IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 28, 2021, accepted February 20, 2021, date of publication March 1, 2021, date of current version March 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063000

Real-World Person Re-Identification via Super-Resolution and Semi-Supervised Methods

LIMIN XIA, JIAHUI ZHU, AND ZHIMIN YU
School of Automation, Central South University, Changsha 410083, China
Corresponding author: Zhimin Yu (yuzhimin@csu.edu.cn)

This work was supported by the Hunan Science and Technology Project under Grant 2017GK2271.

ABSTRACT Person re-identification has made great progress over the years. However, due to the problem of super-resolution and few labeled samples, it is difficult to apply in practice. In this paper, we propose a semi-supervised super-resolution person re-identification method based on soft multi-labels. Firstly, a Mixed-Space Super-Resolution model (MSSR) is constructed based on Generative Adversarial Networks (GAN), which aims to convert low-resolution person images into high-resolution images. Secondly, a Part-based Graph Convolutional Network (PGCN) is proposed to extract discriminative feature by exploring the relationship of local features within person. Finally, to solve the problem of label limitation, we use the PGCN trained with a small amount of labeled samples to predict the soft multi-labels of unlabeled samples, and further train PGCN with unlabeled samples based on a novel multi-label similarity loss. Experiments have been conducted on the Market1501, CUHK03, and MSMT17 datasets to evaluate this method, which show that it outperforms other semi-supervised methods.

INDEX TERMS Person re-identification, GAN, super-resolution, multi-labels, semi-supervised.

I. INTRODUCTION

Person re-identification [1]–[3], which aims to find a given person from non-overlapping cameras, is a research hotspot in the computer vision field with great progress. Due to changes of human pose, viewpoint, illumination condition, resolution and so on, there are still improvements in person re-identification.

Most of the current methods are based on the assumption that all images in the query and gallery are fully high-resolution. A small part of related works use super-resolution models to process images [4]–[8]. However, these models specify the resolution of the input image, making it difficult to apply in practice on a large scale.

In real-world applications, another problem with person re-identification is the lack of labels. A lot of methods try to use semi-supervised or unsupervised learning methods to solve the problem of few or even no labels [9]–[13]. A typical semi-supervised learning method is to use a teacher-student model to predict pseudo labels of unlabeled samples, and then the samples with pseudo labels and samples with ground truth labels are used to train the model jointly. However, the traditional methods require that each class in the labeled dataset contains at least one sample. When this requirement is not met, it is difficult for the teacher model to correctly predict the labels of unlabeled samples.

In this paper, we propose a semi-supervised super-resolution person re-identification method based on soft multi-labels. Firstly, we innovatively proposed a mixed-space super-resolution (MSSR) network. Compared with the traditional Super-Resolution (SR) models, the proposed MSSR network is adapted to input images with different resolutions and can convert them into a high- resolution space, which can complement the details of low-definition images. Secondly, a Part-base Graph Convolutional Network (PGCN) is proposed to extract discriminative features from high-resolution person images. PGCN can extract the local features of the person image and treat them as nodes of undirected graph, which are then dynamically updated through the Graph Convolutional Network (GCN) [14]–[17]. The PGCN is trained in semi-supervised way. Specifically, we first train PGCN with labeled samples by cross entropy loss. Then the trained PGCN is used to predict the soft multi-labeled of unlabeled samples. Finally, we train PGCN with unlabeled samples by a novel multi-label similarity loss. Compared with the traditional semi-supervised person re-identification methods, we do not directly predict the hard pseudo labels of unlabeled samples. Instead, a fully connected layer after the person

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Anisetti.

feature is utilized to predict the soft multi-labels, which can better represent the distribution of unlabeled images. An illustration of soft multi-labels is shown in Figure 1.

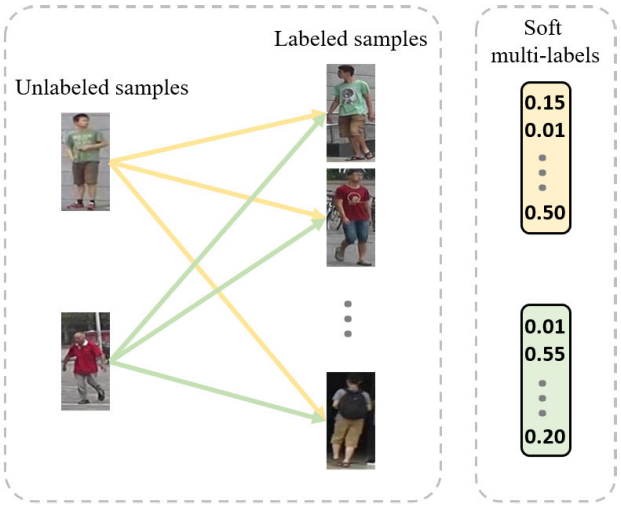

FIGURE 1. Illustration of soft multi-labels. We use the labeled samples to represent the unlabeled samples and regard the similarities as the soft multi-label.

In summary, the main contributions of this paper are as follows:

1) A Mixed-Space Super-Resolution model (MSSR) is proposed. The model has no resolution requirements for the input images, but can generate high-resolution person images as output to compensate for the loss of details in the long-distance images.
2) We propose a Part-based Graph Convolutional Network (PGCN). By exploiting the relationship between local features, PGCN can extract more discriminative features.
3) A semi-supervised person re-identification method is proposed based on soft multi-labels. Compared with the traditional teacher-student model, we use multi-labels in the label prediction stage to preserve the possibility of teacher model classification errors, which is not troubled by unseen person.
4) A multi-label similarity loss function is proposed. This loss function forces person samples with similar multi-labels to be close to each other in the feature space, and person samples with different multi-label are far away from each other in the feature space.
5) Experiments on different datasets show that the proposed semi-supervised person re-identification method based on multi-label has achieved excellent results.

II. RELATED WORKS

A. SUPER-RESOLUTION RE-ID

The main purpose of super-resolution person re-identification is to match the high-resolution images in the query and gallery with low-resolution. In the most person re-identification datasets, the image sizes of the query and gallery are consistent, so there are a few researches focus on super-resolution challenge in person re-identification. Generally, to ensure the size of the query images are consistent with the gallery, most datasets usually crop and enlarge the image, which also leads to the problem of inconsistent image resolution in the dataset. There are two methods of super-resolution person re-identification at present, which can be roughly divided into two aspects: namely extracting resolution invariant person features or constructing image resolution transformation models with different resolutions. In this paper, we call the former as one-stage method and call the latter as two-stage method.

The typical of classic one-stage method is to learn the feature transformation between different resolution images. For example, Jing *et al.* [4] proposed a semi-coupled low-rank discriminant dictionary learning (SLD^2L). They learned their dictionaries and feature maps from a pair of high-resolution images and low-resolution images, aimed to force the positive and negative features to be close to each other after feature transformation. In addition, they also used a regularization term to avoid overfitting. Wang *et al.* [5] considered that the person proportional distance function can distinguish person on different scales, so it is used to learn the person feature representation and classify the person in the corresponding feature space. Li *et al.* [18] also jointly learn person features on different scales. Different from [5], they proposed a joint multi-scale learning framework named joint multi-scale discriminant component analysis (JUDEA). This framework used the heterogeneous class mean discrepancy criterion for cross scale image alignment, and optimized jointly with the discriminant functions of multiple scales.

