

A Deep Multitask Convolutional Neural Network for Remote Sensing Image Super-Resolution and Colorization

Jianan Feng^{ID}, Qian Jiang^{ID}, Ching-Hsun Tseng^{ID}, Xin Jin^{ID}, *Member, IEEE*, Ling Liu^{ID}, Wei Zhou^{ID}, *Member, IEEE*, and Shaowen Yao^{ID}, *Member, IEEE*

Abstract—Remote sensing data have become increasingly vital in target detection, disaster monitoring, and military surveillance. Abundant pan-sharpening and super-resolution (SR) methods based on deep learning have been proposed and have achieved remarkable performance. However, pan-sharpening requires paired panchromatic (PAN) and multispectral (MS) images, and SR cannot increase the spectral resolution of PAN. Thus, we introduce a computational imaging-based method to recover or produce the incomplete data of single PAN or MS. This work also explores the integration of multiple tasks by a single neural network. We start with SR and colorization, study the feasibility of simultaneously finishing SR colorization, and use a model trained in SR colorization to finish pan-sharpening without MS. A generic neural network, remote sensing image improvement network (RSI-Net), is designed for remote sensing image SR, colorization, simultaneous SR colorization, and pan-sharpening. To verify its performance, RSI-Net is compared with the state-of-the-art SR and colorization methods. Experiments show that RSI-Net can be competitive in visual effects and evaluation indexes, and it performs well at simultaneous SR colorization, and RSI-Net finishes pan-sharpening and only needs to input PAN. Our experiments confirm the effect of integrating multiple tasks.

Index Terms—Convolutional neural network (CNN), deep learning (DL), image colorization, image super-resolution (SR), remote sensing image.

I. INTRODUCTION

REMOTE sensing images play a crucial role in disaster monitoring, target detection, military reconnaissance, and other fields. Panchromatic (PAN) and multispectral (MS) images are two vital components in remote sensing images, with information from different bands over the same area. PAN has high resolution (HR) but lacks spectral information,

Manuscript received November 25, 2021; revised January 18, 2022; accepted February 12, 2022. Date of publication February 24, 2022; date of current version April 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62101481, Grant 62002313, Grant 61863036, and Grant 11861071; in part by the Key Areas Research Program of Yunnan Province in China under Grant 202001BB050076; and in part by the Key Laboratory in Software Engineering of Yunnan Province under Grant 2020SE408. (*Corresponding author: Xin Jin.*)

Jianan Feng, Qian Jiang, Xin Jin, Ling Liu, Wei Zhou, and Shaowen Yao are with the Engineering Research Center of Cyberspace, Yunnan University, Kunming 650091, China, and also with the School of Software, Yunnan University, Kunming 650091, China (e-mail: falleneric@163.com; jiangqian_1221@163.com; xinxin_jin@163.com; liuling@mail.ynu.edu.cn; zwei@ynu.edu.cn; yaosw@ynu.edu.cn).

Ching-Hsun Tseng is with the School of Computer Science, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: ching-hsun.tseng@postgrad.manchester.ac.uk).

Digital Object Identifier 10.1109/TGRS.2022.3154435

while MS has spectral information but lacks spatial details. Because of the limitations of satellite sensors, remote sensing images obtained from a onefold sensor cannot have both high spatial resolution and high spectral resolution. Except for boosting the performance of physical imaging equipment of the satellite, the super-resolution (SR) and pan-sharpening methods are effective. The development of efficient computing hardware and advanced algorithms has strengthened the capacity of deep learning (DL) to deal with nonstructural data. DL has shown superiority in computer vision and image processing fields, such as image pan-sharpening [1], [2], SR [3]–[5], and colorization [6], [7].

The pan-sharpening methods can fuse detailed PAN information and MS color information, but will fail when either is missing. Likewise, the SR methods cannot improve the spectral resolution of PAN. We note that pan-sharpening is a process of enlarging the spatial and spectral resolutions of PAN, and this can be equivalent to simultaneous SR and colorization. Therefore, we explore SR and colorization for MS separately and study the feasibility of finishing SR colorization simultaneously, and finally use a model trained in SR colorization to finish pan-sharpening without MS. This work shows that a multitask neural network can be used to recover or produce high-quality remote sensing images with incomplete data.

Image SR aims to restore and recover the HR image from a low-resolution (LR) image via SR algorithms. SR is an ill-posed problem, since multiple HR images may result in the same LR image. Its most common strategy is to artificially construct paired LR-HR datasets. The first simple convolutional neural network (CNN), super resolution convolutional neural network (SRCNN), was proposed by Chao *et al.* [8]. Although it contains only three convolutional blocks, it still performs better than conventional algorithms. The SR methods have been updated regularly, performing impressively in terms of the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). Many deeper DL-based single image super resolution (SISR) neural networks have achieved state-of-the-art performance, such as faster super resolution convolutional neural network (FSRCNN) [9], Laplacian pyramid network (LapSRN) [10], deep recursive residual network (DRRN) [11], residual dense network (RDN) [12], super resolution generative adversarial network (SRGAN) [13], multi-scale residual network (MSRN) [14], super resolution

feedback network (SRFBN) [15], and dual regression network (DRN) [16].

Image colorization aims to obtain a color image with a grayscale image. Analogously, a grayscale image can correspond to a crop of color images. Image colorization is a challenging problem, as two out of three image dimensions are lost. A variety of colorization methods have been successfully applied to colorization problems, ranging from brute-force CNNs [17], [18] to carefully designed generative adversarial networks (GANs) [19], [20], which differ in aspects such as training strategies, network structures, and loss functions. Iizuka *et al.* [17] introduced a classification network to assist in colorizing grayscale images. Yoo *et al.* [20] proposed MemoPainter, which finds color information in training sets. These colorization methods have been applied to natural scenes, animation, cartoons, and image translation.

The pan-sharpening methods aim to inject MS color information into PAN images; hence, they cannot get rid of either. In other words, the pan-sharpening methods will fail when PAN or MS information is missing. However, it is costly to obtain paired PAN and MS.

SR and colorization are generally treated as separate tasks. The model structure must be carefully designed for each task, which requires much effort. Most researchers do not consider potential associations between them, and they combine them as an integrated image generation problem. Both SR and colorization hope to give supplementary information to the input image, but they are diverse in supplying supererogatory pixels to input. Therefore, we wish to design a general neural network structure to adapt to different visual tasks for remote sensing images. To solve the problems above, we propose remote sensing image improvement network (RSI-Net), a neural network to super-resolve and colorize remote sensing images, which is capable of adapting SR, colorization, and simultaneous SR colorization of these three visual missions for remote sensing images with limited data. The main contributions of this work are as follows.

- 1) We propose a multitasking architecture that can complete four tasks for remote sensing images:
 - a) Image SR with scale factor $\times 2$, $\times 4$, $\times 8$;
 - b) Image colorization;
 - c) Image SR colorization simultaneously when given an LR grayscale input;
 - d) Pan-sharpening requiring only PAN input.
- 2) We design a dual attention block to transfer significant feature information in image generation tasks, inspired by convolutional block attention modules (CBAMs).
- 3) Inspired by Inception modules, we propose a multiscale residual block (MRB) for feature extraction and reconstruction in our architecture.

II. RELATED WORK

Recent years have witnessed the rapid development and application of neural networks in image processing. Diverse neural network architectures have been designed for different visual tasks. We present a brief introduction to remote sensing images, DL-based SR algorithms, and colorization algorithms.

