

A Novel SRTSR Model for Cross-Resolution Person Re-Identification

XIAOQI WANG¹, XI YANG¹ , (Member, IEEE), AND DONG YANG²

¹State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

²Xi'an Institute of Space Radio Technology, Xi'an 710100, China

Corresponding author: Xi Yang (yangx@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61976166, in part by the Innovation Capacity Support Plan of Shaanxi Province under Grant 2020KJXX-027, and in part by the Young Talent Fund of University Association for Science and Technology, Shaanxi, China, under Grant 20180104.


ABSTRACT Compared to the general person re-identification (re-id), pedestrian images in cross-resolution person re-id tasks always have variant resolutions, thus the resolution mismatching problem between low-resolution (LR) images and high-resolution (HR) images significantly limits the performance of neural networks. The general solution is to introduce interpolation methods or super-resolution (SR) module to normalize all the images to the same resolution. Unfortunately, current SR methods can only upscale the entire image to a certain scale, but fail to unify the resolution of images with different width-to-height ratio. To solve this problem, we propose a Specific-Resolution-Targeted Super-Resolution (SRTSR) embedded person re-id method. Instead of using deep CNN-based SR methods, our SRTSR network consists of a fast and light-weight feature learning module with a self-adaptive Swish (SA-Swish) as its activation function to improve its performance, and a specific-resolution-targeted upscale module to upscale the images to any specific resolution. Compared to the traditional upscale module which can only accept the upscale factor for the entire image, our upscale module can either accept two independent and arbitrary upscale factors for either height and width, e.g., $2.4\times$ for height and $3.1\times$ for width, or any specific resolution, e.g., 256×128 , helping to unify the resolution of images with different width-to-height ratio. Extensive experiments are conducted and the result shows that the performance of our method is over the SR-based state-of-the-art person re-id methods on common re-id benchmarks.

INDEX TERMS Person re-ID, super-resolution, image retrieval, deep learning.

I. INTRODUCTION

As an image retrieval task, person re-identification (re-id) aims to find a particular pedestrian from the captured images in different time periods and locations. Person re-id tasks are inevitably challenged by problems in practical scenarios, i.e., low-resolution(LR) [1], shelter, pose [2], viewpoint, illumination variation [3] and the image quality. In this paper, we focus on the low-resolution and image quality problem in cross-resolution person re-id datasets.

As Fig. 1 shows, in many occasions, due to the varying physical distance between the camera and pedestrians, the resolution of pedestrian images also varies significantly, leading to the mismatching problem. The most instinct solution is to generate higher resolution images from LR images.

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

For convenience, most of representation learning-based person re-id methods [4]–[6] normalize all the images to a larger and same resolution by interpolation methods before inputting them into network. However, as convolutional neural network (CNN) develops, many SR methods start to show a better performance than the traditional interpolation methods. Thus, a couple of super-resolution (SR) method-based person re-id works have been proposed [1], [7]–[9].

Unfortunately, compared to the traditional interpolation-based SR methods [11], the flexibility of current learning-based SR methods are largely limited. For instance, the Bicubic and Bilinear interpolation, which are the two most common interpolation method used in person re-id area, can upscale or downscale the image to any specific resolution, while the SR methods based on CNN could only upscale the image by particular integral scales, commonly $2\times$, $3\times$ and $4\times$.



FIGURE 1. Illustration of mismatching problem in CAVIAR dataset [10] between Camera A (close to pedestrians) and Camera B (far from pedestrians). It is critical to match the HR and LR pedestrian images appeared in Camera A and B.

To avoid this issue, most of LR person re-id works only show their performance on artificially created multi-resolution datasets which downscale the images by an integral downscale factor, i.e., 1/2, 1/3 and 1/4 of its original resolution [8], [12]. However, on practical scenarios, the resolution of the bounding boxes of pedestrians are not in integral scales. Thus, many SR-based person re-id methods have to normalize all the images to the same resolution before the SR sub-network, which brings significant loss on their performances.

Meta-SR [13] brings us a possible solution. With the help of a weight predictor and a projection mask in upscale module, the upscale factor could be a non-integral type, e.g., 1.1× and 1.2×. However, as Fig. 2 shows, The typical SR methods (the upper half) have to upscale the images to a certain scale, then unify the resolution by interpolation methods. While in practical scenarios of person re-id tasks, it is common to face with the pedestrian images with different width-to-height ratio, and thus the interpolation methods seem to be the only choice in image preprocessing unit.

So, why not create a more flexible learning-based upscale module, making the resolution of output images directly match with each other? And this is exactly the key concept of our Specific-Resolution-Targeted Super Resolution (SRTSR) model. To ensure the time-efficiency and upscale flexibility, we modify the feature learning module, upscale module and the training method of the SR network.



FIGURE 2. Illustration of mechanism differences between the typical SR methods and ours in cross-resolution person re-id dataset, CAVIAR. Compared to the general SR methods, we generate the images of same resolution in one step using a weight matrix of specific resolution.

A. FEATURE LEARNING MODULE

Instead of merely pursuing the reconstruction performance, as a supplementary sub-network in our person re-id network, our SRTSR network is expected to achieve better performance than the interpolation based methods with a minimum time consumption. Based on the fastest SR network, FSR-CNN, we propose a light-weight feature learning module with new concepts of recent SR works, and add shrink and expand convolution layers at either ends of the basic block to achieve better time-efficiency. We also introduce a new activation function SA-Swish [14] to the SR network to improve the performance.

