Article

A New Strategy for One-Example Person re-ID: Exploit the Unlabeled Data Gradually Base on Style-Transferred Images

Qiang Li ^{1,2} and Ling Gao ^{1,2,*}¹ School of Information Science and Engineering, Shandong Normal University, Jinan 250014, Shandong Province, China; 15154738722@163.com² Institute of Data Science and Technology, Shandong Normal University, Jinan 250014, Shandong Province, China

* Correspondence: gaoling@sdnu.edu.cn

Received: 26 March 2020; Accepted: 18 April 2020; Published: 27 April 2020

Abstract: As a research field of symmetry journals, computer vision has received more and more attention. Person re-identification (re-ID) has become a research hotspot in computer vision. We focus on one-example person re-ID, where each person only has one labeled image in the dataset, and other images are unlabeled. There are two main challenges of the task, the insufficient labeled data, and the lack of labeled images cross-cameras. In dealing with the above issue, we propose a new one-example labeling scheme, which generates style-transferred images by CycleGAN (Cycle Generative Adversarial Networks) to ensure that for each person, there is one labeled image under each camera style. Then a self-learning framework is adopted, which iteratively train a CNN (Convolutional Neural Networks) model with labeled images and labeled style-transferred images, and mine the reliable images to assign a pseudo label. The experimental results prove that by integrating the camera style transferred images, we effectively expand the dataset, and the problem of low recognition rate caused by the lack of labeled pedestrian pictures across cameras is effectively solved. Notably, the rank-1 accuracy of our method outperforms the state-of-the-art method by 8.7 points on the Market-1501 dataset, and 6.3 points on the DukeMTMC-ReID dataset.

Keywords: one-example learning; person re-identification; style-transferred images

1. Introduction

Symmetry is an interdisciplinary comprehensive open-access journal that specializes in studying symmetry in various fields such as physics, chemistry, biology, computers, and mathematics. Person re-identification (re-ID) is a technology that uses computer vision-related technologies to find specific pedestrians of interest from the data gallery taken by multiple surveillance cameras. In the research of person re-ID, the most commonly used method is supervised learning. The training set of supervised learning requires input and output pairs, that is, features and labeled targets. This label is manually labeled. Through data training, the CNN [1] model can find the relationship between features and labeled targets. When inputting a new feature, it can judge the label. That is to get a model from the given training data, and then get the prediction value of the new data according to the model. In order to get a relatively good recognition model, a large number of features are needed, and a large number of target labels with manual label are also needed. However, the cost of complete manual label is huge. Compared with supervised learning, the input data of unsupervised learning is not labeled and there is no definite result. Due to the lack of a large number of labeled data, the performance of unsupervised person re-ID is very weak [2–4].

In order to avoid tedious cross-camera labeling and make full use of the easily available supervision information, the weakly supervised learning [5] method has attracted great attention in recent years. Weakly supervised learning is a learning method between supervised and unsupervised learning. Weakly supervised learning not only has a certain target label, but also can use unlabeled data to improve the expression ability of the model. Among the methods based on weakly supervised learning, the most typical one is one-example, that is, only one picture of each pedestrian is labeled, and a large number of pictures are not labeled. The specific method is to initialize a CNN model with labeled pictures first, and then uses the nearest neighbor (NN) classifier [6] to select the nearest unlabeled data in the feature space of the labeled data, and give them pseudo-labels. A pseudo-label is not a real label, but a hypothetical label when we do a one-example task. That is to say, in the process of training, we assume it is a real label. The advantage of this is that we can make full use of the unlabeled data.

Previous works on one-example learning usually focus on the pseudo-labels selection. In [7,8], a static strategy is used to determine the number of pseudo-labels, and then performing the training by fixing the size of the pseudo-labelled training set during the iterations. Wu et al. [9] proposed a progressive learning framework that better utilizes unlabeled data for person re-ID training using limited samples, and dynamically select unlabeled data and assign it a pseudo label by using nearest neighbor (NN) classifier. This algorithm trains a CNN model initially on the data of one-example labeled images, then generates pseudo-labels for all unlabeled samples, and selects some of the most reliable pseudo-labelled data for training. One-example method is widely used in the research of person re-ID, however, none of these methods have considered the problem of pedestrian image retrieval across cameras [10,11]. Person re-ID as an image retrieval problem, where the difficulty depends on the difference between cameras. Under various cameras, the same person could have totally different appearance due to different camera views, occlusion and background, etc. In the traditional one-example setting, for each identity only one image under a certain camera is labeled, and the lack of cross-camera learning leads to low recognition performance.

Figure 1 simply shows the process of style-transferred images. To address these issues, we propose to perform cross-camera similarity mining during model training and finding trusted images. Specifically, we adopt CycleGAN [12] to perform style transformation of different camera images in one dataset. CycleGAN consists of two discriminators (D_A and D_B) and two generators (G_{AB} and G_{BA}), which can realize the style transformation from both A to B and B to A . In this way, the number of labeled images is increased, and the number of labeled cameras is also expanded to have all camera styles. With the generated images from different camera views, we adopt two procedures iteratively: (1) Both of the original training data and the style transferred images are used for training, where different losses for labeled data, pseudo-labeled data and unlabeled data are applied to better learn from training images. (2) Our contributions are as follows:

- A camera style transfer model is adopted to use as an augmentation method, which increases the number of labeled data and provided labeled data under more camera style as well. With these generated training data, more cross-camera information could be exploited for one-example re-ID.
- To better leverage the style-transferred data, during training, both of the original training data and the style-transferred data are used to learn the CNN model with multiple losses. To select trusted images and give pseudo-labels, different types of distances are calculated to measure the similarity between unlabeled images. The selected images are then used for further training.
- The experimental results demonstrate that our approach is superior to the state-of-the-art one-example re-ID methods on two large scale image-based datasets, including Market-1501 and DukeMTMC-reID.