Different from the one-stage method, the two-stage method usually maps the low-resolution image to the high-resolution space at the first step. Since Ledig [6] first proposed SR-GAN, it has been widely used in the clarity model. SR-GAN combined anti-loss and content-loss to restore the lost image details. Jiao *et al.* [7] proposed a joint learning method of Super-resolution and Identity joiNt learninG (SING), which used a hybrid neural network to improve the performance of cross resolution person re-identification. They downsampled the original high-resolution image to reduce the resolution. The resolution-reduced image and the original low resolution image were sent into a two stream super-resolution network. Meanwhile, they built a super-resolution loss to limit the cross-resolution loss of the resolution-reduced images and low-resolution images. Finally, all of the image will be sent to a classification network for person identification. Wang *et al.* [8] cascaded several SR-GANs to improve the processing ability of the framework for different resolution images, and proposed a cascaded super-resolution GAN model (CSRGAN). CSRGAN utilized the cascaded SR-GAN to complete scale adaptive clarification. On this foundation, the super-resolution images and the original images will be used to identification. In addition, they designed an ordinary person loss for high-resolution and low-resolution image pairs with

the same person to ensure that the generated images are more like human, and designed special person loss for the image pairs with different person to generate discriminative images.

However, the traditional methods can only deal with the transformation from low resolution image to high resolution image between two fixed resolutions. In practical applications, since the given person image size is unknown, the resolution requirements of the models are not consistent. It may lead to the need for multiple definition conversion models, which increases the difficulty of training and the complexity of practical application. All in all, traditional method requires multiple resolution transformation models, which makes training harder and more complex.

B. GRAPH CONVOLUTIONAL NETWORKS

Graph convolutional network is a deep learning algorithm that combines graph neural network and convolutional neural network. Graph convolutional networks are widely used in scene graph generation, point cloud classification and segmentation, action recognition, etc.

To estimate the similarities of different query-gallery pair, Shen *et al.* [14] treats different image pairs as the node in the graph and propose the similarity-guided graph neural network (SGGNN). By updating the relationship features between different image pairs dynamically, SGGNN can learn the most suitable similarity for hard sample pairs. Liu *et al.* [15] regard each person image as the node in the graph, so person re-identification has become a connection problem between nodes. He proposed the probability GCN (PrGCN) to solve the deformation problem of person re-identification, which constructs a subgraph for each person to represent the characteristics and then input it into the GCN to assume the link probability between the nodes. Yang *et al.* [16] propose spatial-temporal graph concolutional network (STGAN) to model the different frames of video person re-identification by two GCN streams, the one extract the structural features and the other discriminatory cues. Chen *et al.* [17] propose re-ranking via graph convolution channel attention network (RRGCCAN), which incorporates the attention mechanism into the graph convolutional network so that the graph model on the feature subset is introduced into the initial ranking.

C. SEMI-SUPERVISED RE-ID

At present, the semi-supervised person re-identification method has attracted the attention of many researchers. In 2013, Figueira *et al.* [9] have tried to utilize the semi-supervised method to carry out the person re-identification task. They combined visual features with learning methods, and used multi-classification learning method to fuse the given person features. However, this semi-supervised person re-identification method requires at least one image for each person in training step. In order to eliminate the change of person appearance across perspectives, Liu *et al.* [10] learned two coupling dictionaries from the gallery to the detected image by using the labeled and unlabeled samples in the training phase. They used labeled training images to learn the feature relationship between different perspectives, and introduced unlabeled data to obtain robust feature representation by using geometric edge distribution. Zhu *et al.* [11] proposed a semi-supervised cross-view projection-based dictionary learning method (SCPDL). In this method, they learned a pair of feature projection matrices and a pair of dictionaries jointly by integrating information from labeled and unlabeled images. Then the feature projection matrix was used to reduce the impact of the change of perspective. The learned dictionary transforms the person video from two perspectives into coding coefficients in the common space. In the training step, the labeled data ensures the distinguishability of the dictionary, while the unlabeled data increases the difference between people.

Different from the aforementioned methods using unlabeled data to enhance the robustness of the model, the lack of datasets in person re-identification often requires enhance data. However, the unlabeled samples in the generation model also brings great trouble to researchers. In order to solve this problem, Zheng *et al.* [12] used GAN to generate unlabeled data, and the labeled data together with generated unlabeled data are used to train the model. For the data labels generated by GAN, they used label smoothing regularization of outliers (LSRO) to assign uniform-distribution labels for unlabeled data, which will be together used with the original labeled data to train this model. Zhang *et al.* [13] used similarity-embedded cycle GANs (SECGAN) to expand the data for person re-identification. Compared with the traditional cycle-GAN, SECGAN used a few of labeled samples for cross perspective feature learning. For the person image generation, the similarity module can ensure that the labels of person images will not change after transformation.

Compared with the aforementioned semi-supervised person re-identification method, the proposed semi-supervised method based on multi label can make full use of the information of unlabeled data instead of only use simple information. In addition, it can avoid the single sample problem in the teacher-student model. Although multi-label methods are also used in the literature [19] for unsupervised person re-identification, there is a huge difference between the our method and them: (1) In the use of multi-label data, our method makes full use of all samples and assigns weight to each image pair, while they only uses samples that are easier to distinguish; (2) In [19], the accuracy of multi-label prediction through agents is relatively low, our method uses multi-label for semi-supervised learning, which can ensure the accuracy of multi label prediction.

III. PROPOSED METHOD

Our proposed method mainly consists of two stages. The first stage aims to transform the input images into high resolution. To achieve this purpose, we construct a Mixed-Space Super-Resolution (MSSR) network based on GAN, which can process images with different resolution. The MSSR network consists of both coarse generation module and fine generation module. The coarse generation module takes half size images

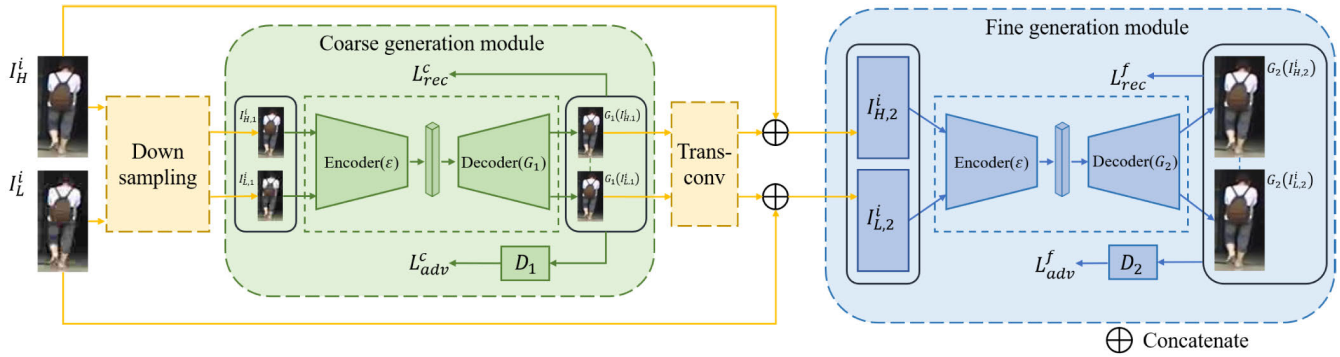

FIGURE 2. Architecture of the proposed MSSR network, which consists of both coarse generation module and fine generation module. The coarse generation module takes half size images as input and outputs coarse restored images. The fine generation module takes the concatenated images as input and outputs the fine restored images.

