IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 14, 2019, accepted October 14, 2019, date of publication October 18, 2019, date of current version October 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948260

SRPRID: Pedestrian Re-Identification Based on Super-Resolution Images

ZHEN QIN[1,2,3], WEI HE[1], FUHU DENG[1,2,3], MENG LI[1,2], AND YAO LIU[1,3]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China
[2]Institute of Electronic and Information Engineering of UESTC, Dongguan 523000, China
[3]Network and Data Security Key Laboratory of Sichuan Province, Chengdu 610000, China

Corresponding author: Fuhu Deng (fuhu.deng@uestc.edu.cn)

This work was supported in part by the Frontier Science and Technology Innovation Projects of National Key Research and Development Program under 2019QY1405, in part by the National Natural Science Foundation of China under Grant 61672135, in part by the Sichuan Science-Technology Support Plan Program under 2018GZ0236 and Grant 2017FZ0004, in part by the Fundamental Research Funds for the Central Universities under Grant 2672018ZYGX2018J057, and in part by the CERNET Innovation Project under Grant NGII20180404.

ABSTRACT Intelligent security expects to avoid the occurrence of robbery, theft, and other undesirable situations through video surveillance. In video surveillance, images of human faces undimmed are not easily available, so pedestrian re-identification (person ReID) is an alternative technique which attracts a mount of researchers attention. Person ReID is a technique used to match pedestrian images across cameras. Due to the interference of shooting angle and camera quality, it is difficult to obtain the images of high resolution, no obstructions, simple backgrounds and similar posture, which brings great challenges to the research of person ReID. Most existing methods of pedestrian re-identification ignore the inconsistency of resolution, and they are based on the assumption that all images have similar and high enough resolution by default. In this paper, we propose a hybrid framework, Super-Recognition of Pedestrian Re-Identification (SRPRID), in order to strengthen pedestrian re-identification based on multi–resolutions images captured by disparate cameras. Particularly, residual dense block (RDB) and Integrated Attention (InnAttn) block are merged to SRPRID. It is worth mentioning that the rank_1 accuracy of our method outperforms the state-of-art method by 17.2 points (86.9% - 69.7%) on CUKH03 dataset of extremely challenging.

INDEX TERMS Person re-identification, super resolution, residual dense block, soft attention, hard attention, convolutional neural networks, video surveillance.

I. INTRODUCTION

The gradual growth of urban population has brought great challenges to social security. In order to ensure the physical security or property security on many public occasions, and prevent or deal with various unsafe incidents in time, the video surveillance network constituted of multiple cameras is deployed in momentous occasions. For a monitoring network complicated and massive image data from monitoring equipment, however, it is bound to take a lot of time cost only relying on human analysis and processing, and even miss the best opportunity. Therefore, it is necessary to use computer vision technology for intelligent video surveillance. The best way to identify a target individual of interest is face recognition. Face identification performs person ReID by analyzing the face images captured by cameras. However, due to many factors such as shooting angle, shooting distance, lighting and equipment quality, face images with extraordinarily high quality are generally not available. When face identification fails to work, person ReID becomes a very important alternative technology. In the field of video surveillance, pedestrian recognition has become an important and indispensable part.

Person ReID is the processing procedure of retrieving all the images of the same identity from a existing and giant sub-dataset called gallery, given a query image of a target person interested. Notably, query image and gallery images are captured by entirely different cameras that may or may not have any overlapping perspectives. In practical applications, the pedestrian images captured by the camera is usually affected by various factors: insufficient lighting, rain, snow, poor quality of the equipment, or combination of factors, which directly result in low resolution of captured pedestrian images. The given gallery library image is of

The associate editor coordinating the review of this manuscript and approving it for publication was Kim-Kwang Raymond Choo.

high resolution (HR), while query library image is of low resolution (LR), and it is desired to find all the images in the gallery library from the same person as query. The low-resolution image directly corresponds to the same person's high-resolution image. It is impossible to accomplish challenging learning tasks well through uncomplicated pedestrian recognition. Different resolution images of the same pedestrian have different feature information, so cross-resolution image matching is more difficult. Therefore, we consider transforming the image in query and gallery into the same resolution through super-resolution reconstruction technology, and then implement pedestrian recognition.

The task of super-resolution image reconstruction is to give a small graph with low resolution and convert it into a large graph with high resolution. Although this task brings good visual effects, this is an irreversible process with uncertainty. Recovering high-resolution images has inveracious pixels and it is not conducive to identifying tasks. Simply converting a low-resolution image to a high-resolution image and re-identifying the pedestrian may bring no performance enhancement, and even reduce the recognition performance.

Therefore, this paper proposes a novel framework SRPRID. This approach introduces the super-resolution reconstruction method into the pedestrian re-recognition task and jointly learning the two tasks of super-resolution reconstruction and pedestrian re-identification. The reconstruction capability promotes the ultimate goal of person ReID better, thereby improving the task of performing pedestrian ReID between the low resolution images and the high resolution images. The contributions are as follows.