B. UPSCALE MODULE

The upscale module in SRTSR provides more flexible upscale choices. To reconstruct fixed-size images from arbitrary size inputs, we build a relation equation between pixels on HR and LR images. With the help of relation equation, the value of any pixel on the reconstructed HR image equals to the value of the corresponding pixel on LR image multiplies an adaptive weight. Specifically, we first generate a weight matrix of which the size is the same with the output HR image, and assign values to every element by the product of the value of pixels on LR image and the weight learned by the feature learning module. In this way, the proposed upscale module is capable to upscale images of varying resolutions to any certain resolution. Besides, the upscale module also provides the general upscale method. You can upscale the image to a certain scale by inputting two independent upscale factors for width and height.

Based on this feature, we use our SRTSR network to replace the interpolation methods as a preprocessing tool to normalize the resolution of pedestrian images in person re-id task. To achieve a better performance, we connect the SRTSR network with Resnet-50, train them with a weighted overall loss, and verify its performance on the common cross-resolution person re-id benchmarks. By using a projection mask and fully connected layer, our SR sub-network can directly normalize images, thus achieving higher precision on person re-id tasks.

II. RELATED WORKS

A. LOW-RESOLUTION PERSON RE-ID

LR person re-id aims to solve the mismatching problem between LR and High Resolution (HR) images. The family of LR person Re-id can be roughly divided into two types: metric learning-based methods and SR-based methods.

By learning the correlated features between the LR images and HR images, the metric learning-based methods can address the LR probes to the HR gallery. For example, JUDEA [5] sets two separate networks, one is especially for HR images and the other is for LR images, which avoids the LR and HR images influencing each other. Li et al. [15] first introduce GAN in person re-id tasks to reconstruct HR

features from the LR images, putting the images with different resolutions to the same feature space.

Besides, there are also a couple of works focusing on SR method. SR method aims to upscale the LR image to HR image, thus it is also the most instinct solution to resolution mismatching problem. CSR-GAN [7] focuses on how to use SR network to unify the resolutions of pedestrian images as close as possible by using the pre-defined threshold to classify the LR images to three levels, and upscale them in different degrees. SING [1] adds a SR network before the feature extraction module of person re-id network and trains the separate networks for LR and HR images jointly. RIPR [8] proposes a Foreground Focus SR method which only upscales the pedestrian body part, reducing the influence of the background clutter. However, all the aforementioned SR methods in person re-id network can only upscale the image to an integral scale, and use additional methods, e.g., parallel network and GAN, to reduce the influence of resolution mismatching problem.

B. SUPER RESOLUTION

As we mentioned above, there are couples of works done to apply the SR methods onto the person re-id task, but as time goes by, now we have more choices in SR area [16], [17]. Since the CNN-based Super Resolution, SRCNN, introduced by Dong *et al.* [18], the deep learning-based SR methods have already achieved better performance than the previous super-resolution methods. However, the SRCNN needs to upscale the input LR image by interpolation methods before it feeds into the network, which means there is an extra step compared to other SR methods, wasting considerable time. Taking a different approach, Shi *et al.* [19] propose a sub-pixel convolution layer to upscale the image by expanding the channel. And the promoted version of SRCNN, FSR-CNN [20], using the deconvolutional layer at the end of the network to upscale the image, dramatically increases the time efficiency. After then, almost all the SR network put their upscale module at the end of the network. Lim *et al.* [21] propose the EDSR and MDSR, where EDSR network adds the activation function in the skip connection connected the either end of basic blocks. And Brifman and Elad [22] use the denoiser to handle SISR problem, which also shows tendency to a high-quality output and fast processing.

Compared to the blooming development of SR networks, the multi-scale upscaling is hardly developed in recent works. MDSR network firstly introduced the Multi-scale Learning to the SR area, which allows multiple upscale factors training in a single model. But the aforementioned networks only consider to upscale the image to an integral scale ($\times 2$, $\times 3$, $\times 4$). Hu *et al.* [13] first proposed Meta-SR, with a new upscale module acting containing a learning-based weight predictor and a mask projection operator to upscale the entire image to arbitrary scale. And in this paper, we make further improvement on the meta upscale module, making the SR methods be capable to completely substitute interpolation

methods to normalize all the images to a uniform resolution in person re-id tasks.

III. OUR APPROACH

The main structure of our network can be divided into two parts, SRTSR sub-network and person re-id sub-network. The SRTSR network consists of the feature learning module which aims to extract features from LR images, and an upscale module which aims to upscale the LR feature to HR image. As for the person re-id part, to compare with other state-of-the-art methods, we use ResNet-50 as the backbone of re-id network to evaluate the performance of our SRTSR network. In this session, we describe the details of our SRTSR network and the combined loss of the entire network.

A. FEATURE LEARNING MODULE FORMULATION

As a supplementary sub-network, SR module in person re-id network is required to be time-efficient, thus the feature learning module of the SR network is expected to be compact and efficient to extract features from the cross-resolution images. So, instead of the RDN [23], ESRGAN [24] and other state-of-the-art deep SR networks, we implement our upscale module on a light-weight feature learning module.

We redesign the feature learning module of the SR network to balance the performance and time efficiency. Skip connection with activation function and the residual in residual strategy [25], [26] is used to improve the performance, meanwhile the expand and shrink 1×1 convolution layer are added at either end of every basic block. As the experiments [21] prove that the batch normalization (BN) layer increases the computational complexity with minor improvement on SR tasks, it is also removed from our basic block.