as input and outputs coarse restored images. Then we resize the coarse restored images and concatenate them with their corresponding original full size images to get the input of fine generation module, which is used to restored the fine-grained information. The second stage is semi-supervised person re-identification. In this stage, we first construct a Part-based Graph Convolutional Network (PGCN) to extract the features of high resolution images generated by MSSR network. Compared with traditional part-based representation models, our PGCN can exploit the contextual information between different local features. Firstly, PGCN is trained with labeled samples by cross entropy loss. Then we adopt the trained PGCN to predict the soft multi-labels of unlabeled samples. Finally, we use labeled samples with cross entropy loss and unlabeled samples with a novel multi-label similarity loss to jointly fine-tune PGCN. The details of each stage is presented in the following subsections, respectively.

A. MIXED-SPACE SUPER-RESOLUTION NETWORK

The Mixed-Space Super-Resolution (MSSR) network is shown in Figure 2. As we can see, the inputs of our MSSR network are consisted of high resolution image and low resolution image. To construct the training set of MSSR, we first downsample the original person images by a local average pooling operation, and then resize the downsampling images to the size of original images, so as to obtain the low resolution image set $I_L = \{I_L^1, I_L^2, \cdots, I_L^N\}$. The original images are taken as the high resolution image set $I_H = \{I_H^1, I_H^2, \cdots, I_H^N\}$.

Given a high resolution image $I_H^i \in \mathbb{R}^{h\times w\times 3}$ from I_H and the corresponding low resolution image $I_L^i \in \mathbb{R}^{h\times w\times 3}$ from I_L, we first downsample them by a 2×2 local average pooling operation to get the half size images $I_{H,1}^i \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$ and $I_{L,1}^i \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$. Then we input $I_{H,1}^i$ and $I_{L,1}^i$ into the coarse generation module and get the generated images $G_1\left(I_{H,1}^i\right)$ and $G_1\left(I_{L,1}^i\right)$. In order to ensure that the generated images are of high resolution, we construct the following loss functions:

$$L_{adv}^c(G_1, D_1) = \mathbb{E}\left[logD_1\left(G_1\left(I_{H,1}^i\right)\right)\right] + \mathbb{E}\left[log\left(1 - D_1\left(G_1\left(I_{L,1}^i\right)\right)\right)\right] \quad (1)$$

$$L_{res}^c(G_1) = \mathbb{E}\left[\left\|G_1\left(I_{H,1}^i\right) - I_{H,1}^i\right\|_1\right] + \mathbb{E}\left[\left\|G_1\left(I_{L,1}^i\right) - I_{H,1}^i\right\|_1\right] \quad (2)$$

Equation (1) is the adversarial loss, which is used to guarantee that $G_1\left(I_{H,1}^i\right)$ and $G_1\left(I_{L,1}^i\right)$ are of the same resolution. Equation (2) is the restoration loss, which is used to ensure that the generated images are of high resolution.

Traditional GAN-based Super-Resolution (SR) models only apply one GAN structure for resolution restoration. However, the details of high resolution images generated by these methods are often disordered. To solve this problem, we adopt a multi-scale architecture for our proposed MSSR network to mimic conventional coarse-to-fine optimization methods. Specifically, we first perform a transposed convolution [20] operation on $G_1\left(I_{H,1}^i\right)$ and $G_1\left(I_{L,1}^i\right)$ to restore them to the size of $h\times w\times 3$, denote as $G_1^R\left(I_{H,1}^i\right) \in \mathbb{R}^{h\times w\times 3}$ and $G_1^R\left(I_{L,1}^i\right) \in \mathbb{R}^{h\times w\times 3}$. Then we concatenate $G_1^R\left(I_{H,1}^i\right)$ and $G_1^R\left(I_{L,1}^i\right)$ with their corresponding original input images respectively, formulated as:

$$I_{H,2}^i = G_1^R\left(I_{H,1}^i\right) \oplus I_H^i \quad (3)$$

$$I_{L,2}^i = G_1^R\left(I_{L,1}^i\right) \oplus I_L^i \quad (4)$$

where $I_{H,2}^i \in \mathbb{R}^{h\times w\times 6}$, $I_{L,2}^i \in \mathbb{R}^{h\times w\times 6}$, $\oplus$ represents concatenation. Then $I_{H,2}^i$ and $I_{L,2}^i$ are input into the fine generation model for finer image generating. The loss functions are as follows:

$$L_{adv}^f(G_2, D_2) = \mathbb{E}\left[logD_2\left(G_2\left(I_{H,2}^i\right)\right)\right] + \mathbb{E}\left[log\left(1 - D_2\left(G_2\left(I_{L,2}^i\right)\right)\right)\right] \quad (5)$$

$$L_{res}^f(G_2) = \mathbb{E}\left[\left\|G_2\left(I_{H,2}^i\right) - I_{H,2}^i\right\|_2\right] + \mathbb{E}\left[\left\|G_2\left(I_{L,2}^i\right) - I_{H,2}^i\right\|_2\right] \quad (6)$$

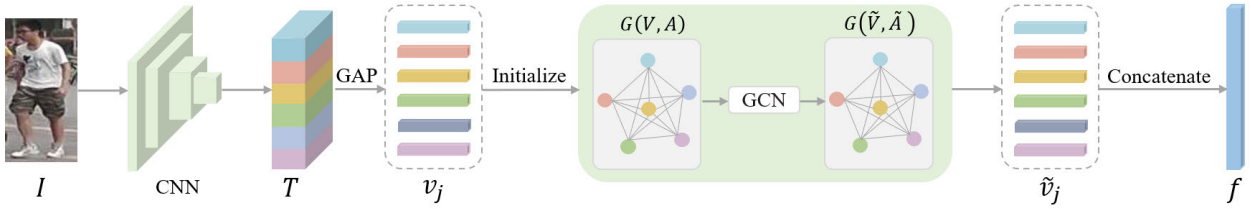

FIGURE 3. Overview of the proposed PGCN. Given a person image, PGCN first extracts its p local features and regard them as nodes of a graph. Then the graph is input into a GCN to exploit the contextual information between different local features. Finally, the local features updated by GCN are concatenated to obtain the final feature, which will be used for ID prediction.

After the coarse-to-fine optimization, we obtain the final resolution restored images $G_2\left(I^i_{H,2}\right) \in \mathbb{R}^{h\times w\times 3}$ and $G_2\left(I^i_{L,2}\right) \in \mathbb{R}^{h\times w\times 3}$. Compared with high resolution images generated by other methods, $G_2\left(I^i_{H,2}\right)$ and $G_2\left(I^i_{L,2}\right)$ contain more details and are closer to the real images. The final loss of our MSSR network is as follow:

$$L = (1-\alpha)\left(L\left(G_1, D_1\right) + L\left(G_1\right)\right) + \alpha\left(L\left(G_2, D_2\right) + L\left(G_2\right)\right) \quad (7)$$

where α is a balance weight, we will discuss its value in experiment section.

