Multi-Scale Fast Fourier Transform Based Attention Network for Remote-Sensing Image Super-Resolution

Zheng Wang, Yanwei Zhao, and Jiacheng Chen

***Abstract*—Recently, with the rise and progress of convolutional neural networks (CNNs), CNN-based remote-sensing image super-resolution (RSSR) methods have gained considerable advancement and showed great power for image reconstruction tasks. However, most of these methods cannot handle well the enormous number of objects with different scales contained in remote-sensing images and thus limits super-resolution performance. To address these issues, we propose a multiscale fast Fourier transform (FFT) based attention network (MSFFTAN), which employs a multiinput U-shape structure as backbone for accurate RSSR. Specifically, we carefully design an FFT-based residual block consisting of an image domain branch and a Fourier domain branch to extract local details and global structures simultaneously. In addition, a local–global channel attention block is developed to further enhance the reconstruction ability of small targets. Finally, we present a branch gated selective block to adaptively explore and aggregate features from multiple scales and depths. Extensive experiments on two public datasets have demonstrated the superiority of MSFFTAN over the state-of-the-art (SOAT) approaches in aspects of both quantitative metrics and visual quality. The peak signal-to-noise ratio of our network is 1.5 dB higher than the SOAT method on the UCMerced LandUse with downscaling factor 2.**

***Index Terms*—Attention mechanism, fast Fourier transform (FFT), multiinput mechanism, remote-sensing image, super-resolution.**

I. Introduction

SINGLE-image super-resolution (SISR) technique aims to generate natural and realistic textures in a high-resolution (HR) image by only utilizing its deteriorated low-resolution (LR) counterpart. SISR has been a hotspot for study in academics and industries thanks to its many applications, including remote-sensing imaging [1], [2], [3], [4], medical imaging [5], [6], [7], and face recognition [8]. SISR is a classic ill-posed problem since numerous distinct HR images can be mapped to the same LR image, which poses a significant challenge to restoration task. In recent years, with the rapid development and popularization of aerospace technology, remote sensing vision has attracted an increasing number of researchers' attention. In the field of remote sensing, the long distance of the imaging device from target objects leads to a small resolution of target objects, which affects performance of subsequent high-level tasks (object detection [9], [10], classification [1], [11], and change detection [12], [13]). The most straightforward solution to this problem is to upgrade the physical equipment to get a HR and clearer image, but this is often unrealistic and requires a significant price. Therefore, the utilization of hardware-agnostic image super-resolution techniques (SISR) for enhancing the resolution of remote sensing images has become the current preferred approach.

To improve the resolution of image, researchers have proposed a variety of approaches, ranging from interpolation-based methods, reconstruction-based methods to example-based methods. Interpolation-based method uses a pixel around an unknown pixel to predict the unknown pixel, which is prone to produce blurred images with artifacts. To solve these problems, reconstruction-based methods often introduce various prior knowledge (sparse prior [14], low-rank prior [15], nonlocal prior [16], and edge prior [12]) to constrain the solution space in pursuit of a better reconstruction. Nevertheless, once the introduced prior knowledge conflicts with the fact, reconstruction performance drops dramatically. In addition, reconstruction-based methods often require long optimization times. Exampled-based methods establish a direct mapping from LR to HR using hand-designed features, but the poor generalization performance of hand-designed features limits its practical application.

Recently, SISR methods based on convolutional neural networks (CNNs) have substantially outperformed traditional method due to the powerful feature extraction capability of deep neural networks. Dong et al. [17] pioneered the introduction of a CNN into an SISR task with unprecedented success. Since then, various kinds of super-resolution networks based on CNN have emerged. Kim et al. [18] constructed a deep network with 20 layers by introducing residual learning. Deeper and larger networks are becoming increasingly frequent in the search of better reconstruction results. Lim et al. [19] constructed a deep network

Manuscript received 31 October 2022; revised 31 January 2023; accepted 13 February 2023. Date of publication 20 February 2023; date of current version 22 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 51875524, in part by the Open Project Program of the State Key Lab of CAD&CG under Grant A2210, and in part by the Zhejiang University, and the Key Research and Development Program of Zhejiang Province under Grant 2023C01168. *(Corresponding author: Jiacheng Chen.)*

Zheng Wang is with the School of Computer Computational Science, Hangzhou City University, Hangzhou 310028, China (e-mail: wangz@zucc.edu.cn).

Yanwei Zhao is with the School of Engineering, Hangzhou City University, Hangzhou 310028, China (e-mail: zhaoyanwei@zucc.edu.cn).

Jiacheng Chen is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: 15968812143@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3246564

with 50 convolution layers by discarding batch normalization [20] and won the NTIRE 2017 challenge. Thanks to the booming development of natural image super-resolution, deep learning-based algorithms for remote-sensing image super-resolution (RSSR) have made great progress. Despite the impressive results obtained by these approaches, the majority of them recover characteristics at a single scale, making it difficult for networks to efficiently extract multiscale information. Therefore, it is important to investigate multiscale feature extraction.

Some recent work has initiated efforts in this direction. Residual aggregation and split attention fusion network [2] uses a UNet-based encoder and decoder structure to extract both shallow semantic information and high-level features. Although this approach is capable of extracting multiscale features, it leads to irreversible information loss through frequent up and down sampling, which will eventually affect the reconstruction results. To minimize this information loss, a dense feature fusion approach is introduced. Specifically, not only the output of the current layer's encoder is taken into account, but also the output of the previous and next layer's encoders. In addition, it is not enough to extract multiscale features only at feature level. We introduce an auxiliary branch to extract features at different scales directly on the picture domain. In this way, we are able to exploit multiscale features in both the image and feature domains. For a super-resolution task, both low-frequency and high-frequency information are critical. Since the normal residual block [21] lacks the ability to integrate high-frequency features, a fast Fourier transform (FFT) is applied on the top of residual block. It is worth noting that each feature value in the frequency domain represents an abstraction of all the values in the original image features, allowing us to easily obtain global dependencies. Therefore, an FFT-based residual block (FFT-RB) can utilize both global and local information.To further strengthen the discriminative power of the network, a novel attention mechanism is introduced called local–global channel attention.

The main contributions of this article can be summarized as follows:

1) For the accurate remote-sensing image super-resolution (RSSR) task, we propose a novel SR approach named multi scale FFT-based attention network (MSFFTAN). MSFFTAN incorporates a multiinput encoder–decoder structure that can capture objects at different sizes in remote-sensing images.
2) To enable efficient extraction of high-frequency features, the FFT is incorporated into the ResBlock. In this way, high and low frequency can aggregate in a comprehensive manner. This operation ensures that our model can obtain rich features to recover texture and edge information efficiently.
3) An effective local–global channel attention block (LGCAB) is elaborately developed in MSFFTAN to enable the network focus on more useful information consistent with a global branch and a local branch which is beneficial to feature learning and model training.

The rest of this article is organized as follows. Section II discusses relevant RSSR research. The MSFFTAN network design is described in full in Section III. In Section IV, the network design and experiment results, including ablation analysis, are presented. Finally, Section V concludes this article.

II. Related Works

In this section, we go through some of the most important approaches for our method, which include CNN-based nature image super-resolution and RSSR. Since CNN-based approaches have shown outstanding performance in recent years, we mainly introduce CNN-based methods.

A. CNN-based Nature Image Super-Resolution

CNN-based techniques have dominated SISR in recent years, thanks to the fast growth of deep convolutional neural networks. Dong et al. [17] introduced the first SISR approach based on CNN (SRCNN). Despite the fact that SRCNN only has three convolutional layers, it outperformed earlier conventional approaches. He et al. [21] used residual connection to build a deep model VDSR [18] with 20 convolutional layers that outperforms the SRCNN significantly. This meant that the higher the network's depth, the greater the performance. To get higher performance, researchers seek to create deeper, wider, and more complicated networks from then on. Following that, EDSR [19] built a network of around 50 layers by eliminating the unnecessary components. Nevertheless, this method treats LR features similarly and overlooks their long-range associations, resulting in inefficient detail retrieval. Thus, several techniques have recently been developed that include an attention mechanism into a CNN-based Super-Resolution (SR) model to rebalance the relevance of various elements. Zhang et al. [22] used residual in a residual structure to build a network with over 400 layers in terms of improving reconstruction performance. The context reasoning attention network was developed by Zhang et al. [23] to adjust the convolution kernel according to the global context adaptively. Mei et al. [16] combined nonlocal operation and sparse representation into an SISR task and proposed a nonlocal sparse attention to alleviate the large computational resources required for nonlocal operation. In addition, using a coarse-to-fine approach, a two-stage attentive network [24] is presented for accurate SISR.

