, our method outperforms the state-of-the-art UDA method [43] by 15.5 points and 21.4 points in rank-1 accuracy and mAP, respectively. On DukeMTMC-reID, comparing with SPGAN [13], the performance gain are 9.2% and 4.2% in rank-1 accuracy and mAP, respectively. On MSMT17, our performances are 6.1 points and 1.4 points higher than ENC [53], which use Market-1501 as the source domain in rank-1 accuracy and mAP, respectively. The major reason is that the UDA methods can not directly learn discriminative information from the target dataset. The learned pattern and relation of the source dataset can not exactly be adopted by the source dataset. In comparison, our method gradually mines the similarity between the training images, that directly learn to discriminate the target dataset.

### D. Ablation Studies

1) Without clustering: We show the result of the baseline without clustering in Table I Ours (w/o clustering). Without clustering, we train the network for 60 iterations. The learning rate is initialized to 0.1 and changed to 0.01 after 40 epochs. We observe a rank-1 accuracy of 39.6% and 29.2% for Market-1501 and DukeMTMC-reID, respectively, which have a large margin with the proposed method. The major reason is that during the bottom-up clustering, the similarity among images is exploited, which provides more supervision information for further network training.

TABLE III
EVALUATION OF THE MERGE PERCENT mp ON MARKET-1501.

| Method | rank-1 | rank-5 | rank-10 | rank-20 | mAP |
|---|---|---|---|---|---|
| mp = 0:01 | 74.1 | 84.6 | 88.1 | 90.6 | 38.4 |
| mp = 0:05 | 73.7 | 84.2 | 87.7 | 90.8 | 38.7 |
| mp = 0:1 | 71.7 | 83.4 | 87.3 | 90.7 | 36.8 |
| mp = 0:2 | 66.8 | 80.5 | 85.0 | 89.0 | 31.5 |
| mp = 0:3 | 53.3 | 72.0 | 78.4 | 84.4 | 22.0 |

2) Without style transfer: We show the result of the baseline without style transfer in Table I Ours (w/o style transfer). We observe a rank-1 accuracy of 66.2% and 47.4% for Market-1501 and DukeMTMC-reID, respectively. In Table II Ours (w/o StyleTrans), we observe an improvements with camera-style transferring of 5 points in rank-1 accuracy. The impressive improvement indicates that the style-transferred images effectively eliminate the variance of different cameras by merging with the origin images during clustering.

3) Without clustering constraint: As shown in Table I Ours (w/o clustering constraint), the rank-1 accuracy is 70.2% and 53.9% for Market and DukeMTMC-reID, respectively, which is about 3 points lower than our method. Note that calculating the distance between "fake-fake" images is time-consuming and will introduce noise into the re-ID system. Our method only considers the "real-real" and "real-fake" relations during clustering that achieve better performance while reducing a large calculation workload.

4) The Impact of Diversity Regularization: The results with and without the diversity regularization item are shown in Table I. The diversity regularization provides a performance improvement on both datasets. Specically, on Market-1501, the diversity regularization item improves the rank-1 accuracy and mAP by 10.1 points and 9.7 points, respectively. On DukeMTMC-reID, the rank-1 accuracy and mAP are improved 7.1 points and 6.5 points, respectively. We suspect that without the diversity regularization, two similar identities may be easily merged into one cluster by mistake. With the diversity regularization term, we tend to merge small clusters rst. The diversity regularization parameter in Eq. (8) balances the cluster size and distance. We evaluate different values for the parameter in Fig. 6. As increases from 0 to 0.003, the rank-1 accuracy on Market-1501 increases from 63.6% to 73.7%. If we set to be greater than 0.003, the too large diversity regularization term would lead to a negative effect.

5) Effect of mp: For each iteration, mp N clusters are merged into a greater cluster for updating the model. To study the effect of the merge percent mp we perform experiments with different mp on Market-1501. The result is shown in Table. III. We observe that our method always achieves better performance when we merge fewer clusters in each iteration. This is because the performance of the trained CNN model highly depends on the reliability of the training label. If we merge the clusters at a slow speed, the merged clusters will tend to be of one identity. However, if we merge the clusters at a fast speed, some images of different identities will be merged into one cluster that can further harm the network. As a result, we use mp = 0:05 in this paper.

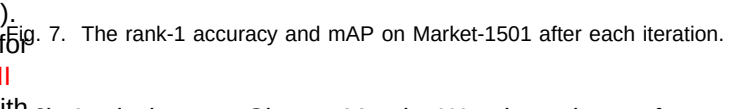

Fig. 7. The rank-1 accuracy and mAP on Market-1501 after each iteration.

6) Analysis over Cluster Merging: We show the performance of re-ID during iterations on Market-1501 in Fig. 7. Through iterations, the rank-1 accuracy and mAP are both increasing in the rst 18 itertions, and we observe a slight performance drop in the last iteration. During the rst 18 iterations, the rank-1 accuracy increases from 36.9% to 73.7%, and mAP increases from 13.2% to 38%. We observe that both the improvement of the performance and the reduction of the clusters are continuous and gradual. It indicates that our method gradually learns from the diversi ed images to generate a more discriminative feature representation.