Like many other SR methods [24], [27], we only process the Y channel in YCbCr color space. Thus the output channel of the upscale module is 1, and the other channels, i.e., Cb and Cr, will be added by the skip connection.

Table 1 illustrates the structure of our feature learning module and some details of our upscale module. Whereas the size of image could be arbitrary, the change of image size will also influence the flops calculation. So in this table, we assume the input image is 256×128 . As Fig.3 shows, to achieve better performance, we add the LR image upscaled

TABLE 1. The architecture of our SR network with the input image size 256×128 and the upscale factors for height and width are both $2.0 \times$. The number of the Basic Blocks is 6.

Stage	Output	Operation	Channels
Conv1	256×128	5×5 Conv	56
Basic Blocks	256×128	1×1 Conv	12
		3×3 Conv	12
		3×3 Conv	12
		1×1 Conv	56
Upscale Blocks	512×256	FC layer	256
	512×256	FC layer	9
Flops	321.39M	-	-
Params	569.9K	-	-

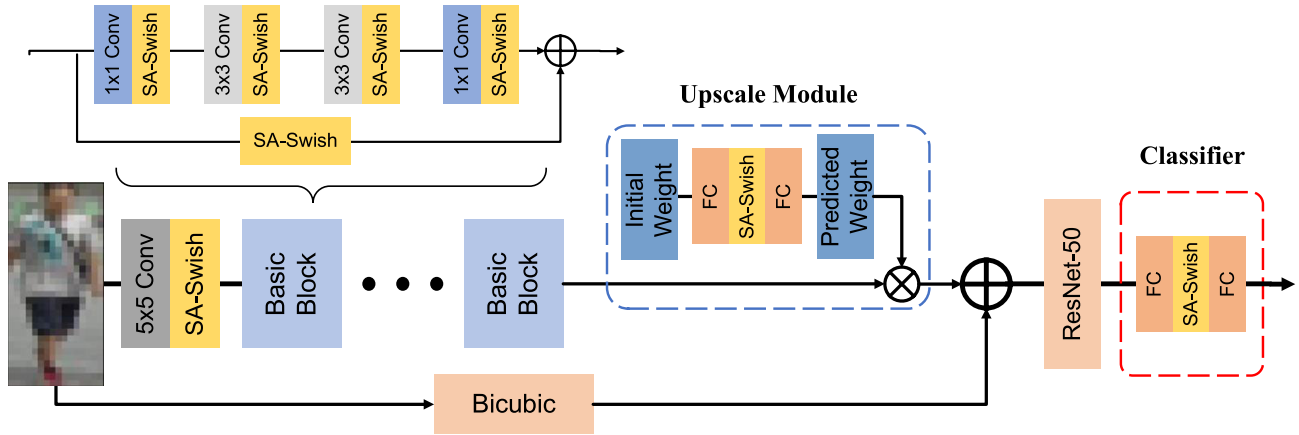


FIGURE 3. A visual diagram of our entire network. Input images are firstly put into the SR sub-network, passing through the feature extractor and upscale module, then the output HR images will be put into the re-id network. The loss function is the combined loss of both SR sub-network and re-id network.

by bicubic interpolation as an additional input at the end of the SR network.

Current SR networks prefer to choose simple and common activation functions like Sigmoid or ReLU. However, as Fig.4 illustrates, a constant upper boundary in Sigmoid raises the possibility of gradient vanishing, and the ReLU is not an ideal smooth activation function. The SR task always requires a smooth output, which could help to achieve a better performance in PSNR evaluation results. Taking all the aspects into consideration, we initially use the Swish as our activation function:

$$Swish(x) = x \cdot sigmoid(x) = x \cdot \frac{1}{1 + e^{-x}} \quad (1)$$

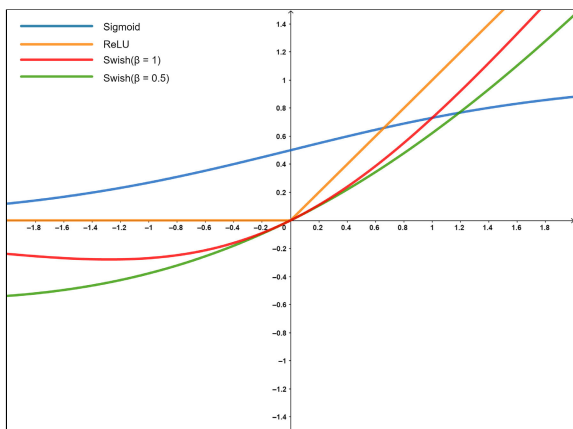


FIGURE 4. The activation comparison between Swish (red and green), Sigmoid (blue) and ReLU (black).

Meanwhile, as Fig.4 indicates, if we add a weight, β , into Sigmoid function in Swish, the activation function could be smoother, i.e.,

$$SA-Swish(x) = x \cdot sigmoid(\beta x) = x \cdot \frac{1}{1 + e^{-\beta x}} \quad (2)$$

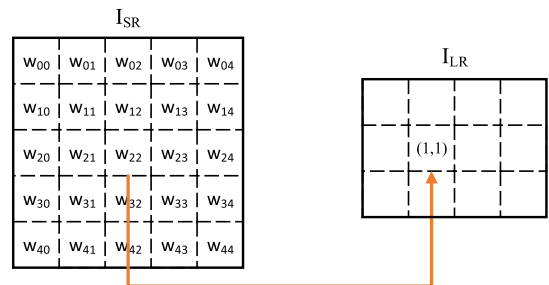


FIGURE 5. An instance of our upscale module. To get a 5×5 SR image from a 3×4 LR image, we first need to generate a 5×5 weight matrix by using a FC layer, and for each pixel in the SR image, we need to find the corresponding pixel in LR image, e.g., the value of pixel (1, 1) in I_{SR} equals to the product of W_{11} and the value of pixel (1, 1) in I_{LR} .