B. Remote-Sensing Image Super-Resolution

Remote-sensing picture SR has recently gained significant attention. Deep learning-based algorithms [25] have recently exceeded these early classical methods considerably. LGCNet [26] is the first CNN-based model for super-resolution in remote-sensing images, using both local and global representations to learn the reconstructed SR image. Dong et al. proposed SMSR [27], which aggregates diverse multiscale characteristics utilizing first-order and second-order learning mechanisms. Meanwhile, during the last few decades, the attention mechanism has made significant advances in a variety of computer vision tasks, such as image classification [11] and object detection [9]. Thus, attention mechanism was introduced to the field of remote-sensing image SR. HSENet [28] exploits the single-scale and

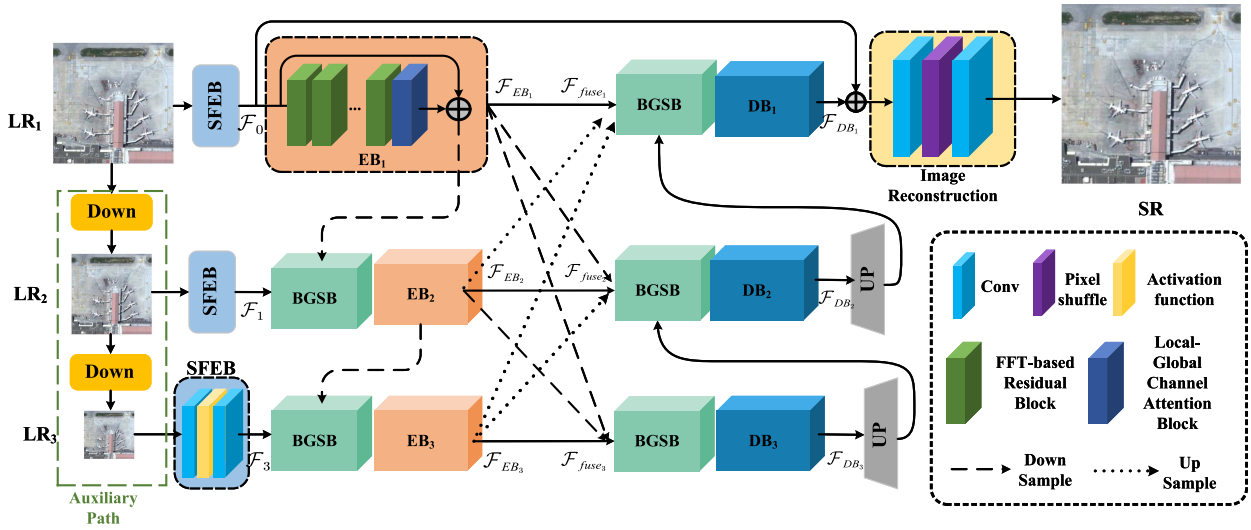


Fig. 1. Architecture of the proposed MSFFTAN network.

cross-scale self-similarity information using multiscale nonlocal attention. A split attention fusion block was established by Chen et al. [2] allowing the method to adapt to varied multiscale land surface reconstructions. Rather than exploring first-order attention (channel or spatial statistics), Zhang et al. [29] advocated a high-order attention block to restore the missing details. Salvetti et al. [25] proposed the residual attention multiimage superresolution network, which leverages feature extraction from multiple LR images of the same scene, resulting in reconstructed images with fine texture details. Hu et al. [30] proposed a network that utilizes a HR, spatially lossless multispectral image to guide the super-resolution reconstruction of a LR hyperspectral image. The experimental results demonstrate that this strategy can effectively preserve spatial detail information in the recovered image. Xu et al. [31] utilized an iterative regularization technique based on tensor subspace representation to amalgamate paired multispectral and hyperspectral images, thereby reconstructing HR hyperspectral images with distinct texture and sharp edges. Hong et al. [32] proposed a decoupled and coupled high-spectral-resolution image super-resolution algorithm that progressively aggregates high-spectral and multispectral information. Through experimentation, it was demonstrated that this fusion method can enhance the quality of reconstruction. In addition, The CUCaNet [33] proposed a cross attention module that is also proposed to efficiently explore the spatial-spectral information. Furthermore, many researchers have introduced generative adversarial networks [34] (GAN) into remote-sensing SR tasks for generating perceptually pleasing remote-sensing images. Pan et al. [35] introduced the concept of back-projection into a generator to further enhance the visual quality. In addition, an attention-based GAN (SRAGAN) was proposed by Li et al. [36], which combined both local and global attention mechanisms to distinguish features at various sizes on different objects. Lei et al. [37] used a transformer to fuse high- and low-frequency information to reconstruct detail-rich pictures, building on the success of transformer in the fields of natural language processing and computer vision.

III. Proposed Methods

In this section, we introduce the MSFFTAN for remote-sensing super-resolution. First, the overall framework of MSFFTNet is presented in Section III-A. Then, branch gated selective block (BGSB), FFT-RB, and LGCAB are described in the following three subsection.

A. Network Architecture

As shown in Fig. 1, MSFFTAN mainly consist of the following four parts:

1) auxiliary path (AP)
2) shallow feature extraction block (SFEB)
3) multiscale deep feature extraction module;
4) reconstruction block.

We present the MSFFTAN, a multiscale feature extraction approach that fully leverages multiscale features retrieved from an input image. The architecture of MSTFFAN is based on a three stage U-shape structure [38] with significant development for efficient multiscale feature representation. Specifically, an MSFFTAN is composed of three encoder blocks (EBs) and decoder blocks (DBs). Each EB or DB is composed of multiple cascaded FFT-RBs. We define $\boldsymbol{I}_{\mathrm{LR}} \in \mathbb{R}^{H\times W\times 3}$, $\boldsymbol{I}_{\mathrm{SR}} \in \mathbb{R}^{sH\times sW\times 3}$, and $\boldsymbol{I}_{\mathrm{HR}} \in \mathbb{R}^{sH\times sW\times 3}$ as the input LR image, the reconstructed SR image, and the corresponding HR image, respectively. In addition, $\boldsymbol{I}_{\mathrm{LR}_2} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$ and $\boldsymbol{I}_{\mathrm{LR}_3} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 3}$ represent the downsampled input image. H and W denote the height and width of the image, respectively, with s representing the upsampling factor.

For $\boldsymbol{I}_{\mathrm{LR}_1}$, the SFEB is used to transform the original LR image to feature domain

$$\boldsymbol{\mathcal{F}}_0 = \mathcal{H}_{\mathrm{SFEB}}(\boldsymbol{I}_{\mathrm{LR}_1}) = Conv_{3\times3}(\delta(Conv_{3\times3}(\boldsymbol{I}_{\mathrm{LR}_1}))) \quad (1)$$

where $\boldsymbol{\mathcal{F}}_0$ represents the shallow feature extracted from the LR image. $\mathcal{H}_{\mathrm{SFEB}}(\cdot)$ denotes the shallow feature extraction block. In detail, SFEB consists of two 3×3 Conv layers with an activation unit. $Conv_{3\times3}(\cdot)$ and $\delta(\cdot)$ denote 3×3 Conv layer and linear

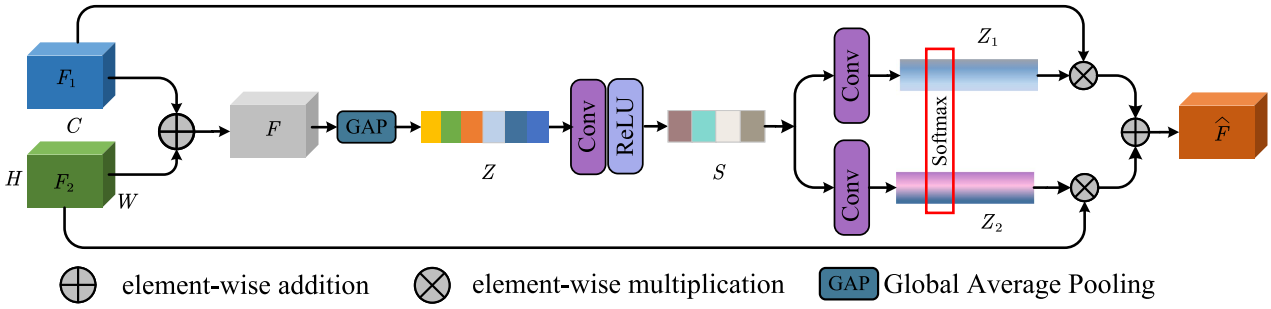


Fig. 2. Structure of the BGSB.

rectification function (ReLU) activation function, respectively. Then, the extracted shallow feature $\boldsymbol{\mathcal{F}}_0$ is fed to the next EBs

$$\boldsymbol{\mathcal{F}}_{\mathrm{EB}_i} = \begin{cases} \mathcal{H}_{\mathrm{EB}_1}\left(\boldsymbol{\mathcal{F}}_0\right), & i = 1 \\ \mathcal{H}_{\mathrm{EB}_i}\left(\mathcal{H}_{\mathrm{BGSB}}\left(\boldsymbol{\mathcal{F}}_{\mathrm{EB}_{i-1}}, \mathcal{H}_{\mathrm{SFEB}}\left(\boldsymbol{LR}_i\right)\right)\right), & i \neq 1 \end{cases} \tag{2}$$

where $\mathcal{H}_{\mathrm{EB}_i}(\cdot)$ stands for the ith EB which consists of multiple FFT-RBs, $\boldsymbol{\mathcal{F}}_{\mathrm{EB}_i}$ represents the deep encoder feature extracted by the ith EB. H_{SFEB} and H_{BGSB} denote SFEB and BGSB, respectively. In addition to the first EB, not only the downsampled features of previous encoder are received, but also the information of corresponding downsampled image. In this way, our EB is anticipated to successfully handle multiscale features by utilizing the complimentary information from the downsampled feature space and the feature available from the image domain. To alleviate the inconsistency in the image and feature domains, we use BGSB for feature selection and feature fusion. In this work, we use a total of three encoder layers. Then, the DB can be described as follows:

