DTCNet: Transformer-CNN Distillation for Super-Resolution of Remote Sensing Image

Cong Lin, Xin Mao, Chenghao Qiu, and Lilan Zou

***Abstract*—Super-resolution reconstruction technology is a crucial approach to enhance the quality of remote sensing optical images. Currently, the mainstream reconstruction methods leverage convolutional neural networks (CNNs). However, they overlook the global information of the images, thereby impacting the reconstruction effectiveness. Methods based on Transformer networks have demonstrated the capability to improve reconstruction quality, but the high model complexity renders them unsuitable for remote sensing devices. To enhance reconstruction performance while maintaining the model lightweight, a distillation Transform-CNN Network is proposed in this article. The strategy employs the Transformer network as a teacher network, guiding its long-range features into a compact CNN, achieving distillation across networks. Simultaneously, to rectify misinformation in the teacher network, prior information is introduced to ensure accurate information transfer. Concerning the student network, a novel upsampling approach is devised, utilizing inherent information in downsampled feature maps for padding, thereby avoiding the introduction of zero-information feature points in the traditional deconvolution process. Experimental evaluations conducted on multiple publicly available remote sensing image datasets demonstrate that the proposed method, while maintaining a smaller parameter count, achieves outstanding reconstruction quality for remote sensing images, surpassing existing approaches.**

***Index Terms*—Gaofen satellite, knowledge distillation (KD), lightweight network, remote sensing image, super-resolution (SR).**

Nomenclature

CNN	Convolutional neural network.
LR	Low resolution.
HR	High resolution.
SR	Super resolution.
KD	Knowledge distillation.
SRRD	Super-resolution reconstruction decoder.
LFE	Low-level feature extractor.
HFE	High-level feature extractor.
HIR	High-quality image reconstruction.
PSNR	Peak signal-to-noise ratio.
SSIM	Structural similarity.
SAM	Spectral angle mapper.
NIQE	Natural image quality evaluator.
NSS	Natural Scene Statistics.
MVG	Multivariate Gaussian.
Ent	Entropy.
Std	Standard deviation.
Grad	Average gradient.

Manuscript received 14 January 2024; revised 3 March 2024 and 1 May 2024; accepted 2 June 2024. Date of publication 5 June 2024; date of current version 14 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62272109, in part by Guangdong University Student Science and Technology Innovation Cultivation Special Fund Support Project under Grant pdjh2023 a0243, and in part by the Undergraduate Innovation Team Project of Guangdong Ocean University under Grant CXTD2024011. *(Corresponding authors: Chenghao Qiu; Lilan Zou.)*

Cong Lin, Xin Mao, and Lilan Zou are with the School of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China (e-mail: lincong07@163.com; 11632119ee@stu.gdou.edu.cn; zoulilan2015@163.com).

Chenghao Qiu is with the School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610000, China (e-mail: qiuch@std.uestc.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3409808

I. Introduction

REMOTE sensing technology is an evolving and continually improving discipline that advances alongside societal progress and technological developments. With widespread applications in various domains, such as agriculture, military, and transportation, remote sensing technology has demonstrated significant value [1], [2], [3]. Optical remote sensing images constitute a crucial component of this technology, offering rich spatial and spectral information. However, due to limitations imposed by imaging devices and signal transmission, remote sensing images often exhibit various types of noise, adversely affecting image clarity and quality. To mitigate noise in remote sensing images, numerous scholars have proposed denoising algorithms based on diverse theories and methods [4], validating their effectiveness through experimentation. Nevertheless, beyond noise reduction, a more challenging issue persists, i.e., how to improve the sharpness of remote sensing images, allowing the visualization of more subtle features and data.

The improvement of the resolution of remote sensing images has been one of the research directions that have attracted much attention in the field of remote sensing. HR remote sensing images are of great significance for applications, such as land cover classification, resource monitoring, and environmental analysis. However, due to the limitations of sensor performance, data transmission, and storage, the resolution of remote sensing images actually acquired often fails to meet the needs of specific applications. An important research direction at present is to utilize multisource data fusion techniques. By integrating data from different sensors, multispectral bands, or different time points, researchers can obtain richer and more comprehensive information while maintaining high quality resolution [5], [6]. This fusion method not only helps to improve the spatial and temporal resolution of the images but also enhances the recognition and classification of feature characteristics in remote

sensing images. In addition, there are also related studies that by reconstructing in the HSV domain within the RGB spatial domain, not only can we achieve spectral SR but also obtain more detailed and accurate information about the motion of the object of interest [7]. Compared to the above two methods, SR reconstruction through a single image can be used to recover a HR image from a single LR image using deep learning methods to improve the spatial details and quality of the image in the case of more limited data [8].

The process of single image SR reconstruction involves restoring a HR image from a lower-resolution counterpart. While this approach enhances image clarity and details, the inherent challenge and uncertainty arise from the insufficient information present in the LR image. To address this problem, three primary categories of methods are commonly employed: 1) interpolation-based methods, such as those discussed in [9] and [10]; 2) reconstruction-based methods, as outlined in [11]; and 3) learning-based methods, exemplified by studies like [12] and [13]. The most straightforward and expeditious approaches fall within the realm of interpolation-based methods. These methods, characterized by their simplicity and speed, involve the estimation of unknown pixels by leveraging information from neighboring pixels. Notable interpolation methods encompass nearest neighbor interpolation, bilinear interpolation, bicubic interpolation, etc. However, these methods often result in low-quality images after reconstruction, which are prone to blurring and aliasing effects. Reconstruction-based methods are usually based on multiple frames of images, which constrain the reconstruction process by combining prior knowledge, such as convex set projection method [11], Bayesian analysis method [14], iterative back-projection method [15], etc. In contrast to interpolation techniques, these methods exhibit superior capabilities in enhancing the quality of reconstructed images. However, it is noteworthy that along with this improvement, they concurrently escalate computational complexity and time requirements. In recent years, learning-based approaches have gained widespread popularity [16], [17]. These methods involve acquiring the mapping relationship from pairs of LR and HR images, ultimately resulting in the attainment of HIR. However, these early learning-based methods require manual feature design, which still face challenges for larger magnification factors and more complex image contents.

The advancement of deep learning technology has catalyzed the rise of SR reconstruction methods employing CNN. These approaches prove effective in restoring intricate texture details within images. However, the early CNN methods had low image quality after reconstruction due to the small number of parameters [18]. To improve the performance of image reconstruction, some scholars increased the depth of the network and enhanced the network's ability to extract image features [19]. Overall, CNN generally have a local receptive field, making them more efficient in terms of computation and memory usage. However, they may struggle to capture global context and long-range dependencies, which are crucial for reconstructing high-quality images from LR inputs [20]. Furthermore, CNN often require deeper or more complex architectures to increase their receptive field, which may lead to increased model complexity and a risk of overfitting [21], [22]. With the introduction of the Transformer model [23], it has achieved great success in the field of natural language processing, and aroused the interest of more and more scholars to apply it to the field of image processing. The Transformer model can handle the dependency relationships between sequence data well, which is beneficial for the model to learn the global features of images [24], [25], [26]. However, the Transformer model also has some problems, such as the computational complexity and memory consumption of the self-attention mechanism increase sharply with the increase of the input sequence length, which brings great pressure on the computing resources and storage space, especially when processing HR images [27]. The self-attention mechanism calculates interactions between all pairs of input positions, leading to a computational cost that has a quadratic relationship with the size of the input. This poses a particular challenge for SR tasks, which typically involve processing large images with fine details. This is undesirable for remote sensing task scenarios with limited device performance, because remote sensing tasks often require fast, accurate, and stable image processing results. Therefore, how to obtain a simple and high-performance model that can adapt to the characteristics and requirements of remote sensing images is an urgent problem to be solved.

Currently, there are numerous strategies aimed at reducing the complexity of neural networks, among which KD has emerged as a prominent focus of research [28]. KD involves the transmission of information between models, facilitating the transfer of knowledge from a complex teacher network to a relatively simplified student network. This approach enables the maintenance of model performance on specific tasks while reducing the overall scale. Notably, KD plays a significant role in improving model inference efficiency and allows for the application of high-performance models in environments with limited computational resources. However, existing KD methods predominantly concentrate on applications between networks of the same type [29], [30], [31]. Meanwhile in the implementation of KD, these architectural paradigms face several challenges, including: how to efficiently transfer knowledge from the teacher network to the student network, especially when there are significant differences between their architectures, this process may encounter difficulties; second, the selection of an appropriate teacher network is crucial for the performance of the student network. The teacher network needs to be sufficiently powerful to provide valuable information and guidance; furthermore, the student network needs to have enough capacity to learn the knowledge from the teacher network, but at the same time, it should not be too large to lose the advantages of KD; finally, how to design effective distillation strategies to ensure that the student network can capture the important features and knowledge from the teacher network.

To address the challenges of cross-architecture knowledge transfer between CNN and Transformer in the KD task, and to improve the efficiency of network distillation in the remote sensing image SR task, while ensuring that the network remains lightweight and the performance of the final model is not significantly affected, this article combines the advantages of CNN and Transformer networks and proposes a teacher–student network called the distillation Transform-CNN network (DTCNet) for use on the remote sensing SR tasks. This method not only helps

to convey the model's deep understanding of the data but also enhances the ability of the student network to generalize to specific tasks. Specifically, first, a pretrained Transformer network is used as the teacher network, which extracts high-dimensional features from the LR remote sensing images and captures the global dependency relationships between the features. Then, a CNN-based SR reconstruction network is designed as the student network, which learns the mapping relationship between the LR and HR images under the guidance of the teacher network, and overcomes the problem of insufficient global information perception ability of pure CNN. Finally, a subpixel convolution layer is introduced in the feature upsampling stage, which better utilizes the inherent information in the feature map and improves the clarity and details of the image. The main contributions of this article are as follows.