Figure 1. Example of style transferred images on Market-1501. Cam1 represents the original real picture taken by Camera 1 in the dataset (Market-1501), and c1s2 represents the pedestrian and Camera 2 shooting style generated pictures of Camera 1.

2. Related Work

2.1. Person re-ID

With the development of deep learning, the research on person re-ID has experienced a transition from traditional machine learning to deep learning. Many deep learning-based methods are effectively applied in the research of person re-ID.

Supervised Person re-ID: In the research of re-ID, the commonly used method is supervised learning. The training set of supervised learning requires input-output pairs. For example, the Siamese model is used as the input for image pairs or triples [13,14]. Li et al. used paired pedestrian images to train CNN models [15]. Zheng et al. proposed the identity embedding (IDE), and achieved good results after fine-tuning the regular dataset [16]. Xiao et al. trained a classification model from multiple fields [17]. Zheng et al. uses joint classification to train using two pairs of images [18].

One-example Person re-ID: For actual surveillance scenarios, such as surveillance video for a city, manually labeling the pedestrian labels for each video from multiple cameras is costly and impractical. So people proposed to annotate only a single labeled sample for each pedestrian and let the network learn from the few labeled samples and the raw images. Liu et al. [19] estimated a pseudo-label on the video datasets DukeMTMC-VideoReID and MARS for unlabeled data together with the partially labeled initial labels to train the model. To a large extent, it solves the limitations of excessive human labeling and image-based pedestrian recognition. However, because the amount of labeled data in the initial test is too small, the trained model is too weak, the number of trusted false labels is too small, and the recognition performance is low. Ye et al. [20] proposed a dynamic graph matching (DGM) method that iteratively updates image and label estimates to learn a better feature space with intermediate estimated labels. Kumar et al. [21] proposed to update the model automatically through self-learning. Ulyanov et al. [22] compared retrieval algorithms (distance in feature space) to the learning mechanism by combining existing SPL (Self-Paced Learning) [23] and CL (Curriculum Learning) [24] algorithms. In this paper we use the same model update strategy as Wu.

2.2. Image Style Transfer

Since GAN (Generative Adversarial Networks) proposed by Goodfellow et al. [25], many GAN-based improvements [26–34] have been made in all aspects, such as natural style transfer, super-resolution, sketch-to-image generation, image-to-image conversion, etc. Among them, image-to-image conversion has attracted a lot of attention. Isola et al. [35] proposed a conditional adversarial network to learn the mapping function from input to output image. However, this method requires paired training data, which is difficult to obtain in many tasks. In doing unpaired image-to-image conversion,

Liand et al. [36] proposed a loss of cyclic consistency to train unpaired data. Wei et al. [37] proposed a model for pedestrian image transformation of existing datasets using PTGAN (Person Transfer Generative Adversarial Networks), which may alleviate the cost of labeling on new datasets and make it easier to train re-identification systems in actual scenes. Another type of image-to-image conversion strategy is neural style transfer [38], which aims to replicate the style of an image. Deng et al. [39] proposed a framework of translation learning to convert the image styles of the images of Market-1501 dataset and DukeMTMC-reID dataset with each other, and realized the classification of the unlabeled dataset images. Zhong et al. [40] used image style transfer to increase the style diversity of the training dataset. Zhong et al. [41] proposed a progressive method for domain adaptation to iteratively update the network. In this research, we use the GAN network to implement data augmentation and cross-camera labeling.

3. Method and Framework

3.1. Preliminary

For a certain person re-ID dataset O , there are N pedestrian images of M identities taken by K cameras. Similar to [9], on the one-example setting, for each person, we only select one image in Camera 1 and give it an initialized label. If there is no image of this identity is taken in Camera 1, then we randomly select an image from the next camera and annotate the image. The purpose of this process is to ensure that each person has one picture labeled for initialization. We denote the labeled dataset as $L = [(x_1, y_1), \dots, (x_M, y_M)]$, where M denotes the number of identities, x denotes the input image data and y denotes the annotated label. The unlabeled dataset is denoted as $U = [(x_{M+1}), \dots, (x_{M+Mu})]$, where Mu denotes the number of unlabeled data. The size of the whole training dataset is then denoted as $N = M + Mu$.

3.2. Cross Camera-Style Transformed Image for re-ID

3.2.1. Review of CycleGAN

Zhu and others proposed CycleGAN. Unlike ordinary GAN models, CycleGAN does not require paired training images, and can generate another scene pattern map from a specific scene pattern map. CycleGAN consists of two discriminators (D_A and D_B) and two generators (G_{AB} and G_{BA}), which can realize the style transformation from both A to B and B to A .