- This paper proposes an end-to-end framework SRPRID, which combines residual-intensive block and attention-intensive mechanism to learn the characteristics of pedestrian images with different resolutions, is proposed by introducing the attention mechanism and residual-intensive block.
- A multi-task architecture is proposed to optimize both super-resolution reconstruction and pedestrian recognition. Multi-task learning can not only optimize and solve multiple related tasks, but also combine their advantages to enforce knowledge sharing, so as to improve the performance of each task.
- The proposed framework reports competitive accuracy on the CUHK03 [1] and SYSU [2] datasets comparing the existing start-of-the-art methods.

The remainder of this paper is as follows. In Section II, we review the related works. The overview of the proposed framework is described in Section III. Section IV provides the experiment results and discussion. Finally, we conclude this paper in Section V.

II. RELATED WORK

In this section, we will now briefly review the state-of-art literature relating to the use of super-resolution images for pedestrian re-identification, broadly categorized into technical approaches of re-identification, low-resolution Images ReID, and super resolution restruction.

A. RE-IDENTIFICATION

In recent years, the research on person ReID has attracted much attention from scholars at domestic and overseas. Since 2016, person ReID has developed rapidly and has been increasing year by year in the conferences of computer vision [3]–[7]. Therefore, it is necessary to study the problem of person ReID. Existing excellent person ReID methods can be roughly divided into two categories: feature representation and metric learning.

The main purpose of person ReID methods based on feature expression is to learn the invariant characteristics of pedestrians under different external conditions. The existing method based on feature expression include: color histogram [8], local binary patterns [9], Gabor features [10], color name [11],etc. The main purpose of person ReID methods based on metric learning is to learn how to measure the similarity of two pictures and make the same person's pictures more similar. Paper [12] used the Siamese network. When the input network is a pair of positive samples (two pictures belonging to the same person), the contrast loss function decreases gradually. On the contrary, it increases until it exceeds the set threshold. By minimizing the loss function, the distance between positive samples becomes smaller and that between negative samples becomes larger. Paper [13], [14] uses triplet loss to train the network. The input of the network is a triple, including a pair of positive samples and a pair of negative samples. This loss function can reduce the distance between positive samples and increase the distance between negative samples, so as to achieve the purpose of person ReID. Conventional triple images are randomly extracted from training data, so there is a problem in the way of sampling, that is, most of the samples are easy to sample pairs. Although this can train the network quickly, the network does not have generalization ability. Paper [15] presents TriHard Loss, which is an online hard sample sampling method based on training batch. There are many pictures in a batch. For each picture in the batch, choose one of the most difficult positive samples and one of the most difficult negative samples, thus forming a sample that meets the requirements of TriHard. Experiments show that TriHard loss outperforms traditional triple loss.

B. LOW-RESOLUTION IMAGES REID

The existing person ReID method focuses on dealing with variability caused by illumination, background clutter, and posture change, thereby causes the attenuation of person ReID performance. There are only a few papers that attempt to address low-resolution person ReID. The paper [16] assumes that the same person's images should be similarly distributed at different resolutions, and proposes a strategy to optimize the cross-resolution image learning framework, namely joint optimization of pedestrian alignment and