1) Introducing a cross-network distillation framework rooted in KD, this approach employs the Transformer network as the teacher network and the CNN as the student network. Utilizing the KD layer, the output of the teacher network serves as an auxiliary supervision signal, guiding the student network to acquire more comprehensive global feature information.
2) A correction method for the erroneous knowledge of the teacher network is proposed, which constrains the erroneous edges in the teacher network by prior information, aiming to improve the correctness of the knowledge learned by the student network.
3) An upsampling method based on subpixel convolution is proposed, which replaces the traditional deconvolution method to improve the image quality in the image SR task. This method uses the subpixel convolution to map the feature map of the LR image to the pixel space of the HR image, thus realizing the upsampling operation. Compared with the traditional deconvolution method, it can make full use of the network's own features, and avoid the blurring problem that may be caused by the zero-padding operation in the deconvolution.

II. Related Work

A. Transformer Network Theoretical Foundations

The Transformer network is an innovative neural network architecture that relies on the self-attention mechanism to achieve sequence-to-sequence mapping and modeling. It has demonstrated notable achievements in various domains, such as natural language processing, speech recognition, and image generation [23]. The self-attention mechanism proves effective in capturing correlations between positions within a sequence, thereby enhancing the model's representational and generalization capabilities. In comparison to traditional recurrent neural networks and CNN, the Transformer network boasts two key advantages: first, it completely eschews recurrent and convolutional operations, opting solely for self-attention mechanisms and feedforward networks, resulting in heightened parallelism and computational efficiency; second, it can handle sequences of arbitrary lengths without being constrained by fixed windows or step sizes, making it more flexible and efficient, particularly in processing lengthy textual data.

Specifically, the self-attention mechanism comprises three components: 1) query vectors, 2) key vectors, and 3) value vectors [32]. Assuming the input sequence is $\mathbf{X}$, for each position i, it can be represented as a d-dimensional matrix $\mathbf{h}_i$, where d is the dimensionality of the matrix. The attention weight for this position can then be calculated using the following formula:

$$\alpha_{i,j} = \frac{\exp(\mathbf{e}_{i,j})}{\sum_{m=1}^{n} \exp(\mathbf{e}_{i,m})} \tag{1}$$

where $\mathbf{e}_{i,j}$ is the similarity between location i and location j, which can be calculated by the dot product of the query matrix $\mathbf{q}_i$ and the key matrix $\mathbf{k}_j$

$$\mathbf{e}_{i,j} = \mathbf{q}_i^T \mathbf{k}_j. \tag{2}$$

Finally, the weighted sum of the attentional weights $\alpha_{i,j}$ and the value matrix $\mathbf{v}_j$ can be used as a representation of position i

$$\mathbf{h}_i = \sum_{j=1}^{n} \alpha_{i,j} \mathbf{v}_j \tag{3}$$

For each position of the input sequence a query matrix $\mathbf{q}$, a key matrix $\mathbf{k}$, and a value matrix $\mathbf{v}$. These matrices are obtained by multiplying the embedding vector $\mathbf{X}$ of the input sequence by different weight matrices

$$\mathbf{q} = \mathbf{X} \cdot \mathbf{W}_q \tag{4}$$

$$\mathbf{k} = \mathbf{X} \cdot \mathbf{W}_k \tag{5}$$

$$\mathbf{v} = \mathbf{X} \cdot \mathbf{W}_v \tag{6}$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are the weight matrices used to compute Query, Key, and Value, respectively. These weight matrices are usually parameters learned through the learning process of the model.

The self-attention mechanism allows for the calculation of attention weights for each position based on the interdependencies among positions within the input sequence. Subsequently, the weighted vectors are utilized as representations for the respective positions. This mechanism establishes connections between different positions, enabling comprehensive modeling of all positions within the sequence.

B. Transformer-Based Super Resolution Model

Compared with CNN, Transformer networks can better capture the global dependency relationships in images and more flexibly handle the information at different positions, which is very important for improving the detail restoration in SR tasks. In recent years, many image SR methods based on Transformer networks have been proposed. To fully exploit the advantages of both, Liang et al. [24] proposed the SwinIR network, which combines the advantages of convolution and Transformer, and achieves efficient feature extraction and reconstruction by using a hierarchical window self-attention mechanism, while reducing a large number of parameters. Yang et al. [25] proposed the TTSR network, which transforms the LR image and the Ref image into query and key, and uses the attention mechanism to achieve cross-image feature alignment and texture transfer, thus solving the image registration and magnification problem. For the SR task of remote sensing images, many methods based on Transformer networks have also emerged. Shang et al.

[33] proposed a hybrid-scale hierarchical Transformer network, which makes full use of the self-similarity and high-dimensional features of different scales after the upsampling layer, and improves the accuracy and efficiency of the reconstruction. Lei et al. [34] proposed a Transformer-based multistage enhancement network, which combines with the traditional SR framework, successfully integrates the multiscale low-dimensional and high-dimensional features, and achieves excellent reconstruction performance. Cai et al. [35] proposed a texture transfer module for shallow texture transfer, which can effectively transfer the texture features of the reference image to the LR image, and proposed a feature fusion scheme to balance the reconstruction effects of texture and smoothness, thus improving the quality of the reconstructed image.

These studies jointly demonstrate the significant advantages and potential of Transformer network-based methods in image SR tasks, but using Transformer networks will incur huge computational overhead, and for their good performance, sufficient and high-quality training data are required. In remote sensing tasks, there are a large number of devices with limited computing power and storage space, such as drones and satellites. For this purpose, the proposed method aims to comprehensively utilize the excellent performance of Transformer networks and the lightweight characteristics of CNN, and design a lightweight but high-performance SR reconstruction model, to adapt to the efficient deployment requirements of resource-constrained remote sensing devices.

III. Proposed Method

In this section, the proposed method is elaborated in detail. First, a mathematical model for remote sensing image SR is established, representing the relationship between LR and HR images as a degradation matrix. Consequently, the SR problem is reformulated as the task of solving this degradation matrix. Subsequently, a cross-structured KD network is designed, employing a Transformer network as the teacher network and a CNN network as the student network. Through KD, the student network is enabled to learn high-level features and global information from the teacher network, enhancing the performance of the student network. Within the student network, subpixel convolution layers are employed as the upsampling module, mitigating the issues of blurring and distortion associated with traditional upsampling methods, such as bilinear interpolation or deconvolution, while simultaneously preserving fine details from the LR image. To further improve the guidance effectiveness of the teacher network, prior information is introduced to correct the output of the teacher network. This correction process eliminates erroneous knowledge within the teacher network, enabling the student network to acquire more accurate and reliable information from the teacher network.

A. Super-Resolution Modeling of Remote Sensing Image

The objective of the model designed for SR reconstruction in remote sensing images is to enhance image quality by restoring the corresponding HR image from its LR counterpart. In this context, the LR image is considered the degraded observation version of the HR image, and its mathematical representation is articulated as follows:

$$\mathbf{I}_{\mathrm{LR}} = \mathbf{K} \cdot \mathbf{I}_{\mathrm{HR}} + \boldsymbol{\eta} \tag{7}$$

where $\mathbf{I}_{\mathrm{LR}}$ represents the observed LR image, $\mathbf{I}_{\mathrm{HR}}$ denotes the input HR image, $\mathbf{K}$ signifies the degradation matrix, and $\boldsymbol{\eta}$ is typically determined by the resolution of the imaging system, representing additive noise. Generally, the LR image is obtained through bicubic downsampling operations.

To address the issue of information loss in remote sensing images caused by downsampling, we introduce a prior term within the CNN. This prior term, imposing constraints on the degradation matrix $\mathbf{K}$, aims to resolve the problem outlined in (7). Through such formulation constraints, the supervised network integrates global characteristics, thereby enhancing the CNN ability to learn the overall structure and global contextual information. In addition, this constraint enables the CNN to more effectively assimilate global knowledge from the teacher network, thereby improving performance in tasks, such as remote sensing image SR. This lays the groundwork for subsequent information distillation between the two networks. The process of addressing the degradation matrix $\mathbf{K}$ to reconstruct HR images can be formalized as the following optimization problem:

$$\mathbf{I}_{RC} = \arg\min\left(\mathbf{I}_{\mathrm{HR}}, \mathbf{I}_{\mathrm{RC}}\right) + N\left(\mathbf{I}_{\mathrm{HR}}\right) \tag{8}$$

where $\mathbf{I}_{\mathrm{RC}}$ represents the reconstructed HR image, $\arg\min(\mathbf{I}_{\mathrm{HR}}, \mathbf{I}_{\mathrm{RC}})$ denotes the operation of minimizing the difference between the two images. $N(\mathbf{I}_{\mathrm{HR}})$ serves as a weighted regularization term, imposing constraints on the solution space derived from prior knowledge. The introduction of an auxiliary variable $\mathbf{g}$ and the decoupling of the data term and regularization term within the equation result in the following expression:

$$\begin{aligned}(\mathbf{I}_{\mathrm{HR}}, \mathbf{g}) &= \arg\min \frac{1}{2}\left\|\mathbf{I}_{\mathrm{RC}} - \mathbf{K} \cdot \mathbf{I}_{\mathrm{HR}}\right\|_2^2 + \omega\varphi(\mathbf{g}) \\ &\text{s.t. } \mathbf{g} = \mathbf{I}_{\mathrm{HR}}.\end{aligned} \tag{9}$$

Further, the problem can be relaxed as an unconstrained problem and the above equation optimization problem can be transformed into two subproblems by the ADMM technique

$$\begin{aligned}(\mathbf{I}_{\mathrm{HR}}, \mathbf{g}) = \frac{1}{2}\left\|\mathbf{I}_{\mathrm{RC}} - \mathbf{K} \cdot \mathbf{I}_{\mathrm{HR}}\right\|_2^2 + \omega\varphi(\mathbf{g}) \\ + u\left\|\mathbf{g} - \mathbf{I}_{\mathrm{HR}}\right\|_2^2\end{aligned} \tag{10}$$

$$\mathbf{I}_{n+1}^{\mathrm{HR}}; \arg\min\left\|\mathbf{I}_{\mathrm{RC}} - \mathbf{K} \cdot \mathbf{I}_{\mathrm{HR}}\right\| + u\left\|\mathbf{g} - \mathbf{I}_{\mathrm{HR}}\right\|_2^2 \tag{11}$$

$$\mathbf{g}_{n+1} = \arg\min u\left\|\mathbf{g} - \mathbf{I}_{\mathrm{HR}}\right\|_2^2 + \omega\varphi(\mathbf{g}) \tag{12}$$

where $\mathbf{g}$ is the penalty parameter, solving the problem for $\mathbf{I}_{n+1}^{\mathrm{HR}}$ is a quadratic optimization problem that can be addressed using a closed-form solution, specifically by computing $\mathbf{I}_{n+1}^{\mathrm{HR}} = \mathbf{D}^{-1}\mathbf{I}_{\mathrm{LR}}$. Here, the matrix $\mathbf{D}$ and the degradation matrix $\mathbf{K}$ typically have high dimensions. Directly computing the inverse matrix involves significant computational complexity, making it less feasible. To overcome this issue, a Newton–Raphson optimization algorithm is proposed for calculating $\mathbf{I}_{n+1}^{\mathrm{HR}}$. The