Where Figure 2a shows the mutual conversion of two style datasets of X and Y , and then the respective discriminators make adversarial discrimination between the real picture and the generated picture. In Figure 2b, after entering a picture x of the dataset X , the generator G_{XY} generates a Y dataset-style picture $\hat{y}$, and then the generator G_{YX} converts the picture into an X dataset-style picture $\hat{x}$. This cycling process realizes the transformation of the dataset picture style. The new image generated by the transformation has the pedestrian subject of the dataset and the image style background of the dataset. In the same way, Figure 2c shows the same operation of dataset transformation. In this way, the dataset can be augmented through the cyclic consistency adversarial network.

Figure 2. Person re-identification (Re-ID) datasets of two different styles. (a) Accomplish the mutual conversion of two data sets X and Y; (b) accomplish the combination of dataset Y's style and dataset X's contents; (c) accomplish the combination of dataset X's style and dataset Y's contents.

CycleGAN is an X to Y unidirectional GAN plus a Y to X unidirectional GAN. The two GANs share two generators, and each has a discriminator, with totally of two generators and two discriminators. Therefore, CycleGAN has four losses. Where the discriminator loss of X to Y is:

$$L_{GAN}(G_{XY}, D_Y, X, Y) = E_{y \sim P_{data}(y)} [\log D_Y(y)] + E_{x \sim P_{data}(x)} [\log(1 - D_Y(G_{XY}(x)))] \quad (1)$$

The discriminator loss of Y to X is:

$$L_{GAN}(G_{YX}, D_X, Y, X) = E_{x \sim P_{data}(x)} [\log D_X(x)] + E_{y \sim P_{data}(y)} [\log(1 - D_X(G_{YX}(y)))] \quad (2)$$

The generator loss of G_{XY} is:

$$L_{CYC}(G_{XY}) = E_{x \sim P_{data}(x)} [\|G_{YX}(G_{XY}(x)) - x\|_1] \quad (3)$$

The generator loss of G_{YX} is:

$$L_{CYC}(G_{YX}) = E_{y \sim P_{data}(y)} [\|G_{XY}(G_{YX}(y)) - y\|_1] \quad (4)$$

So the final loss of CycleGAN is the sum of the four losses expressed as:

$$L(G_{XY}, G_{YX}, D_X, D_Y) = L_{GAN}(G_{XY}, D_Y, X, Y) + L_{GAN}(G_{YX}, D_X, Y, X) + L_{CYC}(G_{XY}) + L_{CYC}(G_{YX}) \quad (5)$$

3.2.2. CycleGAN for Pedestrian Camera Style Transformation

In this paper, the image collections are represented by two different camera styles. Specifically, the same steps as in [10] are used to learn each camera style using CycleGAN and complete the conversion between the image-image styles of the two cameras. We use two different style cameras as two different domain spaces, and the image set P_{ij} is converted to each other by CycleGAN.

$$P_{ij} = \begin{cases} real & i = j \\ fake & i \neq j \end{cases} \quad i, j \leq K \quad (6)$$

where i represents the current camera and j represents the remaining cameras. After the camera style transformation, the number of the whole training images is $K \times N$, where consists M pedestrians captured by K cameras.

After cross-camera style transformation, all the pictures are divided into the original picture dataset O and the style-transferred image dataset E . Then they are respectively set as pairs of labeled datasets $L = \{(x_1, y_1), \dots, (x_{n_l}, y_{n_l})\}$ and $L^e = \{(x_{n_l+1}^e, y_{n_l+1}^e), \dots, (x_{n_l+k}^e, y_{n_l+k}^e)\}$, where x is the input image data and y is the identity label information representing the image data. Similarly set the unlabeled dataset based on the original real image $U = \{(x_{n_l+k+1}), \dots, (x_{n_l+k+u})\}$ and the generated image $U^e = \{(x_{n_l+k+u+1}^e), \dots, (x_{n_l+k+2u+1}^e)\}$. The unpaired unlabeled datasets have only input data and no label information. As shown in Figure 3, we have generated style-transferred images.

Figure 3. Generated images of cross-camera style transferred in the dataset (Market-1501).

3.3. Framework

The goal of our method is to make full use of the few labeled images and the large amount of unlabeled data. There are two steps when learning the CNN model. The first step is to train the CNN model on four parts: L , L^e , S^t and M^t denote label set, style-transferred label set, pseudo-label set and unlabeled set. The second step is to select some reliable unlabeled candidates from a large amount of unlabeled data according to their similarity. Specifically, as shown in Figure 4, we first train an initial model on the original labeled data and the labeled style-transferred data. Then we use this initialed model to extract features for all the unlabeled data. Similarities between images are calculated and the most reliable ones are selected into the candidate set, which is continuously updated during iterations. Then we use the labeled data and the pseudo-labeled data to learn a more robust CNN model, which is updated during iterations to get more and more robust.

Figure 4. The framework of our method.

As shown in Figure 4, the original dataset is first expanded to K times by CycleGAN (K is the number of cameras in the dataset). After expansion, the CNN network is initialized. The unlabeled data is used to perform sample prediction through the initial CNN network. A highly reliable sample is selected into the candidate set and then a part is selected as a false label. After the pseudo-label selection is completed, the CNN model is retrained with the original label data. The entire process is dynamically changed. After the completion of a round of iteration, the remaining unlabeled sample dataset is automatically updated. The number of pseudo-labels is gradually increased as the iteration progresses.