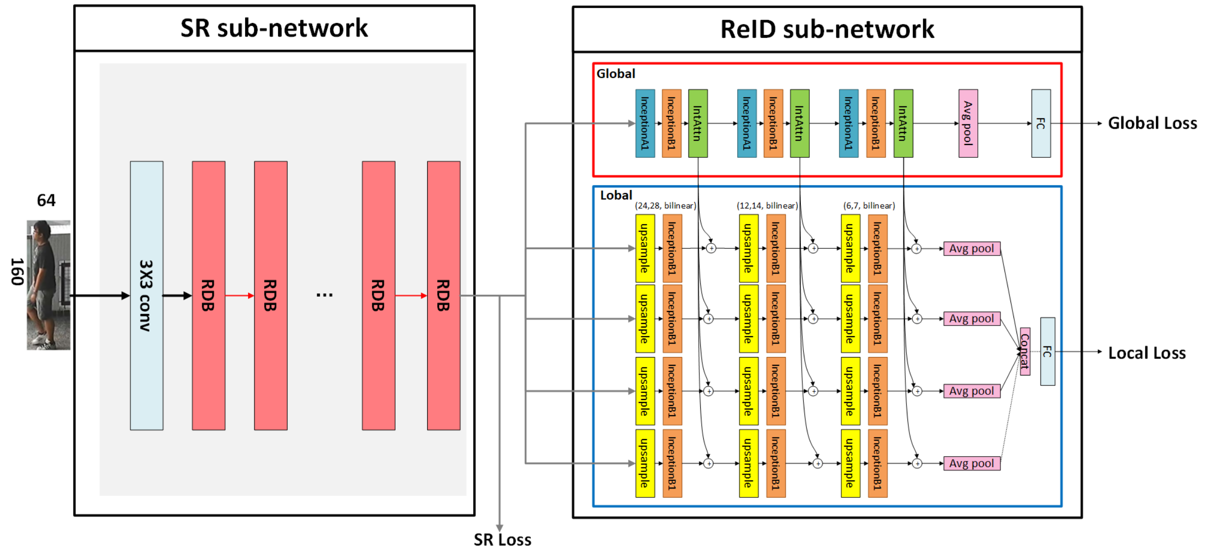


FIGURE 1. An overview of the proposed SRPRID deep hybrid framework for cross-cameras cross-resolutions person re-identification. SRPRID consists of two sub-networks: (1) super resolution (SR) sub-network; (2) person re-identification (ReID) sub-network. The main components are distinguished by different colors: red for residual dense block (*RDB*) [19], blue for *InceptionA*1 fine-tuned *InceptionA* shown in Fig.2, orange for *InceptionB*1 fine-tuned InceptionB shown in Fig.3, green for Integrated Attention (*IntAttn*) block shown in Fig.4.

distance measurement models. The paper [17] proposes a semi-coupling low rank discriminant dictionary learning method for super-resolution pedestrian recognition. Paper [18] compares the relationship between too many scales and distance function spaces by changing the scale of low-resolution images with matching high-resolution images. The above three methods alleviate the adverse effects of low-resolution images on person ReID to a certain extent, but they are unable to synthesize discriminative appearance information lost during image acquisition, that is, they fail to solve low resolution and the difference in the amount of information that exists between a low-resolution image and a high-resolution image.

C. SUPER RESOLUTION RESTRUCTION

Deep learning plays an important role in various fields. Dong et al. propose SRCNN [20] that first introduces deep convolutional neural network into super-resolution reconstruction. Kim et al. introduce a deeper convolution network VDSR [21] inspired by VGG 16. And the idea of residual learning is introduced into VDSR, in order to reduce the complexity of network learning and speed up the learning. Lim et al.use multiple residual blocks to build a better network EDSR [22]. With the deepening of the network, the problem of gradient disappearance will occur in the process of task learning. The reason of gradient disappearance is that the input information and gradient information are transferred between many layers. The characteristic of dense connection is that each layer is directly connected, so as to avoid gradient disappearance in the deep network. Inspired by this, Gao et al. proposes SRDenseNet [23] that is used dense blocks for super resolution restruction. Residual blocks make network deeper and wider so that higher image features can be learned. And dense blocks can slow down the problem of gradient disappearance in deep network learning to some extent. In order to achieve better performance, Zhang et al. proposed residual dense block (RDB) [19] combined residual blocks and dense blocks.

III. PROPOSED FRAMEWORK

The majorities of the existing excellent pedestrian recognition approaches are based on a default assumption that all individuals' pictures have semblable and high-quality enough resolutions. The characteristic variability triggered by illumination, occlusion and background interference can be solved by means of feature expression and metric learning. And the captured photographs have different resolutions due to various external factors, which is also one of the difficult issues to be solved in pedestrian recognition. This paper aims to address the difficulty of cross-camera and cross-resolution pedestrian recognition. Zhang et al. proposed residual dense block (RDB) [19] that can effectively establish the low-resolution images to the high-resolution images as more as possible. And Wei etc. proposed the Harmounious Attention Network [24] that can effectively copy with the challenge of pedestrian recognition. This paper proposes an end-to-end hybrid network SRPRID shown in Fig. 1 for cross-resolution pedestrian weight recognition based on mature super-resolution technology and pedestrian weight recognition technology.

Our SRPRID consists of two sub-networks: SR(super resolution) sub-network and ReID(re-identification) sub-network. SR-network is mainly used for reconstruction and extraction of distinguishing characteristics, while ReID is principally intended to accurately match cross-resolution images of a common individual.

A. SR SUB-NETWORK

The first sub-network is a residual dense network composed of multiple residual dense blocks. The input to the network are pedestrian images of different resolutions. First, a convolution layer of 3x3 kernel is used to extract shallow features and used as input to the subsequent residual intensive layers. The output of the first layer is F_0 shown in equ.1.

$$F_0 = H_{conv}(I_{lr}) \tag{1}$$

where H_{conv} represents convolution operation, and I_{lr} denote the images of input. After the first RDB layer, the output F_1 can be calculated by equ.2.

$$F_1 = H_{RDB}(F_0) \tag{2}$$

where H_{RDB} denotes a sequence of operations performed on a residual block and it has been detailed in Section 2.3. Provided that M residual dense blocks were used continuously, the output F_M can be obtained by equ.3.

$$\begin{aligned} F_M &= H_{RDB,M}(H_{RDB,M-1}(H_{RDB,M-2(\ldots F_1\ldots)})) \\ &= H_{RDB,M}(H_{RDB,M-1}(H_{RDB,M-2(\ldots H_{RDB}(F_0)\ldots)})) \end{aligned} \tag{3}$$

where $H_{RDB,M}$ represents the operational process of $m-th$ RDB. The features of different levels are extracted by a series of residual intensive blocks. In order to make full use of the features in different levels, these features need to be integrated. After the RDBs, we further perform the operation of features fusion(FF), and it can be obtained by equ.4.