$$\begin{aligned} &\mathcal{F}_{\mathrm{fuse}_i} \\ &= \begin{cases} \mathcal{H}_{\mathrm{BGSB}}\left(\mathcal{F}_{\mathrm{EB}_1}, UP_2\left(\mathcal{F}_{\mathrm{EB}_2}\right), UP_4\left(\mathcal{F}_{\mathrm{EB}_3}\right)\right), i=1 \\ \mathcal{H}_{\mathrm{BGSB}}\left(Down_2\left(\mathcal{F}_{\mathrm{EB}_1}\right), \mathcal{F}_{\mathrm{EB}_2}, UP_2\left(\mathcal{F}_{\mathrm{EB}_3}\right)\right), i=2 \\ \mathcal{H}_{\mathrm{BGSB}}\left(Down_4\left(\mathcal{F}_{\mathrm{EB}_1}\right), Down_2\left(\mathcal{F}_{\mathrm{EB}_2}\right), \mathcal{F}_{\mathrm{EB}_3}\right), i=3 \end{cases} \end{aligned} \tag{3}$$

$$\boldsymbol{\mathcal{F}}_{\mathrm{DB}_i} = \begin{cases} \mathcal{H}_{\mathrm{DB}_1}\left(\boldsymbol{\mathcal{F}}_{\mathrm{fuse}_i}\right) & i = 3 \\ \mathcal{H}_{\mathrm{DB}_i}\left(\left[\boldsymbol{\mathcal{F}}_{\mathrm{fuse}_i}, UP_2\left(\boldsymbol{\mathcal{F}}_{\mathrm{DB}_{i-1}}\right)\right]\right) & i \neq 3 \end{cases} \tag{4}$$

where H_{DB_i} represents the ith DB which consists of multiple FFT-RBs, $\boldsymbol{\mathcal{F}}_{\mathrm{DB}_i}$ stands for the deep decoder feature extracted by the ith decoder layer. Upsampling operation or downsampling operation is indicated by $UP_{\mathrm{factor}}(\cdot)$ or $Down_{\mathrm{factor}}(\cdot)$ and factor represents the magnification factor. Notably, to further enhance the network's ability to extract multiscale features, we aggregate features of different sizes and dimensions using BGSB module denoted as $\boldsymbol{F}_{\mathrm{fuse\_}i}$. Thus, $\boldsymbol{F}_{\mathrm{fuse\_}i}$ contains rich structural information. Finally, deep decoder feature is fed into the reconstruction block which is consistent with a Conv layer, a subpixel layer, and a Conv layer as

$$\boldsymbol{I}_{\mathrm{SR}} = Conv\_2\left(\mathcal{H}_{\uparrow}\left(Conv\_1\left(\boldsymbol{DB}_1 + \boldsymbol{\mathcal{F}}_0\right)\right)\right) \tag{5}$$

where $\mathcal{H}_{\uparrow}(\cdot)$ is the function of upscale operation and SR represents the recovered HR image. We further adopt residual connection between a shallow feature and a deep feature to alleviate the training difficulty. In this way, we are able to force the network to focus on the lost high-frequency information, thus accelerating the convergence of the network.

B. Branch Gated Selective Block

Simple concatenation or summation are the most frequent strategies for feature aggregation. However, these choices hinder the representation capability of the network. Based on the fact that visual cortical neurons can adaptively change their receptive fields depending on the intensity of the stimulus [39], we propose a novel multiscale multiresolution feature fusion block named BGSB (see Fig. 2), which is composed of branch aggregation (BA) and gate selective fusion (GSF).

1) Branch Aggregation: The BA generates global feature descriptors by combining the information from multiresolution branches. Specifically, the downsized feature $\boldsymbol{\mathcal{F}}_1 \in \mathbb{R}^{H\times W\times C}$ and the feature obtained from the downsampled image $\boldsymbol{\mathcal{F}}_2 \in \mathbb{R}^{H\times W\times C}$ are summed as the input $\boldsymbol{\mathcal{F}}$, and the global average pooling (GAP) is utilized to squeeze the global spatial information into a channel descriptor $\boldsymbol{\mathcal{Z}}$, which can be expressed as

$$\boldsymbol{\mathcal{F}} = \boldsymbol{\mathcal{F}_1} + \boldsymbol{\mathcal{F}_2} \tag{6}$$

$$\boldsymbol{\mathcal{Z}_c} = \mathcal{H}_{\mathrm{GAP}}\left(\boldsymbol{\mathcal{F}_c}\right) = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} \boldsymbol{\mathcal{F}}_c(i,j) \tag{7}$$

where $\mathcal{H}_{\mathrm{GAP}}(\cdot)$ denotes the global average pooling operation. $\boldsymbol{\mathcal{F}}_c$ and $\boldsymbol{\mathcal{Z}}_c$ denote cth channel input feature and output feature of $\mathcal{H}_{\mathrm{GAP}}$, respectively. $\boldsymbol{\mathcal{F}}_c(i,j)$ is the value at the position (i,j) of cth channel of input feature $\boldsymbol{\mathcal{F}}$.

2) Gate Selective Fusion: The channel statistic $\boldsymbol{\mathcal{Z}}$ may be thought of as a grouping of local descriptors whose statistics can be utilized to represent the entire image. To make full use of the multiresolution feature interdependences, we employ a gating mechanism by the simple softmax function

$$\boldsymbol{\mathcal{S}} = \delta\left(Conv_{1\times1}(\boldsymbol{\mathcal{Z}})\right) \tag{8}$$

$$\boldsymbol{\mathcal{S}}_1 = Conv_{1\times1}(\boldsymbol{\mathcal{S}}), \boldsymbol{\mathcal{S}}_2 = Conv_{1\times1}(\boldsymbol{\mathcal{S}}) \tag{9}$$

where $\delta(\cdot)$ denotes activation function and $Conv_{1\times1}(\cdot)$ denotes 1×1 convolution. Then, we use softmax function to obtain

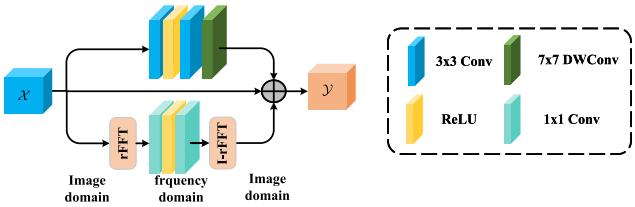


Fig. 3. Architecture of the FFT-RB.

attention weights belonging to each branch

$$\mathcal{Z}_1 = \frac{e^{\mathcal{S}_1}}{e^{\mathcal{S}_1} + e^{\mathcal{S}_2}}, \quad \mathcal{Z}_2 = \frac{e^{\mathcal{S}_2}}{e^{\mathcal{S}_1} + e^{\mathcal{S}_2}} \tag{10}$$

where $\mathcal{Z}_1$ and $\mathcal{Z}_2$ represent attention weight of different resolution branch. These descriptors are used by the GSF operator to recalibrate the feature map after aggregation. In this way, it is possible to adaptively aggregate different resolutions branches that carry information at different scales.

C. FFT-Based Residual Block

Image recovery task requires both low-frequency and high-frequency information, however, the standard ResBlock lacks the capacity to integrate high-frequency characteristics. Inspired by Mao et al.[40], we propose an FFT-RB as shown in Fig. 3 which consists of a conventional spatial domain Conv branch and a frequency domain branch. Specifically, to convert information to frequency domain space and extract complementary features for the space domain, we employ the discrete Fourier transform. Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{H\times W\times C}$ be the input volume, where H, W, and C indicate the height, width, and channel of the feature, respectively. The bottom branch is processed as follows:

$$\widehat{\boldsymbol{\mathcal{X}}} = \mathcal{H}_{rFFT2D}(\boldsymbol{\mathcal{X}}) \tag{11}$$

where $\mathcal{H}_r FFT2D(\cdot)$ represents 2-D discrete FFT and $\widehat{\boldsymbol{\mathcal{X}}}$ represents the result of 2-D real FFT. Then, the real part and imaginary part are concatenated along the channel dimension

$$\boldsymbol{\mathcal{X}}_{\text{imag}+\text{real}} = \left[\mathcal{I}\left(\widehat{\boldsymbol{\mathcal{X}}}\right), \mathcal{R}\left(\widehat{\boldsymbol{\mathcal{X}}}\right)\right] \tag{12}$$

where $\mathcal{I}(\cdot)$ and $\mathcal{R}(\cdot)$ get real and imaginary parts, respectively. $[\cdot]$ denotes the concatenate operation. We use two 1×1 Conv to extract high-frequency features

$$\boldsymbol{\mathcal{X}}_{\text{high}} = Conv_{1\times 1}\left(\delta\left(Conv_{1\times 1}\left(\boldsymbol{\mathcal{X}}_{\text{image}+\text{real}}\right)\right)\right). \tag{13}$$