B. PART-BASED GRAPH CONVOLUTIONAL NETWORK

To exploit the contextual information between different human body parts, we propose a Part-based Graph Convolutional Network (PGCN) for person re-identification, which is shown in Figure 3.

Given a person image I, PGCN extracts its global feature map T by a CNN at first. Then we partition T into p horizontal stripes and average all the column vectors in a same stripe into a single part-level column vector v_j by global average pooling, where $j = \{1, 2, \cdots, p\}$.

Based on above local features, we construct a graph $G(V, A)$ to represent the relationships among different parts. Where $V = \{v_1, v_2, \cdots, v_p\}$ is the graph nodes and A is the adjacent matrix denoting the connections of nodes. Then we input $G(V, A)$ to a GCN that includes L hidden layers to propagate node information, formulated as:

$$H^l = \sigma\left(\hat{A}H^{l-1}W^{l-1}\right) \quad (8)$$

where $\hat{A}$ is the adjacent matrix initialized with A, W^{l-1} is the network parameter of layer $l-1$, σ is the ReLU activation function, H is the feature matrix initialized with V. After L graph convolution layers, the nodes are interacted fully and the new graph $G(\tilde{V}, \tilde{A})$ is obtained. Different from traditional GCN, we will jointly updated $\hat{A}$ during training to learn better connections.

Finally, we take the nodes of $G(\tilde{V}, \tilde{A})$ as the final local features $\tilde{v}_j$ and concatenate them to obtain the final feature f of input person image, formulated as:

$$f = \tilde{v}_1 \oplus \tilde{v}_2 \oplus \cdots \oplus \tilde{v}_p \quad (9)$$

In the test phase, PGCN will be used to extract features of probe and gallery person images. Then we will introduce how to train PGCN in the next subsection.

C. SEMI-SUPERVISED PERSON RE-IDENTIFICATION BASED ON SOFT MULTI-LABELS

In real-world applications, labeled samples usually are limited. In this subsection, we introduce how to train our PGCN in semi-supervised way.

1) TRAINING WITH LABELED SAMPLES

In semi-supervised person re-identification, we have a small person image set with annotations, denoted as $\{I^i_a, y^i_a\}_{i=1}^{N_a}$, where I^i_a represents the labeled image, y^i_a is the ID label of image I^i_a, N_a is the number of labeled images. Firstly, we train our PGCN with these labeled samples, which is shown in Figure 4 step 1 (blue box). Input an image I^i_a, PGCN extracts its feature f^i_a and inputs it into a FC layer to get the predict identity $\hat{y}^i_a$. Then a cross entropy loss is followed, formulated as:

$$L_{CE} = \frac{1}{N_a}\sum_{i=1}^{N_a}\sum_{c=1}^{M} y^{ic}_a log\left(\hat{y}^{ic}_a\right) \quad (10)$$

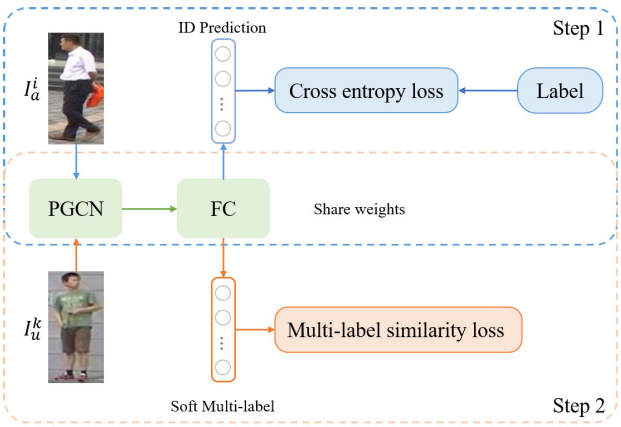

FIGURE 4. The training process of PGCN, which can be divided into two steps. Step 1 trains PGCN with labeled samples by a cross entropy loss. Step 2 trains PGCN with unlabeled samples by a novel multi-label similarity loss.

where M is the number of identities of labeled images, $\hat{y}_a^{ic}$ is the predicted probability that image I_a^i belongs to identity c, y_a^{ic} is the ground truth that image I_a^i belongs to identity c.

2) TRAINING WITH UNLABELED SAMPLES

After training with labeled samples, PGCN has certain person re-identification capability. When input a new person image (unlabeled sample), we believe that PGCN can determine which labeled pedestrians the image is similar to. With this in mind, we use the trained PGCN (including FC layer) to predict the soft multi-labels of unlabeled samples at first. As shown in Figure 4 step 2 (original box), given an unlabeled person image I_u^k, we input it into the trained PGCN+FC to get the soft multi-label, denoted as y_s^k.

In order to train the model with unlabeled samples, we propose a novel multi-label similarity loss based on above soft multi-labels. Specifically, we first define the similarity between different multi-labels, formulated as:

$$S(y_s^{k_1}, y_s^{k_2}) = 1 - \frac{\left\| y_s^{k_1} - y_s^{k_2} \right\|_1}{2} \tag{11}$$

where $\|\cdot\|_1$ represents the L1 norm. $S(y_s^{k_1}, y_s^{k_2}) \in [0, 1]$, the greater the difference between multi-labels, the smaller the similarity $S(y_s^{k_1}, y_s^{k_2})$. When two multi-labels are identical, $S(y_s^{k_1}, y_s^{k_2}) = 1$.

The similarity of multi-labels represents the proximity of pedestrians to a certain extent. Our goal is to make sample features with similar multi-labels close to each other, and sample features with large multi-label differences are far away from each other. And the multi-label similarity loss is defined as:

$$L_S = \frac{\sum_{k=1}^{N_u} \sigma\left(S\left(y_s^{k_1}, y_s^{k_2}\right) - \tau\right) \left\| f^{k_1} - f^{k_2} \right\|_2}{\sum_{k=1}^{N_u} \left\| S\left(y_s^{k_1}, y_s^{k_2}\right) - \tau \right\|_1 \left\| f^{k_1} - f^{k_2} \right\|_2} \tag{12}$$

where N_u is the batchsize, f^{k_1} and f^{k_2} represent the features of samples $I_u^{k_1}$ and $I_u^{k_2}$ respectively, σ is the ReLU function, τ is a hyperparameter. When $S(y_s^{k_1}, y_s^{k_2}) > \tau$, we think $I_u^{k_1}$ and $I_u^{k_2}$ are positive sample pair, otherwise are negative sample pair. Our multi-label similarity loss aims to reduce the feature distance between positive samples.

3) TRAINING AND TEST DETAILS

After obtaining the soft multi-labels of the unlabeled samples, we keep them unchanged during training. In addition, PGCN is trained by jointly optimized the cross entropy loss of labeled samples and the multi-label similarity loss of unlabeled samples.

In the testing phase, given a query image, we first use PGCN to extract its feature. Then we compute the cosine distance between the query image feature and all gallery image features. The final ranking list is sorted according to the distance from small to large.

IV. EXPERIMENTS

A. DATASETS AND SETTINGS

To demonstrate the effectiveness of proposed method, we evaluate it on three datasets.

1) Market-1501 [21]

The Market-1501 dataset is composed of 32668 detected images taken by 6 cameras and 1501 person in these images. Each person is captured by at least 2 cameras, leading to multiple images in one camera.