$$F_{FF} = H_{FF}(F_0, F_1, F_2, F_3, \ldots F_{M-1}, F_{M-2}) \tag{4}$$

The previous series of operations are carried out for low-resolution images. In order to better reconstruct the high-resolution images corresponding to low-resolution images, we adopted a up-sampling network ESPCN [25] fine-tuned. The output of SR sb-network can be obtained by equ.5.

$$F_O = H_{SR}(I_{LR}) \tag{5}$$

B. REID SUB-NETWORK

ReID sub-network is mainly divided into two branches: local branch and global branch. The main task of local branch is to learn the characteristics of different local blocks, and global branch is to learn the global features. For both branches, we adopt some basic blocks that are *InceptionA*1, *InceptionB*1 and *IntAttn*, respectively. In order to simplify the network without intolerable loss of network performance, we use *InceptionA*1 and *InceptionB*1 inspired by InceptionA/B units [26], [27], shown in Fig.2 and Fig.3. And *IntAttn* is illustrated in Fig.4.

1) InceptionA1

This module adopts several different filters to extract different features, namely three 3x3 convolution kernel and one 1x1 convolution kernels. In order to speed up the calculation, an additional 1x1 convolution is added before the convolution operation, which limits the number of input channels and thus

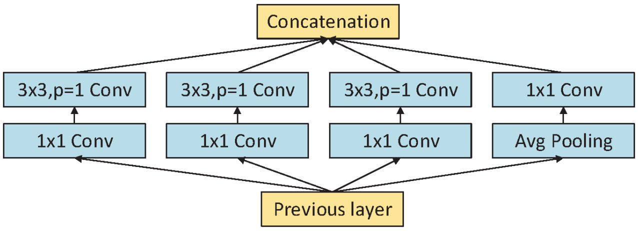


FIGURE 2. InceptionA1 block.

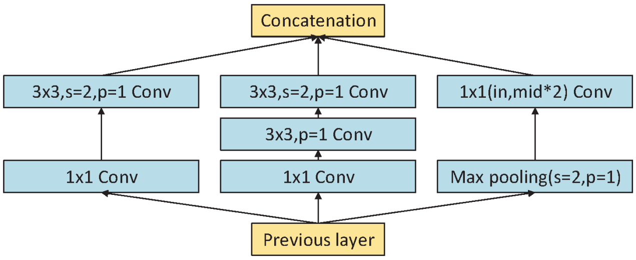


FIGURE 3. InceptionB1 block.

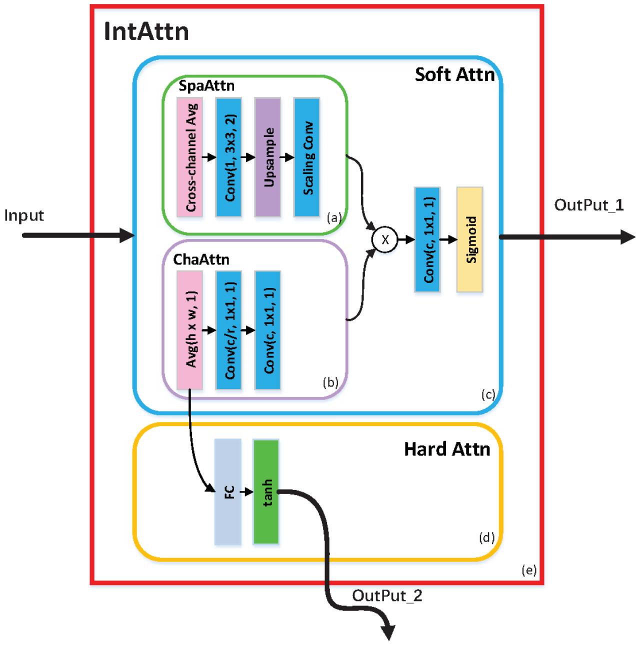


FIGURE 4. ***IntAttn block:*** **The complete structure of IntAttn(e) is composed by Soft Spatial Attention(a), Soft Channel Attention(b), and Hard regional Attention(d). It has one input that is the output of the upper layer (*InceptionB*$_1$), while two output. Different types of concrete network layers are distinguished by rectangles with different background colors: pink for average pooling, blue for convolution layer, purple for upsample with bilinear operation, yellow for sigmoid layer, cyan for full connection layer, orange for tanh layer. For the sake of understanding, different submodules are also represented by rectangles with rounded corners with different border colors.**

plays the role of dimensionality reduction. In addition, average pooling is used to better extract the global information of images.