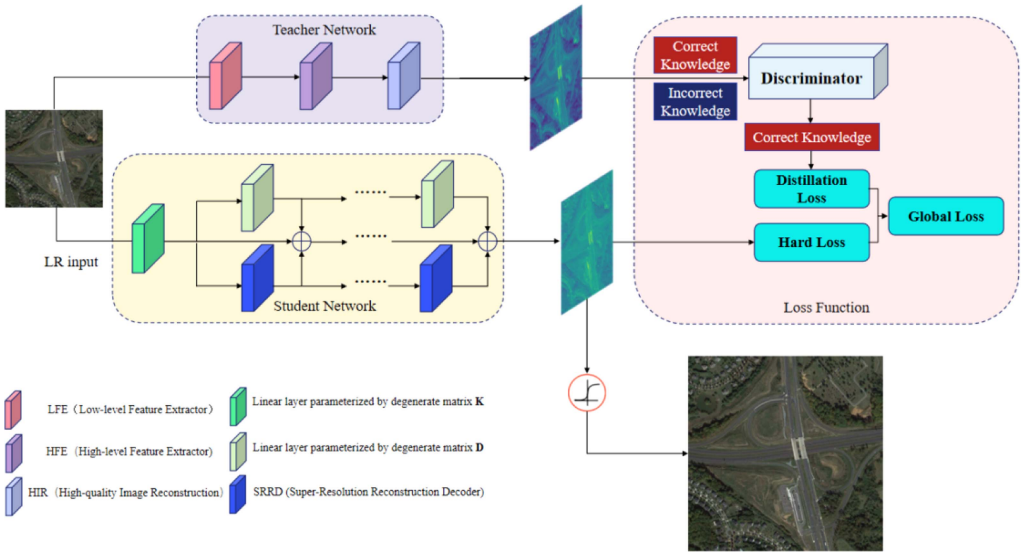


Fig. 1. General structure of the proposed distillation cross network distillation architecture.

equation can be reformulated as

$$\mathbf{I}_{n+1}^{\mathrm{HR}} = \mathbf{I}_{n}^{\mathrm{HR}} - \left[\frac{\mathbf{D}}{\mathbf{K}^T\mathbf{K} + u}\right] \tag{13}$$

$$\mathbf{D} = \mathbf{K}^T\left(\mathbf{K}\cdot\mathbf{I}_{n}^{\mathrm{HR}} - \mathbf{I}_{\mathrm{HR}}\right) + u\left(\mathbf{I}_{n}^{\mathrm{HR}} - \mathbf{g}\right) + \mathbf{P} \tag{14}$$

$$\mathbf{P} = r\sum_{\mathrm{i}=0}^{\mathrm{n}}\left(\mathbf{I}_{n+1}^{\mathrm{HR}} - \mathbf{I}_{m}^{\mathrm{HR}}\right) \tag{15}$$

where r is the scale factor, $\mathbf{P}$ is the integral term, and $\mathbf{g}$ is the proximity operator computed on $\varphi(\mathbf{g})$. After the transformation, we successfully implement the reconstructed HR image $\mathbf{I}_{n+1}^{\mathrm{HR}}$ by solving the degeneracy matrix $\mathbf{K}$.

B. Knowledge Distillation for DTCNet

To solve the degradation model described in the previous section, we propose a KD-based network which is named DTCNet in this section, as shown in Fig. 1. The DTCNet is divided into two main components: 1) the upper segment represents the teacher network based on the Transformer architecture, while 2) the lower segment corresponds to the student network based on CNN. The initial LR image undergoes processing through both networks, yielding two sets of logits results. Initially, we rectify potential errors in the teacher network by leveraging prior edge information. Subsequently, employing KD, we guide the student network to acquire more accurate knowledge representations. Through iterative backpropagation, we continuously optimize the student network, aiming to simplify model complexity while maintaining high performance. Finally, applying the sigmoid activation function to the logits results from the student network yields the ultimate HR image output.

The teacher network consists of three modules: LFE, HFE, and HIR. The upper part of Fig. 1 shows the details of the teacher network. The LFE module uses a 3×3 convolution layer to perform LFE on the input LR image. The HFE module contains multiple cascaded residual Swin Transformer blocks and convolution blocks, which split the feature map output by the LFE module into multiple nonoverlapping patch embeddings, and then perform deep feature extraction through layerwise residual Swin Transformer blocks. Then, the multiple nonoverlapping patch embeddings are recombined into a feature map. Finally, through residual connection, the shallow feature map and the deep feature map are fused and input to the HIR module to obtain a high-quality reconstructed image.

Due to the presence of attention mechanisms in the residual Swin Transformer blocks of the teacher network, students can focus on specific local structures of the image during the training phase. In addition, given the distinct emphases of Transformer and CNN, employing a conventional CNN as the student network may encounter challenges in fully assimilating knowledge. Therefore, modifications were made to the attention mechanisms and certain convolutional layers of the student network, and an SRRD along with a linear layer parameterized by matrix $\mathbf{D}$ were designed.

The specific details of the student network are illustrated in the lower part of Fig. 1. The student network consists of two parts: 1) feature extraction and 2) image reconstruction, specifically feature encoding and average pooling, feature decoding, and upsampling, as shown in Fig. 2. First, the degraded image $\mathbf{I}_{\mathrm{LR}}$ passes through three feature encoding layers sequentially. Each feature encoding layer includes multiple convolutional layers and ReLU functions. After each round of feature encoding and downsampling, the extracted feature maps undergo average pooling to retain essential features and alleviate overfitting. Similarly, the feature maps obtained through feature extraction serve as inputs for image reconstruction. After three rounds of upsampling and feature decoding, the final HR image is obtained through hierarchical residual connections.

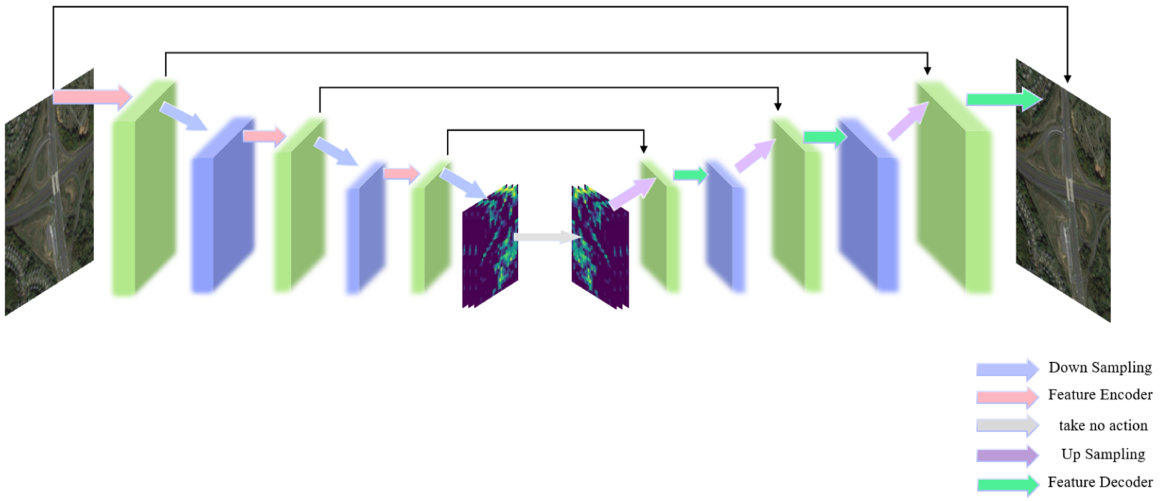


Fig. 2. SR reconstruction decoder.

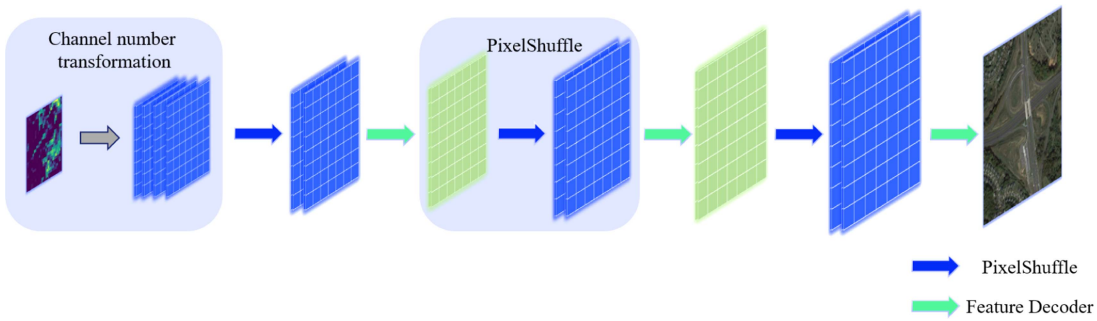


Fig. 3. Feature map upsampling reconstruction flowchart.

C. Subpixel Convolution for Image Super-Resolution

To enhance image resolution, the upsampling process typically employs deconvolution methods, enlarging the image size by padding the blank pixel regions with zeros. However, the zero-padded regions do not contain meaningful information, and the use of conventional deconvolution methods for upsampling can impede gradient optimization. In addition, since the image reconstruction segment of the teacher network utilizes subpixel convolution methods, the application of deconvolution methods may interfere with the alignment between the student and teacher networks. Therefore, we have improved the upsampling segment of the student network by introducing subpixel convolution methods.