Here, $Conv_{1\times 1}(\cdot)$ and $\delta(\cdot)$ denote the 1×1 Conv and ReLU activation function, respectively. Finally, inverse 2-D real FFT operations are used to transform frequency features back to spatial domain. It is worth noting that due to the intrinsic characteristics of the Fourier transform, FFT can easily obtain the global field of perception without adding any additional burden. Influenced by ConvNest [41], we added a large convolution kernel to expand the perceptual field in the spatial branch

$$\boldsymbol{\mathcal{X}}_{\text{space}} = Conv_{7\times 7DW}\left(Conv_{3\times 3}\left(\delta\left(Conv_{3\times 3}(\boldsymbol{\mathcal{X}})\right)\right)\right) \tag{14}$$

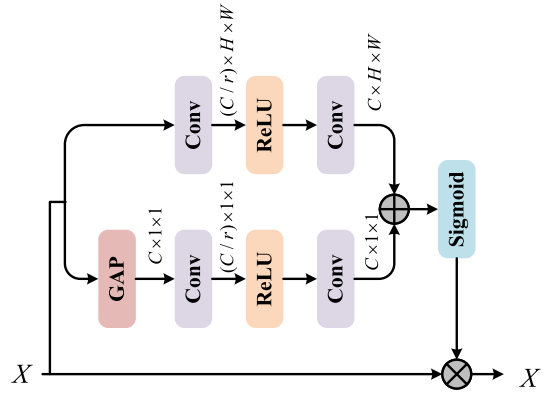


Fig. 4. Structure of the LGCAB.

where $Conv_{3\times 3}(\cdot)$ and $\delta(\cdot)$ denote the 3×3 Conv and ReLU activation function, respectively. $Conv_{7\times 7DW}(\cdot)$ denotes depth-separable convolution [42] with kernel size 7. Then, the final output $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} + \boldsymbol{\mathcal{X}}_{\text{space}} + \boldsymbol{\mathcal{X}}_{\text{high}}$ of FFT-RB is calculated through LGCAB to further refine features.

D. Local–Global Channel Attention Block

Existing channel attention mechanisms [43] typically build channel descriptors via a global average pooling operation, which overlooks many beneficial little objects that play a vital role in RSSR. Hence, to be capable of assessing both informative large and tiny target objects, an LGCAB is proposed, as shown in Fig. 4. It allows the network to concentrate on significant features while still paying attention to minor target details. Consider an input feature $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{H\times W\times C}$, where C, W, and H indicate channel number, width, and height, respectively. The top branch of the LGCAB is in charge of characterizing little items, whereas the bottom branch is responsible for detecting global essential foundational features. The top branch can be expressed as

$$\boldsymbol{A}_{\text{local}} = W_U(\delta\left(W_D(\boldsymbol{\mathcal{X}})\right)) \tag{15}$$

where $\boldsymbol{A}_{\text{local}}$ denotes the local channel attention map. $\delta(\cdot)$ denotes activation function.$W_U(\cdot)$ and $W_D(\cdot)$ denote the weights of two 1×1 Conv layers to increase and decrease the number of channels by reducing factor r, respectively. This branch does not use global average pooling, preserving the original resolution of the features and enabling the capture of fine-grained information. In this way, it is possible to concentrate on the attributes of the whole features. The GAP operation can be expressed as

$$\mathcal{Z} = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{F}_c(i,j) \tag{16}$$

where $\mathcal{F}_C(i,j)$ is the value at the position (i,j) of cth channel of input feature $\boldsymbol{\mathcal{X}}$. Then, the bottom branch can be expressed as

$$\boldsymbol{A}_{\text{global}} = W_U\left(\delta\left(W_D(\mathcal{Z})\right)\right) \tag{17}$$

where $\boldsymbol{A}_{\text{global}}$ denotes the global channel attention map. $\delta(\cdot)$ denotes activation function. $W_U(\cdot)$ and $W_D(\cdot)$ denote the weights of two 1×1 Conv layers to increase and decrease the number

Fig. 5. Examples of the different categories of different scenes in the UCMerced LandUse and AID datasets.

of channels by reducing factor r, respectively. As this branch uses global average pooling, it allows the network to focus on large objects that occupy a significant portion of the image. Next, global and local attention maps are be used to rescale the input feature $\mathcal{X}$

$$\tilde{\boldsymbol{\mathcal{X}}} = \sigma\left(\boldsymbol{A}_{\text{local}} + \boldsymbol{A}_{\text{global}}\right) \otimes \boldsymbol{\mathcal{X}} \tag{18}$$

where $\tilde{\boldsymbol{\mathcal{X}}}$ indicates the refined output features. $\sigma(\cdot)$ and $\otimes(\cdot)$ represent a sigmoid function and element-wise multiplication between feature maps, respectively. By using the above steps, we enable to emphasize important information and suppress irrelevant features using a global and local manner, thus enhancing the discriminative capacity of the network.

E. Loss Function

To optimize the RSSR network, various loss functions have been investigated, such as $L1$ loss [19], $L2$ loss [44], perceptual loss [45], and adversarial loss [34]. As stated by Lim et al. [19], $L2$ loss can maximize peak signal-to-noise (PSNR) metrics, but it is prone to produce blurry images. Therefore, $L1$ loss is chosen as our optimization function for training MSFFTAN. Then, MSFFTAN is optimized by minimizing the pixel-wise dissimilarity between estimated super-resolved image SR and corresponding ground truth HR. The optimization function L_{L1} is formulated as

$$\boldsymbol{\mathcal{L}}_{L1}(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{H}_{\mathrm{MSFFTAN}}\left(\boldsymbol{I}^{i}_{\mathrm{LR}};\Theta\right) - \boldsymbol{I}^{i}_{\mathrm{HR}}\right\|_1 \tag{19}$$

where Θ denotes trainable parameters of MSFFTAN network, and deep MSFFTAN is trained by using a training set $\{\boldsymbol{I}^{i}_{LR}, \boldsymbol{I}^{i}_{HR}\}_{i=1}^{N}$, which contains N LR images patches and their HR counterparts. Auxiliary loss terms, in addition to the $L1$ loss, has been suggested in recent research for performance enhancement. Auxiliary loss terms that reduce the distance between the input and output in the feature space have been frequently employed in image restoration tasks and have shown promising results. Since the primary objective of super-resolution is to recover the lost high-frequency characteristic, it is critical to minimize the frequency space comparison. To this end, we introduce an FFT-based frequency reconstruction loss L_{FFT} function. The L_{FFT} loss measures the Euclidean distance between HR images and SR images in the Fourier entity as follows:

$$\boldsymbol{\mathcal{L}}_{\mathrm{FFT}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{F}\left(H_{\mathrm{MSFFTAN}}\left(\boldsymbol{I}^{i}_{\mathrm{LR}}\right) - \mathcal{F}\left(\boldsymbol{I}^{i}_{\mathrm{HR}}\right)\right)\right\| \tag{20}$$

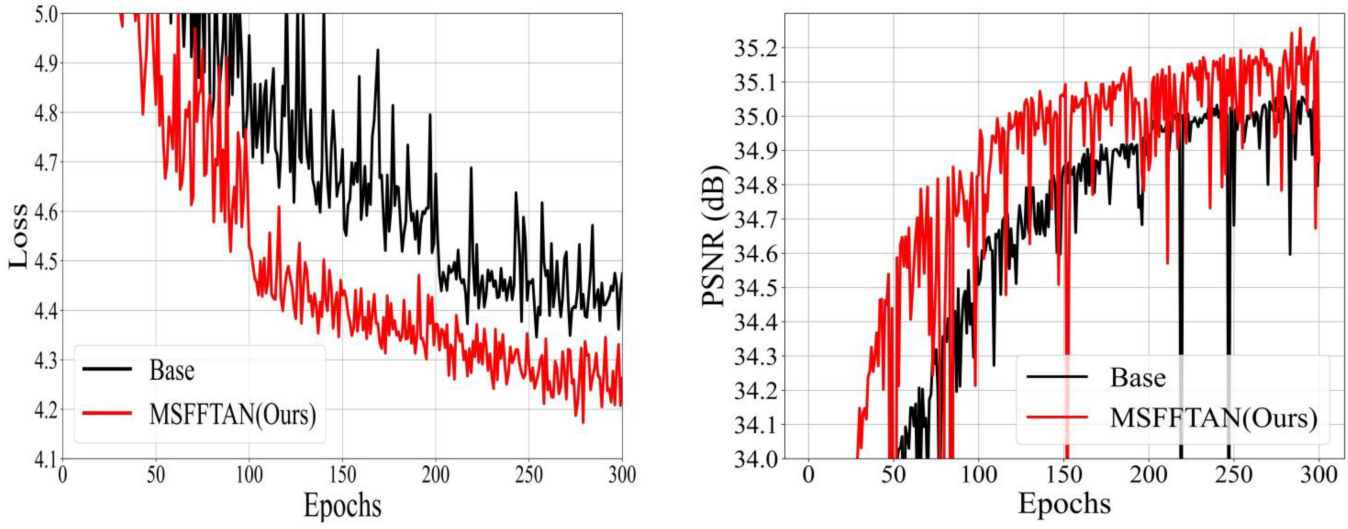
where $\mathcal{F}$ denotes the FFT that transfers image domain to the frequency domain. The following is the final loss function for training our network:

$$\boldsymbol{\mathcal{L}}_{\text{total}} = \boldsymbol{\mathcal{L}}_{L1} + \tau\boldsymbol{\mathcal{L}}_{\mathrm{FFT}} \tag{21}$$

where we experimentally set $\tau = 0.01$.