However, when β reduces to 0, SA-Swish becomes a monotonic and linear, and when β is larger than 1, it becomes no longer smooth, so it is necessary to set a clear boundary to avoid gradient exploding and vanishing. Thus, we decided to design the self-adaptive Swish (SA-Swish) to let itself to determine the value of β by changing it to a self-adaptive parameter and set a boundary $\beta \in (0, 1]$ in function swish, and set the initial value of β to 0.9.

B. UPSCALE MODULE FORMULATION

Showing an outstanding performance among various upscale modules [19], [28], [29], deconvolutional layer is still the most popular upscale module in SR networks [30], [31]. On the contrary, the stride of deconvolutional layer, which determines the upsampling scale, has to be an integer number. Thus for SR method based person re-id tasks, the common solution is to normalize images to one or several resolutions before one or cascaded SR networks [7]. Inspired by Meta-SR [13], to better utilize the effectiveness of SR network, we propose a novel upscale module which firstly creates the weight matrix with the same size of the HR output and establishes the relationship between pixels.

As Fig. 5 shows, the core concept of our upscale module is that the pixel value at (i, j) in the output SR image equals to the weighted pixel value at (i', j') in input LR image. To establish the relationship between pixels on HR and LR images, we introduce a instinct projection equation:

$$(i, j) = f(i', j') = f\left(\left\lfloor \frac{i}{r_0} \right\rfloor, \left\lfloor \frac{j}{r_1} \right\rfloor\right), \quad (3)$$

where symbol $\lfloor \cdot \rfloor$ represents the floor function, the r_0 denotes upscale factor for height and r_1 denotes that for width. The floor function here is to avoid fatal errors as the upscale factors could be any float numbers.

Person re-id task always requires the image preprocessing unit before the network to unify the input images to a certain output resolution, instead of giving a specific upscale factor. So, during the testing phase, the upscale factors are no longer the given parameters, but calculated by

$$r_0 = \frac{outH}{inH} \text{ and } r_1 = \frac{outW}{inW}. \quad (4)$$

The *outH* and *outW* stand for output height and width, while *inH* and *inW* stand for input height and width. Especially, different from the testing process, in the training process, the network randomly downscale the images to lower resolutions (1/4 to 1 of its original resolution), and learn how to upscale them back. In a short word, the r_0 and r_1 are randomly generated values in the training phase, while it is determined by the target resolution and input resolution in the testing phase.

Given the weight matrix and the projection equation, the pixel value at (i, j) in the output SR image can be written as the product of corresponding weight for (i, j) and the pixel value at (i', j') in the LR feature F^{LR} :

$$PV^{SR}(i, j) = W(i, j) F^{LR}(i', j') \quad (5)$$

In this way, we can find the corresponding pixel for every pixel in the output SR image. Then, we need to generate a weight matrix $W(i, j)$ of which the size is exactly the same with the output SR image.

To achieve better performance, initial value in the weight matrix cannot be idle. As the floor function in fact causes the loss of precision in projection process, and to reduce its influence, we set the initial weight as the offset between the real value and the ideal value of the correspondence:

$$W_{init}(i, j) = \left(\frac{i}{r_0} - \left\lfloor \frac{i}{r_0} \right\rfloor, \frac{j}{r_1} - \left\lfloor \frac{j}{r_1} \right\rfloor \right) \quad (6)$$

With initial weights, we can create a weight matrix with the same size of the output SR image. Passing through two fully connected (FC) layers, every single element in the weight matrix becomes an adaptive weight, and by multiplying the adaptive weight and the value of corresponding pixels on LR images, the values of pixels on SR images are obtained.

C. OVERALL LOSS FUNCTION

The overall loss function consists of two parts: the MAE loss for SR network,

$$L_1(I_i - \hat{I}_i) = \frac{1}{N} \sum_{i=1}^N \|I_i - \hat{I}_i\|_1 \quad (7)$$

Here, the I_i and \hat{I}_i refers to the HR images and our output SR images, respectively. The cross entropy loss for the person re-id network,

$$CrossEntropy(FC(f_i) - L_i) = - \sum_{i=1}^N FC(f_i) \log(L_i) \quad (8)$$

In practice, we first solely train our SR sub-network by L1 loss, obtain the pre-trained model, and then train the entire network by the combined loss. Thus, as for the overall loss, we add a tiny weight α to balance the losses. So the combined loss is,

$$L = \alpha L_1(I_i - \hat{I}_i) + CrossEntropy(FC(f_i) - L_i) \quad (9)$$

IV. EXPERIMENTS

A. DATASETS AND IMPLEMENTATION DETAILS

We evaluate our person re-id network on four cross-resolution re-id benchmarks, VR-Market, VR-MSMT17 [8] and self-created VR-Market1501-Hard by accepting non-integral resolution proportions, and we remain the original division settings of training and testing sets.

The VR-Market1501 contains 12,936 images of 750 pedestrians in the training set, and 19,732 images of 751 pedestrians in the testing set. And in VR-MSMT17 dataset, there are 32,621 images of 1041 persons in training set and 93,820 images in testing set. Compared to their original dataset, all the images in VR-Market1501 and VR-MSMT17 are downscaled to $(\frac{1}{4}, 1]$ time of their original scale.