3.3.1. Network Training

During iterations, there are four types of data for training, namely the original dataset's label set L , the style-transferred dataset's label set L^e , the pseudo-labeled set S^t , and the unlabeled label set M^t . The original labeled datasets and the labeled style-transferred datasets contain reliable discriminative information, while the pseudo-labeled set contains the selected reliable unlabeled data, which is relatively reliable. By using the identity classifier, these labeled and pseudo-labeled dataset are optimized. There are no reliable pseudo-labels or reliable utilization information in the index set. It is used to sort out the remaining unlabeled data in the iteration process. As the iteration progresses, the size of the candidate set changes dynamically. Finally, the Exclusive Loss is used to optimize the CNN model.

Training with Labeled Images

The most critical step in the entire transfer learning and one-example learning process is how to make full use of the labeled dataset. There are real identity labels in the label datasets L and L^e . Cross-entropy loss is used for optimization. Unlike the work of WU [6], our dataset was cross-camera style-transferred image, so that the original dataset was enlarged to K times of the current one. In the training process, for each identity, we randomly select one image from a camera for training. The objective function is then calculated as:

$$\min_{\theta, \omega} \sum_{i=1}^{Kn_l} l_{CE}(f(\omega; \phi(\theta; x_i)), y_i) \quad (7)$$

where x_i denote randomly picking an image from a camera as input for i -th person and y_i denote the identification label for i -th input image. f is an identity classifier parameterized by ω , which is used to classify feature embedding functions ϕ (parameterize by θ). l_{CE} denote cross-entropy loss:

$$l_{CE} = -[y \log(p) + (1 - y) \log(1 - p)] \quad (8)$$

where y denotes sample label, positive class is one and negative class is zero. p denotes the probability that the sample is predicted to be positive. l_{CE} represents the loss of the predicted label f and the real identity y_i , the smaller of this value the more similar to the real identity.

This process is particularly important for the iterative model training because a good initial training model is critical for subsequent training with pseudo-label data. This process will make full use of unlabeled data.

Training with Pseudo-Labeled Images

To make more use of the unlabeled images, we select the most reliable ones and assign them pseudo-labels for training. By using these images, whether or not a good optimization model can be obtained depends on the credibility of the pseudo-label. Selecting the most reliable unlabeled images for training will increase the robustness of the model, otherwise it can seriously damage the model.

Therefore, the processing of the label set is also processed using cross-entropy loss to optimize the training by:

$$\min_{\theta, \omega} \sum_{i=Kn_l+1}^{Kn_l+n_u} s_i l_{CE}(f(\omega; \phi(\theta; x_i)), \hat{y}_i) \quad (9)$$

where $s_i \in \{0, 1\}$ is the selection index of x_i (similarly, x_i denotes randomly picking an image from a camera as input for i -th person), x_i indicates the unlabeled samples, s_i is generated during the previous labeling of the former residence, and it determines whether to select pseudo-label data for identity classification training.

Training with Unlabeled Images

To better make use of the large amount of unlabeled data, we adopt the exclusive loss as an auxiliary loss for self-supervise. Use the next function to push some data x_i in the unlabeled set away from other data x_j in the feature map ($i \neq j$):

$$\max_{\theta} \sum_{x_i, x_j \in M^t} \|\phi(\theta; x_i) - \phi(\theta; x_j)\|, (x_i \neq x_j), \quad (10)$$

Feature memory $F \in R^{|\mathcal{F}^t| \times n_{\phi}}$ stores all target image features and update after each iteration. F_i represents the i -th column of F . Set $F_i = \tilde{\phi}(\theta; x_i)$ is L2-normalized feature embedding for the data x_i . Since $\|F_i - F_j\|_2 = 2 - F_i^T F_j$, maximizing the Euclidean distance between the data x_i and x_j equals minimizing the cosine similarity $F_i^T F_j$. Therefore, Equation (10) can be optimized by a softmax-like loss:

$$l_e(F; \tilde{\phi}(\theta; x_i)) = -\log \frac{\exp(F_i^T \tilde{\phi}(\theta; x_i) / \tau)}{\sum_{j=1}^{|\mathcal{F}^t|} \exp(F_j^T \tilde{\phi}(\theta; x_i) / \tau)} \quad (11)$$

where $\phi(\theta; \cdot)$ is our model, used to extract D -dimensional features for each picture; θ denotes weights of re-ID model; the hyperparameters τ represents the temperature fact of the softmax function. Higher temperature τ leads to weaker probability distributions. F has the following updates after each iteration:

$$F_i \leftarrow \mu F_i + (1 - \mu) \tilde{\phi}(\theta; x_i) \quad (12)$$

The hyperparameter μ is the update rate of F , which will not be fixed to a constant value, because it needs a smaller update to accelerate at the beginning of training. As the epochs increase, F must become gradually stable. At this time, a larger value of μ is required, so the entire process is gradually increasing.

Exclusive loss, as a self-supervised auxiliary loss, is mainly used to learn to distinguish representations from unlabeled data without identifying identity labels. In the iterative optimization of the model, the exclusivity loss mainly keep two different input images apart by CNN model. Therefore, in this process, more attention needs to be paid to the details of the input identity. This provides useful supervising information through the differences between images during iteration.