The core concept of subpixel convolution involves taking a LR image of size $H \times W \times C$ as input. After extracting features from the LR image, n feature maps are generated. In the upsampling stage, the last convolutional layer adjusts the channel count to $C \times r^2$, and through subpixel operation, the values at corresponding positions of $r \times r$ feature maps are used to fill and reorganize subpixels, transforming it into a HR image of size $r \cdot H \times r \cdot W \times C$.

Fig. 3 illustrates the image reconstruction process. Initially, the feature maps undergo upsampling through the PixelShuffle layer, breaking down the features of each channel into the required multiple of the sampling factor, yielding the preliminary upsampling result. Subsequently, the upsampling result is added to the corresponding downsampled feature map, serving as the output of the decoding layer. This upsampling and decoding process is repeated until the target resolution is reached. Finally, through residual connections, the output of the last layer is added to the original LR image, yielding the reconstructed HR image.

D. Teacher Knowledge Optimization via Loss Functions

Typically, in a large number of studies, traditional knowledge extraction techniques are formulated with the common goal of minimizing the objective function

$$\mathcal{L} = l\{f_T(\mathbf{x}), f_S(\mathbf{x})\} \tag{16}$$

where l is a loss function used to penalize the differences between the teacher and student networks, $f_T(\mathbf{x})$ represents the logits output of the teacher network, and $f_S(\mathbf{x})$ represents the logits output of the student network.

Through extensive research on KD, various KD methods tailored for different tasks have been identified, such as the approach proposed by Chen et al. [36] for image retrieval and Re-ID. Therefore, for KD based on image reconstruction tasks, a suitable loss function is needed to ensure optimal network

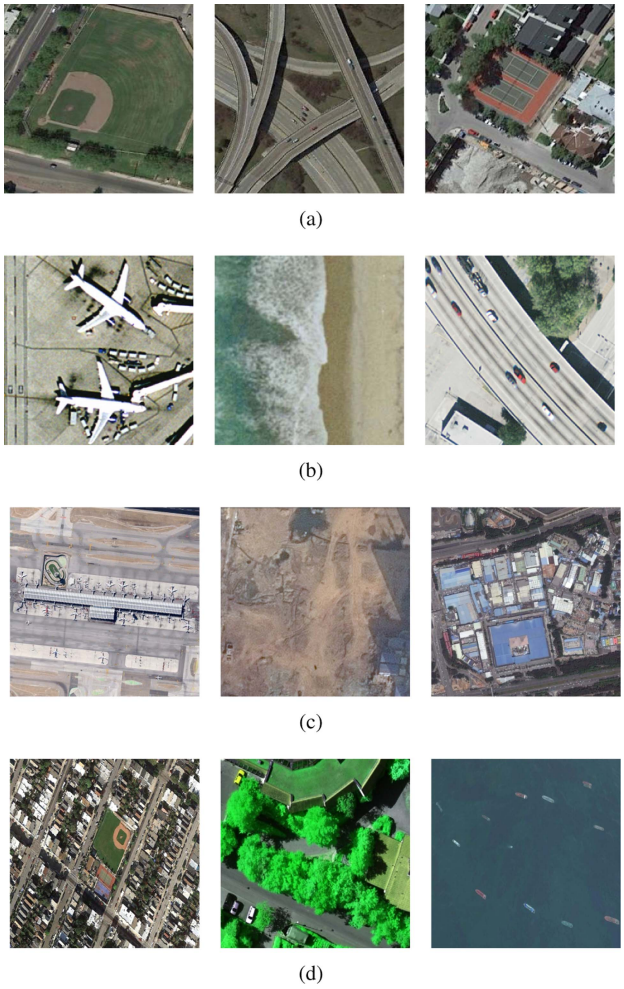


Fig. 4. Selection of images from (a) NWPU-RESISC45, (b) UCMerced, (c) AID, and (d) NWPU VHR-10.

performance. In this study, we utilize softmax (with temperature τ) and the Kullback–Leibler divergence for l, aiming to extract logit features from the teacher and student networks during the training phase. This compels the student network to closely emulate the teacher network at corresponding feature layers, ultimately achieving the goal of KD. Specifically, the distillation loss is computed as follows:

$$\mathcal{L}_{\varsigma_R} = \mathrm{KL}(\mathrm{softmax}(f_T(\mathbf{x})/\tau), \mathrm{softmax}(f_S(\mathbf{x})/\tau)) \quad (17)$$

where KL represents the Kullback–Leibler divergence loss, and τ denotes the distillation temperature.

Due to the complexity of remote sensing images, images reconstructed through SR are prone to containing errors in the edges. Therefore, further optimization of the loss function $\mathcal{L}_{SR}$ is necessary to prevent erroneous edges in the HR images. In this context, we introduce an edge correction function to constrain $\mathcal{L}_{SR}$, with the specific formula given as follows:

$$\mathcal{L}_{SR} = \mathrm{KL}\left(\mathrm{softmax}\left(\frac{f_T(\mathbf{x})}{\tau}\right)\mathrm{softmax}\left(\frac{f_S(\mathbf{x})}{\tau}\right)\right) - \mathrm{KL}\left(\mathrm{ECF}(\mathrm{softmax}(f_T(\mathbf{x}))), \mathrm{ECF}(\mathbf{y})\right) \quad (18)$$

where y represents the ground truth, and ECF denotes the edge correction function.

Remote sensing images often involve complex terrain structures, making the accurate restoration of edge information crucial for achieving valuable objectives in land classification, environmental monitoring, and related scientific research [37]. To effectively drive the learning and optimization of the remote sensing image SR reconstruction model and better preserve key edge features in the image, the final loss function is defined as follows:

$$\mathcal{L}_{\text{hard}} = \mathrm{MSE}(\mathrm{softmax}(f_S(\mathbf{x})), \mathbf{y}) + \mathrm{MSE}(\mathrm{ECF}(\mathrm{softmax}(f_T(\mathbf{x}))), \mathrm{ECF}(\mathbf{y})) \quad (19)$$

$$\mathcal{L}_{\text{global}} = \alpha \cdot \mathcal{L}_{\text{hard}} + (1-\alpha) \cdot \mathcal{L}_{sR} \quad (20)$$

which denotes the proportion of weight given to distillation loss and standard loss; in this article $\alpha = 0.3$.

IV. Experimental Results and Discussion

A. Experimental Data and Settings

In this section, experiments were undertaken to verify the efficacy and versatility of the proposed model using four publicly available remote sensing image datasets: 1) NWPU-RESISC45 [38], 2) UCMerced [39], 3) AID [40], and 4) NWPU VHR-10 [41]. These datasets encompass diverse scenes and types of remote sensing images, showcasing notable complexity and variability. Fig. 4 illustrates some of the original and unique images selected from these datasets.

The NWPU-RESISC45 dataset, with 45 categories of remote sensing images, comprises 700 images per category. The images are sized at 256×256 pixels, with pixel resolutions ranging from 30–0.2 m. For training, a random selection of 20 images was made from each category, resulting in a total of 900 images.

The UCMerced dataset encompasses 21 categories of scene images, with each category comprising 100 images, totaling 2100 images. The pixel dimensions of the images are 256×256, and the spatial resolution is 0.3 m. For the test set in this section, 50% of the images were randomly chosen from this dataset, amounting to a total of 1050 images.

The AID dataset comprises 30 categories of scene images, totaling 10 000 images, each sized at 600×600 pixels with a spatial resolution of 0.5 m. Specifically for this section, 2000 images were selected from the dataset for testing purposes.

The NWPU VHR-10 contains 800 HR color images with resolutions in the range of 0.08–2 m, covering ten different scene types, each containing multiple target objects. To evaluate the performance of the method proposed in this article, 30 images were randomly selected from this dataset as a test set.

To obtain LR data, bicubic interpolation was applied to downsample the original images, and SR reconstruction evaluations were conducted at three different upscaling factors (×2, ×3, and ×4). During training, the Adam optimizer was utilized, with β_1 and β_2 set to 0.9 and 0.999, respectively. The initial learning rate was set to 0.0005, and it was halved after every 20 000 mini-batch updates. The training and testing were performed in

an environment using PyTorch 2.1, CUDA 12.2, cuDNN 8.3.0, and the NVIDIA Tesla A800.

B. Comparison With State-of-the-Art Methods and Evaluation Indicators

In this article, five state-of-the-art methods have been selected for comparative tests as well as double cubic interpolation and six learning-based methods are shown as follows.

1) Remote sensing single-image SR based on a deep compendium model (DCM) [42].
2) Transformer for single image SR (ESRT) [26].
3) Enriched CNN-Transformer feature aggregation networks for SR (ACT) [20].
4) Transformer-based multistage enhancement for remote sensing image SR (TransENet) [34].
5) Hybrid-scale hierarchical transformer for remote sensing image SR (HSTNet) [33].

The DCM method is based on a CNN model, while ESRT, ACT, TransEnet, and HSTNet are methods based on Transformer networks. The selection of these two different types of models aims to comprehensively compare their performance in image SR tasks, analyze their strengths and weaknesses, and identify suitable scenarios for each. Two types of image SR experiments were conducted in this section. One set of experiments utilized LR images simulated through bicubic downsampling, while the other set used real remote sensing images. For the evaluation of results in both experiments, reference PSNR, SSIM [43], and SAM [44] were employed as assessment metrics for simulated images. In addition, no-reference metrics, such as NIQE, Ent, Std, and Grad [45] were used to test the images reconstructed from real remote sensing data. Through these two categories of experiments, a comprehensive assessment of the algorithmic performance in different scenarios could be achieved. The following will elucidate the meanings of various evaluation metrics and demonstrate how these metrics effectively reflect the performance of images after SR reconstruction processing.

1) *PSNR:* The PSNR is one of the commonly used metrics for assessing image reconstruction quality. It evaluates the degree of image distortion by comparing the mean square error between the pixel values of the original image and the reconstructed image. A higher PSNR value indicates less difference between the reconstructed image and the original image, thus indicating higher reconstruction quality. The calculation is as follows:

$$\text{PSNR} = 10\log_{10}\left(\frac{255^2}{\text{MSE}}\right) \tag{21}$$

where MSE is the mean square error which can be calculated as follows:

$$\text{MSE} = \frac{1}{H \times W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}[\mathbf{I}(i,j) - \mathbf{R}(i,j)]^2 \tag{22}$$

where $H \times W$ denotes the size of the image, $\mathbf{I}$ and $\mathbf{R}$ denote the original and reconstructed image.