IV. Experiment

In this section, experiments are conducted to evaluate our proposed model. The datasets and metrics we employed in our experiments are described in Section IV-A. Then, the implementation details are presented in Section IV-B. Section IV-C compares our model to state-of-the-art (SOAT) methods on several datasets to show that our proposed approach is superior. Finally, an ablation study is performed in Section IV-D to analyze the contribution of each component to our MSFFTAN network.



(a) (b)

Fig. 6. Convergence analysis. (a) Loss analysis on UCMerced LandUse with upscale factor 2. (b) PSNR analysis on UCMerced LandUse with upscale factor 2.

A. Dataset and Implementation Details

To test the efficiency of the proposed approach, we utilize the following two publicly available (some examples of these two datasets are shown in Fig. 5): 1) UCMerced LandUse [46]; and 2) AID [47] datasets. These datasets have seen a lot of application in the field of remote-sensing super-resolution [28], [36], [37]. The HR images were downsampled with a scale factor using a bicubic interpolation operation in the MATLAB setting to produce LR images.

UCMerced LandUse dataset: This collection includes agricultural, runway, sparseresidential, storagetanks, and other remote-sensing types. Each class has 100 pictures, each of which is 256 $\times$ 256 pixels in size and has a spatial resolution of 0.3 m/pixel. This dataset was divided into two halves: Train and test, with 20% of the training set used as validation.

AID dataset: This dataset contains 10 000 photos from 30 different types of remote-sensing scenarios, such as airports, bareland, churches, dense-residential areas, and so on. All of the photos are 600×600 pixels, with a spatial resolution of up to 0.5 m/pixel. According to TransNet [37], 80% of the whole dataset is randomly selected to be the training set, and the remaining images are used as the test set in the AID dataset. Moreover, we randomly select five images per class in a total of 150 images to construct the corresponding validation.

Metrics: Peak signal to noise ratio (PSNR) and structural similarity (SSIM) [48] are chosen as the common image super-resolution evaluation metrics, and all super-resolution results are evaluated on the RGB space. Besides, we further introduce the learned perceptual image patch similarity (LPIPS) [49] to evaluate the reconstruction quality of the competing methods. A lower LPIPS value indicates a higher perceptual quality. We also analyze the floating point operations (FLOPS) and runtime of the models. Note that the FLOPs is calculated corresponding to a 48×48 image.

B. Implementation Details

To obtain better generalization performance, we use data augmentation, which includes random rotation by $90°$, random horizontal flipping and vertical flipping. We use Pytorch framework to implement and train the proposed MSFFTAN, and the model is trained using one NVIDIA GeForce GTX 3090. We train different models to super-resolve the remote-sensing images for scale factors 2, 3, and 4 with random initialization. The ADAM [50] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used. The learning rate is initialized as 2×10^{-4} and halved every 400 epochs. For training, we randomly crop 16 48 $\times$ 48 LR patches as a training batch while HR image size corresponding to the scaling factor. In our MSFFTAN, all convolution layers contain 64 filters except 1×1 convolution layers. Specifically, the number of FFT-RB included in our backbone of different depths is 3, 2, and 1.

C. Comparisons With the SOAT Methods

1) Quantitative Results: To demonstrate the superiority of MSFFTAN, eight SOAT super-resolution methods, including Bicubic, SRCNN [17], FSRCNN [51], VDSR [18], LGCNet [26], DCM [52], HSENet [28], and CTN [53], are compared in terms of quantitative and visual quality on the UCMerced LandUse dataset. Among them, SRCANN, FSRCNN, and VDSR are the approaches proposed for nature image SR task, while LGCNet, DCM, HSENet, and CTN are currently leading SR methods exclusively developed for remote-sensing images. It should be noted that, we analyze various comparison methods using the open-source code, and all of these methods are trained and evaluated in the same environment. Specifically, quantitative evaluations are made in two datasets for three scale ($\times$2, $\times$3,and$\times$4). Table I displays the average results of different approaches on the UCMerced Landuse test dataset which clearly

TABLE I
PSNR/SSIM RESULTS ON UCMERCED LANDUSE DATASET OF SCALE X2, X3, AND X4

Scale	Bicubic PSNR/SSIM	SRCNN PSNR/SSIM	FSRCNN PSNR/SSIM	VDSR PSNR/SSIM	LGCNet PSNR/SSIM	DCM PSNR/SSIM	HSENet PSNR/SSIM	TransENet PSNR/SSIM	Ours PSNR/SSIM M
2	30.76/0.8789	32.84/0.9152	33.18/0.9196	33.38/0.9220	33.48/0.9235	33.65/0.9274	34.22/0.9327	35.43/0.9355	**35.88/0.9721**
3	27.46/0.7631	28.66/0.8038	29.09/0.8167	29.28/0.8232	29.28/0.8238	29.52/0.8349	30.00/0.8420	31.03/0.8526	**31.17/0.8815**
4	25.65/0.6725	26.78/0.7219	26.93/0.7267	26.85/0.7317	27.02/0.7333	27.22/0.7528	27.73/0.7623	28.74/0.7694	**28.97/0.8009**

The best and second results are highlighted and underlined.

TABLE II
PSNR/SSIM RESULTS ON AID DATASET OF SCALE X2, X3, AND X4

Scale	Bicubic PSNR/SSIM	SRCNN PSNR/SSIM	FSRCNN PSNR/SSIM	VDSR PSNR/SSIM	LGCNet PSNR/SSIM	DCM PSNR/SSIM	HSENet PSNR/SSIM	TransENet PSNR/SSIM	Ours PSNR/SSIM
2	32.39/0.8906	34.49/0.9286	34.73/0.933	35.05/0.9346	34.80/0.9320	35.21/0.9366	35.24/0.9368	35.28/0.9374	**37.04/0.9626**
3	29.08/0.7863	30.55/0.8372	30.98/0.840	31.15/0.8522	30.73/0.8417	31.31/0.8561	31.39/0.8572	31.45/0.8595	**33.00/0.8895**
4	27.30/0.7036	28.40/0.7561	28.77/0.772	28.99/0.7753	28.61/0.7626	29.17/0.7824	29.21/0.7850	29.38/0.7909	**30.78/0.8185**

The best and second results are highlighted and underlined.

reveal that MSFFTAN outperforms other advanced methods by a wide margin, offering the best restoration results in all three upscale factors. Specifically, our model achieves 1.66, 1.17, and 1.24 dB improvement over the second-best method (HSENet) on all three upscale factors. Furthermore, for the SSIM metric, our model outperforms HSENet by 0.0394, 0.0395, and 0.0386, respectively. However, the complexity of MSFFTAN is half of HSENet, which is attribute to the ability of our designed network to fully exploit and explore multiscale information. The AID dataset is utilized to evaluate the generality and generalization performance further since the images in this dataset contain more categories and a higher disparity than those in UCMerced Landuse dataset. In this dataset, we evaluate the developed MSFFTAN against several SR algorithms, including Bicubic, SRCNN, LGCNet, VDSR, DCM, and TransENet [37]. According to Table II, it can be seen that MSFFAN has the greatest average PSNR and SSIM score in all three upscale factors. More specifically, compared to the currently leading method TransENet, we improve the PSNR (SSIM) from 35.28 (0.9374) to 37.04 (0.9626) for upscale factor 2 and from 29.38 (0.7909) to 30.78 (0.8185) for upscale factor 4. The results reveal that in most circumstances, the designed MSFFTAN exceeds the existing leading approaches, confirming the stronger generalization ability of MSFFTAN. In addition, Table III provides comprehensive discovery of several approaches for all 30 scene classes of the AID dataset at an upscale factor of 4. MSFFTAN yields the highest PSNR scores in 14 scene classes, while TransENet scored better in the remaining scene categories. It is worth mentioning, however, that MSFFTAN obtains a good result that is 1.4 dB higher than TransENet. To further demonstrate the superiority of our proposed method, we employed the LPIPS metric. The lower the image quality is, the higher LPIPS is. As seen in Table IV, MSFFTAN outperforms other approaches by a significant margin. Specifically, MSFFTAN is 0.0009 lower than the current SOTA method HSENet on scale factor 2. This reveals that the reconstructed images generated by our method exhibit a higher degree of aesthetic appeal to the human visual system. Finally, as demonstrated in Fig. 6, it can be observed that the MSFFTAN exhibits faster convergence, further highlighting