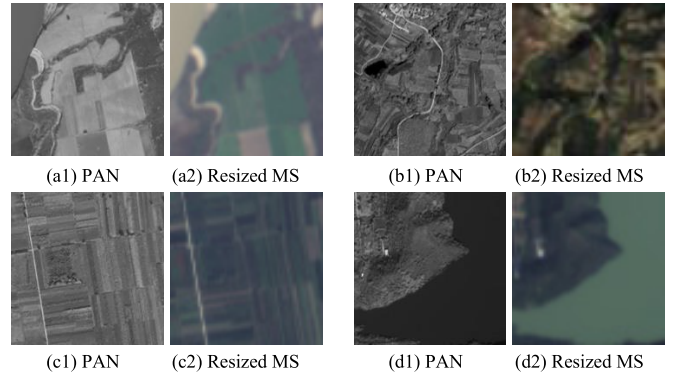


Fig. 1. PAN and MS images.

A. Remote Sensing Image

Remote sensing images are photographs that record the magnitudes of electromagnetic waves of various features and are mainly divided into PAN, MS, and hyperspectral images. PAN and MS are widely studied by virtue of their properties, but high-quality remote sensing data are normally not accessible or affordable in practical scenarios. They are not easily obtained by drones for large-scale applications, such as mapping, and drone use is legally restricted; hence, image acquisition often requires human intervention. Due to these problems, the use of LR satellite images is common, and this requires the enhancement of image quality.

PAN $\in \mathbb{R}^{4H \times 4W \times 1}$ and MS $\in \mathbb{R}^{H \times W \times 3}$ have different wavebands of information over the same area. There are multiple bands of spectral information in MS, and only red, green, and blue bands are extracted to synthesize truecolor images. PAN is the image capture of the panchromatic waveband in ground object radiation, which has high spatial resolution. But it contains no color information because it is single-channel. PAN and MS are complementary in spatial and spectral resolutions, as shown in Fig. 1.

Many DL-based SR [21]–[24] and pan-sharpening algorithms, exploiting different network structures, have been proposed to acquire high-quality satellite images. The sensors used to acquire remote sensing images differ from those used for natural scenes, which results in different image textures. The SR methods for remote sensing images need precise details. Pan-sharpening is realized by the fusion of PAN and MS. This method considers fusing an LR MS and an HR PAN to obtain an HR MS. To obtain a pair of corresponding PAN and MS images is not always easy. Pan-sharpening models come into play when models are designed for different inputs.

B. Image Super-Resolution

In most cases, SR is regarded as a supervised learning task, whose key is to build proper LR–HR mapping relationships. The first DL-based SR method, SRCNN [8], overmatches most conventional methods and has performed impressively in terms of PSNR and SSIM. Different from the preupsampling SR, the upsampling layer is placed at the end of the network in post-upsampling SR. Chao *et al.* [9] designed a fast SRCNN, FSRCNN, using a small-size kernel convolutional block and replacing preupsampling at the beginning of the network with a deconvolutional block at the end. Profiting from

this post-upsampling architecture, intermediate feature maps remain the same size as the input image, which decreases computational costs. However, the limitations of parameters result in imperfect results.

Lai *et al.* [10] introduced LapSRN, a Laplacian pyramid framework in a deep neural network, to gradually generate HR images. Ying *et al.* [11] proposed DRRN, which combines several local residual blocks with a global residual block, using multichannel residual blocks to avoid vanishing or exploding gradients. Recursive learning improves performance, but the model's complexity will increase sharply with increasing recursion. With a dense and residual structure, RDN [12] obtained successful results, removing batch normal layers to decrease computational costs and removing pooling layers to retain pixel-level feature information. SRGAN [13] used a GAN to generate photo-realistic images. Adversarial loss encouraged a natural image through a discriminator network to differentiate between SR results and original realistic labels. SRGAN performed poorly on numerical indexes and detail textures due to its GAN-based model. Inspired by residual blocks and Inception models, MSRN [14] used an MRB to fuse multiscale feature information and original input information. Feature information was concatenated before the last reconstruction network. SRFBN [15] used a recurrent structure and parameter sharing to deepen the network. A dual regression scheme, DRN [16], used closed-loop mapping to enhance performance.

Scholars [25]–[27] have recently proposed well-designed models for remote sensing image SR.

C. Image Colorization

Image colorization is receiving increased attention. There are two main categories of colorization: automatic and user-guided.

User-guided image colorization uses humans to influence factors such as color strokes, reference images, color palettes, and language. This can make full use of human intervention, but it can be burdensome, and uncorrelated guidance can produce absurd results. It is not a small sum of its burden when given a variety of grayscale images through user-guided image colorization. It is not lighthearted to find a related reference image when an input contains multiple instances.

Automatic colorization methods based on DL models predict color values for every pixel from a grayscale image without human intervention. Iizuka *et al.* [17] combined global priors with local image features to colorize grayscale images, but their model was constrained by classification labels. Isola *et al.* [19] realized image-to-image translation with a conditional GAN. Unfortunately, the GAN-based methods perform poorly with respect to PSNR and SSIM. Su *et al.* [18] used semantic features of the object instance as prior knowledge to guide a deep neural network in colorization, which led to unnatural transitions between the instance and background. Once its objection detection algorithm fails, colorization results will be influenced. Yoo *et al.* [20] proposed a GAN-based MemoPainter to realize high-quality colorization

with limited data, but it could not pay attention to details when several subjects appeared in diagrams.

SR and colorization are commonly acknowledged to be ill-posed problems, and they are distinguishing in the direction of replenishing additional information. The purpose of SR is to expand the pixels surrounding an original pixel. This takes place independently on each channel. Colorization predicts two or three color channels on the basis of a luminance channel. We focus on a universal architecture for various visual missions in remote sensing images, which can learn mapping from LR to HR and finish the transformation from grayscale to color.

III. PROPOSED METHOD

We present the details of our approach and an architecture that can complete four tasks for remote sensing images

$$I_{HQ} = \mathcal{M}(I_{LQ}) \quad (1)$$

where I_{LQ} is the low-quality input (LR, grayscale, LR grayscale, and LR PAN), I_{HQ} is the high-quality input (HR, color, HR color, and HR MS), and \mathcal{M} denotes the proposed RSINet.

As shown in Fig. 2, our model can be explained in terms of two main elements:

- 1) feature extraction network-based MRBs
- 2) information recovery architecture (IRA)
 - a) involution-based downsample block (IDB)
 - b) intensive multiscale upsample block (IMUB)
 - c) dual-stream attention block (DAB).

A. Multiscale Residual Block

Inception [28]–[30] models have shown success at image classification tasks, with the extensive use of special Inception blocks. Inception blocks are multibranch blocks, where branches extract features by themselves, with the number of convolution blocks set differently to acquire diverse receptive fields. A larger number of convolution blocks and a larger convolution kernel generally led to border receptive fields. The locations and sizes of subjects vary by image, so the whole and parts of images must be considered comprehensively. In this process, feature maps are often one-off, and most significant tensors will be selected before being sent to the next block. Representative image features are passed on in this continuous process, and many original details are filtered.

Different from image classification, the image generation task aims to restore missing information while retaining input information. To adapt to image generation missions, an MRB for feature extraction based on Inception ResNet blocks is proposed.

As depicted in Fig. 3(d), an MRB differs from Inception blocks in the following aspects:

- 1) the global residual structure is removed and local residual structures are scattered over each branch;
- 2) the depth of a block is deepened to expand the receptive field.