2) *SSIM:* The SSIM is a perceptual model used to evaluate the structural similarity between images. It considers three aspects: 1) luminance, 2) contrast, and 3) structure, and integrates their similarities into a single score. The SSIM values range from -1–1, with a value closer to 1 indicating a higher structural similarity between the reconstructed image and the original image, and therefore closer to human visual perception. Consequently, higher SSIM values correspond to better reconstruction quality. The calculation formula is as follows:

$$\text{SSIM} = \frac{(2\mu_{\mathbf{I}}\mu_{\mathbf{R}} + c_1)(2\sigma_{\mathbf{IR}} + c_2)}{(\mu_{\mathbf{I}}^2 + \mu_{\mathbf{R}}^2 + c_1)(\sigma_{\mathbf{I}}^2 + \sigma_{\mathbf{R}}^2 + c_2)} \tag{23}$$

where $\mu_{\mathbf{I}}$ and $\mu_{\mathbf{R}}$ denote the mean of the pixel intensities of the original and reconstructed images, $\sigma_{\mathbf{I}}$ and $\sigma_{\mathbf{R}}$ denote the std of the pixel intensities of the original and reconstructed images, and $\sigma_{\mathbf{IR}}$ denotes the covariance of the pixel intensities of the original and reconstructed images. The constants c_1 and c_2 are used to maintain stability.

3) *SAM:* The SAM is employed to measure the spectral differences between pixels in images. It assesses the disparity between the spectral accuracy of the reconstructed image and the original image. Lower SAM values indicate a closer spectral resemblance between the reconstructed image and the original image, thus indicating higher reconstruction quality. The calculation formula is as follows:

$$\text{SAM} = \cos^{-1}\left(\frac{\sum_{i=1}^{n}(\chi_i \cdot \hat{\chi}_i)}{\sqrt{\sum_{i=1}^{n}\chi_i^2}\cdot\sqrt{\sum_{i=1}^{n}\hat{\chi}_i^2}}\right) \tag{24}$$

where χ_i and $\hat{\chi}_i$ represent the spectral vectors of the corresponding pixels in the two images, respectively, and n is the dimension of the spectral vector.

4) *NIQE:* The NIQE is a metric used for assessing the quality of natural images. It performs quality feature extraction through an NSS model and models the features using a MVG model, representing the evaluated image quality as the distance between the two MVG distributions. Lower NIQE scores indicate higher image quality. The calculation formula is as follows:

$$\text{NIQE} = \sqrt{\left[(v_1 - v_2)^T\left(\frac{\sum_1 - \sum_2}{}\right)^{-1}(v_1 - v_2)\right]} \tag{25}$$

where v_1 and v_2 represent the MVG model for natural and distorted images, respectively, and $\sum_1$ and $\sum_2$ denote the variance matrix for natural and distorted images, respectively.

5) *Ent:* The Ent of an image measures its complexity or uncertainty. In image reconstruction, a lower Ent value may indicate that the image has more structure and predictability. The calculation formula is as follows:

$$\text{Ent} = -\sum_{i=0}^{255} p_i \log p_i \tag{26}$$

where i denotes the grey value of the pixel and p_i denotes the probability of occurrence of a pixel point with grey value i.

TABLE I
PSNR AND SSIM RESULTS FOR ×2, ×3, AND ×4 UPSAMPLING RATIOS ON THE UCMERCED DATASET

Methods	Scale	PSNR	SSIM	SAM
Bicubic	×2	30.76	0.8789	4.05
DCM	×2	33.65	0.9274	2.92
ESRT	×2	33.70	0.9270	2.84
ACT	×2	33.88	0.9283	2.68
TransENet	×2	34.03	0.9301	2.66
HSTNet	×2	34.19	0.9338	-
Proposed	×2	**35.43**	**0.9443**	**2.57**
Bicubic	×3	27.46	0.7631	5.88
DCM	×3	29.52	0.8394	4.71
ESRT	×3	29.52	0.8318	4.91
ACT	×3	29.80	0.8395	4.17
TransENet	×3	29.92	0.8408	4.31
HSTNet	×3	30.07	0.8421	-
Proposed	×3	**30.75**	**0.8602**	**4.04**
Bicubic	×4	25.65	0.6725	7.90
DCM	×4	27.22	0.7528	6.03
ESRT	×4	27.41	0.7485	6.08
ACT	×4	27.54	0.7531	5.31
TransENet	×4	27.77	0.7630	5.75
HSTNet	×4	27.89	0.7694	-
Proposed	×4	**28.73**	**0.7904**	**5.13**

With optimal results in bold and suboptimal results highlighted in red.

TABLE II
PSNR AND SSIM RESULTS FOR ×2, ×3, AND ×4 UPSAMPLING RATIOS ON THE AID DATASET

Method	Scale	PSNR	SSIM	SAM
Bicubic	×2	32.39	0.8906	5.39
DCM	×2	35.21	0.9366	3.10
ESRT	×2	35.15	0.9358	3.53
ACT	×2	35.17	0.9362	3.12
TransENet	×2	35.28	0.9374	3.04
HSTNet	×2	35.35	0.9387	-
Proposed	×2	**36.11**	**0.9436**	**2.40**
Bicubic	×3	29.08	0.7863	6.13
DCM	×3	31.31	0.8561	4.97
ESRT	×3	31.34	0.8562	4.86
ACT	×3	31.39	0.8579	4.90
TransENet	×3	31.45	0.8595	4.64
HSTNet	×3	31.61	0.8613	-
Proposed	×3	**32.22**	**0.8745**	**4.53**
Bicubic	×4	27.30	0.7036	9.57
DCM	×4	29.17	0.7824	7.19
ESRT	×4	29.18	0.7831	7.06
ACT	×4	29.19	0.7836	7.04
TransENet	×4	29.38	0.7909	6.79
HSTNet	×4	29.57	0.7983	-
Proposed	×4	**30.13**	**0.8082**	**6.17**

With optimal results in bold and suboptimal results highlighted in red.

6) *Std:* The std measures the degree of variation in pixel values within an image. A higher std typically indicates that the image has greater contrast or more details. The calculation formula is as follows:

$$\text{Std} = \sqrt{\frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} [\mathbf{R}(i,j) - \mu]} \quad (27)$$

where μ is the mean value of pixel intensity in the image.

7) *Grad:* The mean gradient represents the average rate of change of pixel values within an image. A higher mean gradient typically indicates that the image has more edges and textures. The calculation formula is as follows:

$$\text{Grad} = \frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \sqrt{\frac{\partial_h^2 + \partial_v^2}{2}} \quad (28)$$

where ∂_h and ∂_v are the gradient of the image in horizontal and vertical directions, respectively. In this paper, for all tabulated data, the indicator parameters for the optimal method have been highlighted using bold font, while for the sub-optimal results, red font has been used for labelling.

C. Super-Resolution Reconstruction Results on Different Datasets

1) UCMerced Dataset Test Result: Table I presents the quantitative evaluation results of remote sensing SR methods on the UCMerced dataset. From the data analysis, a significant difference is observed between traditional interpolation methods and current learning-based techniques. It is noteworthy that the proposed remote sensing SR method achieved optimal PSNR, SSIM, and SAM results at three different upscaling factors compared to other methods. Specifically, compared to the second-best method, the proposed method exhibited an average PSNR improvement of 0.92 dB and an average SSIM improvement of 0.0165 across the three upscaling factors, both of which are statistically significant. SAM is employed for assessing the spectral similarity between two images, where a smaller SAM value indicates a higher spectral similarity between the images. The data results presented in the table demonstrate that the images reconstructed through the proposed algorithm exhibit superior spectral quality.

This indicates the successful restoration of image details and textures by the proposed method, further enhancing visual quality and perceptual effects. In comparison to existing SR methods, the proposed approach demonstrates more robustness and generalization capability, suitable for various image scenes and content.

For further analysis of the proposed method's SR performance in different scenarios, Table III provides the average PSNR values for 21 scene categories in the UCMerced dataset. The table reveals that the proposed method achieved the highest PSNR values in 18 out of 21 scene categories, indicating its adaptability to diverse image content and features, resulting in high-quality image restoration. In the remaining four scene categories, the proposed method ranked just below the DCM method, with the PSNR difference between them not exceeding 1 dB, demonstrating the continued competitiveness of the proposed method.

Fig. 5 illustrates the reconstruction results of the proposed method and six other SR methods on the UCMerced dataset. It is evident from the two images that the proposed method

TABLE III
AVERAGE PSNR FOR EACH CATEGORY OF THE UCMERCED DATASET WITH A SCALING FACTOR OF ×3

Class	Bicubic	DCM	ESRT	ACT	TransENet	HSTNet	Proposed
Agricultural	26.86	**29.06**	28.13	27.86	28.02	27.93	28.47
Airplane	26.71	30.77	29.45	29.78	29.94	29.98	**31.45**
Baseballdiamond	33.33	33.76	34.88	35.05	35.04	35.13	**36.38**
Beach	36.14	36.38	37.45	37.55	37.53	37.76	**38.86**
Buildings	25.09	28.51	28.18	28.66	28.81	29.12	**29.41**
Chaparral	25.21	26.81	26.43	26.62	26.69	26.78	**27.57**
Denseresidential	25.76	28.79	28.53	28.97	29.11	29.27	**29.72**
Forest	27.53	28.16	28.47	28.56	28.59	28.65	**29.64**
Freeway	27.36	30.45	29.87	30.25	30.38	30.65	**31.24**
Golfcourse	35.21	34.43	36.54	36.63	36.68	36.69	**38.30**
Harbor	21.25	**26.55**	23.87	24.42	24.72	24.91	25.45
Intersection	26.48	29.28	28.53	28.85	29.03	29.32	**29.74**
Mediumresidential	25.68	27.21	27.93	28.30	28.47	28.64	**29.26**
Mobilehomepark	22.25	**26.05**	24.92	25.32	25.64	25.74	26.03
Overpass	24.59	27.77	27.17	27.76	27.83	28.31	**28.86**
Parkinglot	21.75	**24.95**	23.72	24.11	24.45	24.53	24.93
River	28.12	28.89	29.14	29.28	29.25	29.32	**30.39**
Runway	29.30	32.53	30.98	31.21	31.25	31.21	**33.08**
Sparseresidential	28.34	29.81	31.35	31.55	31.57	31.71	**32.68**
Storagetanks	29.97	29.02	32.42	32.74	32.71	32.98	**34.02**
Tenniscourt	29.75	30.76	31.99	32.40	32.51	32.77	**33.09**

The best and second best results are marked in bold and red, respectively.