2) CUHK03 [22]

The CUHK03 dataset is the first large-scale person re-identification dataset sufficient for deep learning, which contains 13,164 images of 1,360 person taken by 6 cameras.

3) MSMT17 [23]

The MSMT17 dataset is currently the closest dataset to the real scene. Previous person re-identification datasets are often collected in a short period of time (ie. the Market1501 dataset is collected in summer), which limits the appearance and wear of person. In addition, in order to enhance the difference in lighting in the images, the MSMT17 dataset was taken indoors and outdoors. The MSMT17 dataset contains 126,441 images of 4104 person.

4) EVALUATION METRICS

Following previous works, we adopt the cumulative matching characteristic (CMC) and mean Average Precision (mAP) to evaluate the methods' performances.

B. IMPLEMENTATION DETAILS

To construct the low resolution training set for MSSR, we downsample the original person images by a local average pooling operation and then resize the downsampling images to the size of original images. In this paper, we use three down-sampling rates ($2\times$, $3\times$ and $4\times$) to ensure that the resulting low-resolution space is a mixed space and the principle of resizing is bilinear interpolation. The basis skeleton of MSSR's generator is U-net-variant [24] that uses a slightly modified residual block to replace the traditional 3×3 convolutional kernel. Compared with original U-net, utilizing residual block can better retain the detailed information, so as to promote the generation of HR images. The difference between the modified residual block and the original residual block is that the BN layer and the ReLU layer have been removed, so as to accelerate convergence during the training process. In addition, the last ReLU activation function in residual blocks has also been modified to LeakyReLU activation function. In the construction of the discriminator, we adopted AlexNet and modified its input size to 256×128. The size of MSSR's input/output image is 256×128 and batchsize is 16. During training, we use the Adam optimizer and set the learning rate as 0.001, beta1 as 0.9, beta2 as 0.999, epsilon as 1e-9. In addition, we also adopted the

following data enhancement methods:(1) Label Smoothing, the real image will be set to a random value between 0.9∼1.1; (2) Each batch contains only real samples or fake samples. The whole training process lasts approximately 3 days.

The CNN backbone of PGCN is ResNet50-variant [25] that consists of four block layers and initialized with the parameters pre-trained on ImageNet [26]. The size of input images is 256 × 128 and the dimension of local features is 512 ($v_j \in \mathbb{R}^{512}$). The number of GCN's layers is set to 2 (i.e. $L = 2$), the sizes of parameter matrices W^0 and W^1 are 512 × 1024 and 1024 × 512 respectively, and thus the dimension of $\tilde{v}_j$ is 512. The identities of labeled samples are set to 250 (i.e. $M = 250$) and these labeled identities are randomly selected from the original training set. Firstly, we train PGCN 20 epochs with labeled samples, the optimizer is SGD with a learning rate of 0.001, a momentum of 0.9 and a weight decay of 5e-4. Then the PGCN is trained with both labeled and unlabeled samples, lasting 40 epochs. The optimizer is SGD with an initial learning rate of 0.01, a momentum of 0.9, a weight decay of 5e-4, and the learning rate is divided by 10 after every 20 epochs. The batchsize of both training processes is 64.

All the experiments are implemented in PyTorch with a NVIDIA RTX 2080Ti GPU and 2.6Ghz Intel Core i5 CPUs.

C. COMPARISON TO THE-STATE-OF-THE-ARTS

In order to prove the effectiveness of the method in this paper, we compare it with following methods:

1) HAND-CRAFTED FEATURE REPRESENTATION BASED METHODS

Local Maximal Occurrence Representation (LOMO) [2], Bag-of-words (BoW) [21], Iterative Sparse Re-Weighting (ISR) [27].

2) LABEL PREDICTION METHODS

Cross-View

Asymmetric Metric Learning (CAMEL) [28], Progressive Unsupervised Learning (PUL) [29], MultiAble Learning (MAR) [19], Multi-view Clustering Method (MCM) [30], Transductive Semi-Supervised Metric (TSSML) Learning [31], Unsupervised Graph Association (UGA) [32], Memory-based Multi-label Classification Loss method (MMCL) [33].

3) UNSUPERVISED DOMAIN-ADAPTIVE METHODS

Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) [34], Person Transfer Generative Adversarial Network (PTGAN) [23], Similarity Preserving Generative Adversarial Network (SPGAN) [35], Hetero-Homogeneous Learning (HHL) [36], Generation of Latent Attribute-correlated Visual Features (GLAVF) [37], Exemplar-invariance, Camera-invariance and Neighborhood-invariance (ECN) [38], Adaptive Attention-Aware Network (AAAN) [39], Adaptive Exploration (AE) [40].

The experimental performance in the three datasets are listed in the Table 1, Table 2 and Table 3. We observe that our model significantly outperform other models. Specifically, we reach the rank-1 of 67.7% on the Market1501, 52.9% on the CUHK03 and 48.7% on the MSMT17.

TABLE 1. Comparison to the state-of-the-art unsupervised/ semi-supervised results in the Market-1501 dataset.

Methods	Market-1501		
	Rank-1	Rank-5	mAP
LOMO [2]	27.2	41.6	8.0
BoW [21]	35.8	52.4	14.8
ISR [27]	40.3	62.2	14.3
CAMEL [28]	54.5	73.1	26.3
PUL [29]	30.0	43.4	16.4
TJ-AIDL [34]	58.2	74.8	26.5
PTGAN [23]	38.6	57.3	15.7
SPGAN [35]	51.5	70.1	27.1
HHL [36]	62.2	78.8	31.4
MAR [19]	67.7	81.9	40.0
Ours	73.2	85.6	49.8

TABLE 2. Comparison to the state-of-the-art unsupervised/ semi-supervised results in the CUHK03 dataset.

Methods	CUHK03		
	Rank-1	Rank-5	Rank-10
BoW [21]	23.0	52.4	-
CAMEL [28]	31.9	57.2	-
PUL [29]	24.8	41.1	-
PTGAN [23]	37.5	-	-
MCM [30]	44.1	73.2	82.9
TSSML [31]	43.7	64.1	-
GLAVF [37]	44.5	63.8	74.0
Ours	52.9	77.6	88.1

TABLE 3. Comparison to the state-of-the-art unsupervised/ semi-supervised results in the MSMT17 dataset.