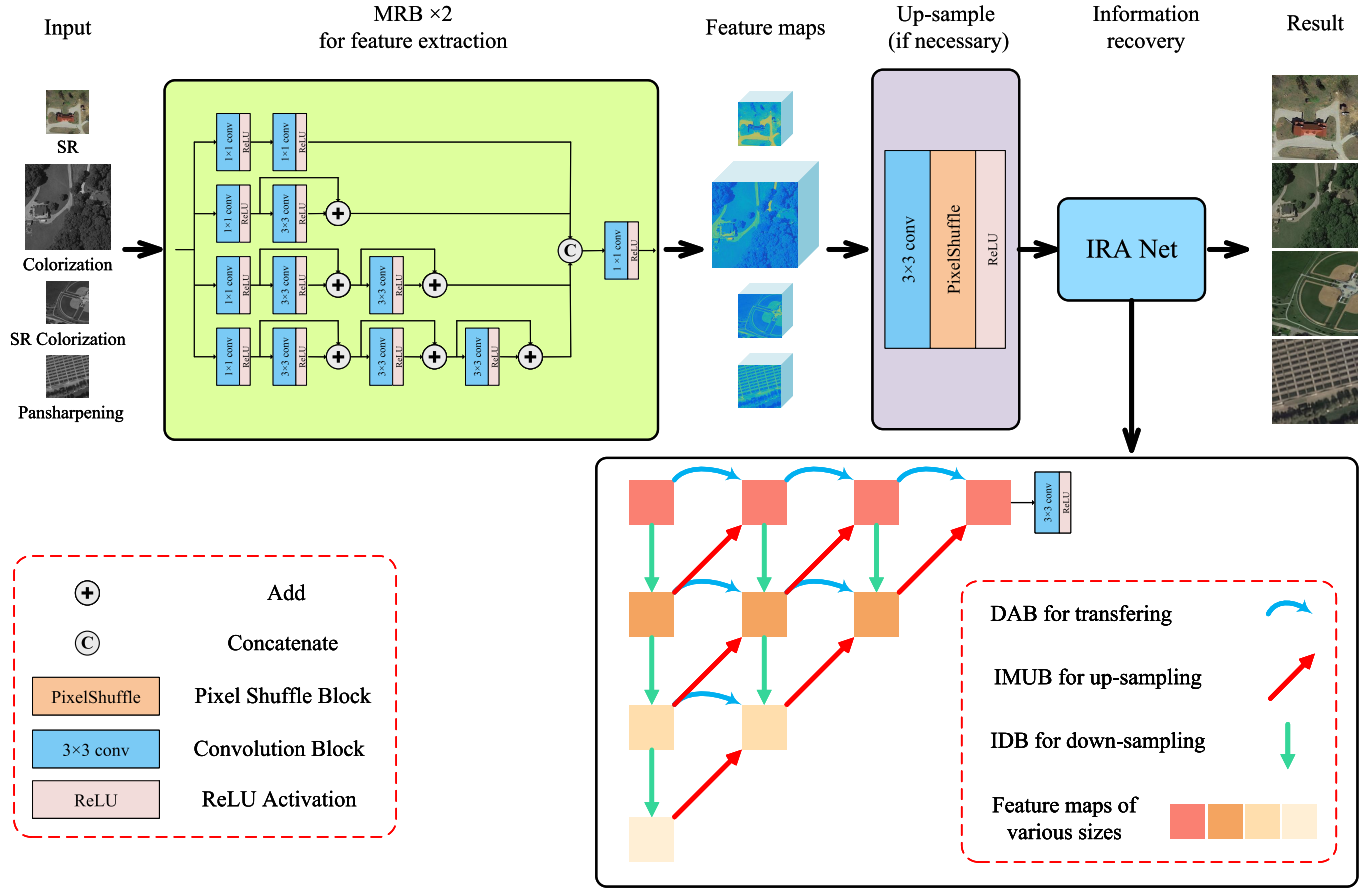


Fig. 2. Structure of the proposed algorithm.

MRB can be represented as

$$\text{Res}^k(\mathbf{X}) = C^k(\mathbf{X}) + \mathbf{X} \quad (2)$$

$$\text{feats} = \begin{cases} C^1(C^1(\mathbf{X})) \\ \text{Res}^3(C^1(\mathbf{X})) \\ \text{Res}^3(\text{Res}^3(C^1(\mathbf{X}))) \\ \text{Res}^3(\text{Res}^3(\text{Res}^3(C^1(\mathbf{X})))) \end{cases} \quad (3)$$

$$\text{MRB}(\mathbf{X}) = C^1(\text{concat}[\text{feats}]) \quad (4)$$

where \mathbf{X} is the input, C^k is a $k \times k$ convolution operation with an activation function, Res^k is the residual block, a group of feats is obtained by diverse residual blocks, and concat denotes concatenation. The most significant feature maps are acquired by a convolution operation after concatenation of feats in the channel dimension.

Multilocal residual structures help MRB convey the information of the image and avoid the vanishing gradient problem. MRB expands the multibranch Inception blocks and deepens the branch depth to obtain wider receptive fields. A remote sensing image contains many targets, such as roads, houses, grasslands, and soil, and the objects can be easier to distinguish in our improved remote sensing image, which may be beneficial to its applications.

While the image generation task must output a final complete image, feature maps cannot always be reduced like that

in image classification tasks, so pooling layers are removed in MRB. MRB is placed at the beginning of the model because:

- 1) MRB has a powerful feature extraction capability, which can directly obtain abstract features from the input;
- 2) We desire a low feature dimension of the input MRB, which will reduce computation.

B. Information Recovery Architecture

U-net [31], with its contracting path and symmetric expanding path, was first used in image segmentation and has subsequently been applied in all kinds of computer vision tasks. Zhou *et al.* [32] proposed U-net++, based on the original U-net, which redesigns skip pathways between the encoder and the decoder, and it generates full-resolution feature maps. U-net++ is applied in our IRA due to its strong capacity for information reconstruction. To finish four remote sensing image tasks, we improved U-net++ from the following aspects:

- 1) the dense block is removed;
- 2) the skip pathway consists of a dual attention block and a residual structure;
- 3) the root and end blocks are designed for different visual missions.