Particularly, to demonstrate the effectiveness of the meta upscale module, we create a new dataset VR-Market1501-Hard, where the images are randomly downscaled from 1/4 to 1, of which the downscale factor is generated by $\text{random.uniform}(1.01, 4)$. To keep the images from serve distortion, the difference of downscale factor for width and for height is restricted to less than or equal to 0.4.

All the time-related experiments are run on the computer with an I7-9700K CPU and a 2070S GPU. To pursue better performance on either image quality assessments (PSNR/SSIM) and recognition accuracy, the SRTSR network will be trained by the corresponding person re-id dataset first, and then the entire network will be trained together with the help of the pretrained model of SRTSR network. During the holistic training process, weight of L1 loss for SRTSR network is 0.05.

Due to the previous LR person re-id works only present Rank-1 and Rank-5 results, in this paper, all the person re-id results are also present in Rank-1 and Rank-5 form. In person re-id results, we fixed the size of the weight matrix while in

image quality assessments (IQA), we use the fixed upscale factors for height and width. Since the other SR methods, e.g. RCAN and Meta-SR, costs much time once trained with person re-id network, all the IQA(PSNR/SSIM) results are obtained by training the SR network individually.

B. EFFECTIVENESS OF SA-SWISH

In this session, we compare the effectiveness of three activation functions: ReLU, LeakyReLU, P-ReLU, Swish and our SA-Swish. To better solely compare the difference between the activation functions, we substitute our activation function to the other, use the traditional upscale method, i.e., by accepting a upscale factor for the entire image, and run the program for 20 epochs on Market-1501 dataset.

The average time in Table 2 indicates total time for one entire epoch, and as the result of an additional self-adaptive parameter, SA-Swish and P-ReLU cost longer but acceptable time to achieve better performance than other activation functions. However, generally, the activation function will not cost too much time, while in this experiment, the processing time varies a lot. We think it is mainly because the convolution part is so light-weight and fast, thus the time-consuming proportion of activation functions with multiple self-adaptive parameters increases.

TABLE 2. Comparison between activation functions, the bold denotes the best result and underline denotes the second.

Method	Scale	PSNR	SSIM	Average Time
ReLU	2×	37.57	0.9312	7min38s
	3×	32.59	0.8384	
	4×	30.12	0.7504	
LeakyReLU	2×	37.71	0.9302	<u>7min41s</u>
	3×	32.91	0.8401	
	4×	30.21	0.7522	
P-ReLU	2×	<u>37.97</u>	<u>0.9371</u>	9min12s
	3×	<u>33.16</u>	<u>0.8436</u>	
	4×	30.40	0.7524	
Swish	2×	37.50	0.9348	8min03s
	3×	32.64	0.8421	
	4×	<u>30.41</u>	<u>0.7543</u>	
SA-Swish	2×	38.01	0.9413	8min41s
	3×	33.27	0.8439	
	4×	30.71	0.7640	

For higher upscale factors, i.e., 3× and 4×, SA-Swish achieves the best results among the other. To show the effectiveness of SR methods better, in the following experiments, we run our program on these cross-resolution datasets mostly consists of very low resolution images. Thus, we use the proposed SA-Swish as the activation function of our SRTSR network and re-id network.

C. EFFECTIVENESS OF OUR SR MODULE

To comprehensively verify the effectiveness of SRTSR, in this session, we first validate the effectiveness of each part of our SR network by the ablation study, and then compare it with the common state-of-the-art SR methods, i.e., ESRGAN [24], RCAN [25] and Meta-SR [13] from two perspectives:

1) Upscale images by a certain scale, and 2) Upscale images to a certain resolution.

In our ablation study, we set the FSRCNN [20] as the basic network, which is named as A, and divide our contribution by three parts, B.the meta upscale module; C. substitute the 3 × 3 convolutional layer to the basic block of our network; D. the number of basic block. Also, in order to evaluate the effectiveness of the skip connection, named as E, we test the network with and without the skip connection, respectively. In the following experiment, we evaluate the effectiveness of each change by objective indexes, i.e., PSNR, SSIM and person re-id precision, rank-1 and rank-5. For example, A+B means FSRCNN with the new upscale layer, but without the basic block. And it is noticeable that A+B+C means we have substituted two convolutional layer to our basic block, and D indicates the number of extra basic blocks besides them.

Table 3 illustrates the IQA results, PSNR and SSIM, and the person re-id accuracy, rank-1 and rank-5, of the ablation study. The person re-id accuracy results are obtained by training the joint network on VR-Market1501 dataset, while the IQA results are obtained by solely training the SR network.

TABLE 3. PSNR, SSIM evaluation and time consumption of different network structures: A. FSRCNN, B. the meta upscale module; C. substitute the 3 × 3 convolutional layer to the basic block of our network; D. the extra number of basic blocks; E. the skip connection.

Methods	PSNR(dB)	SSIM	Rank-1	Rank-5	Time
A	32.09	0.8534	54.1	76.8	5m41s
A+E	32.86	0.8605	54.3	76.8	5m43s
A+B+E	33.61	0.8619	56.7	78.1	6m24s
A+B+C+E	34.06	0.8653	57.0	80.0	7m38s
A+B+C+2D+E	<u>34.84</u>	0.8739	58.6	<u>80.6</u>	8m31s
A+B+C+4D+E	35.07	0.8774	59.0	81.2	10m46s
A+B+C+6D+E	34.82	0.8741	<u>58.9</u>	80.1	12m10s

As Table 3 shows, the person re-id accuracy is highly related with the IQA results, PSNR and SSIM. With the highest PSNR/SSIM results, A+B+C+4D+E achieves the best person re-id accuracy. The table also indicates the SR network with 6 basic blocks (A+B+C+4D+E) is deep enough to handle deep LR images in person re-id datasets, as its performance reaches the peak compared to either 4 or 8 blocks. Thus, in the following experiments, the number of basic blocks in the SR sub-network is 6.