2) InceptionB1

This module is similar to *InceptionA*1, with the following differences: (a) the initial design of this module adopts filters of different sizes for feature extraction, namely 1x1, 3x3 and

5x5, but considering that the larger the convolution kernel is, the computational efficiency would reduce exponentially. Therefore, we decompose the 5x5 convolution kernel into two 3x3 convolution kernels with the same effect, so as to accelerate the training speed without reducing the number of extracted features. In many high-performance networks proposed by predecessors, this operation is used, which can not only improve the depth of the network, but also reduce the number of parameters by about 30% while ensuring the same receptive field. (b) In addition, this module replaces the average pooling in $InceptionA1$ with maximum pooling, so as to extract more local and most differentiated features.

3) INTEGRATED ATTENTION

Visual attention mechanism is the special brain signal processing mechanism of human vision. The human eye can quickly scan the global image to obtain the target area that needs to be focused on (namely the focus of attention), and then pay more attention to the target area to obtain more details that need special attention.The attention mechanism in deep learning is proposed by referring to the unique visual attention mechanism of human beings.The attention mechanism can help the model to assign different weights to each dimension of the input image matrix and extract more critical and important information, so as to improve the performance of the model while reducing the calculation and storage overhead. Since the publication of the 2017 Google paper Attention Is All You Need [28], experts and scholars in various fields of deep learning have realized the importance of attention mechanism and introduce it into their work [29]–[31]. Therefore, we also introduce the attention mechanism into our model, and propose a novel integrated attention module, which is called Integrated-Attention(*IntAttn*).

IntAttn is composed of soft spatial attention [29], soft channel attention [32] and hard regional attention [33]. The specific structure of IntAttn is shown in Fig. 4. In order to simulate the dorsal and ventral functions of the human brain, we combine the soft attention mechanism with the hard attention mechanism. The soft attention mechanism helps to filter fine-grained and important pixels, while the hard attention mechanism has great advantages in selecting coarse-grained and potentially discernable localities.

a: SOFT SPATIAL-CHANNEL ATTENTION

The input of IntAttn is a three dimensional tensor $X^{\ell}\ \varepsilon\ M^{c\times h\times w}$, where X represents the input of pedestrain image, d shows the current layer, c means channel of image, h expresses height, and w indicates width. The goal of soft attention is to learn a weight matrix with the size equivalent to X_{ℓ}, that is the $A^{\ell}\ \varepsilon\ M^{c\times h\times w}$. Soft attention is divided into spatial attention and channel attention.Spatial attention can be understood as paying attention to different areas of an image. It means that for all channels, a weight is learned for the graph of size H, x and W on a two-dimensional plane, and a weight is learned for each pixel.While channel attention pays attention to the different features of images, that is, for each C, in the channel dimension, different weights are learned, and the weights are the same in the plane dimension. According to the principle of matrix factorization, A^{ℓ} is decomposed into two tensors shown in equ.6.

$$A^{\ell} = S^{\ell}\ \times\ C^{\ell} \tag{6}$$

where $S^{\ell}\ \varepsilon\ M^{1\times h\times w}$ represents spatial attention tensor, and $S^{\ell}\ \varepsilon\ M^{c\times 1\times 1}$ denotes channel attention tensor.

Soft Spatial Attention: Soft Spatial Attention (SSA) is constituted by four successive layers, that is cross channel average pooling layer, convolution layer with kernel of 3x3, upsampling layers by bilinear operation and scaling convolution respectively.

Soft Channel Attention: Soft Channel Attention (SCA) is made up of three layers that are average pooling layer and two different convolution layers. The output of SSA and SCA are multiplied, then it goes through a convolution layer and a sigmoid layer, and lastly, we get the output of the soft attention which is the first output ($OutPut_1$).

b: HARD ATTENTION

The input of Hard Regional Attention (HRA) is the first layer output of SCA. In fact, HRA consists of three layers. They are the average pooling of SCA, full connection layer and tanh layer, respectively. The output result of the average pooling is taken as the input of the full connection layer, which goes through a full connection layer and finally through a tanh layer to get the second output of IntAttn ($Output_2$), which is sent to the corresponding local branch.

Approach Summary: The most critical components of the network have been described detailedly in Section 3.1 and 3.2, respectively. We briefly review the structure of the model proposed in this paper. Our entirely network is primarily divided into super-resolution reconstruction sub-network and pedestrian recognition sub-network. Super-resolution reconstruction sub-network combines residual-intensive block and continuous memory mechanism to extract and recover the most distinguishing features beneficial to the identification task. The input of SR sub-network is images in different resolution belonging to various individuals. Feature extraction and continuous memory mechanism are used to avoid that important feature information is not lost in the process of information transmission at all levels. The output of SR sub-network is features extracted from high-resolution images, which will be the input of the next sub-network.