### E. Visualization

To further understand the discriminative ability of our unsupervised learned feature, we utilize t-SNE [56] to visualize the feature embeddings of the merged clusters by plotting them to the 2-dimension map. As illustrated in Fig. 8, from iteration 0 to iteration 6, we observe an obvious and constantly points gathering of the same color, which indicates that we gradually learn a more discriminative feature representation. We can also observe that the number of clusters is continuously decreasing. On iteration 6, the images of the same identity usually gather together, which represents the learned similarity within identities.

## V. CONCLUSIONS

In this paper, we propose to tackle the unsupervised re-ID task via cross-camera similarity exploration. It jointly optimizes a CNN model and the relationship among the cross-camera individual samples. Specically, we apply an unsupervised style transfer model on the training images to get style-transferred images under different camera style. Then the network training starts by treating each individual image as an identity. Then, bottom-up clustering is applied to the feature embedding extracted from the network to reduce the number of classes. During the whole process, the network gradually exploits similarity from across-camera unlabeled images. In experiments, our method not only achieves higher performance than the state-of-the-art methods in three large-scale re-ID datasets but also performs favorably compared with other UDA and semi-supervised learning methods.

Acknowledgment. This work is supported by National Nature Science Foundation of China (61931008, 61671196,

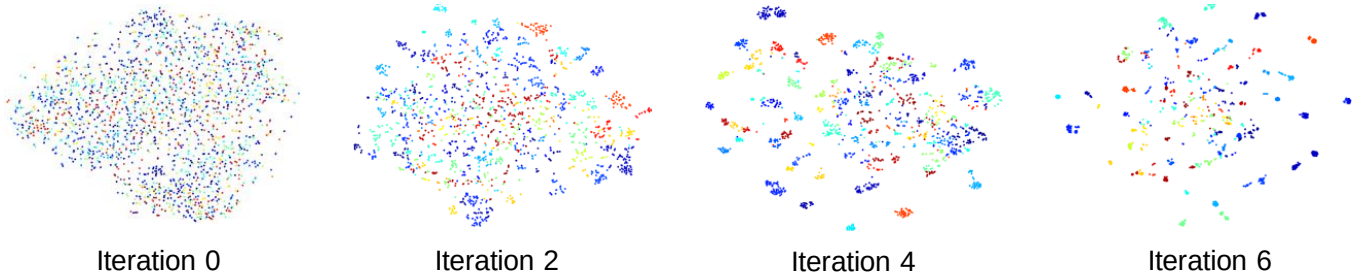


Fig. 8. T-SNE visualization of the learned feature embeddings on a part of the Market-1501 training set (100 identities, 1,867 images). Points of the same color represent images from the same identity.

61701149, 61801157, 61971268, 61901145, 61901150, 61972123), National Natural Science Major Foundation of Research Instrumentation of PR China under Grants 61427808 , Zhejiang Province Nature Science Foundation of China (LR17F030006, Q19F010030), 111 Project, No. D17019.

## REFERENCES

[1] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014. 1

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

[3] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016. 1, 2

[4] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *CVPR*, 2019. 1, 2

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010. 1, 2

[6] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015. 1, 2

[7] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1629–1642, 2014. 1, 2

[8] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013. 1, 2

[9] H. Wang, S. Gong, and T. Xiang, "Unsupervised learning of generative topic saliency for person re-identification," in *BMVC*, 2014. 1, 2

[10] E. Kodirov, T. Xiang, and S. Gong, "Dictionary learning with iterative laplacian regularisation for unsupervised person re-identification." in *BMVC*, 2015. 1, 2

[11] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 4, pp. 1–18, 2018. 1, 3, 7

[12] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *CVPR*, 2016. 1, 7

[13] 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 reidentification," in *CVPR*, 2018. 1, 3, 7

[14] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang, "Unsupervised domain adaptive re-identification: Theory and practice," *Pattern Recognition*, p. 107173, 2020. 1, 3

[15] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," in *CVPR*, 2018. 1

[16] Z. Li, J. Tang, and T. Mei, "Deep collaborative embedding for social image understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 9, pp. 2070–2083, 2018. 1

[17] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *AAAI*, 2019. 1, 2

[18] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *ECCV*, 2018. 1, 2, 6

[19] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018. 1, 2, 4

[20] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014. 2

[21] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011. 2

[22] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015. 2, 6, 7

[23] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 1, pp. 1–20, 2017. 2

[24] C. Yan, M. Luo, W. Liu, and Q. Zheng, "Robust dictionary learning with graph regularization for unsupervised person re-identification," *Multimedia Tools and Applications*, vol. 77, no. 3, pp. 3553–3577, 2018. 2

[25] H.-X. Yu and W.-S. Wu, Ancong anfd Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *ICCV*, 2017. 2

[26] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *ICCV*, 2017. 2

[27] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," *ICCV*, 2017. 2

[28] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *ECCV*, 2018. 2

[29] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2976–2990, 2019. 2

[30] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *CVPR*, 2018. 2, 7

[31] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li, "Unsupervised graph association for person re-identification," in *ICCV*, 2019. 2

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014. 2

[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017. 2

[34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017. 2, 3, 4

[35] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016. 2

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016. 2

[37] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *CVPR*, 2017. 2

[38] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *NIPS*, 2016. 2

[39] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2015. 2, 6

[40] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016. 2