3.3.2. Selecting Reliable Images

The process of selecting reliable images and assigning pseudo-labels plays a vital role in making full use of unlabeled data. In the process of label estimation and screening, the most commonly used method is to extract unlabeled data according to classification loss, but the predicted value of classification loss cannot be well adapted to the detection evaluation in practical applications, and for single-label sample data, the classifier can easily lead to overfitting. The nearest method uses the distance in the feature space as a reference standard for the false label reliability, and uses the nearest neighbor (NN) classifier to assign false labels to unlabeled data according to the nearest labeled

neighbor. Euclidean distance is calculated according to the input features of the original dataset, and evaluate all unlabeled data $x_i \in U \cup U^e$ by Equation (13):

$$x^*, y^* = \underset{(x_l, y_l) \in L}{\operatorname{argmin}} \|\phi(\theta; x_i) - \phi(\theta; x_l)\| \quad (13)$$

Dissimilarity cost is evaluated according to Equation (14):

$$d(\theta; x_i) = \|\phi(\theta; x_i) - \phi(\theta; x^*)\| \quad (14)$$

In the t -th iteration, we use Equation (15) to select the unlabeled samples that are closest to the labeled samples:

$$s^t = \underset{\|s^t\|_0 = m_t}{\operatorname{argmin}} \sum_{i=Kn_l+1}^{K(n_l+n_u)} s_i d(\theta; x_i) \quad (15)$$

where K denotes the number of cameras in a certain original image dataset, and m^t denotes the size of the pseudo-label set selected in the t -th iteration. Then the best predicted value y^* is given to a pseudo-identity identification mark $\hat{y}_i$, and put it into the iterative training optimization model by setting the pseudo-label $(x_i, \hat{y}_i)$.

Regarding the iterative scheme, Equations (9) and (11) are optimized at each iteration. By estimating labels for unlabeled data by $\hat{y}_i = y^*$, reliable samples are selected for training by Equation (15). m^t is updated by Equation $m_t \leftarrow m_{t-1} + p \times (n_{l+k} + 2u)$, where $p \in (0, 1)$ and p is the enlarging factor which controls the sample size of the candidate set during the t -th iteration.

4. Experimental Analysis

4.1. Datasets

Market-1501 [42]: The dataset contains a training set and a test set, where the training set has 12,936 pictures of 751 pedestrians, and the test set has 19732 pictures of 750 pedestrians.

DukeMTMC-reID [43]: The data set contains a training set and a test set, where the training set has 16,522 images and the test set has 17,661 images.

4.2. Experiment Settings

For our one-example experiment based on data augmentation, we use the same protocol as [44]. In all datasets, in addition to randomly selecting an image from Camera 1 for initialization for each identity, we also added pictures that convert images under different styles of cameras. If there is no record of an identity under one camera, we will randomly select a sample from the next camera to ensure that there is one sample for each identity to initialize.

4.3. Experiment Details

We first remove the last classification layer of ResNet-50 to act as a feature embedding model. Initialization is made by using ImageNet [34] pre-trained model. Attach an additional fully connected layer with a batch normalization function and a classification layer on top of the CNN feature extractor to optimize the model through label and pseudo-label loss. For the loss of exclusivity, we process the unlabeled features through batch connection normalization of the fully connected layer, and then perform the normalization operation.

4.4. Comparison with the State-of-the-Art Methods

In Table 1, the Baseline (Supervised) shows the best performance of 100% labeling. Baseline (One-Example) shows the initial model trained by the single label method. "TJOF" indicates that the

joint objective function [6] is the latest method in the study of single labeling of image datasets, and “Ours” is our method.

Table 1. Compared with the state-of-the-art methods on Market-1501.

Settings	Methods	Market-1501				
		Rank-1	Rank-5	Rank-10	Rank-20	mAP
Supervised	Baseline [35]	83.1	92.5	95.0	96.9	63.7
	Baseline [35]	26.0	41.4	49.2	59.6	9.0
One-Example	DGM+IDE [12]	36.8	54.0	59.6	68.5	16.9
	Stepwise [9]	41.2	55.6	62.2	66.8	19.7
	TJOF [6]	55.8	72.3	78.4	83.5	26.2
	Ours	64.5	78.5	83.7	87.7	31.4

We compare our method to two baseline methods. The Baseline (Supervised) is conducted in a fully supervised environment, all data is labeled and used in training. Baseline (One-Example) uses only one labeled image as training set for each person. Specifically, we achieve 38.5, 37.1, 34.5 and 28.1 points of rank-1, rank-5, rank-10 and rank-20 accuracy improvement over the Baseline (one-example) on Market-1501, respectively. On the mAP accuracy, we achieve 22.4 points improvement over the Baseline (one-example).

The joint learning method (TJOF) is proposed by Wu et al, which use one-example data, pseudo-labeled data and index data for training. We achieve 8.7, 6.2, 5.3 and 4.2 points of rank-1, rank-5, rank-10 and rank-20 accuracy improvement over the TJOF on Market-1501, respectively. On the mAP accuracy, we achieve 5.2 points improvement over the TJOF.