Person re-identification sub-network integrates $Inception-A1$, $InceptionB1$, and $IntAttn$ based on multiple attention mechanisms. From the general framework diagram in Fig.1, it is clear that the person re-identification sub-network is divided into two branches: local branch and global branch. (a) The global branch considers the global pixels of the pedestrian picture and learns entire pictures to extract the global person characteristics. The input of the global branch is the output of the previous sub-network. Then the output of global branch is input into the integrated attention module. Then, after the operations of $InceptionA1/B1$, the results are input into the integrated attention module, and the soft attention output of the attention module is transmitted to the next group inceptions (including *InceptionA*1 and *InceptionB*1). After repeating the above hierarchical structure, the average pooling operation of the maps is performed before the full connection layer to achieve dimensionality reduction. (b) Local branches consider and learn local features separately from T local blocks of person images. The input of local branches is consistent with the global branch. Firstly, through an up-sampling layer, the image is processed into fixed size block by bilinear interpolation. The output up-sampling on different local branches is added to another output of the *IntAttn* in global branch at the same level. The output is then fed into the *InceptionB*1 module. After repeating the previous hierarchy, each sub-local branch passes through an average pooling layer, and cascades all sub-local branches. Lastly, a full connection layer is performed.

C. TRAINING

We use the Adam optimizer Algorithm [34] with the default hyper-parameters that are 3×10^{-4} initial learning rate and weight decay 5×10^{-4} for most experiments. In order to prevent the learning rate from getting too large, it swings back and forth when it converges to the global optimal point. In the learning process, we make the learning rate decrease exponentially with the number of training rounds, that is, the learning rate decays.The attenuation strategy is: if epoch is less than 20, the learning rate is 3×10^{-4}. If the epoch is between 20 and 40, its learning rate is 3×10^{-5}. If the epoch is greater than or equal to 40, its learning rate is 3×10^{-6}.

Generally, the problem of classification and recognition is that the number of sample classes is limited, so softmax and cross-entropy [35] can be used in training network. Our total loss function is given by Equ.7.

$$L_C(Y, Y') = -\left[\sum_i^N Y_s logY'_s \sum_i^N Y_\ell logY'_\ell + \sum_i^N Y_g logY'_g\right] \qquad (7)$$

In order to make our network more robust, the loss is mainly composed of three parts, namely, the loss of sr-subnetwork, reid-subnetwork local branch and global branch. Therefore, the cross entropy loss is constituted by the three parts of the above Equ.7. Where Y denotes true ID of the current image, and Y' is predicted ID. And the angle of s, ι, g represent different values belonging to SR-subnetwork, reid-sunetwork local branch ans global branch. The i shows $i - th$ of images in a batch, and N is the size of a batch.

IV. EXPERIMENT

A. DATASET

In the field of person recognition, there are many public and available datasets. However few datasets is used in the existing work of pedestrian matching with different resolutions, and some datasets is too small to be suitable for training and evaluation of deep learning model. To evaluate the performance of the proposed framework, we conducted experiment with two simulated LR person re-id datasets that are based on the large datasets CUHK03 [1] and SYSU [2]. The comparative experiment we provided is only on the two datasets mentioned above. This is not to show that our framework is just applicable to them. Considering that in real life, images captured by cameras have multiple resolutions, we use multiple different sampling rates to simulate LR images with multiple resolutions corresponding to the same HR image.

1) MLR-CUHK03

CUHK03 dataset [1] was collected from The Chinese University of Hong Kong campus and taken from 5 different pairs of camera views. It has more than 14,000 images of 1,467 pedestrians. MLR-CUHK03 is constructed from CUHK03 dataset. Like the paper [36], for each camera pair, we randomly selected one as LR probe image source by performing multi-resolution down-sampling by a ratio randomly picked from $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$.

2) MLR-SYSU

SYSU dataset [2] was established and organized by Professor Lai Jianhuang, Dr. Guo Chunchao and Chen Shizhe from the School of Data Science and Computer Science, Sun Yat-sen University, and used for academic research in the fields of pedestrian recognition and pedestrian image retrieval. It has totally 24,446 images of 502 people captured by two different cameras. In order to apply the SYSU data set to the application scenario (low-resolution pedestrian recognition) in this paper, we select 3 images from each camera perspective to perform the resolution reduction operation. The selected images were randomly sampled at $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}\}$ ratio. The preprocessed data set is the MLR-SYSU dataset suitable for this article.

The two datasets we used in our experiment are different in the test partitioning protocol. MLR-CUHK03 data sets adopt traditional segmentation method. Specifically, we randomly selected 100 as test sets and 1160 as training sets, which were repeated 20 times. Among them, query is everyone's all low-resolution images, and gallery is everyone's choice of all high-resolution images. MLR-SYSU uses different segmentation methods from MLR-CUHK03. The original data set SYSU contains 502 pedestrians. We divide all the pictures in this data set into two parts, one is the training set, the other is the test set. The test set is divided into query and gallery. query is all the low-resolution pictures of everyone (each has three pictures in query), and gallery is one of all the high-resolution pictures of every pedestrian (each has only one picture in gallery).

In order to evaluate the performance of the proposed pedestrian re-identification method, in our experiments, we adhibited the most frequently used Cumulative Matching Characteristics (CMC) [42] *top_k* accuracy to evaluate our proposed method and other comparative state of the art technique. All our experimental evaluation results are based on the single-gallery-shot standard, that is, each gallery

TABLE 1. Performance comparison combinations of super-resolution reconstruction and person re-identification methods.