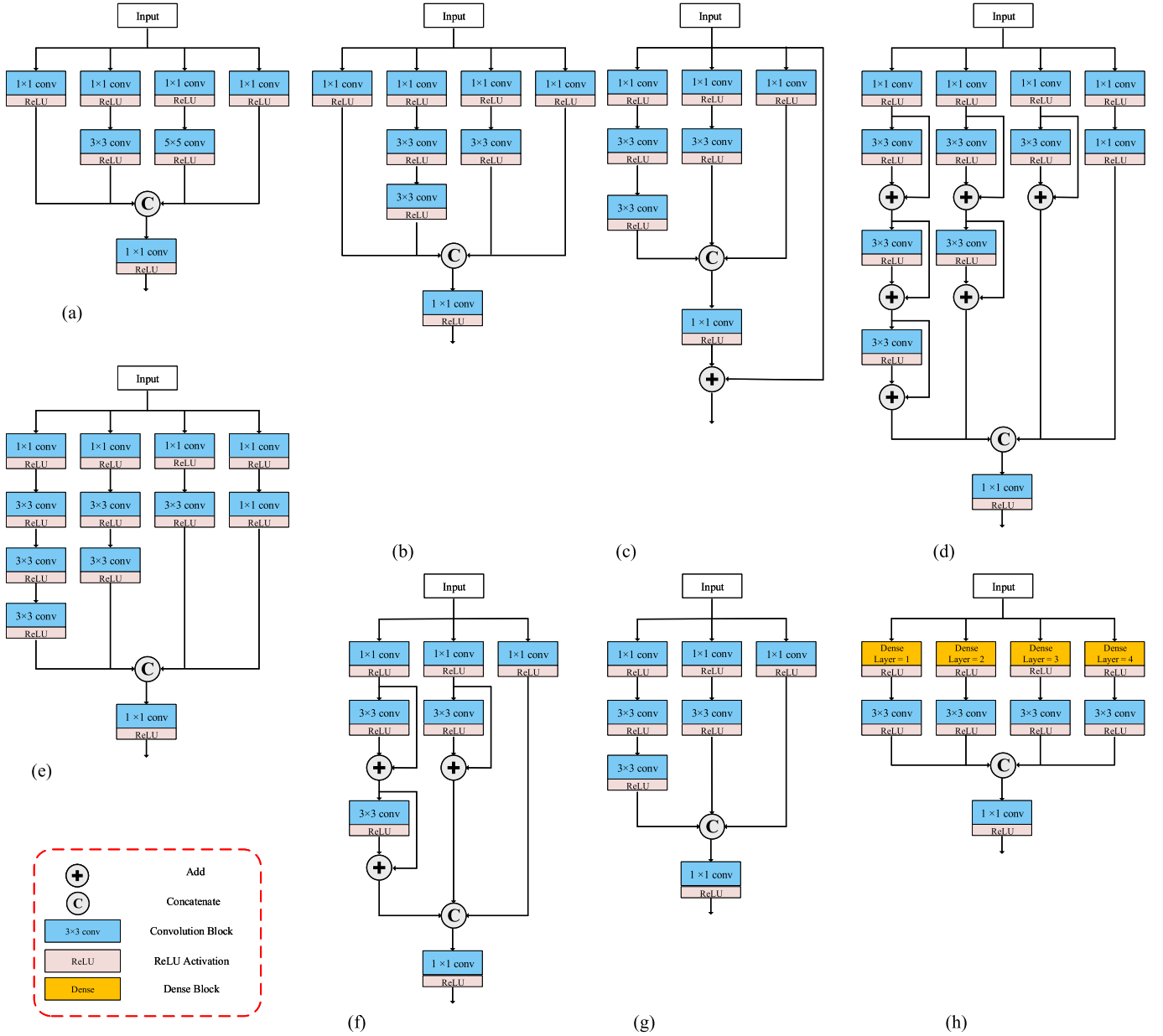


Fig. 3. Inception modules and the proposed MRB.

The stack of feature maps in IRA represented by $X_{i,j}$ is computed as

$$\mathbf{X}_{i,j} = \begin{cases} \text{fea}, & i = 0, j = 0 \\ \mathcal{D}(\mathbf{X}_{i-1,0}), & i > 1, j = 0 \\ \mathcal{A}(\mathbf{X}_{i,j-1}) \\ + \mathcal{U}(\mathbf{X}_{i-1,j-1}) + \mathbf{X}_{i-1,j-1}, & \text{otherwise} \end{cases} \quad (5)$$

where \mathcal{D} is IDB, \mathcal{A} is DAB, and \mathcal{U} is IMUB.

IRA can correct previous feature information and recover high-frequency or color information. In IRA, the first-layer network can stably deliver features, and the networks of other layers generate new information at different scales, which is continuously supplied to the first layer.

1) Involution-Based Downsample Block: Pooling and convolution with stride are commonly applied in downsampling in U-net. Through a downsample block, features are changed to half the size of the original, and the number of channels will increase occasionally. Features obtained in this way are still part of the previous features, and most U-net-based neural networks pay no attention to downsampling. We introduce an attention mechanism to this process, which can reconstruct more significant information.

Li *et al.* [33] inverted the design principles of convolution, and their so-called involution block could be powered by different visual networks. The involution block is used in conjunction with a convolution block, and it can aggregate contextual semantic information and be adaptively assigned to different positions in a model.

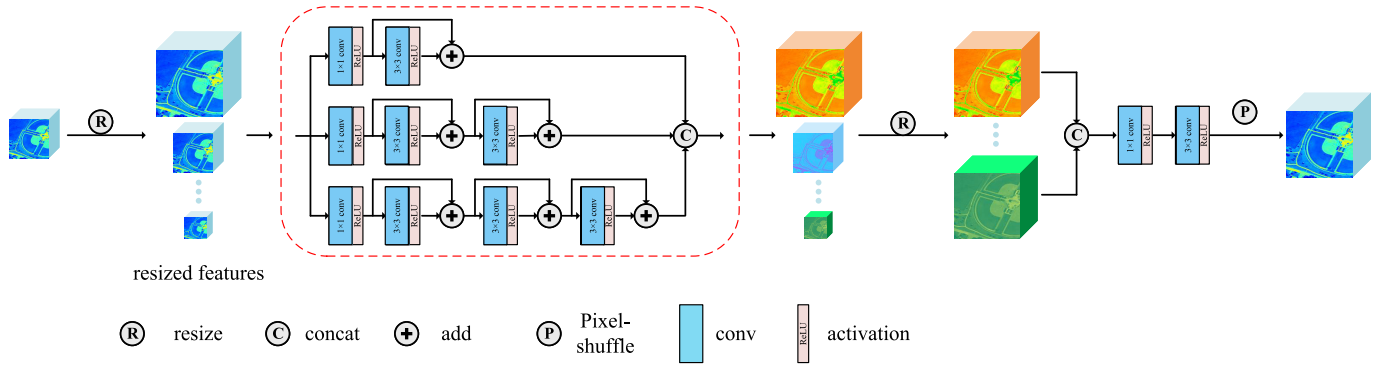


Fig. 4. Intensive multiscale residual upsample block (IMRUB).

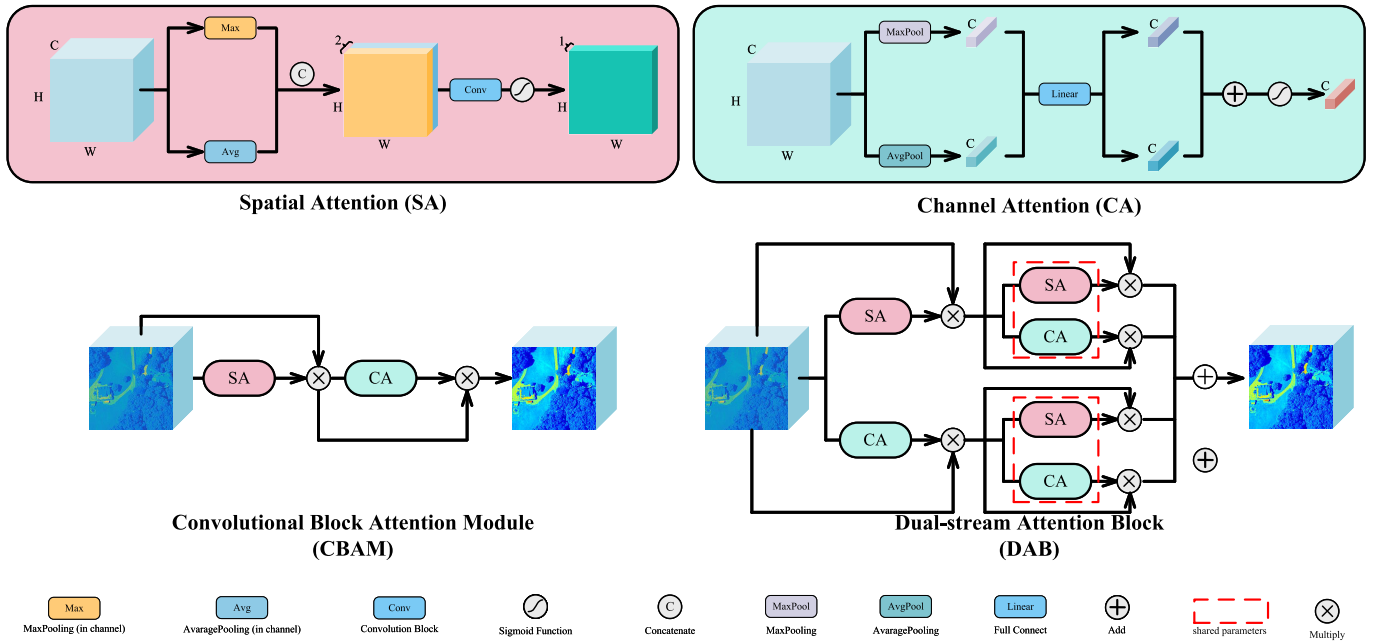


Fig. 5. CBAM and dual attention block.

The designed IDB includes an involution block and two convolution blocks. A 1×1 convolution block determines the output dimension, a 3×3 involution block (stride = 2) generates small-size features, and it ends with a 3×3 convolution block.

After passing a convolution block, a feature does not change much. After passing an involution block, the feature information can be reorganized to extract $1/4$ of the original feature, and another convolution block outputs results.