In Fig. 6, we prove the effectiveness of our SR network by training all the SR methods by using the training set of Market-1501 and evaluate their SR performance in the query set of Market-1501 dataset by PSNR and SSIM on the Y channel in YCbCr color space. All the chosen images will firstly downscale to $\frac{1}{r_0}$ by $\frac{1}{r_1}$ time of its original resolution, where r_0 and r_1 are upscale factors for height and width, respectively. And then upscale back to the original resolution, then compare the PSNR and SSIM value with the original HR images. Since the other SR methods cannot upscale the image to arbitrary size, so when we are using the other SR methods, we firstly upscale the image to $\lceil rH, rW \rceil$, where symbol $\lceil \cdot \rceil$ denotes ceil function, and then downscale it to the original resolution by bicubic interpolation.

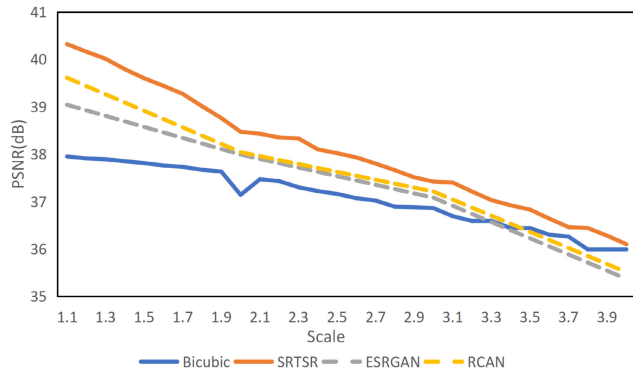


FIGURE 6. PSNR comparison between our SRTSR network and other SR methods. The upscale factor for height is 2.0 and that for width varies from 1.1 to 4.0.

Fig. 6 shows the performance of SR methods, and the solid line indicates that this method can upscale the input image to any scale directly, while the dotted line indicates this method need the help of interpolation methods to upscale the image to the sample points.

Though the performance of our network reduces faster than any other SR methods, it still performs better than them when the upscale factor for width is between 1.0 and 3.9, when 1.0 to 3.0 are always considered as the most useful range. The rapid descend may caused by the projection relationship. When the difference of upscale factors is large enough, there will be multiple pixels in SR image corresponds to the same pixel in LR image, which significantly limit its performance.

Meanwhile, Fig. 6 also illustrates that when the upscale factor for width is smaller than 2x, our SR method performs much better than the other. In this range, our projection relationship is still strong while the other SR methods can only upscale the image to 2x and then downscale them by interpolation method. However, the range from 1x to 2x is exactly the most common upscale factors the person re-id requires, so ours will be a more suitable SR method for re-id task.

When accepting the integral upscale factors, our SR module does not lead much compared to the state-of-the-art SR methods as the result of pursuing time efficiency, e.g., the total processing time of upscaling a single image (including the time of read file and etc., on Market-1501 dataset [32]) for RCAN is 1.51s and for ESRGAN is 132ms, while ours is 34ms, which could be even much faster in end-to-end network.

The visual results shows a couple of example of the SR results and aims to evaluate the single image SR performance of the SR models. We compare our SR method with the traditional bicubic interpolation, and other common SR networks. The upscale factors for height and width are 2.7x and 2.4x respectively. When we are using other SR methods, we first upscale the image to 3x, and then downscale to the same resolution by bicubic interpolation.

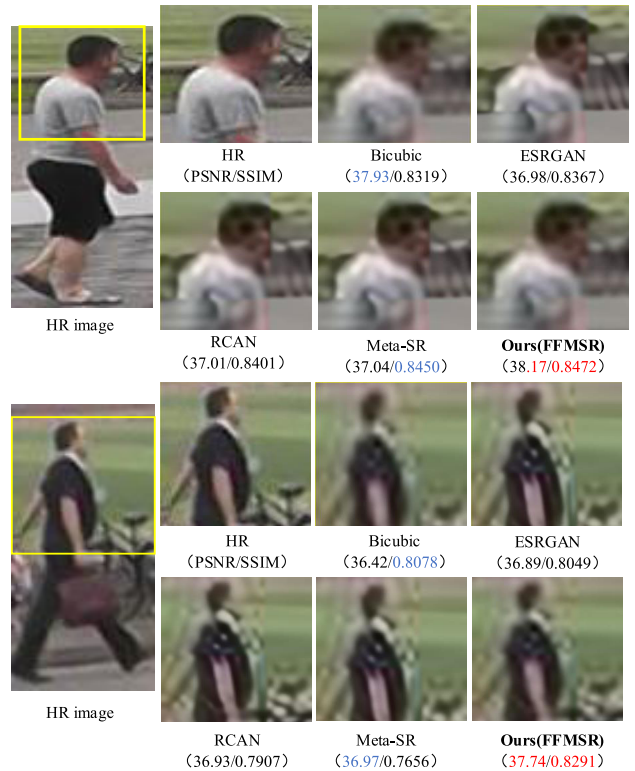


FIGURE 7. Visual results and PSNR/SSIM comparison between our SRTSR network and other SR methods. The upscale factor for height is 2.4 and that for width is 2.7.