TABLE III
MEAN PSNR (DB) OF EACH CLASS FOR UPSCALING FACTOR 4 ON AID TEST DATASET

Class	Bicubic PSNR	SRCNN PSNR	LGCNet PSNR	VDSR PSNR	DCM PSNR	HSENet PSNR	TransENet PSNR	Ours PSNR
Airport	27.03	28.17	28.39	28.82	28.99	29.03	29.23	**29.43**
Bareland	34.88	35.63	35.78	35.98	36.17	36.21	36.20	**36.22**
Baseball field	29.06	30.51	30.75	31.18	31.36	31.23	**31.59**	31.34
Beach	31.07	31.92	32.08	32.29	32.45	32.76	32.55	**33.42**
Bridge	28.98	30.41	30.67	31.19	31.39	31.30	**31.63**	30.85
Center	25.26	26.59	26.92	27.48	27.72	27.84	**28.03**	27.56
Church	22.15	23.41	23.68	24.12	24.29	24.39	24.51	**24.52**
Commercial	25.83	27.05	27.24	27.62	27.78	27.99	27.97	**28.20**
Dense residential	23.05	24.13	24.33	24.70	24.87	24.44	**25.13**	24.73
Desert	38.49	38.84	39.06	39.13	39.27	39.37	**39.31**	38.96
Farmland	32.30	33.48	33.77	34.20	34.42	33.90	**34.58**	34.23
Forest	27.39	28.15	28.20	28.36	28.47	38.31	**28.56**	28.45
Industrial	24.75	26.00	26.24	26.72	26.92	26.99	27.21	**27.36**
Meadow	32.06	32.57	32.65	32.77	32.88	32.74	32.94	**33.29**
Medium residential	26.09	27.37	27.63	28.06	28.25	28.11	**28.45**	26.84
Mountain	28.04	28.90	28.97	29.11	29.18	29.26	**29.28**	28.97
Park	26.23	27.25	27.37	27.69	27.82	28.23	28.01	**28.59**
Parking	22.33	24.01	24.40	25.21	25.74	26.17	**26.40**	26.36
Playground	27.27	28.72	29.04	29.62	29.92	31.18	30.30	**32.12**
Pond	28.94	29.85	30.00	30.26	30.39	30.40	**30.53**	30.37
Port	24.69	25.82	26.02	26.43	26.62	26.92	26.91	**27.00**
Railway station	26.31	27.55	27.76	28.19	28.38	28.47	**28.61**	28.16
Resort	25.98	27.12	27.32	27.71	27.88	27.99	**28.08**	27.36
River	29.61	30.48	30.60	30.82	30.91	30.88	**31.00**	30.41
School	24.91	26.13	26.34	26.78	26.94	27.51	27.22	**27.47**
Sparse residential	25.41	26.16	26.27	26.46	26.53	26.44	26.43	**26.51**
Square	26.75	28.13	28.39	28.91	29.13	29.05	**29.39**	28.88
Stadium	24.81	26.10	26.37	26.88	27.10	27.28	27.41	**27.84**
Storage tanks	24.18	25.27	25.48	25.86	26.00	26.07	26.20	**26.82**
Viaduct	25.86	27.03	27.26	27.74	27.93	28.12	**28.21**	28.05
AVG	27.30	28.40	28.61	28.99	29.17	29.21	29.38	**30.78**

The best results are highlighted.

the effectiveness and superiority of the proposed module. These positive results support the efficacy of our method.

2) Visual Comparison: We assess the visual quality of the given MSFFTAN to current leading approaches to further validate its efficacy. Figs. 7–11 display multiple example super-resolution results from the testset acquired utilizing various approaches, as well as the HR images for convenient comparison. It is worth noting that a close-up region denoted by red rectangle is displayed below the related image for convenient comparison. According to Fig. 7, we can observe that MSFFTAN is able to reconstruct images closest to HR. It is noteworthy

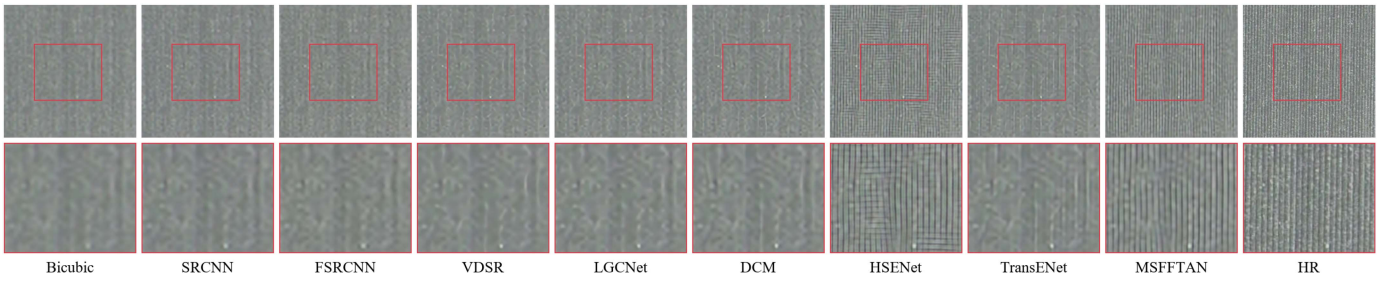


Fig. 7. Visual comparison on UCMerced Landuse with scale factor 2.

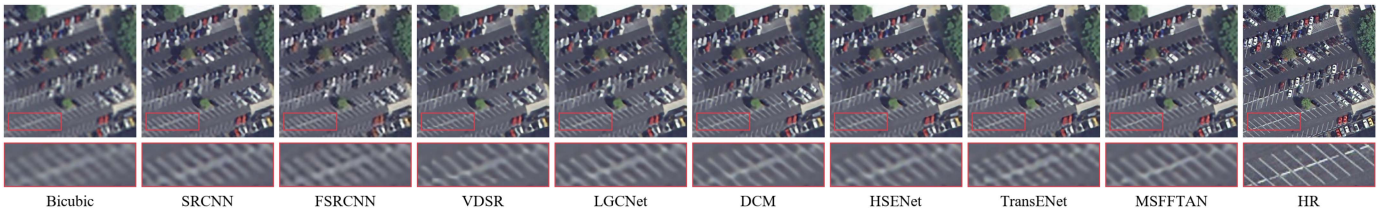


Fig. 8. Visual comparison on the UCMerced Landuse with scale factor 3.

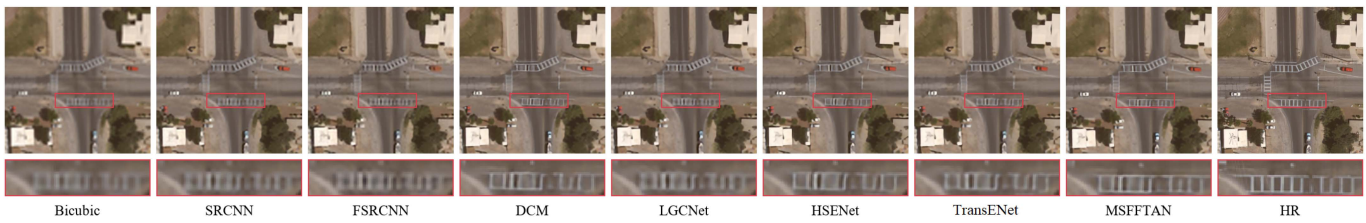


Fig. 9. Visual comparison on the UCMerced Landuse with scale factor 3.

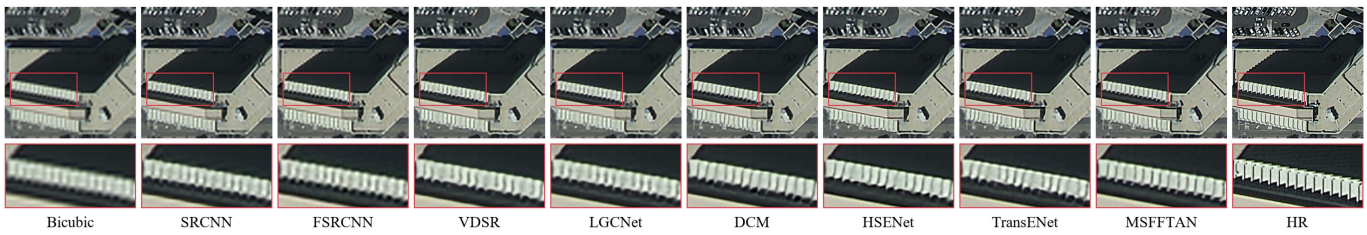


Fig. 10. Visual comparison on the UCMerced Landuse with scale factor 4.

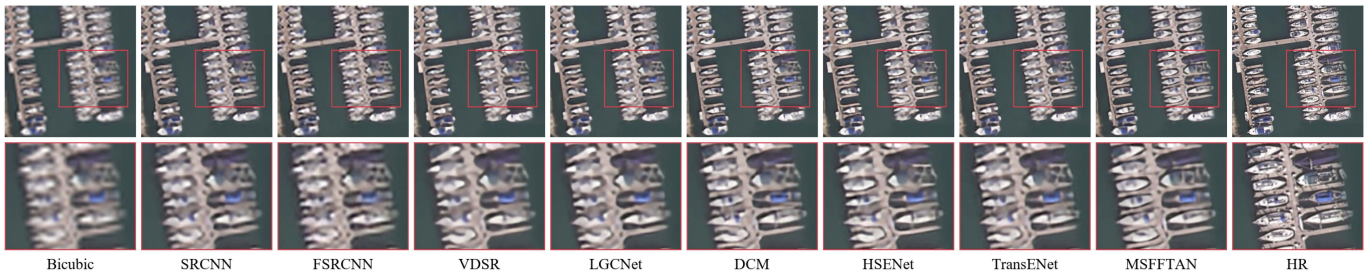


Fig. 11. Visual comparison on the UCMerced Landuse with scale factor 4.