2) *Intensive Multiscale Residual Upsample Block*: On the basis of MRB, we develop an intensive multiscale residual block (IMRB) with a pyramid structure. Several convolution block groups are used to extract features in various receptive fields and features of varying sizes, rather than single feature maps. As shown in Fig. 4, a group of multiscale features are obtained by resizing input features with the idea of a Laplace pyramid network. They are sent into the same feature reconstruction structure and adjusted to the desired shape. The aggregation of reshaped features is not added directly, and they are concatenated before passing through a convolution block.

3) *Dual-Stream Attention Block*: Jie *et al.* [34] proposed squeeze-and-excitation (SE) Net, a network based on attention mechanisms that emphasizes informative features and pays less attention to other features. The SE block consists of global average pooling layers, fully connected layers, and a sigmoid function. CBAM [35] was proposed to strengthen the spatial and channel attention (CA) mechanisms. Given an intermediate feature map, a CBAM block gets spatial and CA maps and multiplies input feature maps by attention-based maps to optimize adaptive features.

There are two attention mechanisms in CBAM: CA and spatial attention (SA):

In CA, two 1-D vectors are obtained by compressing the feature map using average and max pooling in the spatial dimension. These are converted into a CA map by a weight-sharing network and an add operation, and the final feature is the product of the input feature and the attention map.

In SA, the channel is compressed to obtain average and max feature maps in the channel dimension, and the concatenated



Fig. 6. SR $\times 4$ results of different methods on the NWPU-45 datasets.

maps are sent to a convolution block to obtain an SA map. Since the obtained feature map contains insufficient information, it must be multiplied by the input. It is worth mentioning that CA maps are normalized by a sigmoid function.

CBAM generates a finer attention map sequentially than in parallel for image classification tasks. But these fine attention maps are insufficient for the task of restoring image information.

For this reason, we designed a DAB, as shown in Fig. 5, with three stages:

- 1) SA and CA are used to obtain two extracted feature maps;

- 2) a set of weight-sharing SA and CA is applied to extract deeper abstract features from previous features;
- 3) features obtained in the second stage are added up.

This is equivalent to using four kinds of attention mechanisms: reinforced CA, SA, and alternate combinations, which helps it strengthen important features and weaken irrelevant ones.

IV. EXPERIMENTS

A. Dataset and Evaluation Measures

Remote sensing information includes roads, bushes, houses, and other surface information, which is of great significance

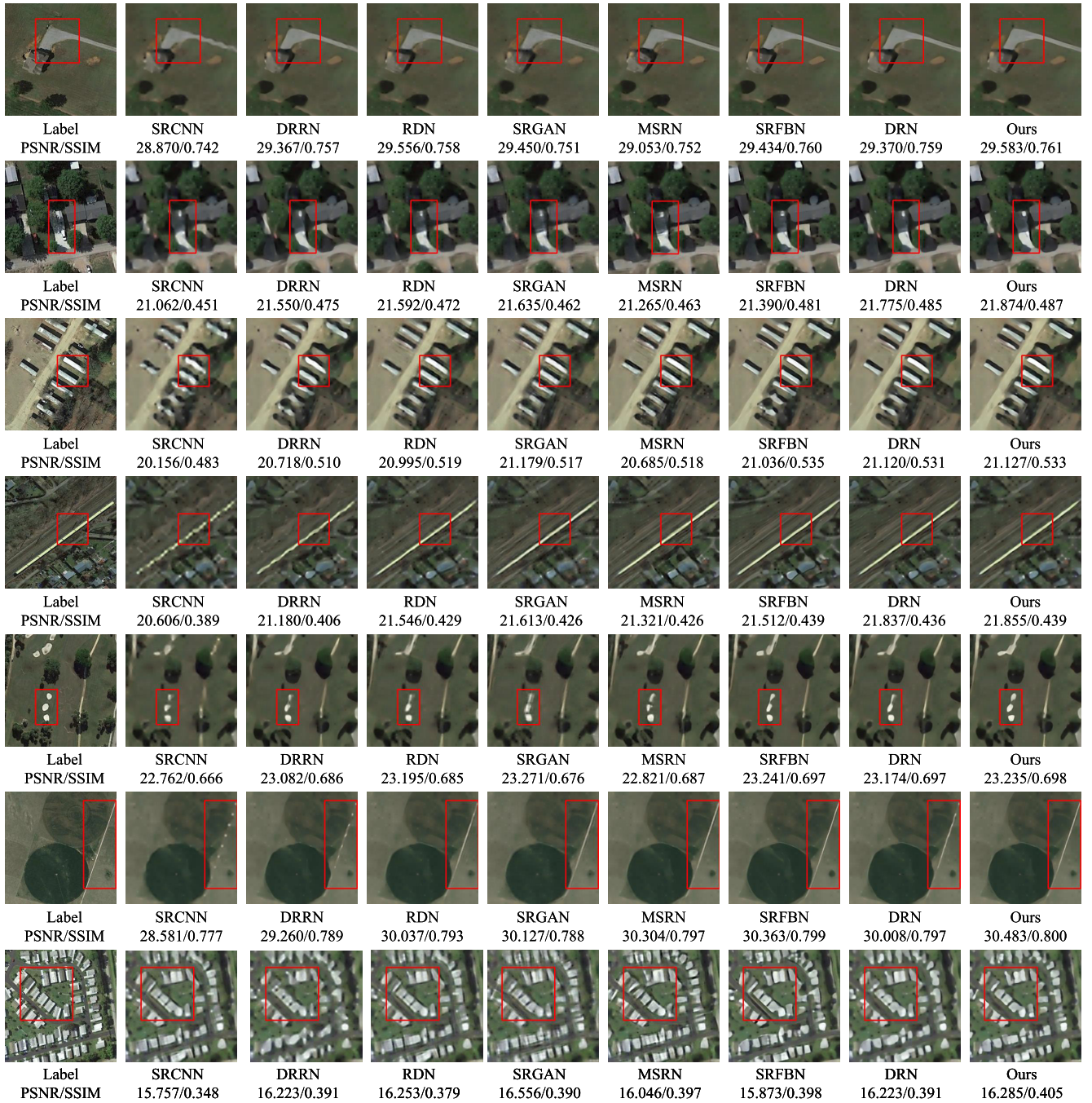


Fig. 7. SR $\times 8$ results of different methods on the NWPU-45 datasets.

in monitoring and military affairs. In SR, the accuracy of lines and patterns constitutes important attributes, and incorrect results will seriously affect the above applications. In colorization, since high-frequency detail information is already available, we only pay attention to whether the color is accurate.

Northwestern Polytechnical University reported on a large-scale dataset, NWPU-45 [36], which covered 45 types of scenes with 700 images, each of size 256×256 pixels. We randomly selected 9920 images in 16 classes of remote sensing scenes as training samples and 800 as test samples.

To verify the capacity of our model, we also tested some examples on the RSSCN7 [37] and aerial image dataset (AID) [38] datasets. Besides, the PANs that come from Gao Fen (GF)-2 satellite are used for pan-sharpening.