Surprisingly, the visual results (Fig. 7) and the objective IQAs suggests bicubic interpolation, which is demonstrated as the least competitive SR method among the other, could rank the second or the third. The results could possibly mean the SR + interpolation sometimes performs less competitive than directly using interpolation methods.

We can see that the distortion happens when normalizing the images upscaled by the other SR methods, which also explains the reason why precision reduces in person re-id tasks when we simply substitute the interpolation methods by SR methods. Meanwhile, the distortion brings severe influence on the evaluation results, especially on SSIM values. Our SRTSR method uses two upscale factors for height and width respectively, so the SRTSR can directly upscale to image to the desired resolution. In this way, the SRTSR avoids the distortion and performs better than the other SR methods.

D. COMPARISON WITH STATE-OF-THE-ART PERSON RE-ID METHODS

In this session, we compare our person Re-id result with other SR method-based person Re-id solutions.

Table 4 shows the Rank-1/Rank-5 results of the SR methods-based person re-id in three cross-resolution datasets. Compared to the baseline, our SRTSR raises the Rank-1/Rank-5 significantly. And our SRTSR performs the best among the other SR methods in person re-id area in the three cross-resolution dataset. As our SR module can upscale the

TABLE 4. Person re-id performance on four cross-resolution re-id benchmarks, the bold font denotes the best result.

Methods	VR-Market1501		VR-MSMT17		MLR-CUHK-03	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
Baseline	57.3	79.2	49.2	66.8	56.3	79.2
FFSR [8]	59.2	80.1	52.8	69.0	70.5	92.3
CSR-GAN [7]	59.8	81.3	51.9	67.5	68.9	92.2
SING [1]	60.5	81.8	52.1	68.3	67.7	90.7
SRTSR	60.3	82.9	52.9	70.2	71.9	92.8

input image to arbitrary size, our re-id network can achieve better performance on cross-resolution person re-id dataset compared to other SR network embedded re-id network. Also, it is noticeable that most the SR solutions performs well in MLR-CUHK-03 dataset. Especially, our SRTSR outperforms the baseline by 15.6%/13.6%, which may indicates that the SR methods can achieve ideal result on png format images rather than jpg format images.

Similarly, Table 5 shows the same situation, the SR methods performs better in MLR-CUHK-03 dataset than the other two jpg format datasets. In this part, we use our SR network to generate a clearer and HR dataset from the original one, and substitute bicubic interpolation as a new normalization tool. And Table 5 proves our SR sub-network is capable to increase the accuracy of other state-of-the-art person re-id methods on cross-resolution datasets.

Table 5 also shows the result of person re-id precision on self-created dataset, VR-Market1501-Hard. The increasing number of varying resolutions does not bring much difficulty for bicubic interpolation, as the precision only drops 2.4%, which could attribute to the difference between the upscale factors for width and height. One noticeable thing is that the precision of FSRCNN + bicubic interpolation drops significantly, and in some cases, it is even lower than

TABLE 5. Person re-id performance of four different re-id methods on three cross-resolution benchmarks (Normalized by Bicubic interpolation), FSRCNN, ESRGAN, and SRTSR, respectively. The bold font denotes the best result.

Bicubic	VR-Market1501		VR-MSMT17		MLR-CUHK-03		VR-Market-Hard	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
ResNet50	57.3	79.2	49.2	66.8	56.3	79.2	54.9	77.4
BoT [33]	66.0	84.3	63.2	76.5	62.9	81.3	63.2	83.8
ABD-Net [34]	68.8	84.9	66.1	79.5	64.3	82.2	64.9	84.2
AGW [35]	69.1	86.2	67.6	81.7	63.6	83.7	65.4	84.3
FSRCNN	VR-Market1501		VR-MSMT17		MLR-CUHK-03		VR-Market-Hard	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
ResNet50	57.6	79.8	49.4	67.6	60.7	80.6	54.1	76.8
BoT	66.8	85.0	64.1	76.9	66.4	84.7	62.7	83.5
ABD-Net	70.0	88.2	68.2	81.7	67.2	85.6	64.2	83.9
AGW	70.5	88.8	69.3	83.4	67.7	90.7	64.6	84.3
ESRGAN	VR-Market1501		VR-MSMT17		MLR-CUHK-03		VR-Market-Hard	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
ResNet50	58.0	80.8	50.2	68.5	63.2	80.6	54.7	77.3
BoT	68.2	86.4	64.2	77.2	66.9	87.1	64.0	85.1
ABD-Net	70.2	88.2	68.7	82.6	68.0	88.4	65.9	86.4
AGW	70.4	89.0	69.6	83.5	68.5	90.9	67.2	86.8
SRTSR	VR-Market1501		VR-MSMT17		MLR-CUHK-03		VR-Market-Hard	
	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5	Rank-1	Rank-5
ResNet50	59.7	81.6	51.0	69.8	67.4	90.7	59.0	81.2
BoT	68.0	86.5	64.9	79.0	69.9	90.4	67.9	86.2
ABD-Net	71.8	90.3	69.3	82.2	70.8	92.0	71.2	89.7
AGW	71.6	90.8	70.7	84.2	72.1	92.6	71.2	90.0

solely using bicubic interpolation, and so does the ESRGAN. Compared to FSRCNN, ESRGAN + bicubic only provides minor improvements, which is around 0.4% to 2.5%. By contrast, our SRTSR achieve the best result, and almost remains the same level on VR-Market1501 and VR-Market1501-Hard datasets, which could indicate ours can achieve ideal performance on datasets of which the resolutions are in non-integral proportion.