TABLE IV
LPIPS RESULTS ON UCMERCED LANDUSE DATASET OF SCALE X2, X3, AND X4

Scale	Bicubic LPIPS	SRCNN LPIPS	FSRCNN LPIPS	VDSR LPIPS	LGCNet LPIPS	DCM LPIPS	HSENet LPIPS	TransENet LPIPS	Ours
2	0.0721	0.0444	0.0471	0.0287	0.0293	0.0284	<u>0.0266</u>	0.0279	**0.0257**
3	0.1281	0.0945	0.1062	0.0801	0.0752	0.0698	0.0654	<u>0.0649</u>	**0.0645**
4	0.1650	0.1260	0.1395	0.1102	0.1093	0.1046	0.1081	<u>0.1030</u>	**0.1020**

The best and second results are highlighted and underlined.

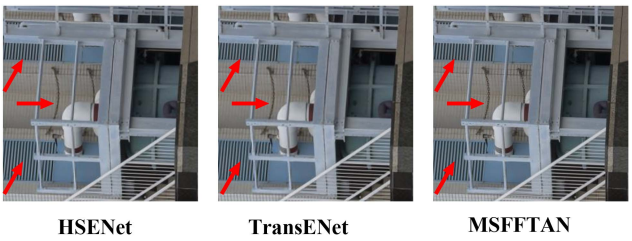


Fig. 12. Visual comparison on a real remote-sensing sample.

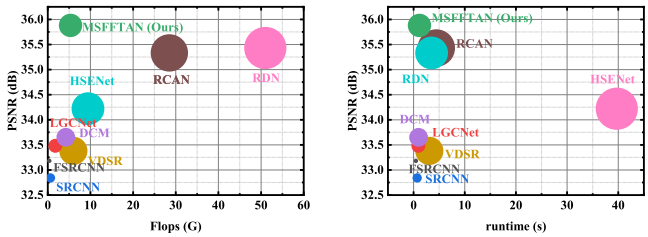


Fig. 13. Performance and Complexity. Results are evaluated on UCMerced LandUse dataset with scale factor 2.

that the self attention-based HSENet and TransENet produced significant checkerboard effects and artifacts. We conjecture that this is due to the self-attention mechanism being influenced by noise and degradation, aggregating incorrect information. As displayed in Fig. 8, MSFFTAN produces the clearest parking places at a large magnification, whereas other approaches yield variable degrees of blurring, distortion, and warping, which further demonstrates the superiority of our method. The second-best network recovered by the zebra line loses a lot of lines, as seen in Fig. 9, but our MSFFTAN can provide the closest image to the HR. Furthermore, other approaches cause artifacts in the most challenging situation (magnification scale of 4×), however our method produces good visual results, as shown in Figs. 10 and 11. As depicted in Fig. 11, our proposed MSFFTAN ensures the maximum preservation of the yacht's authenticity, while other methods exhibit varying degrees of distortion and degradation. To further prove the generalization performance of the proposed method, we tested it on real remote-sensing images. As shown in Fig. 12, MSFFTAN has better reconstruction performance than the leading RSSR method. Specifically, MSFFTAN is able to recover better lines (as shown by the red arrow in the figure), while HSENet and TransENet's recovered image lines distorted. From the above analysis, we can conclude that MSFFTAN can produce visually satisfying HR images, which have rich and real textures, sharp edges, and clear boundaries.

D. Ablation Study

1) Study the number of FFT-RB in Encoder and Decoder: The number of basic blocks on network performance is investigated in this subsection, as network depth has a substantial impact on model reconstruction properties. As a result, we perform a series of experiments to investigate this point. Table V compares the reconstruction using the UCMerced LandUse dataset with different basic block settings when the upscale factor is set to 2. Specifically, MSFFTAN_abc stands for the

TABLE V
RESULTS WITH DIFFERENT EBS AND DBS SETTINGS FOR UPSCALING FACTOR 2 ON UCMERCED LANDUSE DATASET

Methods	Params	FLOPs	PSNR	SSIM
MSCAN_111	6.727M	2.68G	35.486	0.9682
MSCAN_211	6.947M	3.11G	35.476	0.9683
MSCAN_221	7.813M	3.54G	35.573	0.9694
MSCAN_222	11.250M	3.97G	35.476	0.9681
MSCAN_311	7.167M	3.54G	35.510	0.9681
MSCAN_411	7.387M	3.98G	35.524	0.9681
MSCAN_321	**8.034M**	**3.97G**	**35.597**	**0.9696**
MSCAN_333	15.7M	5.27G	35.538	0.9688
MSCAN_511	7.608M	4.41G	35.488	0.9683

The best results are highlighted.

different depths of backbone number of basic module settings, which are a, b, and c. We can see that when the number of FFT-RBs of MSFFTAN of the encoder and decoder are set to 321, MSFFTAN can obtain the highest PSNR and SSIM. It is worth noticing that when we increase the number of blocks in the network to reach MSFFTAN_333, the performance of network drops, which we ascribe to parameter overfitting. Finally, this also demonstrates that using an appropriate blocks setting may further enhance the reconstruction quality.

2) Effectiveness of MultiInput Mechanism: Multiinput strategy is an essential part of multiscale information exploration and aggregation. To achieve improved performance, the multiinput technique is designed to permit as much origin multiscale information in remote-sensing images as feasible. Here, we investigate the effect of this design with different inputs. According to the Table VI, when we add the Input-2 AP, our model achieves a 0.007-dB improvement. In addition, by adding Input-3 AP, we get a 0.06-dB improvement. We discover that the benefit of providing extra auxiliary pathways grows as the network deepens, owing to the increasing loss of shallow information as MSFFTAN grows. As a result, we may conclude that using a multiinput strategy can lead to improved performance.

TABLE VI
RESULTS WITH DIFFERENT INPUT SETTINGS FOR UPSCALING FACTOR 2 ON UCMERCED LANDUSE DATASET

Input-1	Input-2	Input-3	Params	FLOPs	PSNR	SSIM
✔	✗	✗	7.171M	3.78G	35.528	0.9687
✔	✔	✗	7.333M	3.87G	35.535	0.9686
✔	✔	✔	**8.034M**	**3.97G**	**35.597**	**0.9696**

The best results are highlighted.

TABLE VII
RESULTS WITH DIFFERENT AGGREGATION SETTINGS FOR UPSCALING FACTOR 2 ON UCMERCED LANDUSE DATASET

Method	Params	FLOPs	PSNR	SSIM
SUM	10.083 M	5.35G	35.818	0.9718
CONCAT	10.505 M	5.47G	35.752	0.9721
BGSB (ours)	**10.156 M**	**5.35G**	**35.827**	**0.9724**

The best results are highlighted.

TABLE VIII
RESULTS WITH DIFFERENT CHANNEL ATTENTION SETTINGS FOR UPSCALING FACTOR 4 ON UCMERCED LANDUSE DATASET

Method	Params	FLOPs	PSNR	SSIM
w/o CA	7.946 M	3.96G	35.511	0.9683
w/ SE [43]	7.969 M	3.96G	35.498	0.9685
w/ CBAM [54]	7.948 M	3.96G	35.340	0.9671
w/ LGCAB (ours)	**8.034 M**	**3.97G**	**35.578**	**0.9694**

The best results are highlighted.

3) Study of BGSB: BGSB is specifically designed for multiscale and hierarchical feature exploration and aggregation. In this part, we perform a series of experiments to illustrate the efficacy of BGSB comparing with the SUM operation and CONCAT operation (as shown in Table VII). In comparison to the SUM operation, our BGSB improves PSNR and SSIM by 0.019 dB and 0.000016, respectively, with near little increase in Parameters. In addition, the FLOPS for these two operations are almost identical, but BGSB achieves better performance. More importantly, when compared to CONCAT operation, BGSB has a significant performance and complexity advantage. Specifically, BGSB obtains a boost of 0.116 dB and 0.0013, but only takes up 96% of the parameters and 97% of FLOPS. These positive results support the efficacy of our BGSB. Finally, this also demonstrates that using an appropriate multiscale feature fusion approach may show considerable future reconstruction effort.

4) Study of LGCAB: In LGCAB, we use the dual-branch structure to better extract small- and large-size information simultaneously. To prove the effectiveness of using LGCAB, we remove LGCAB or add other commonly used channel attention blocks (e.g., SE or CBAM) to perform ablation experiments. As shown in Table VIII, we show the results of these modified networks. If we do not employ any channel attention mechanism, the super-resolution performance will drop dramatically, and the usage of LGCAB raises the PSNR and SSIM scores by 0.067 dB and 0.001, respectively. It is worth noting that the use of the widely employed SE and CBAM modules resulted in a rapid decline in network performance. We hypothesize that this is due to the fact that the SE and CBAM modules employ global average pooling and global max pooling to compress spatial information, resulting in the loss of a large number of small-scale features that are also crucial for the final reconstruction. Additionally, they only capture global peak signals that do not accurately reflect texture and structural information, which is another reason for the decrease in network performance. Furthermore, compared with one branch channel attention block (all spatial information is discarded), using the dual-branch structure promotes the average PSNR and SSIM values by 0.159 dB and 0.0016, respectively. Therefore, we can draw a conclusion that we can get better performance by applying the LGCAB which can capture both large- and micro-scale characteristics.

TABLE IX
RESULTS WITH DIFFERENT LOSSES FOR UPSCALING FACTOR 2 ON UCMERCED LANDUSE DATASET AND AID DATASET

Methods	UCMerced LandUse PSNR/SSIM	AID PSNR/SSIM
MSFFTAN w/o FFT loss	31.414/0.8181	31.138/0.8790
MSFFTAN	**31.432/0.8185**	**31.173/0.8798**

The best results are highlighted.