As shown in Table 2, we achieve 29.1, 26.1, 22.6 and 16.3 points of rank-1, rank-5, rank-10 and rank-20 accuracy improvement over the Baseline(one-example) on Duke-MTMC-reID, respectively. On the mAP accuracy, we achieve 21.0 points improvement over the Baseline(one-example) on DukeMTMC-reID. We achieve 6.3, 4.1, 3.4 and 2.8 points of rank-1, rank-5, rank-10 and rank-20 accuracy improvement over the TJOF, respectively. On the mAP accuracy, we achieve 1.5 points improvement over the TJOF. Our method shows superior performance in both large-scale datasets.

Table 2. Compared with the state-of-the-art methods on DukeMTMC-reID.

Settings	Methods	DukeMTMC-reID				
		Rank-1	Rank-5	Rank-10	Rank-20	mAP
Supervised	Baseline [35]	83.1	92.5	95.0	96.9	63.7
	Baseline [35]	26.0	41.4	49.2	59.6	9.0
One-Example	DGM+IDE [12]	32.4	47.9	53.8	65.3	11.2
	Stepwise [9]	46.3	60.4	64.6	70.2	20.8
	TJOF [6]	48.8	63.4	68.4	73.1	28.5
	Ours	55.1	67.5	71.8	75.9	30.0

4.5. Ablation Studies

In Table 3, we list the current state-of-the-art methods compared with ours. “Ours (W.O/T)” indicates that our method does not use the results of style-transferred image data during training. “Ours (W/T)” represents the results of our method in training using style-transferred image.

Table 3. Comparison of Ours (W.O/T) and Ours (W/T).

Method	Market-1501					DukeMTMC-reID				
	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP
TJOF	55.8	72.3	78.4	83.5	26.2	48.8	63.4	68.4	73.1	28.5
Ours(W.O/T)	58.9	75.2	80.8	85.7	30.2	51.3	65.6	70.8	75.0	29.3
Ours(W/T)	64.5	78.5	83.7	87.7	31.4	55.1	67.5	71.8	75.9	30.0

Using original image for training and using style-transferred image for training: When not using style-transferred image for training, we achieve 3.1 and 4.0 points of rank-1 accuracy and mAP accuracy improvement over the TJOF on Market-1501. There is a better performance when we using style-transferred image for training. We achieve 5.6 and 1.2 points of rank-1 accuracy and mAP accuracy improvement over the Ours(W.O/T) on Market-1501. Similarly, on DukeMTMC-reID, we achieve 3.8 and 0.7 points of rank-1 accuracy and mAP accuracy improvement over the Ours (W.O/T).

In Figures 5 and 6, the enlarge factor ϵ is set to 0.05. First of all, from the set of control experiments in Figure 6, we can clearly conclude that the introduction of style-transferred image data during training will be more robust than the model trained with only the original dataset. Secondly, it can be clearly seen that as a higher and higher proportion of unlabeled data is added to training, the Rank-1 accuracy and average accuracy mAP will increase slowly to a certain extent, or even decline. This is because as the unlabeled data gradually increases during the iterative training process, there is less and less reliable pseudo-labeled data to choose from in the later iterative process, so that more noise is introduced during the training process. As a result, there is a certain damage to the robustness of the model.

Figure 5. Ablation studies on Market-1501. Rank-1 accuracy on the evaluation set during iteration. “Ours (W.O/T)” indicates that cross-camera-style transformation data is not added during training, and “Ours (W/T)” indicates that style-transfers image data is added during training. The X-axis indicates the percentage of the selected data to the entire unlabeled data, and the filled dot represents the result of one iteration.

Figure 6. Ablation studies on Market-1501. Average accuracy mAP on the evaluation set during iteration. “Ours (W.O/T)” indicates that cross-camera-style transformation data is not added during training, and “Ours (W/T)” indicates that style-transfers image data is added during training. The X-axis indicates the percentage of the selected data to the entire unlabeled data, and the filled dot represents the result of one iteration. Each point indicates an iteration step.

Analysis over iterations: To further analysis over iterations, we list more detailed data in Figures 7 and 8.

Figure 7 illustrates the label estimation performance and re-ID performance over iterations. As the iteration goes on, the precision score decreases gradually, because there are more and more unlabeled data in the training process. However, with more correctly estimated pseudo-labeled data are used, the recall score of label estimation gradually increases. Comparison with Figures 5, 6 and 7b, the rank-1, mAP and recall score curve: Although the curves are all rising, they are different. Rank-1 and mAP score are rising first and then falling, while recall score is always rising.

(a)

(b)

Figure 7. Ablation studies on Market-1501. The enlarging factor p is set to 0.05 in comparison. “Ours (W.O/T)” indicates that cross-camera-style transformation data is not added during training, and “Ours (W/T)” indicates that style-transfers image data is added during training. (a): Precision and recall of the label prediction of selected pseudo-labeled data. (b): The x-axis represents the selected data as a percentage of the entire unlabeled data.

On DukeMTMC-reID, we have achieved an accuracy of nearly 94.7% higher than Wu’s method of 70.0%, and the recall accuracy reached 53.6% at the last iteration, which is also higher than any

other one-example labeling methods. This result shows that the use of style transferred data can well solve the problem of over-fitting caused by training pictures under a single camera. To some extent, it overcomes the difficulty in pedestrian recognition under different cameras.