Methods	MSMT17		
	Rank-1	Rank-5	Rank-10
PTGAN [23]	11.8	-	27.4
UGA [32]	49.5	-	-
MMCL [33]	40.8	54.3	58.9
ECN [38]	30.2	41.5	46.8
AAAN [39]	35.4	44.5	48.5
AE [40]	32.3	44.4	50.1
Ours	48.7	62.6	70.4

a: COMPARISON TO THE HAND-CRAFTED FEATURE REPRESENTATION BASED MODELS

When compared with hand-crafted feature representation based models, our performance is greatly better than them. For example, compared with the best hand-crafted feature representation based method ISR [27], our method achieves 32.9% gain in Rank-1 and 35.5% gain in mAP on Market-1501 dataset as shown in Table 1. The main reason is that head-crafted feature are mostly based on color histograms, which makes it impossible to obtain the most discriminative features. In addition, it is difficult to achieve better robustness

because hand-crafted features are often designed based on a certain area of interest.

b: COMPARISON TO THE LABEL PREDICTION METHODS

When compared with the label prediction methods, our method also shows great advantages. For example, as shown in Table 2, our method increases the Rank-1 from 44.1% to 52.9% compared with MCM [30] on CUHK03 dataset. One reason is that our multi-label method has higher fault tolerance than hard-label prediction methods, even if the label prediction is not accurate enough, it can also effectively guide the training of the model. Another reason is that we propose a novel multi-label similarity loss, which allows us to make full use of all unlabeled samples. Traditional multi-label methods only take the positive and negative sample pairs with high confidence for training and ignore the hard samples. Our multi-label similarity loss gives the weight of feature distance between different sample pairs according to the similarity between multi-labels, so that the model can be trained more effectively.

c: COMPARISON TO THE UNSUPERVISED DOMAIN-ADAPTIVE METHODS

When compared with unsupervised domain-adaptive methods, our method outperforms them significantly. As shown in Table 3, even compared with the best unsupervised domain-adaptive method AAAN [39], we still achieve 13.3% gain in Rank-1 on MSMT17 dataset. We think this advantage is mainly because there is no domain gaps in our method. We use a small number of labeled samples from the same domain to pre-train the model, which is acceptable in practical applications.

D. ABLATION STUDY

In this subsection, to analyze the effectiveness of each component in our method, we perform ablation studies to demonstrate: (1)The effectiveness of the proposed MSSR; (2)The effectiveness of adversarial loss L_{adv}; (3)The effectiveness of restoration loss L_{res}; (4)The effectiveness of multi-scale methods; (5)The effectiveness of the proposed PGCN; (6)The effectiveness of the proposed soft multi-label method; (7)The effectiveness of multi-label similarity loss L_S.

1) THE EFFECTIVENESS OF THE PROPOSED MSSR NETWORK

A major contribution of this paper lies in the proposed MSSR network. To show the effectiveness of the proposed model, we conducted experiments on four down-sampling scales of 2, 3, 4 and 8. Among them, the first three scales are used as the scales of the training data, and the last scale does not appear in the training data. We use Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [41] for evaluation and the result is show in Table 4. When computing LPIPS, the ImageNet-pretrained AlexNet is utilized.

As is observed, for the appeared scales, our method is significantly better than methods SING and CSR-GAN on the LPIPS index, and slightly weaker than CSRGAN on the SSIM index. However, for the unknown scale (8 in the experiment), our method shows greater performance superiority. Experimental results show that our method can significantly restore the lost details of the image at a resolution that has not yet appeared.

TABLE 4. Quantitative results of MSSR on the CUHK03 dataset.

Method	Down-sampling rate = 2,3,4		Down-sampling rate = 8	
	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
SING [7]	0.65	0.18	0.52	0.34
CSR-GAN [8]	0.76	0.13	0.67	0.25
Ours	0.75	0.07	0.7	0.11

Figure 5 shows part of the image we restored. For each person, we have adopted four kinds of down-sampled rate $r = \{1, 2, 4, 8\}$ as input, where $r = 1$ represents that the image is the original input, and it is also the ground truth. It can be observed that the images generated by our model have good visual quality for undetermined scales.

In addition to verifying that our model has the function of restoring detailed information, we also verified its contribution to person re-identification. As is said in related works, the datasets for person re-identification have been resized from the original images, so the richness of the detailed information contained in them is also different. Using MSSR to process all the images in the dataset can ensure that all images are in the same space, so it is completely beneficial to the image matching work. We use PCB [3] as the baseline network for two person re-identification experiments. The two datasets were slightly different. One used the original Market1501 dataset, and the other used the Market1501 dataset processed by MSSR network. Note that both PCB networks are experimented in their respective training set and testing set. The experimental results are shown in the Table 6. From the Table 6, we can obverse that the images processed by MSSR have achieved a certain lead in both rank-1 and mAP with a fixed person re-identification method. We also carried out two ablation experiments, which deleted the two loss functions in the MSSR, and we will discuss these two loss functions later.

2) THE EFFECTIVENESS OF ADVERSARIAL LOSS L_{adv}

L_{adv} helps the model to correctly unify the input of the mixed space into the HR space. In order to verify the effectiveness of the source loss function, we deleted it from MSSR and train our model according to the original training strategy. The results of image restoration and the corresponding supervised person re-identification results are listed in Table 5 and Table 6, respectively. We can observe that when the loss function is removed, the improvement effect of the model on the task of person re-identification will become extremely small and the detail restoration information in the image will be lost a lot. This is because the loss function is designed to distinguish high-resolution and low-resolution images. After deleting the loss function, the model cannot guarantee that

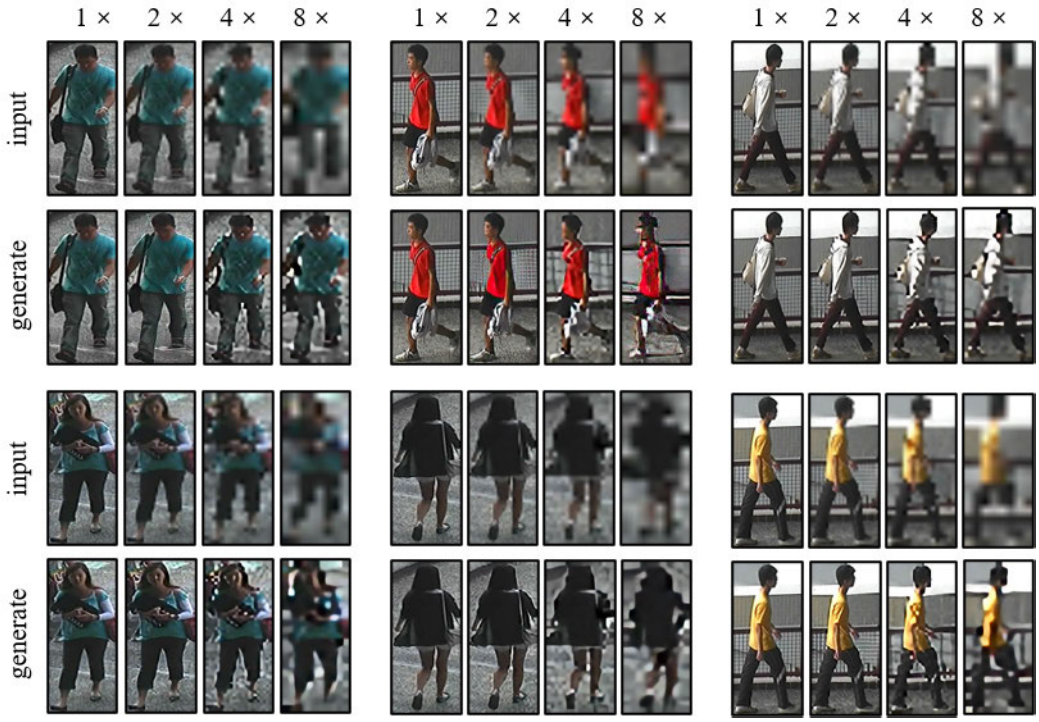

FIGURE 5. Visual results of the MSSR on the CUHK03 dataset. Note that we only trained the model on down-sample rate 2,3,4, and the images of down-sample rate 8 did not appear in the training set.