TABLE X
ABLATION INVESTIGATION OF DIFFERENT COMPONENTS ON UCMERCED LANDUSE DATASET WITH UPSCALING FACTOR 2

Case index	Model 0	Model 1	Model 2	Model3
AP	✗	✔	✔	✔
FFT-RB	✗	✗	✔	✔
LGCAB	✗	✗	✗	✔
PSNR (dB) ↑	35.110	35.196	35.231	**35.305**
SSIM↑	0.9314	0.9319	0.9323	**0.9332**

The best results are highlighted.

5) Study of FFT Loss: In this section, we investigate the impact of the loss function on the final performance of the model as shown in Table IX. In order to balance the distribution of the FFT loss and L1 Pixel, we use a relatively small weight on the FFT loss term, which helps to optimize the network. In addition, through experimentation, we found that adding the FFT loss can improve the quality of the reconstructed image, specifically resulting in a 0.018 and 0.035-dB increase in PSNR on the UCMerced LandUse and AID datasets, respectively.

6) Effectiveness of Our Proposed Components: In this subsection, we investigate the individual contributions of various components of our proposed model through ablation experiments. We use a baseline model consisting of the main path without an AP, FFT-RB, and LGCAB. All comparative models are trained for 500 epochs on the UCMerced LandUse dataset under consistent experimental conditions. From Table X, we can conduct that using AP (Model 1) can improve 0.086 dB compared with the baseline model (Model 0). The AP module can automatically supplement a variety of origin shallow multiscale information which play important roles in the reconstruction of degraded remote-sensing images. Compared

with the baseline model (Model 0), the FFT-RB model (Model 2) achieves an improvement of 0.121 dB and 0.001 in terms of PSNR and SSIM. The incorporation of the FFT-RB module, which utilizes FFT, enables efficient capture of global information, which is crucial for the reconstruction of high-spatial resolution remote-sensing imagery. Furthermore, the proposed LGCAB module derives a numerical gain of 0.24 dB and 0.002 for PSNR and SSIM, respectively. The LGCAB model employs a resource allocation strategy that prioritizes the allocation of resources to regions of higher criticality, while concurrently implementing mechanisms to suppress irrelevant information from both a local and global perspective. In summary, the overall performance of the network is notably superior when incorporating our proposed components, thereby demonstrating the effectiveness of our proposed modules.

E. Model Complexity Analysis

The tradeoff between PSNR and FLOPS is examined as shown in Fig. 13 in this section. FLOPs stands for floating point operations, which is defined as the number of computations and can be used to measure the complexity of models. Obviously, MSFFTAN achieves competitive results with fewer FLOPS. Despite CTN having fewer FLOPS than MSFFTAN, its performance is 2.29 dB worse. MSFFTAN, one the other hand, achieves a 1.66-dB enhancement while only requiring half of HSENet's FLOPS, indicating that MSFFTAN can reach a reasonable balance between model complexity and performance. In conclusion, MSFFTAN has fewer FLOPS and produces excellent super-resolution results than previous approaches, demonstrating that our method has achieved a satisfactory balance between network complexity and image super-resolution quality.

V. Conclusion

In this work, a novel FFT-based multiscale attention network, referred to as MSFFTAN, is proposed for the task of RSSR. The MSFFTAN utilizes a multiinput encoder–decoder structure to extract multiscale information and enhance features, resulting in superior reconstruction capabilities. In particular, a FFT-RB, containing a convolution operation and FFT operation, is elaborately designed to extract and aggregate both local details and global structures. To enhance the ability of MSFFTAN to utilize both large and small target information, an LGCAB is constructed. More importantly, a BGSB is presented to make full use of middle features from multiple scales and depths in order to increase the quality of the reconstructed results. Extensive experiments on both two public datasets indicate that MSFFTAN outperformed other presently leading approaches in quantitative and qualitative evaluations.

References

[1] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, 2021.

[2] L. Chen, H. Liu, M. Yang, Y. Qian, Z. Xiao, and X. Zhong, "Remote sensing image super-resolution via residual aggregation and split attentional fusion network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9546–9556, 2021.

[3] B. Liu et al., "Saliency-guided remote sensing image super-resolution," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5144.

[4] Y. Ma, P. Lv, H. Liu, X. Sun, and Y. Zhong, "Remote sensing image super-resolution based on dense channel attention network," *Remote Sens.*, vol. 13, no. 15, 2021, Art. no. 2966.

[5] C.-M. Feng, H. Fu, S. Yuan, and Y. Xu, "Multi-contrast MRI super-resolution via a multi-stage integration network," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 140–149.

[6] C.-M. Feng, Y. Yan, H. Fu, L. Chen, and Y. Xu, "Task transformer network for joint MRI reconstruction and super-resolution," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 307–317.

[7] C.-M. Feng et al., "Multi-modal transformer for accelerated MR imaging," *IEEE Trans. Med. Imag.*, early access, doi: 10.1109/TMI.2022.3180228.

[8] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2492–2501.

[9] S. Shivappriya, M. J. P. Priyadarsini, A. Stateczny, C. Puttamadappa, and B. Parameshachari, "Cascade object detection and remote sensing object detection method based on trainable activation function," *Remote Sens.*, vol. 13, no. 2, 2021.

[10] S. Zhang, Q. Yuan, J. Li, J. Sun, and X. Zhang, "Scene-adaptive remote sensing image super-resolution using a multiscale attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4764–4779, Jul. 2020.

[11] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[12] L. Xia, J. Chen, J. Luo, J. Zhang, D. Yang, and Z. Shen, "Building change detection based on an edge-guided convolutional neural network combined with a transformer," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4524.

[13] Z. Cai, Z. Jiang, and Y. Yuan, "Task-related self-supervised learning for remote sensing image change detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 1535–1539.

[14] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[15] S. Sarkar and R. R. Sahay, "A non-local superpatch-based algorithm exploiting low rank prior for restoration of hyperspectral images," *IEEE Trans. Image Process.*, vol. 30, pp. 6335–6348, 2021.

[16] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3517–3526.

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

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

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

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[22] 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. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[23] Y. Zhang, D. Wei, C. Qin, H. Wang, H. Pfister, and Y. Fu, "Context reasoning attention network for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4278–4287.

[24] J. Zhang et al., "A two-stage attentive network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1020–1033, Mar. 2022.

[25] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2207.

[26] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.

[27] X. Dong, X. Sun, X. Jia, Z. Xi, L. Gao, and B. Zhang, "Remote sensing image super-resolution using novel dense-sampling networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1618–1633, Feb. 2020.

[28] 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, 2021.
[29] 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.
[30] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.
[31] T. Xu, T.-Z. Huang, L.-J. Deng, and N. Yokoya, "An iterative regularization method based on tensor subspace representation for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 5529316.
[32] D. Hong, J. Yao, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," 2022, *arXiv:2205.03742*.
[33] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Comput. Vis.–ECCV: 16th Eur. Conf., Glasgow, U.K., Aug. 23–28, 2020, Proc., Part XXIX 16*, Springer, 2020, pp. 208–224.
[34] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[35] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7918–7933, Oct. 2019.
[36] Y. Li et al., "Single-image super-resolution for remote sensing images using a deep generative adversarial network with local and global attention mechanisms," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–24, 2022, Art. no. 3000224.
[37] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 5615611.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv.–MICCAI: 18th Int. Conf.*, Munich, Germany, Springer, 2015, pp. 234–241.
[39] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[40] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual fourier transformation for single image deblurring," 2021, *arXiv:2111.11745*.
[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11976–11986.
[42] A. G. Howard et al., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[43] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[44] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3147–3155.
[45] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[46] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[47] G.-S. Xia et al., "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[51] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 391–407.
[52] J. M. Haut, M. E. Paoletti, R. Fernandez-Beltran, J. Plaza, A. Plaza, and J. Li, "Remote sensing single-image superresolution based on a deep compendium model," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 9, pp. 1432–1436, Sep. 2019.
[53] S. Wang, T. Zhou, Y. Lu, and H. Di, "Contextual transformation network for lightweight remote-sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.
[54] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

Zheng Wang received the B.S. degree in biological engineering from the Zhejiang University of Technology, Hangzhou, China, in 2008, the M.S. degree in computational biology from the University of East Anglia, Norwich, U.K., in 2011, and the Ph.D. degree in control science from the Zhejiang University of Technology, in 2020.

He is currently a Lecturer with the School of Computer and Computational Sciences, Hangzhou City University, Hangzhou, China. He has authored or coauthored over ten journal and conference papers. His research interests include pattern recognition, intelligent computing, and optimization dispatch complex systems.

Yanwei Zhao received the B.S. degree in mechanical engineering and automation from the Henan University Of Science And Technology, Luoyang, China, in 1982, and the Ph.D. degree in mechanical and electronic engineering from Shanghai University, Shanghai, China, in 2005.

He is currently a Professor with the College of Engineering, Hangzhou City University, Hangzhou, China. His research interests include manufacturing, control technology, artificial intelligence, and intelligent scheduling technology.

Jiacheng Chen is currently working toward the Ph.D degree in Computer Science and Technology with the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China.

His research interests include image restoration and image processing.