For image restoration, we used PSNR and SSIM as evaluation metrics. Given images I and \hat{I} , both with N pixels, they are defined as

$$\text{PSNR} = 10 \times \log_{10} \left[\frac{(L)^2}{\text{MSE}} \right], \quad \text{MSE} = \frac{1}{N} \|I - \hat{I}\|_F^2 \quad (6)$$

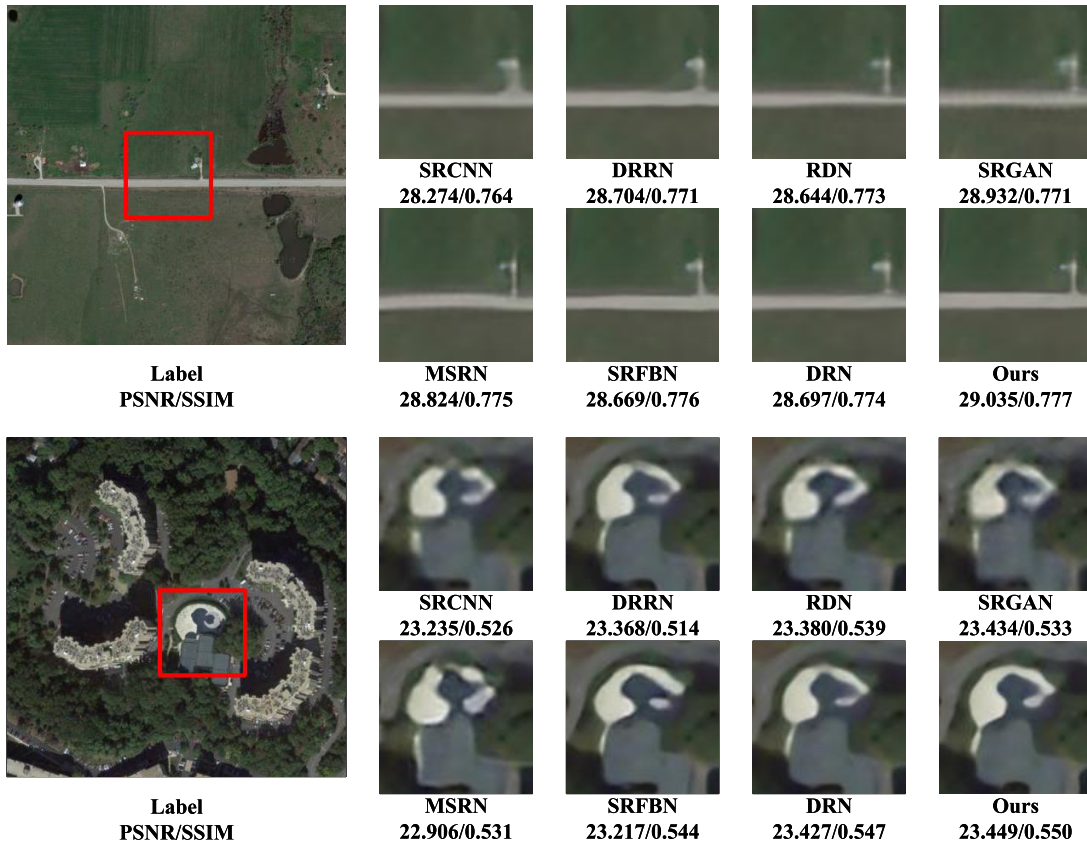


Fig. 8. SR $\times 8$ results of different methods on the RSCNN7 datasets.

where $\|\cdot\|_F^2$ is the Frobenius norm, and L is the dynamic range of pixel values, usually equal to 255, and

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad c_1 = (k_1L)^2, \quad c_2 = (k_2L)^2 \quad (7)$$

where μ_x is the mean of image x , μ_y is the mean of image y , σ_x^2 is the variance of image x , σ_y^2 is the variance of image y , σ_{xy} is the covariance of images x and y , and k_2 is a constant for stability, equal to 0.03.

B. Implementation

The proposed method can complete SR, colorization, SR colorization, and pan-sharpening tasks. Colorization will not change the shape of an image and does not require an upsampling layer. Two MRBs are joined sequentially for feature extraction. The extracted features are upsampled by a group of conv and pixel shuffle [39] operations. In IRA, the depth d is set to 3 or 4, and the growth rate g is 32. The last convolution block helps export the desired results. The Mish [40] activation function is applied instead of rectified linear unit (ReLU).

In SR, colorization, and SR colorization, we compute the \mathcal{L}_1 loss between the ground-truth HR image \hat{y}_i and the

reconstructed image y_i as

$$\mathcal{L}_1 = \frac{1}{N_B} \sum_{i=1}^{N_B} |\hat{y}_i - y_i| \quad (8)$$

where N_B is the batch size, which is equal to 4. This loss is optimized by Adam with an initial learning rate of 1×10^{-4} . Cosine annealing LR [41] is chosen to assist in optimizing the model parameters. Different from the usual SR training, we do not cut the label into small pieces for training. We use 256×256 images as labels and get corresponding inputs by bicubic interpolation. For image colorization, we use the OpenCV algorithm to transfer red, green, blue (RGB) labels to gray inputs.

Training and testing are carried out in PyTorch, which is boosted using a 24G Nvidia RTX 3090. The proposed method devotes more time to model training than part of other methods due to its complex structure, while also performing better.

C. Ablation Studies

To investigate the impact of the proposed modules, we performed experiments on MRB, IMRUB, and DAB for remote sensing image SR $\times 2$. All evaluations used the same configurations. PSNR (6) and SSIM (7) were the judging criteria. Higher PSNR and SSIM means that generated results are of better quality.

Table I gives the PSNR and SSIM values of the images of several Inception blocks and the proposed MRB, from

Fig. 9. SR $\times 8$ results of different methods on the AID datasets.

TABLE I

ABLATION STUDY ON SR $\times 2$. BEST RESULTS ARE IN RED; SECOND BEST ARE IN BOLD

location	Structure	PSNR	SSIM
root	Inception v1 (Fig.3(a))	32.3482	0.9185
	Inception v2 (Fig.3(b))	32.3504	0.9184
	Inception resnet (Fig.3(c))	32.3304	0.9187
	MRB-1 (Fig.3(e))	32.3320	0.9178
	MRB-2 (Fig.3(f))	32.3613	0.9186
	MRB-3 (Fig.3(g))	32.3296	0.9183
	MRB-4 (Fig.3(h))	32.3462	0.9186
	MRB (Fig.3(d)) (ours)	32.3629	0.9189
up-sample	Conv + Pixel-Shuffle	32.2800	0.9170
	Deconv	32.2990	0.9173
	IMUB(ours)	32.3629	0.9189
attention block	Shortcut	32.3450	0.9179
	CBAM	14.5500	0.6131
	DAB(ours)	32.3629	0.9189

which it is seen that the performance of MRB is enhanced by local residual structure and greater depth. The results of other altered MRB structure validated our idea.

Table I also shows the results of various attention mechanisms. Because of the particularity of the image generation task, each type of attention block is combined with the residual structure so as to convey the original characteristics. CBAM obviously has a counter effect in the generation process, and a single-stream attention block is improper for generating

images. This result confirms the distinction between image generation and classification tasks. The combination of a dual-stream attention structure and residual can help recover high-quality images.

It is seen from Table I that a well-designed IMUB surpasses the other two upsample blocks. IMUB uses pyramid and multibranches to construct a convolution group which is similar to MRB to accumulate various features before a pixel shuffle block, which is used to rearrange the feature

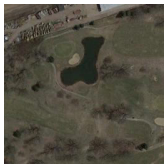


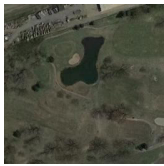


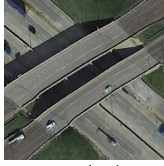
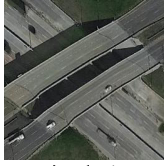
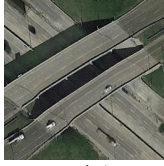
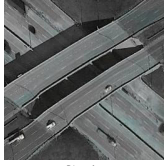


























					
Label PSNR/SSIM	Iizuka's 26.0879	Isola's 35.528/0.989	Su's 32.995/0.988	Yoo's 38.635/0.989	Ours 40.782/0.992
					
Label PSNR/SSIM	Iizuka's 29.086/0.985	Isola's 35.985/0.985	Su's 33.075/0.987	Yoo's 31.699/0.985	Ours 37.853/0.993
					
Label PSNR/SSIM	Iizuka's 30.075/0.976	Isola's 35.649/0.987	Su's 34.501/0.988	Yoo's 32.808/0.983	Ours 36.480/0.990
					
Label PSNR/SSIM	Iizuka's 26.293/0.969	Isola's 34.048/0.988	Su's 30.823/0.990	Yoo's 29.666/0.979	Ours 37.118/0.993
					
Label PSNR/SSIM	Iizuka's 28.607/0.980	Isola's 31.570/0.983	Su's 28.022/0.980	Yoo's 30.508/0.983	Ours 34.221/0.991
					
Label PSNR/SSIM	Iizuka's 27.400/0.983	Isola's 33.777/0.988	Su's 34.350/0.991	Yoo's 34.144/0.987	Ours 36.475/0.993

Fig. 10. Colorization results of methods on the NWPU-45 datasets.

maps, regardless of whether the values are useful. IMUB accumulates feature maps that can offer more valued pixels. The ablation experiments demonstrate the usefulness of the designed modules.