The effects of enlarging factor on the experiments: the enlarging factor p controls the speed of expanding the pseudo-label candidate set during the iterative process. It can be seen from Table 4 that the smaller the enlarging factor p , the better the experimental results. Smaller magnification results in greatly improved model stability in iteratively model training, but it takes longer and more iterative steps. Therefore, in this paper, we set p to 0.05.

(a)

(b)

Figure 8. Ablation studies on DukeMTMC-reID. The enlarging factor p is set to 0.05 in comparison. “Ours (W.O/T)” indicates that cross-camera-style transformation data is not added during training, and “Ours (W/T)” indicates that style-transfers image data is added during training. (a): Precision and recall of the label prediction of selected pseudo-labeled data. (b): The x-axis represents the selected data as a percentage of the entire unlabeled data.

4.6. Discussion

In this paper, CycleGAN is used to transfer the style of the images and generate pedestrian images under different cameras. After the style transfer, according to the setting of one-example person re-ID, two large person re-ID data sets, Market-1501 and DukeMTMC-reID, were tested. In this paper, through the way of cross camera style-transferred, the sample expansion of the recent one-example experiment is completed. The experimental results in two large datasets show the effectiveness of this method. In the process of training, through the contrast experiment of whether to use style-transferred data, it shows that using style-transferred in the process of training can improve the performance of the model.

Compared with the recent one-example person re-ID experiments, our method is much better than other methods in performance even if it does not use the style-transferred for training. Compared with Wu's work in the same data set, our accuracy is nearly 89.33%, higher than 70%, and 45.6% in the last iteration of recall rate, which is also higher than other methods based on one-example improvement. When using migration data for training, almost all of the performance will be superior to the performance on the original data set. Such results show that our method can solve the problem of over fitting of training images under a single camera, and also solve the problem of identifying a pedestrian under different cameras to a certain extent.

Table 4. p , the enlarging factor.

p	Methods	Market-1501		DukeMTMC-reID	
		Rank-1	mAP	Rank-1	mAP
$p = 0.20$	Ours(W.O/T)	43.1	19.3	40.0	17.6
	Ours(W/T)	50.2	22.5	42.1	19.5
$p = 0.10$	Ours(W.O/T)	53.3	27.2	49.3	22.6
	Ours(W/T)	60.1	28.6	50.2	25.3
$p = 0.05$	Ours(W.O/T)	58.9	30.2	51.3	29.3
	Ours(W/T)	64.5	31.4	55.1	30.0

5. Conclusions

In the research of person re-ID, it is not realistic to rely on complete manual label. This paper mainly studies person re-ID based on one-example. One-example person re-ID only provides less label data and only one image of a pedestrian is labeled, so the performance of the initialization model is not high. First of all, through the transformation of camera style, we make the amount of data become K times of the original data set, so that the labeled samples used for initialization training also become K times of the original. Then, the initial training of the model is carried out by the original images (real) and the style-transferred images (generated). This initial training method makes our model more robust, and the effectiveness of this method is verified by experiments. This method provides a good initialization model for subsequent iterations. After getting a good initialization model, in the next experiment, we mainly make full use of the unlabeled data and the style-transferred data by the way of step-by-step iteration. We focus on the analysis of the factors that affect the effect of the iteration, and study the changes of each evaluation index with the iteration process. In the process of training, the style-transferred data is put into training, which greatly improves the recognition performance. Our method not only reduces the risk of over-fitting when using one camera takes pictures, but also provides a new idea of using unlabeled data efficiently.

According to the main work of this paper, there are some improvements that can be made: the features used for classification in this paper are calculated based on the original images (real). This method might not be able to classify the style-transferred images (generated) efficiently. In subsequent experiments, you can use the average of the original data feature and the style-transferred data feature as the feature, and then use the average feature to calculate the distance to perform image

classification. In addition, the calculation characteristics of transferred data may make the model more stable and improve the re-ID performance.

Author Contributions: Supervision: L.G.; validation: Q.L.; writing—original draft, Q.L.; writing—review and editing: L.G. All authors have read and agreed to the published version of the manuscript.

Funding: The work is partially supported by the National Natural Science Foundation of China (Nos.61873151, 61672329).