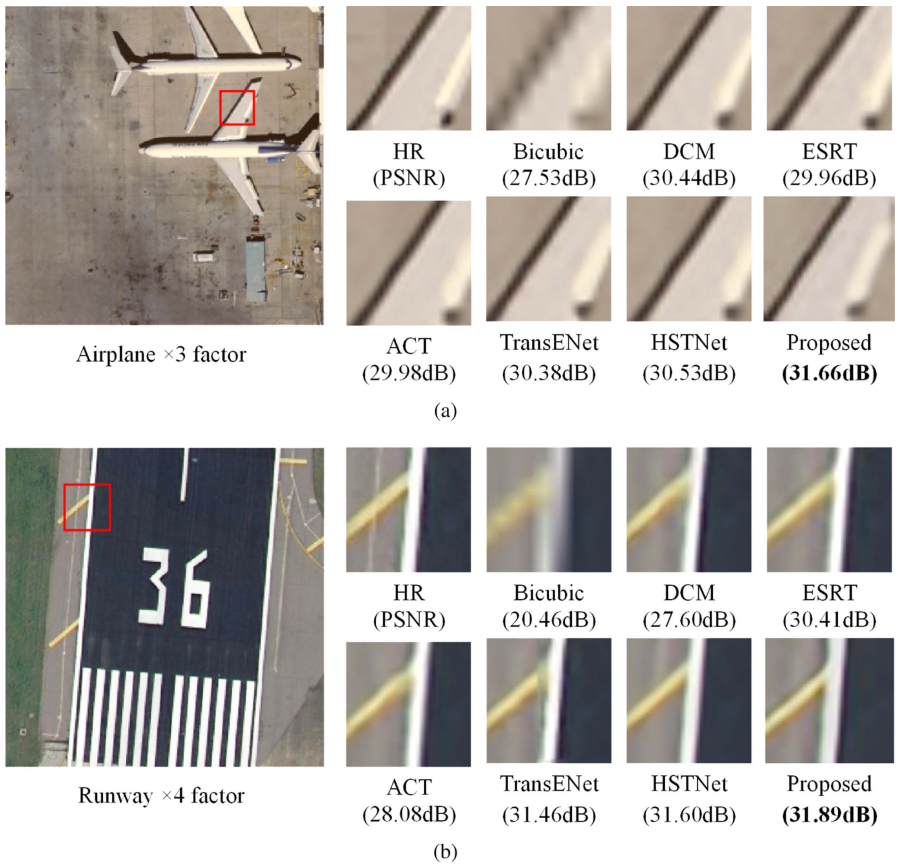


Fig. 5. Visualization results of (a) airplane image with ×3 factor, (b) runway image with ×4 factor.

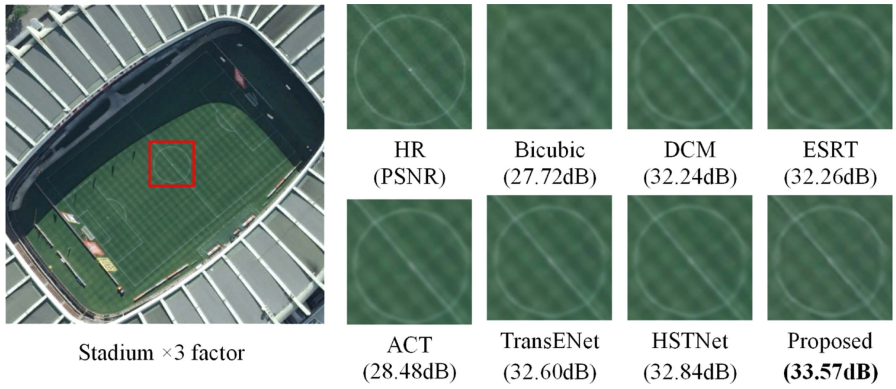


Fig. 6. Visualization results of stadium image with ×3 factor.

TABLE IV
MEAN PSNR, SSIM AND SAM METRICS OVER THE NWPU VHR-10 TEST DATASET

Scale	Bicubic PSNR/SSIM/SAM	DCM PSNR/SSIM/SAM	ESRT PSNR/SSIM/SAM	ACT PSNR/SSIM/SAM	TransENet PSNR/SSIM/SAM	Proposed PSNR/SSIM/SAM
2	33.51/0.9212/4.70	34.50/0.9382/3.62	35.24/0.9451/3.38	35.64/0.9488/3.22	35.01/0.9261/4.37	**35.99/0.9522/2.81**
3	30.54/0.8481/6.17	31.07/0.8686/5.40	31.01/0.8686/5.41	31.88/0.8855/4.95	30.03/0.8417/6.08	**32.47/0.8974/4.18**
4	28.96/0.7844/7.20	29.13/0.8034/6.77	29.36/0.8093/6.63	29.89/0.8275/6.23	29.67/0.8128/6.58	**30.52/0.8402/5.23**

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.

can more accurately restore the details and textures of the images, making them closely resemble the original HR images. In contrast, other methods, especially traditional interpolation methods, result in significantly blurred reconstructed images. Among the other methods, there are noticeable blurry transitions at the wing edges and ground edges, as well as edge blurring issues in the runway images, particularly in Bicubic, DCM, ACT, and TransEnet methods. In comparison, the images reconstructed using the proposed method exhibit clear and sharp edges, indicating a significant advantage in enhancing the visual quality and perceptual effects of the images. Whether from qualitative or quantitative results, the proposed algorithm demonstrates outstanding remote sensing image SR reconstruction capabilities.

Both qualitatively and quantitatively, the proposed algorithm exhibits excellent remote sensing image SR reconstruction capabilities.

2) AID Dataset Test Results: To validate the robustness of the proposed method, experiments were conducted not only on the UCMerced dataset for comparative analysis but also on the AID dataset for testing. The results are presented in Table II and Fig. 6. Quantitatively, the proposed method achieved optimal PSNR, SSIM, and SAM results across all three upscaling factors in the AID dataset, highlighting its competitiveness across different datasets. In terms of visual comparisons, the HR images reconstructed by ESRT, ACT, and HSTNet resulted in a blurred central white spot in the soccer field. Conversely, the images reconstructed using the proposed method distinctly exhibit recognizable lawn textures and prominent white circular patterns. These findings provide compelling evidence of the efficacy and resilience of the proposed method, suggesting its applicability to diverse SR reconstruction tasks in various remote sensing scenarios.

3) NWPU VHR-10 Dataset Test Results: Compared to the UCMerced and AID datasets, the NWPU VHR-10 dataset possesses higher spatial resolution and richer image semantic information. Consequently, it can capture finer image details, and the reconstruction testing on this dataset provides a better evaluation of the various methods' performance in reconstructing target information in remote sensing images. The quantitative and qualitative results are presented in Table IV and Fig. 7, respectively.

The results in Table IV indicate that the proposed method continues to exhibit optimal reconstruction performance on higher resolution datasets. Specifically, the proposed method achieves higher PSNR, SSIM, and SAM values by 0.63 dB, 0.0127, and 1.00, respectively, compared to the second-best method. In contrast, TransENet, while demonstrating good performance on the other two datasets, exhibits a noticeable decline on the HR dataset. From the Fig. 7, it can be seen that the reconstructed image by the proposed algorithm is very close to the original HR image in terms of detail, clarity, and color, while the reconstructed image by other methods suffers from a certain degree of blurring, distortion, and off-color. All these results indicate that the proposed algorithm is able to effectively preserve and recover the information of the HR image and improve the reconstruction quality.

D. Real Remote Sensing Image Super-Resolution Test Results

In the previous section, the reconstruction experiments of remote sensing images were conducted by simulating LR images

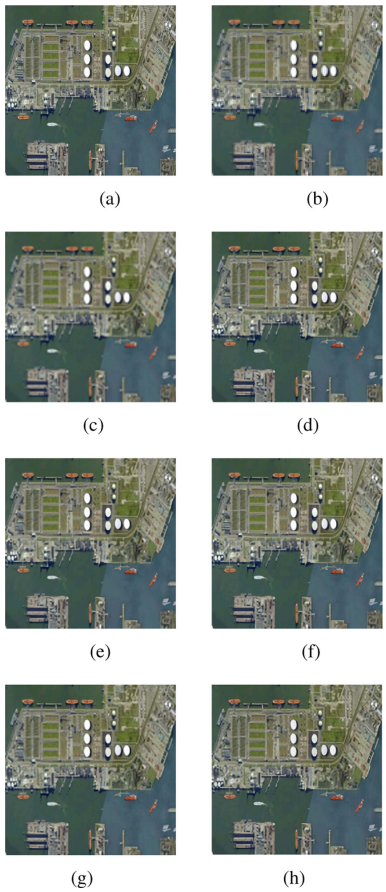


Fig. 7. Results comparisons of ×4 reconstruction on the NWPU VHR-10 dataset for different methods. (a) HR. (b) LR. (c) Bicubic. (d) DCM. (e) ESRT. (f) ACT. (g) TransENet. (h) Proposed.

through bicubic downsampling, and the performance of various algorithms in the reconstruction of simulated LR images was compared. To comprehensively evaluate the performance of these algorithms in real-world applications, this section employed authentic remote sensing images with resolutions of 8 and 3.2 m/pixel, obtained from the Gaofen-1 and Gaofen-2 satellites. These real remote sensing images were used to test the reconstruction performance of each algorithm, and no-reference metrics were employed to objectively evaluate the performance of each algorithm.