SR Methods	ReID Methods	MLR-CUHK03				MLR-SYSU			
		rank1	rank5	rank10	rank20	rank1	rank5	rank10	rank20
Bilinear	XQDA [37]	45.5	78.0	87.8	93.7	39.3	67.4	77.2	85.6
Bicubic	XQDA	45.1	78.1	87.7	93.3	40.0	66.9	77.2	85.5
SRCNN [38]	XQDA	44.7	77.8	87.5	93.1	40.3	67.3	77.4	85.6
Bilinear	NFST [39]	48.0	47.9	46.2	49.0	41.6	69.0	79.5	87.7
Bicubic	NFST	47.9	74.8	83.6	92.8	42.4	69.3	80.5	88.0
SRCNN	NFST	49.0	74.8	85.0	92.1	43.2	69.6	80.3	88.0
Bilinear	DGD [27]	58.5	86.0	92.2	96.0	39.6	66.4	74.8	82.5
Bicubic	DGD	62.5	88.7	93.7	96.5	41.5	67.4	76.9	84.7
SRCNN	DGD	63.8	89.3	93.9	96.8	42.6	68.2	77.1	85.5
SDF [18]		22.2	48.0	64.0	80.0	13.3	26.7	42.9	66.7
SLD^2L [40]		-	-	-	-	20.3	34.8	43.4	55.4
JUDEA [16]		26.2	58.0	73.4	87.0	18.3	41.9	54.5	68.0
SING [41]		69.7	90.7	94.7	97.4	50.7	75.4	83.1	88.1
Ours		**86.9**	**98.0**	**99.0**	**99.5**	45.9	74.4	**84.1**	**91.2**

identity has only one instance. In each batch, for each query, the proposed algorithm framework will extract the feature matrix of query and all gallery samples, calculate the distance between query and all gallery images, and then arrange them in ascending order from small to large. The specific calculation method of *CMC top_k* accuracy is shown Equ.8: if the gallery samples of the *top_k* contain the target person to be checked, the Acc_k value is 1; otherwise, it is 0.

$$Acc_k = \begin{cases} 1 & r_i \leq k \\ 0 & r_i > k \end{cases} \tag{8}$$

where r_i express top-i ranked gallery samples concluding the query identity. The final *CMC* evaluation result is obtained by summing the Acc_k value of each query and then taking the average value.

B. EXPERIMENT SETTING

Inspired by [41], we propose an integrated and hybrid deep convolutional neural network. All the basic convolution layers used in this network are composed of three operations: convolution [43] layer, Batch Normalization(BN) [44] layer and ReLU [45] layer. A full connection layers consist of a Full Connection, Batch Normalization, and ReLU operation.

Our algorithm is implemented based on the deep learning framework Pytorch and runs on a workstation configured with two pieces GPU(Graphics Processing Unit) of GTX1080Ti [46]. All images used in our experiments are resized to (160, 64). Train batch size is set to 32, while test batch size is 100.

C. COMPARISON WITH STATE OF THE ART METHODS

1) EVALUATION ON CUHK03

In order to verify the effectiveness of the proposed algorithm, this method was applied to the cuhk03 dataset, and the evaluation results are shown in the Table 1. We can find apparent performance of SRPRID compared with other state-of-the-arts. The proposed SRPRID framwork outperforms the second best model SING by 17.2% (86.9% - 69.7%) in *rank*_1, 7.3% (98.0% - 90.7%) in *rank*_5, 4.3% (99.0% - 94.7%) in *rank*_10, and 2.1% (99.5% - 97.4%) in *rank*_20.

2) EVALUATION ON SYSU

Table 1 shows the evaluation results of the experiments on the SYSU dataset. The results demonstrate that the proposed SRPRID framework achieves compatitive performance over the existing methods. *SRPRID_C*(*Ours*) indicates that only the cross entropy loss function is used in the training process.

3) COMPARING COMBINATION OF SR AND REID METHODS

In order to prove the advantages of our proposed effective combination method of SR and ReID, we directly combined the existing SR and ReID with the mixed model proposed in this paper for comparison. The comparison results are also shown in Table 1. The comparison results obtained by direct combination of SR and ReID method are referred to paper [41]. In order to be fair, all comparative experiments should be carried out under the same training data and conditions. The SR methods used in our compared experiments are: Bilinear, Bilinear, Bicubic and SRCNN [38]. The ReID methods used are: XQDA [37], NFST [39] and DGD [27]. The experimental results shows the dramatic benefit of the proposed combination method.