D. Comparison With State-of-the-Art Models

1) *Super-Resolution*: For image SR, the proposed method is compared with bicubic interpolation, which is used as a baseline, as well as with the state-of-the-art methods SRCNN [8], FSRCNN [9], LapSRN [10], DRRN [11], RDN [12], SRGAN [13], MSRN [14], SRFBN [15], DRN [16], mixed high-order attention network (MHAN) [25], dense

deepback-projection network (D-DBPN) [26], and hybrid-scale self-similarity exploitation network (HSENet) [27]. We evaluate all the methods on the $2\times$, $4\times$, and $8\times$ SR task based on PSNR and SSIM.

The comparison results are shown in Table II, which shows that our proposed method outperforms the baselines on the NWPU-45 datasets. For image $2\times$ SR, our method (depth = 3) performs better than our method (depth = 4) due to a deeper network that is unsuitable for relatively simple tasks. The outputs of $4\times$ and $8\times$ enlargement are visualized in Figs. 6 and 7, from which we can find that PSNR and SSIM are higher for our method. However, SRFBN gets the

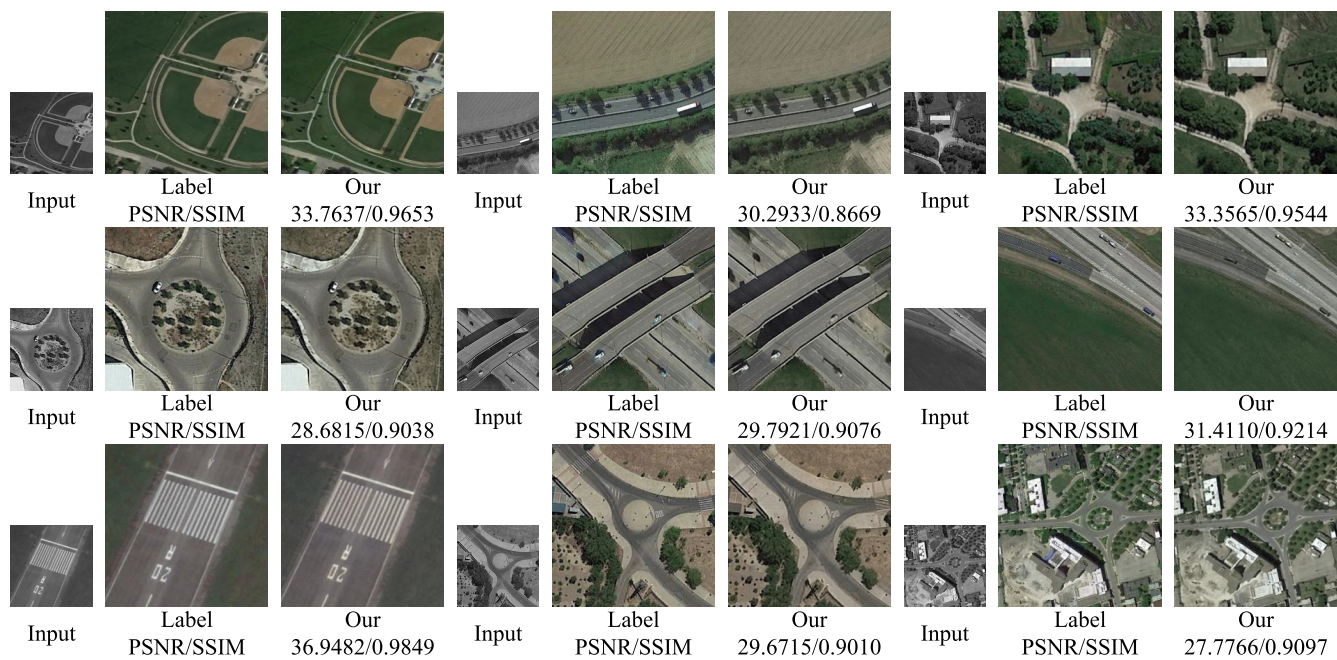
Fig. 11. SR $\times 2$ and colorization results on the NWPU-45 datasets.Fig. 12. SR $\times 2$ and colorization results on PANs.

TABLE II
COMPARISON FOR IMAGE SR. BEST RESULTS ARE IN RED; SECOND BEST ARE IN **BOLD**

Method	$\times 2$		$\times 4$		$\times 8$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BCUBIC	30.4602	0.8887	24.5316	0.6571	21.5072	0.4638
SRCNN [8]	31.4082	0.9038	25.7599	0.6975	22.6340	0.5014
FSRCNN [9]	29.3601	0.8882	24.3885	0.6073	20.2738	0.4755
LapSRN [10]	31.5990	0.9065	26.0077	0.7069	-	-
DRRN [11]	32.0968	0.9139	26.4221	0.7259	22.9645	0.5244
RDN [12]	32.0938	0.9138	26.5325	0.7301	23.2093	0.5372
SRGAN [13]	31.8913	0.9104	26.3774	0.7184	23.2024	0.5270
MSRN [14]	32.3103	0.9185	26.4886	0.7313	23.0960	0.5401
SRFBN [15]	32.3155	0.9173	26.1116	0.7115	23.2704	0.5525
DRN [16]	32.1850	0.9154	26.5217	0.7299	23.2583	0.5431
MHAN [25]	32.0519	0.9134	26.3893	0.7241	23.1429	0.5332
D-DBPN [26]	32.1470	0.9147	26.5752	0.7338	23.1236	0.5468
HSENet [27]	31.9700	0.9124	26.2178	0.7177	22.9353	0.5195
Ours (depth=3)	32.3791	0.9189	26.5761	0.7331	23.1970	0.5419
Ours (depth=4)	32.3629	0.9189	26.5964	0.7335	23.2832	0.5469

TABLE III
COMPARISONS FOR IMAGE COLORIZATION. BEST RESULTS ARE IN RED; SECOND BEST ARE IN **BOLD**

Method	PSNR	SSIM
Iizukas [17]	28.547	0.978
Isolas [19]	33.366	0.984
Sus [18]	32.972	0.988
Yoos [20]	31.068	0.979
Ours (depth=4)	34.597	0.989

highest SSIM in SR SR $8\times$. Input is sparse in SR SR $8\times$, and the loop structure of SRFBN help learn a few times more. The experiments shown that the image of our method is better than that of SRFBN, visually. Besides, our method can be used for four tasks, simultaneously.

For a precise and fair comparison, all the methods were trained on the NWPU-45 datasets and some examples of the AID and RSCNN7 datasets. The $8\times$ visual results are shown in Figs. 8 and 9. Our model can adapt to remote sensing images of different resolutions. Although it was not trained on the RSCNN7 and AID datasets, it could still achieve good results. As shown in Figs. 8 and 9, our method has the highest PSNR and SSIM and also has accurate texture information, making it closest to the original image.

2) *Colorization*: The image colorization of the proposed method was compared with the state-of-the-art methods, including those of Iizukas [17], Isolas [19], Sus [18], and Yoos [20]. As shown in Table III, our method can obtain higher PSNRs than the others. Our results have visual advantages, as shown in Fig. 10. Iizukas [17] cannot accurately finish colorization, as seen by its lower PSNR and SSIM. Isolas [19] and Yoos [20] have color bleeding problems, such as a road turning green. Sus [18] sometimes produces grayscale images because image segmentation algorithms are leveraged, and terrible colorization occurs once segmentation algorithms fail.