Figs. 8–11 display images of scenes, such as rivers, factories, elevated roads, and paddy fields captured by the Gaofen satellites. By zooming in on the corresponding regions, the performance of each algorithm in SR reconstruction is compared more clearly. The data captured by the Gaofen satellites consist of 600×600-sized images, providing richer information compared to the 256×256-sized images used for training. It is observed from the four images that the proposed method exhibits a significant advantage in handling complex real-world scenarios, effectively utilizing available information to generate clearer and more accurate HR images.

In Fig. 9, the proposed method accurately reconstructs the texture information of the factory roof. In Fig. 10, compared to

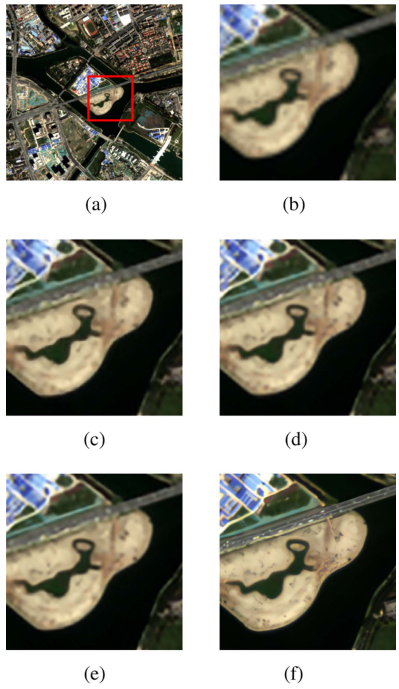


Fig. 8. River image acquired by Gaofen-1 satellite and its ×3 upsampling factor result figure. (a) Input. (b) Bicubic. (c) DCM. (d) ACT. (e) HSTNet. (f) Proposed.

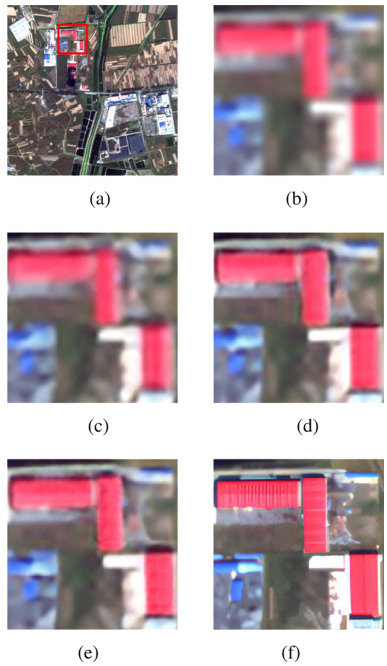


Fig. 9. Factory image acquired by Gaofen-1 satellite and its ×4 upsampling factor result figure. (a) Input. (b) Bicubic. (c) DCM. (d) ACT. (e) HSTNet. (f) Proposed.

other methods showing blurry textures on the highway and vehicles, the proposed method reconstructs these important details more accurately. This is crucial for subsequent tasks, such as image segmentation and object recognition.

In this section, four no-reference evaluation metrics were employed to assess the quality of the reconstructed images:

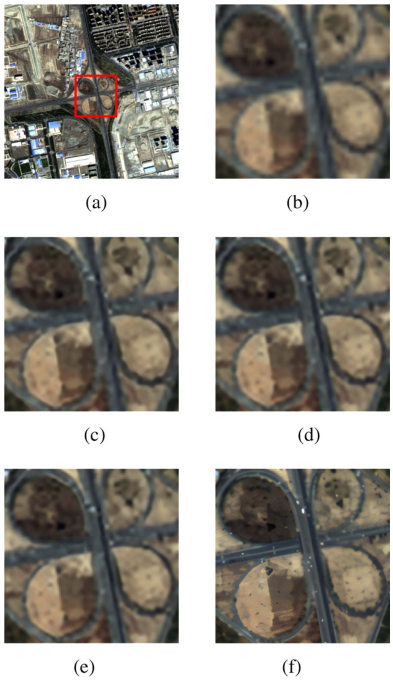


Fig. 10. Overpass image acquired by Gaofen-2 satellite and its ×3 upsampling factor result figure. (a) Input. (b) Bicubic. (c) DCM. (d) ACT. (e) HSTNet. (f) Proposed.

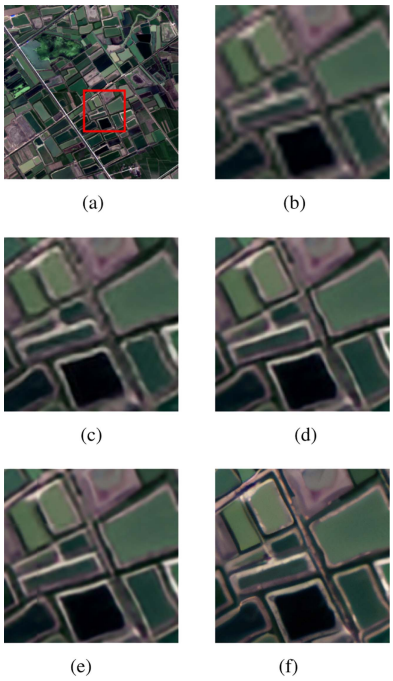


Fig. 11. Paddy fields image acquired by Gaofen-2 satellite and its ×4 upsampling factor result figure. (a) Input. (b) Bicubic. (c) DCM. (d) ACT. (e) HSTNet. (f) Proposed.

1) NIQE, 2) Ent, 3) Std, and 4) Grad. NIQE gauges the overall naturalness of the image, with lower values indicative of images resembling natural scenes more closely. Ent measures the information Ent of the image, with higher values suggesting a greater diversity of information in the image. Std quantifies the std of the image, with higher values pointing to a broader distribution of pixel values and increased image contrast. Grad assesses the average gradient of the image, with higher values indicating more prominent edges, details, and enhanced image clarity. Tables V–VIII present the numerical values of these four metrics for the reconstructed images. It can be observed from the tables that the proposed method obtained optimal results for all images and upscaling factors. In addition, it is found that in images captured by the Gaofen-2 satellite, the proposed method averaged 13.47% better than the second-best method across all metrics.

TABLE V
NONREFERENCE METRICS RESULTS FOR RIVER IMAGE SUPER-RESOLUTION

Method	NIQE(↓)	Ent(↑)	Std(↑)	Grad(↑)
Bicubic	6.58	7.11	69.63	9.73
DCM	6.30	7.16	71.81	16.79
ACT	6.18	7.17	72.14	16.18
HSTNet	5.58	7.18	72.60	23.19
Proposed	**5.05**	**7.24**	**73.24**	**31.54**

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.

TABLE VI
NONREFERENCE METRICS RESULTS FOR FACTORY IMAGE SUPER-RESOLUTION

Method	NIQE(↓)	Ent(↑)	Std(↑)	Grad(↑)
Bicubic	7.91	7.14	37.97	9.52
DCM	7.79	7.20	40.60	11.26
ACT	6.65	7.29	44.22	13.40
HSTNet	5.90	7.32	48.63	22.00
Proposed	**5.32**	**7.44**	**54.85**	**26.56**

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.

TABLE VII
NONREFERENCE METRICS RESULTS FOR OVERPASS IMAGE SUPER-RESOLUTION

Method	NIQE(↓)	Ent(↑)	Std(↑)	Grad(↑)
Bicubic	9.33	7.01	33.93	10.25
DCM	7.67	7.06	35.03	12.11
ACT	7.37	7.07	35.33	12.05
HSTNet	7.23	7.07	35.47	13.26
Proposed	**6.31**	**7.84**	**39.26**	**18.49**

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.

TABLE VIII
NONREFERENCE METRICS RESULTS FOR PADDY FIELDS IMAGE SUPER-RESOLUTION

Method	NIQE(↓)	Ent(↑)	Std(↑)	Grad(↑)
Bicubic	8.67	6.82	31.53	9.02
DCM	8.49	7.03	35.83	13.70
ACT	7.80	7.06	37.06	15.54
HSTNet	7.73	7.07	37.10	15.03
Proposed	**6.86**	**7.20**	**38.92**	**21.32**

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.

E. Comparison of Different Model Sizes

Existing SR reconstruction methods often require the use of deep neural networks to improve reconstruction quality, leading

TABLE IX
RESULTS OF THE ABLATION EXPERIMENTS FOR THE UCMERCED DATASET WITH SCALE FACTOR ×4

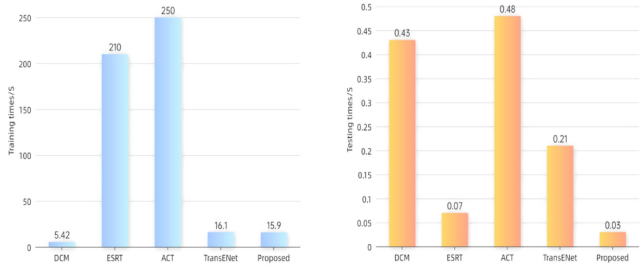
Model	Params/M	FLOPs/G	PSNR/dB	SSIM
Proposed without teachers' network	1.73	9.07	27.51	0.7553
Proposed without subpixel	1.28	8.34	28.58	0.7870
Proposed without priori information	1.73	9.07	28.66	0.7874
Proposed	1.73	9.07	28.73	0.7904

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.

TABLE X
COMPLEXITY COMPARISON OF VARIOUS MODELS AT ×4 UPSAMPLING FACTOR

Method	Params/M	FLOPs/G
Bicubic	-	-
DCM	2.18	13.00
ESRT	**0.75**	**2.35**
ACT	46.00	22.02
TransENet	37.38	12.04
HSTNet	43.40	194.40
Proposed	1.73	9.07

The bold values represent the optimal results for each indicator, while red represents the suboptimal results.