TABLE 5. Ablation study on the CUHK03 dataset for MSSR.

Method	SSIM↑	LPIPS↓
Ours	0.75	0.07
Ours w/o L_{adv}	0.56	0.30
Ours w/o L_{res}	0.48	0.37
Ours w/o multi-scale	0.68	0.15

TABLE 6. The contribution to person re-identification of MSSR on the Market-1501 dataset.

Preprocessing of Market1501	PCB	
	Rank-1	mAP
None	92.3	77.4
MSSR	93.5	78.6
MSSR w/o L_{adv}	92.6	77.5
MSSR w/o L_{res}	89.6	70.8

the input and output can be mapped to the same space. However, due to the existence of the reconstruction loss function, the model can still restore a small number of detailed features so the indicators in the person re-identification will also have a tiny improvement.

3) THE EFFECTIVENESS OF RESTORATION LOSS L_{res}

When the restoration loss L_{res} is missing, it cannot guarantee that the images generated by the model are high-resolution images, and the detailed information of the images cannot be restored. As shown in Table 5, without L_{res}, both SSIM and LPIPS indexes of images generated by MSSR network are developing in a bad direction. Furthermore, from Table 6, we can find that the performance of person re-identification model also has a corresponding decline when L_{res} is missing. This proves that the restoration loss can indeed restore the lost details of the images and further improve person re-identification effectively.

4) THE EFFECTIVENESS OF MULTI-SCALE STRATEGY

As is said before, the multi-scale strategy can help the model generate more refined images. To prove the necessity of the course generation stage, we changed the input of the finer generation model to 3 channels and input the person images directly. The output results of the experiment are shown in the Table 5 and some experimental results are posted in Figure 6, we can obverse that when the down-sampling rate is small, both the single-scale method and the multi-scale method can achieve good results. However, when the down-sampling rate is large, the single-scale method loses more detailed information and the performance is not satisfactory while the multi-scale method still maintained good effect.

5) THE EFFECTIVENESS OF GRAPH CONVOLUTIONAL NETWORK

In our proposed PGCN, graph convolutional network is mainly used to exploit the contextual information between different human body parts. To demonstrate the benefit of GCN, we remove GCN from our PGCN and directly concatenate the local features output from CNN as the final feature of input image. The experiment result is shown in Table 7.

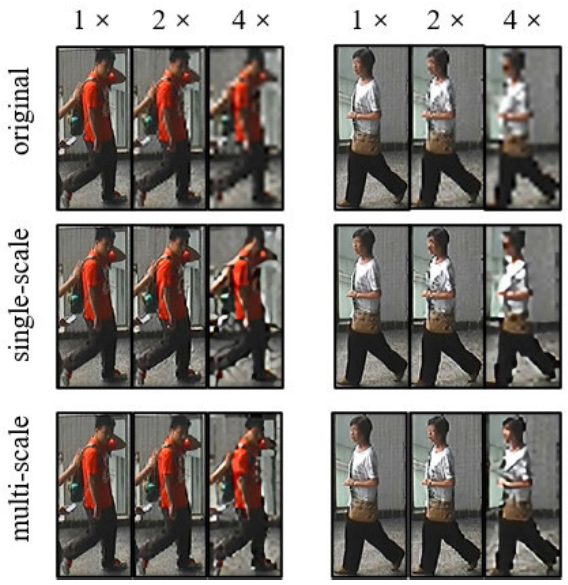

FIGURE 6. The comparison between single-scale model and our multi-scale MSSR on the CUHK03 dataset. The number above the image represents the downsampling ratio.

TABLE 7. Ablation study on the CUHK03 dataset for Re-ID.

Method	Rank-1	Rank-5	Rank-10
Ours	52.9	77.6	88.1
Ours w/o GCN	37.1	60.3	78.6
Ours w/o soft multi-lable	30.1	56.7	70.5
Ours w/o multi-lable similarity loss	48.2	71.5	80.6

As we can see, when GCN is removed, the Rank-1 accuracy on CUHK03 drops from 52.9% to 37.1%. This significant decline shows the desired advantage of local features with contextual information in person re-identification task.

6) THE EFFECTIVENESS OF PROPOSED SOFT MULTI-LABELS

Before conducting the ablation experiment of the multi-label method, we define the non-multi-label method at first: for the fused person features, we directly predict a hard label, then the traditional teacher-student model is used for semi-supervised training to get the final result. In the setting of parameters, the ablation experiments are consistent. The result is shown in Table 7. It can be observed that our multi-label method significantly outperform hard label method. The key reason is that our multi-label method gives the model sufficient fault tolerance, even if some predicted labels are inaccurate, the model can still be trained effectively.

7) THE EFFECTIVENESS OF MULTI-LABEL SIMILARITY LOSS

Once loss L_S is excluded, we use the triplet loss instead. In the identification of positive and negative pairs of samples, we take the top 10% and bottom 10% in the similarity ranking as positive and negative pairs respectively. The result is shown in Table 7. It can be observed that the performance of the model has a certain degree of decline. This is mainly caused by the utilization of samples. Soft multi-label loss can apply all unlabeled samples to model training, while the triplet loss only use part of unlabeled samples.

E. HYPERPARAMETER EVALUATIONS

We evaluate how α (influence the importance of multiple scales in the MSSR) and τ (affect the identification of positive and negative pairs in the sample) affect our model. The results are shown in Figure 7 and Figure 8. We can obverse that α specifies the weights of the two stages in image generation. If it is too low, it will cause a significant decrease in image quality, while it is too high, the image quality is similar. This is because our refined generative model can actually complete the work of image restoration, while the rough generative model mainly serves to extract and restore high-level information. As for the threshold parameter τ, when it is small, the wrong sample identification will mislead the training of the model. When it is large, the number of positive samples drops rapidly, resulting in an imbalance in the loss function, which will also influence the model.

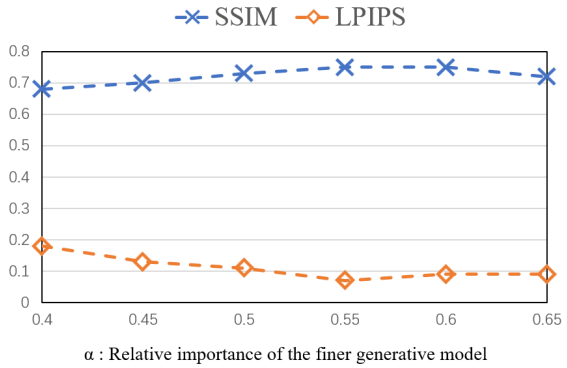

FIGURE 7. Evaluation on hyperparameter α.

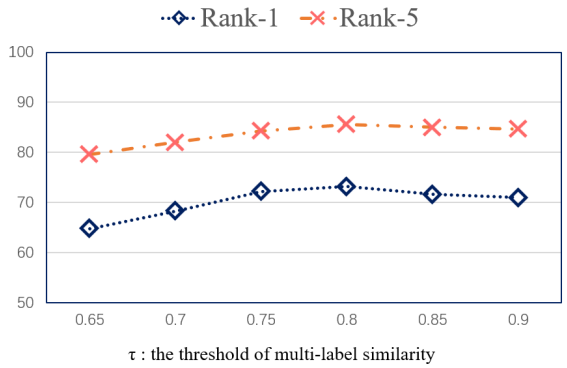

FIGURE 8. Evaluation on hyperparameter τ.