V. CONCLUSION

This paper proposes a novel framework of multi-resolution person re-identification for Super-Recognition of Pedestrian Re-identification. The proposed framework is not a hard and fast combination of super-resolution and person re-identification, but combines the advantages of state-of-the-arts super-resolution and person re-identification methods. In particular, the features learned by the first task, namely super-resolution reconstruction task, are conducive to subsequent recognition. This paper proves that, the introduction of dense residuals into the framework is conducive to the hierarchical extraction of distinguishing features, and the introduction of attention mechanism into the pedestrian recognition subnetwork makes the recognition better. Experiments have shown that it has obvious advantages over the existing methods.

REFERENCES

[1] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, Jun. 2014, pp. 152–159.

[2] C.-C. Guo, S.-Z. Chen, J.-H. Lai, X.-J. Hu, and S.-C. Shi, "Multi-shot person re-identification with automatic ambiguity inference and removal," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 3540–3545.

[3] H. Zhao, M. Tian, S. Sun, S. Jing, J. Yan, Y. Shuai, X. Wang, and X. Tang, "Spindle Net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 907–915.

[4] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. CVPR*, Jun. 2017, pp. 420–429.

[5] L. He, L. Jian, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7073–7082.

[6] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1062–1071.

[7] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.

[8] S. Khamis, C. H. Kuo, V. K. Singh, V. D. Shet, and L. S. Davis, "Joint learning for attribute-consistent person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 134–146.

[9] X. Fei, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

[10] W. Li and X. Wang , "Locally aligned feature transforms across views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3594–3601.

[11] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 536–551.

[12] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.

[13] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[14] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[15] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737

[16] X. Li, W.-S. Zheng, X. Wang, and T. Xiang, "Multi-scale learning for low-resolution person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3765–3773.

[17] 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.

[18] Z. Wang, R. Hu, J. Jiang, J. Jiang, C. Liang, and J. Wang, "Scale-adaptive low-resolution person re-identification via learning a discriminating surface," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2016, pp. 2669–2675.

[19] Y. Zhang, Y. Tian, K. Yu, B. Zhong, and F. Yun, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2472–2481.

[20] D. Chao, C. L. Chen, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[22] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2017, pp. 136–144.

[23] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[24] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2285–2294.

[25] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[26] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 1–7.

[27] X. Tong, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1249–1258.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[29] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[30] X. Zhou, Y. Shen, Y. Zhu, and L. Huang, "Predicting multi-step citywide passenger demands using attention-based neural networks," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 736–744 .

[31] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

[32] J. Hu, L. Shen, S. Albanie, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 7132–7141.

[33] C.-H. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2568–2576.

[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[35] F. Farahnak-Ghazani and M. S. Baghshah, "Multi-label classification with feature-aware implicit encoding and generalized cross-entropy loss," in *Proc. 24th Iranian Conf. Elect. Eng. (ICEE)*, May 2016, pp. 1574–1579.

[36] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, Apr. 2018, pp. 1–8.

[37] 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.

[38] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[39] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1239–1248.

[40] X.-Y. Jing, X. Zhu, F. Wu, R. Hu, X. You, Y. Wang, H. Feng, and J. Y. Yang, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1363–1378, Mar. 2017.

[41] X. Yin and X. Liu, "Multi-task convolutional neural network for face recognition," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 964–975, 2017.

[42] H. Moon and P. J. Phillips, "Computational and performance aspects of PCA-based face-recognition algorithms," *Perception*, vol. 30, no. 3, pp. 21–303, 2001.

[43] T. N. Sainath, B. Kingsbury, G. Saon, H. Soltau, A. R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Netw.*, vol. 64, pp. 39–48, Apr. 2015.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 1–11.

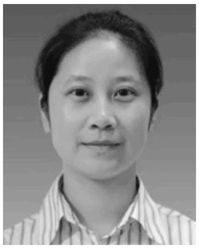
[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[46] S. Chen, Z. Jin, Y. Zhao, Z. Fu, R. Jiang, Y. Chen, and X. S. Hua, "Deep siamese network with multi-level similarity perception for person re-identification," in *Proc. ACM Multimedia Conf.*, Oct. 2017, pp. 1942–1950.

ZHEN QIN received the Ph.D. degree from the University of Electronic Science and Technology of China, in 2012, and the M.Sc. degree from the Queen Mary University of London, in 2008. He was a Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University. He is currently an Associate Professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China. His research interests include network measurement, mobile social networks, data fusion and analysis, and image processing.

WEI HE is currently pursuing the master's degree with the School of Information and Software Engineering, University of Electronic Science and Technology of China. Her research interests include data analysis and image processing.

FUHU DENG received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2006 and 2009, respectively, and the Ph.D. degree from the Dublin Institute of Technology, Ireland, in 2014. He currently holds a postdoctoral position at the University of Electronic Science and Technology of China. His research interests include wireless communications, WLAN, and big data analysis.



MENG LI received the Ph.D. degree from the University of Electronic Science and Technology of China, in 2014. Her research interests include machine learning, medical image processing, and computer-aided diagnosis.

YAO LIU received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), in 2016. She is currently an Associate Professor with the School of Information and Software Engineering, UESTC. Her research interests include machine learning, social networks, network security, and data mining.