V. CONCLUSION

We propose a supplementary super resolution network with a novel activation function, SA-Swish, to upscale the pedestrian images to the same resolution directly. The proposed light-weight feature learning module is time efficient and accurate, and for each pair of upscale factors, the proposed upscale module generates a weight matrix and a new correspondence between LR and SR images. The upscale module can generate the SR pedestrian images of arbitrary size.

The extensive experiments show that the SA-Swish achieves better performance than Swish, and our SR module performs the best especially on datasets with varies of resolutions. And the experiments on datasets of different formats may indicates that the SR methods could show better performance on png format images than jpg format images. Similarly, with the help of our SR network, the person re-id network achieves the best Rank-1/Rank-5 results among the other SR method-based are-id networks.

REFERENCES

- [1] J. Jiao, W. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. AAAI*, 2018, pp. 6967–6974.
- [2] C. Zhang, L. Zhu, S. Zhang, and W. Yu, "PAC-GAN: An effective pose augmentation scheme for unsupervised cross-view person re-identification," *Neurocomputing*, vol. 387, pp. 22–39, Apr. 2020.
- [3] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 608–617.
- [4] 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.
- [5] 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.
- [6] S. Zhu, X. Gong, Z. Kuang, and J. Du, "Partial person re-identification with two-stream network and reconstruction," *Neurocomputing*, vol. 398, pp. 453–459, Jul. 2020.
- [7] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded SR-GAN for scale-adaptive low resolution person re-identification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3891–3897.
- [8] S. Mao, S. Zhang, and M. Yang, "Resolution-invariant person re-identification," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 883–889.
- [9] Y. Ha, J. Tian, Q. Miao, Q. Yang, J. Guo, and R. Jiang, "Part-based enhanced super resolution network for low-resolution person re-identification," *IEEE Access*, vol. 8, pp. 57594–57605, 2020.
- [10] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification," in *Proc. Brit. Mach. Vis. Conf.*, 2011, p. 6.
- [11] Y. Liu, Y. Yang, Y. Shu, T. Zhou, J. Luo, and X. Liu, "Super-resolution ultrasound imaging by sparse Bayesian learning method," *IEEE Access*, vol. 7, pp. 47197–47205, 2019.

- [12] T. Xiao, 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.
- [13] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-SR: A magnification-arbitrary network for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1575–1584.
- [14] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*. [Online]. Available: <http://arxiv.org/abs/1710.05941>
- [15] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C.-F. Wang, "Recover and identify: A generative dual model for cross-resolution person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8090–8099.
- [16] Y. Zhang, Q. Fan, F. Bao, Y. Liu, and C. Zhang, "Single-image super-resolution based on rational fractal interpolation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3782–3797, Aug. 2018.
- [17] T.-A. Song, S. R. Chowdhury, F. Yang, and J. Dutta, "PET image super-resolution using generative adversarial networks," *Neural Netw.*, vol. 125, pp. 83–91, May 2020.
- [18] C. Dong, C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. ECCV*, Sep. 2014, pp. 184–199.
- [19] W. Shi, J. Caballero, F. Huszar, 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.
- [20] C. Dong, C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. ECCV*, Oct. 2016, pp. 391–407.
- [21] 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. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.
- [22] A. Brifman, Y. Romano, and M. Elad, "Turning a denoiser into a super-resolver using plug and play priors," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1404–1408.
- [23] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
- [24] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Chen, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. ECCV Workshops*, Sep. 2018, pp. 63–79.
- [25] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. ECCV*, Sep. 2018, pp. 286–301.
- [26] B. Singh, D. Toshniwal, and S. K. Allur, "Shunt connection: An intelligent skipping of contiguous blocks for optimizing MobileNet-V2," *Neural Netw.*, vol. 118, pp. 192–203, Oct. 2019.
- [27] H. M. Kasem, M. M. Selim, E. M. Mohamed, and A. H. Hussein, "DRCS-SR: Deep robust compressed sensing for single image super-resolution," *IEEE Access*, vol. 8, pp. 170618–170634, 2020.
- [28] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," 2018, *arXiv:1808.08718*. [Online]. Available: <http://arxiv.org/abs/1808.08718>
- [29] 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.
- [30] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.
- [31] B.-C. Yang, "Super resolution using dual path connections," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1552–1560.
- [32] 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.
- [33] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1487–1495.
- [34] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8350–8360.
- [35] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: <http://arxiv.org/abs/2001.04193>



XIAOQI WANG received the B.E. degree in communications engineering from Xidian University, Xi'an, China, in 2019, where he is currently pursuing the M.S. degree in communication and information system. His research interests include deep learning, image processing, and person re-identification.



XI YANG (Member, IEEE) received the B.Eng. degree in electronic information engineering and the Ph.D. degree in pattern recognition and intelligence system from Xidian University, Xi'an, China, in 2010 and 2015, respectively. From 2013 to 2014, she was a Visiting Ph.D. Student with the Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX, USA. In 2015, she joined the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, where she is currently an Associate Professor in communications and information systems. Her current research interests include image/video processing, computer vision, and multimedia information retrieval.



DONG YANG received the B.Eng. degree in electronic information engineering and the Ph.D. degree in signal processing from Xidian University, Xi'an, China, in 2010 and 2015, respectively. He is currently with the Xi'an Institute of Space Radio Technology. His current research interests include image processing and spaceborne radar systems.

...