Acknowledgments: We gratefully acknowledge the assistance of Vana and Hua-Xiang Zhang's laboratory.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, L.; Wang, G.; et al. Recent advances in convolutional neural networks. *arXiv* **2015**, arXiv:1512.07108. [[CrossRef](#)]
- Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 1097–1105.
- Tetsu, M.; Takahiro, O.; Einoshin, S.; Yoichi, S. Hierarchical gaussian descriptor for person reidentification. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1631–1642.
- Liao, S.; Hu, Y.; Zhu, X.Y.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
- Zhou, Z. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 48–57. [[CrossRef](#)]
- Wu, Y.; Lin, Y.; Dong, X. Progressive Learning for Person Re-Identification with One Example. *IEEE Trans. Image Process.* **2019**, *28*, 2872–2881. [[CrossRef](#)] [[PubMed](#)]
- Yu, H.X.; Wu, A.; Zheng, W.S. Unsupervised person reidentification by deep asymmetric metric embedding. *TPAMI* **2019**. [[CrossRef](#)]
- Fan, H.; Zheng, L.; Yan, C.; Yang, Y. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 1–18. [[CrossRef](#)]
- Wu, Y.; Lin, Y.; Dong, X. Exploit the Unknown Gradually: One-Shot Video-Based Person Re-identification by Stepwise Learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018.
- Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camera style adaptation for person re-identification. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018; pp. 654–664.
- Dong, X.; Meng, D.; Ma, F.; Yang, Y. A dual-network progressive approach to weakly supervised object detection. In Proceedings of the ACM Multi-Media Conference, Santiago, Chile, 11–18 October 2017; pp. 279–287.
- Zhu, J.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017; pp. 3661–3671.
- Dong, X.; Yan, Y.; Tan, M.; Yang, Y.; Tsang, I.W. Late fusion via subspace search with consistency preservation. *IEEE Trans. Image Process.* **2019**, *28*, 518–528. [[CrossRef](#)] [[PubMed](#)]
- Deng, C.; Chen, Z.; Liu, X.; Gao, X.; Tao, D. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.* **2018**, *27*, 3893–3903. [[CrossRef](#)] [[PubMed](#)]
- Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
- Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the 2015 International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 1116–1124.
- Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3376–3385.

18. Zheng, L.; Zheng, L.; Yang, Y. A discriminatively learned CNN embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 13.
19. Liu, Z.; Wang, D.; Lu, H. Stepwise metric promotion for unsupervised video person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2448–2457.
20. Ye, M.; Ma, A.J.; Zheng, L.; Yuen, P.C. Dynamic label graph matching for unsupervised video re-identification. In Proceedings of the IEEE International conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5152–5160.
21. Kumar, M.P.; Packer, B.; Koller, D. Self-paced learning for latent variable models. In Proceedings of the Advances in Neural Information Processing Systems, Boston, MA, USA, 7–12 June 2010; pp. 1189–1197.
22. Ulyanov, D.; Lebedev, V.; Vedaldi, A.; Lempitsky, V.S. Texture networks: Feed-forward synthesis of textures and stylized images. In Proceedings of the ICML, Taipei, Taiwan, 20–24 November 2016; pp. 791–808.
23. Roy, S.; Paul, S.; Young, N.E.; Roy-Chowdhury, A.K. Exploiting transitivity for learning person re-identification models on a budget. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5265–5274.
24. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Yang, Y. Improving person Re-Identification by Attribute and Identity Learning. Available online: <https://arxiv.org/abs/1703.07220> (accessed on 6 June 2017).
25. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
26. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
27. Wang, X.; Gupta, A. Generative image modeling using style and structure adversarial networks. In Proceedings of the European Conference on Computer Vision, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1363–1372.
28. Wang, J.; Zhu, X.; Gong, S.; Wei, L. Transferable joint attribute-identity deep learning for unsupervised person reidentification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6738–6746.
29. Shu, R.; Bui, H.H.; Narui, H.; Ermon, S. A dirt-t approach to unsupervised domain adaptation. In Proceedings of the ICLR, Munich, Germany, 8–14 September 2018; pp. 480–496.
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, San Francisco, CA, USA, 13–18 June 2012; pp. 1097–1105.
31. Supancic, J.S.; Ramanan, D. Self-paced learning for long-term tracking. In Proceedings of the CVPR, Portland, OR, USA, 23–28 June 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 2379–2386.
32. Zhong, Z.; Zheng, L.; Zheng, Z.; Li, S.; Yang, Y. Camstyle: A novel data augmentation method for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 1176–1190. [[CrossRef](#)] [[PubMed](#)]
33. Li, J.; Ma, A.J.; Yuen, P.C. Semi-supervised region metric learning for person re-identification. *Int. J. Comput. Vis.* **2018**, *126*, 855–874. [[CrossRef](#)]
34. Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Style aggregated network for facial landmark detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 379–388.
35. Isola, P.; Zhu, J.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. *arXiv* **2016**, arXiv:1611.07004.
36. Liand, C.; Wand, M. Precomputed real-time texture synthesis with markovian generative adversarial networks. In Proceedings of the ECCV, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1349–1358.
37. Wei, L.; Zhang, S.; Gao, W. Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3117–3129.
38. Gatys, L.A.; Ecker, A.S.; Bethge, M. Imagestyletransfer using convolutional neural networks. In Proceedings of the CVPR, Barcelona, Spain, 5–10 December 2016; pp. 2667–2675.
39. Deng, W.; Zheng, L.; Ye, Q. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2631–2641.

40. Zhong, Z.; Zheng, L.; Li, S.; Yang, Y. Generalizing a person retrieval model hetero-and homogeneously. In Proceedings of the ECCV, Salt Lake City, UT, USA, 18–22 June 2018; pp. 420–429.
41. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3652–3661.
42. Liu, X.; Liu, W.; Mei, T.; Ma, H. PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. *IEEE Trans. Multimed.* **2018**, *20*, 645–658. [[CrossRef](#)]
43. Ma, H.; Liu, W. A progressive search paradigm for the Internet of things. *IEEE Multimed.* **2017**, *25*, 76–86. [[CrossRef](#)]
44. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. *arXiv* **2016**, arXiv:1605.05396.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).