F. COMPARISON OF COMPUTATIONAL COMPLEXITY

Computational complexity is an important concern of real-world person re-identification. In this subsection, we use

the query time (Q.Time: the time for querying each image) to evaluate the computational complexity of our model and compare it with PCB [3], SING [7] and CSR-GAN [8]. PCB only has person re-identification model, SING, CSR-GAN and our method consist of super-resolution generation model and person re-identification model. The results are shown in Table 8, where Q.time is composed of two parts, the former represents the execution time of super-resolution generation model, and the latter represents the execution time of person re-identification model. As we can see, the Q.Time of PCB is 6.9s, which is much longer than that of other methods in person re-identification stage. This is because the feature dimension of PCB is 12,288, which leads to a huge amount of calculation when comparing person features. SING's super-resolution generation model and person re-identification model have the shortest Q.Time (0.09s and 1.2s), because SING's generation model is only composed of three convolutional layers and the dimension of person feature is only 256. The total Q.Time of CSR-GAN is 3.6s, which is longer than our 3.5s. This is because the super-resolution generation model of CSR-GAN is composed of three cascaded SR-GANs, which is large-scale and time-consuming. On the whole, the Q.Time of our model is acceptable.

TABLE 8. Analysis on the efficiency of our model on Market-1501.

Method	Feature Length	Q.Time(s)
PCB [3]	12,288	$0 + 6.9$
SING [7]	256	$0.09 + 1.2$
CSR-GAN [8]	2,048	$1.4 + 2.2$
MSSR+PGCN (Ours)	3,072	$0.8 + 2.7$

V. CONCLUSION

This paper focuses on the application of person re-identification in real-world scenarios. In order to solve the problems of super-resolution and label limit in person re-identification, we proposed a Mixed-Space Super-Resolution (MSSR) network and a soft multi-label based semi-supervised person re-identification method. Our MSSR network can convert person images of different resolutions to high resolution. Furthermore, in order to better restore the details of images, MSSR network introduces a multi-scale strategy for high resolution image generation. The soft multi-labeled based semi-supervised person re-identification method propose a PGCN to learn more discriminative features by exploiting the contextual information between different human body parts at first. Then the PGCN trained with a small amount of labeled samples is used to predict the soft multi-labels for unlabeled samples. Finally, we train PGCN with the unlabeled samples based on a novel multi-label similarity loss. Experimental results in three benchmarks validate the effectiveness of the proposed method.

REFERENCES

[1] M. O. Almasawa, L. A. Elrefaei, and K. Moria, "A survey on deep learning-based person re-identification systems," *IEEE Access*, vol. 7, pp. 175228–175247, 2019.
[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[3] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 480–496.
[4] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 695–704.
[5] Z. Wang, R. Hu, Y. Yu, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. IJCAI*, vol. 2, 2016, p. 6.
[6] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[7] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–8.
[8] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded SR-GAN for scale-adaptive low resolution person re-identification," in *Proc. IJCAI*, 2018, vol. 1, no. 2, p. 4.
[9] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino, "Semi-supervised multi-feature learning for person re-identification," in *Proc. 10th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug. 2013, pp. 111–116.
[10] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3550–3557.
[11] X. Zhu, X.-Y. Jing, L. Yang, X. You, D. Chen, G. Gao, and Y. Wang, "Semi-supervised cross-view projection-based dictionary learning for video-based person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2599–2611, Oct. 2018.
[12] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.
[13] X. Zhang, X.-Y. Jing, X. Zhu, and F. Ma, "Semi-supervised person re-identification by similarity-embedded cycle gans," *Neural Comput. Appl.*, vol. 32, pp. 14143–14152, Mar. 2020.
[14] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 486–504.
[15] H. Liu, Z. Xiao, B. Fan, H. Zeng, Y. Zhang, and G. Jiang, "PrGCN: Probability prediction with graph convolutional network for person re-identification," *Neurocomputing*, vol. 423, pp. 57–70, Jan. 2021.
[16] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3286–3296.
[17] X. Chen, L. Zheng, C. Zhao, Q. Wang, and M. Li, "RRGCCAN: Re-ranking via graph convolution channel attention network for person re-identification," *IEEE Access*, vol. 8, pp. 131352–131360, 2020.
[18] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.
[19] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2148–2157.
[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[22] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, Jun. 2014, pp. 152–159.
[23] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[25] J. Zhao, J. Li, X. Tu, F. Zhao, Y. Xin, J. Xing, H. Liu, S. Yan, and J. Feng, "Multi-prototype networks for unconstrained set-based face recognition," 2019, *arXiv:1902.04755*. [Online]. Available: http://arxiv.org/abs/1902.04755
[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[27] G. Lisanti, I. Masi, A. D. Bagdanov, and A. D. Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1629–1642, Aug. 2015.
[28] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 994–1002.
[29] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 4, pp. 1–18, Nov. 2018.
[30] X. Xin, J. Wang, R. Xie, S. Zhou, W. Huang, and N. Zheng, "Semi-supervised person re-identification using multi-view clustering," *Pattern Recognit.*, vol. 88, pp. 285–297, Apr. 2019.
[31] X. Chang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Transductive semi-supervised metric learning for person re-identification," *Pattern Recognit.*, vol. 108, Dec. 2020, Art. no. 107569.
[32] J. Wu, H. Liu, Y. Yang, Z. Lei, S. Liao, and S. Li, "Unsupervised graph association for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8321–8330.
[33] D. Wang and S. Zhang, "Unsupervised person re-identification via multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10981–10990.
[34] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2275–2284.
[35] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.
[36] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 172–188.
[37] H. Li, Y. Chen, D. Tao, Z. Yu, and G. Qi, "Attribute-aligned domain-invariant feature learning for unsupervised domain adaptation person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1480–1494, 2021.
[38] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 598–607.
[39] W. Zhang, Z. Wei, L. Huang, K. Xie, and Q. Qin, "Adaptive attention-aware network for unsupervised person re-identification," *Neurocomputing*, vol. 411, pp. 20–31, Oct. 2020.
[40] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–19, Apr. 2020.
[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

LIMIN XIA was born in Hunan, China, in 1963. He received the M.S. and Ph.D. degrees in control science and engineering from Central South University, China, in 1995 and 2000, respectively.

From 2000 to 2004, he was an Associate Professor with Central South University, where he has been a Professor with the School of Automation, since 2005. He has published more than 150 articles in these fields. His research interests include the areas of computer vision, pattern recognition, and behavior analysis.

JIAHUI ZHU was born in Zhejiang, China, in 1996. He received the B.S. degree in automation from Central South University, Hunan, China, in 2018, where he is currently pursuing the M.S. degree in control science and engineering.

His research interests include the areas of computer vision, pattern recognition, and behavior analysis.

ZHIMIN YU was born in Jiangxi, China, in 1996. He received the B.S. degree in automation from Central South University, Hunan, China, in 2019, where he is currently pursuing the M.S. degree in control science and engineering.

His research interests include the areas of computer vision, pattern recognition, and behavior analysis.