E. Combination of SR and Colorization

After verifying that the proposed method can independently finish SR and colorization, we combined these tasks, which

proved feasible. The results are shown in Fig. 11, which indicates that the proposed model can extract sufficient features from limited input and produce color information while reconstructing detailed information. Given an LR grayscale MS, our model can produce reasonable details and color information.

F. Pan-Sharpening

There are three reasons to use a trained model in SR colorization to finish pan-sharpening:

- 1) pan-sharpening can be regarded as the union of SR and colorization, which are consistent in the transformation of the image shape;
- 2) the information contained in PAN and MS covers the same area; hence, their texture information is close;
- 3) the spatial resolution of training datasets is close to that of PAN;
- 4) there are no realistic color labels for PAN images, which results in an inability to provide training datasets.

From Fig. 12, a pretrained model can give reasonable color information while improving the spatial resolution. The colors of roads and grass are natural, and detailed information has been restored.

V. CONCLUSION

We proposed a neural network to enhance the spatial and spectral resolutions of remote sensing images. Based on the Inception modules, we proposed an MRB and an IRA consisting of IDB, IMUB, and DAB. MRB can help convey the detailed information of an image and expand multibranches of Inception blocks. IMUB extracts features in various receptive fields for upsampling. DAB strengthens important features and weakens irrelevant features.

Comparative experiments on the NWPU-45 dataset demonstrated that our method not only obtains the best SR results on different resolution images but also outperforms other colorization methods. We combined SR and colorization into one mission to finish remote image SR colorization. Finally, we used a model trained in SR colorization to finish pan-sharpening without MS.

As a result, our method will be beneficial in recovering high-quality remote sensing images, and the results can be applied in object identification, disaster monitoring, military reconnaissance, and other fields.

REFERENCES

- [1] Y. Yang, W. Tu, S. Huang, H. Lu, W. Wan, and L. Gan, "Dual-stream convolutional neural network with residual information enhancement for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [2] H. Yin, "PSCSC-Net: A deep coupled convolutional sparse coding network for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [3] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-based super-resolution for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
- [4] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.
- [5] D.-L. Chen, L. Zhang, and H. Huang, "Robust extraction and super-resolution of low-resolution flying airplane from satellite video," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.
- [6] S. Wan, Y. Xia, L. Qi, Y.-H. Yang, and M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1756–1768, Jul. 2020.
- [7] B. Li, Y.-K. Lai, M. John, and P. L. Rosin, "Automatic example-based image colorization using location-aware cross-scale matching," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4606–4619, Sep. 2019.
- [8] 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. (ECCV)*, 2014, pp. 184–199.
- [9] D. Chao, C. L. Chen, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 391–407.
- [10] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5835–5843.
- [11] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3147–3155.
- [12] 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.
- [13] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
- [14] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution: 15th European conference Munich, Germany, September 8–14, 2018, proceedings, Part VIII," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 527–542.
- [15] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3867–3876.
- [16] Y. Guo *et al.*, "Closed-loop matters: Dual regression networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5407–5416.
- [17] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 110.1–110.11, 2016.
- [18] J.-W. Su, H.-K. Chu, and J.-B. Huang, "Instance-aware image colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7968–7977.
- [19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
- [20] S. Yoo, H. Bahng, S. Chung, J. Lee, J. Chang, and J. Choo, "Coloring with limited data: Few-shot colorization via memory augmented networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11283–11292.
- [21] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
- [22] S. Wang, T. Zhou, Y. Lu, and H. Di, "Contextual transformation network for lightweight remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, early access, Dec. 1, 2021, doi: 10.1109/TGRS.2021.3132093.
- [23] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
- [24] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-based super-resolution for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
- [25] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5183–5196, Jun. 2021.
- [26] M. B. Pereira and J. A. D. Santos, "An end-to-end framework for low-resolution remote sensing semantic segmentation," in *Proc. IEEE Latin Amer. GRSS ISPRS Remote Sens. Conf. (LAGIRS)*, Mar. 2020, pp. 6–11.
- [27] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2022.
- [28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1–9.
- [29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
- [30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*. Palo Alto, CA, USA: AAAI Press, 2017, pp. 4278–4284.
- [31] O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Cham, Switzerland: Springer, 2015.
- [32] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. 4th Deep Learn. Med. Image Anal. (DLMIA) Workshop*, 2018, pp. 11045:3–11045:11.
- [33] D. Li *et al.*, "Involution: Inverting the inference of convolution for visual recognition," 2021, *arXiv:2103.06255*.
- [34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020, doi: 10.1109/TPAMI.2019.2913372.
- [35] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham, Switzerland: Springer, 2018.
- [36] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [37] Z. Qin, L. Ni, Z. Tong, and W. Qian, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 1–5, Nov. 2015.
- [38] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Apr. 2017.
- [39] W. Shi *et al.*, "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.
- [40] D. Misra, "Mish: A self regularized non-monotonic neural activation function," 2019, *arXiv:1908.08681*.
- [41] L. Ilya and H. Frank, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–16.



Jianan Feng received the B.E. degree from the Information School, Yunnan University of Finance and Economics, Kunming, China, in 2019. He is pursuing the master's degree with the School of Software, Yunnan University, Kunming. His research interests include neural networks, image super-resolution, image colorization, and image classification.



Qian Jiang received the B.S. degree in thermal energy and power engineering and the M.S. degree in power engineering and engineering thermo-physics from Central South University (CSU), Changsha, China, in 2012 and 2015, respectively, and the Ph.D. degree in communication and information systems from Yunnan University, Kunming, China, in 2019.

She is an Associate Professor with the School of Software, Yunnan University. Her research interests include deep neural networks, fuzzy set theory, bio-informatics, image processing, and information fusion.



Ling Liu received the B.E. degree in software engineering from Yunnan University, Kunming, China, in 2019, where she is pursuing the master's degree with the School of Software.

Her research interests include deep neural networks, image colorization, and image super-resolution.



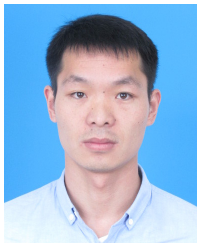
Ching-Hsun Tseng received the master's degree from the Institute of Management of Technology, NCTU, Hsinchu, Taiwan, in 2019. He is pursuing the Ph.D. degree in computer science with The University of Manchester, Manchester, U.K.

Before that, he worked as a Data Engineer with SparkAmplify, Taipei, Taiwan, a U.S. Pattern Recognition (PR) Artificial Intelligence (AI) company. He focuses on fields of computer vision and natural language processing (NLP). He has published a series of international journals and conference papers. His master dissertation based on GAN was selected as a fine work in Fubon Life Management Thesis Award.



Wei Zhou (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Science, Beijing, China, in 2008.

He is a Full Professor with the School of Software, Yunnan University, Kunming, China. His research interests include distributed data intensive computing and bio-informatics.



Xin Jin (Member, IEEE) received the B.S. degree in electronics and information engineering from Henan Normal University, Xinxiang, China, in 2013, and the Ph.D. degree in communication and information systems from Yunnan University, Kunming, China, in 2018.

He is an Associate Professor with the School of Software, Yunnan University. His research interests include pulse coupled neural networks and its applications, image processing, information fusion, optimization algorithm, and fuzzy set theory.



Shaowen Yao (Member, IEEE) received the B.S. and M.S. degrees in telecommunication engineering from Yunnan University, Kunming, China, in 1988 and 1991, respectively, and the Ph.D. degree in computer application technology from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2002.

He is a Professor with the School of Software, Yunnan University. His research interests include neural networks, fuzzy set theory, image processing, bio-informatics, and data mining.