(a) (b)

Fig. 12. Average training time and test time comparison for each model. (a) Training time comparison. (b) Testing time comparison.

to significant computational expenses and performance overhead. This is deemed unacceptable for remote sensing tasks with limited computational resources. Therefore, a lightweight SR reconstruction model is proposed, aiming to reduce model complexity and resource consumption while ensuring reconstruction quality. Table X presents a comparison of resource consumption between the proposed model and several state-of-the-art SR reconstruction models. From the table, it can be observed that the proposed model's parameter count and floating-point operations are second only to the ESRT model. However, in the results from the previous section, images reconstructed using the proposed method surpassed all other models in terms of PSNR and SSIM. Taking the ×4 upsampling factor on the AID dataset as an example, the proposed model achieved a PSNR of 28.73 dB and SSIM of 0.7904, outperforming the ESRT model by 1.32 dB and 0.0419, respectively. On the other hand, HSTNet, which has comparable performance to the proposed method, incurs computational costs tens of times higher. This indicates that the proposed model strikes a favorable balance between performance and efficiency, highlighting its advantages in efficient resource utilization and exceptional performance with a relatively small model size. This further underscores the positive implications of the proposed method for achieving outstanding SR reconstruction in resource-constrained environments.

Fig. 12 illustrates, on the same platform, the training times per epoch for a ×4 task and the testing reconstruction times on the 64×64 UCMerced dataset for the same ×4 task. The times are measured in seconds. The training time results indicate that the training efficiency of ESRT and ACT algorithms is relatively low, with an average training time exceeding 3 min per epoch. In contrast, the proposed algorithm exhibits training times only slightly inferior to the DCM algorithm, ranking second with an average training time of 15.9 s per epoch. The testing time results further emphasize the superiority of the proposed algorithm. In the same testing environment on the identical platform, due to the smaller model size of the proposed algorithm, the processing time for each image is only 0.025 s, significantly outperforming other reconstruction models. Overall, considering both training and testing times, the proposed algorithm demonstrates a pronounced efficiency advantage, accomplishing the training and testing processes in a shorter timeframe while ensuring a high level of reconstruction quality.

F. Ablation Experiment

This section evaluates the performance of the proposed model by performing a number of ablation experiments, which examine how different components affect the reconstruction results. We employ metrics, such as PSNR, SSIM, parameter count, and floating-point operations to gauge the quality of reconstructed images and the complexity of the model. The outcomes of the ablation experiments are summarized in Table IX. As shown in the table:

1) the teacher network significantly enhances the reconstruction results. When there is no teacher network, the PSNR and SSIM of the reconstructed images are reduced by 1.22 dB and 0.0351, respectively, indicating that both the naturalness and structural similarity of the images are significantly weakened;
2) subpixel convolution positively affected the reconstruction results. When the inverse convolution was used to replace the subpixel convolution for upsampling, the PSNR and SSIM of the reconstructed image were also reduced by both 0.15 dB and 0.0034, although the number of parameters decreased slightly by 0.45 M, which indicates that the clarity of the image and the detail reproduction were reduced to some extent;

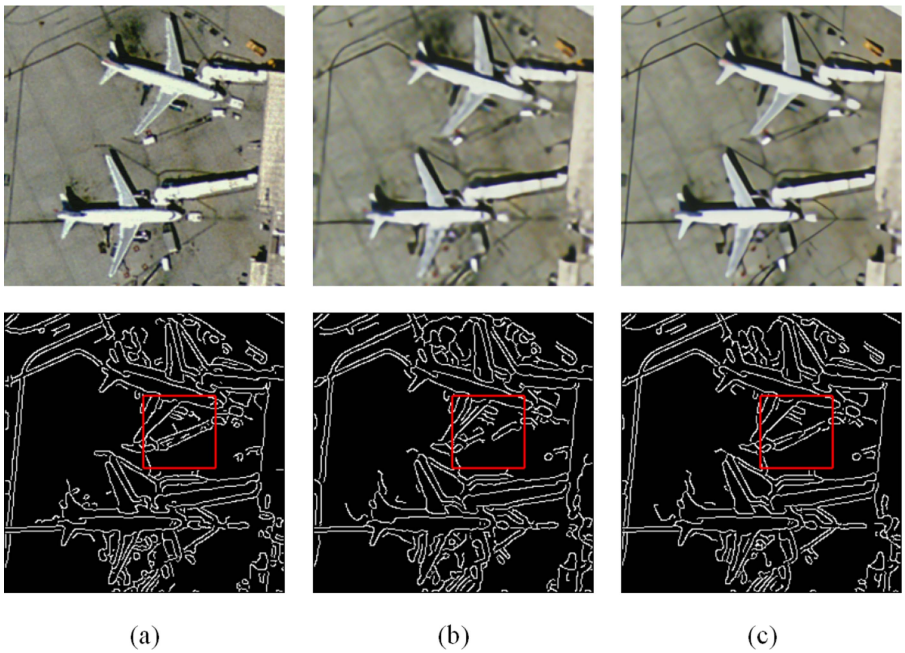


Fig. 13. Effect of a priori information on edge detection in airport images. (a) Real edges. (b) Edge detection results without a priori correction. (c) Edge detection results with a priori correction.

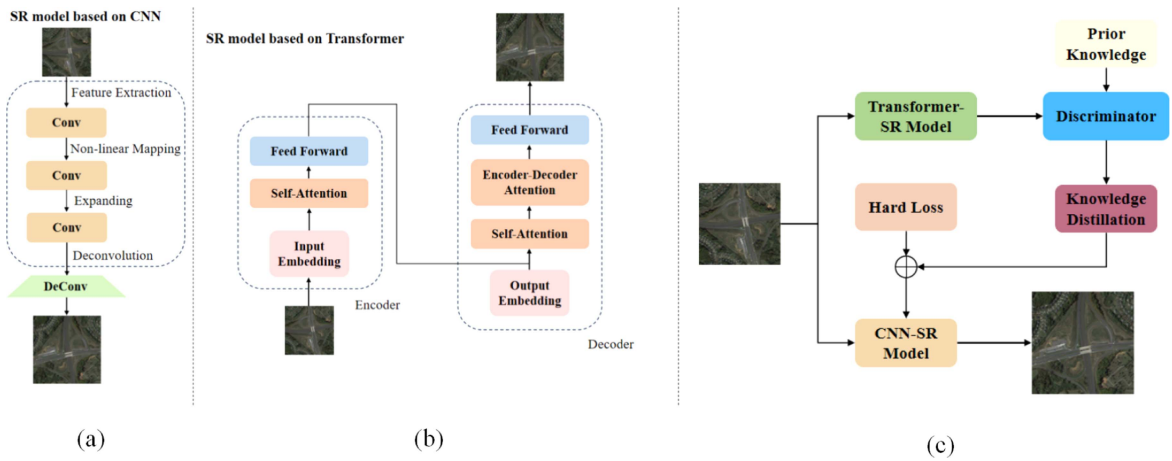


Fig. 14. Illustration of different SR frameworks. (a) CNN-based framework. (b) Transformer-based framework. (c) Proposed framework.

3) a priori information can be effective in correcting erroneous knowledge. From the results presented in the table, it can be observed that introducing prior information during the distillation process to correct the knowledge of the teacher network leads to a significant improvement in the performance of reconstruction, without affecting the complexity of the network. Meanwhile Fig. 13 shows the results of the edges of the reconstructed airport image before and after the addition of a priori information to the proposed algorithm. As can be seen from the right wing of the aircraft above in the image, the reconstructed image without the network corrected with a priori information shows broken edges, while the network corrected with a priori information improves the edge completeness of the image, thus improving the quality of the final image reconstruction.

V. Discussion

Distillation, CNN, and Transformer represent relatively mature technologies in the field of remote sensing. However, their integration poses a nontrivial challenge. Fig. 1 illustrates the SR reconstruction frameworks based on CNN and Transformer networks. The SR method based on CNN, depicted in Fig. 14(a), employs convolutional and upsampling layers to progressively enhance the image resolution. On the other hand, the SR method based on Transformer, illustrated in Fig. 14(b), utilizes self-attention mechanisms to capture long-range dependencies in the image, thereby enhancing fine details and textures. Therefore, the intelligent fusion of their respective advantages is a research question of significant interest.

Currently, KD methods effectively enable small models to mimic the outputs of pretrained large models within the same

network architecture. Most existing KD methods assume identical network architectures between teacher and student models, such as ResNet-34 and ResNet-18 [29], or Swin-S and Swin-T [30]. However, this assumption restricts the applicability of KD, as in practical scenarios, teacher and student models may possess different network architectures. Cross-architecture KD presents a challenging problem since models from different architectures may exhibit significant feature disparities, leading to difficulties in feature alignment. Thus, the design of effective methods for cross-architecture KD is a topic worthy of investigation.

In this article, for the cross-architecture KD problem in neural network architecture search, a CNN network based on mathematical model constraints is proposed, and the network structure is shown in Fig. 14(c), which is able to efficiently extract the global features of an image and achieve the feature alignment between different architecture networks under the condition of satisfying certain linear constraints. In addition, this article uses prior knowledge to correct the output of the teacher network so as to eliminate the error of the teacher network and deliver more accurate and reliable knowledge to the student network. Through experiments on multiple datasets and tasks, this article demonstrates that the proposed approach can significantly reduce the complexity of the model while ensuring high performance.

VI. Conclusion

This article introduces a lightweight CNN model, which is meticulously crafted and fine-tuned for the SR reconstruction of remote sensing images. To enhance the representational capacity of the CNN model, a KD approach is employed. This involves using an advanced Transformer network as the teacher model, transmitting the high-level semantic features it extracts to the CNN model as the student model. This guidance enables the CNN model to more effectively capture global and structural information in the images. Following this, experiments were conducted on publicly accessible remote sensing datasets to compare the proposed method with various popular SR reconstruction methods. The experimental findings illustrate that the proposed network attains superior image quality and visual effects, all the while upholding reduced computational complexity and memory consumption.

Through the proposed framework, it can be extended to other resource-constrained scenarios, such as real-time SR reconstruction. However, this extension introduces new challenges and issues that require in-depth research and exploration for viable solutions. First, facing lower computational resources and higher real-time requirements necessitates the exploration of methods to further compress and accelerate the proposed CNN network without compromising reconstruction quality. It is imperative to ensure that critical feature extraction and knowledge absorption capabilities are maintained even under limited computational resources. Second, different image application scenarios demand customized Transformer and CNN network structures to achieve optimal distillation and reconstruction effects. Selecting appropriate structures and parameter configurations, and considering the specificity of different tasks and scenes, will be a crucial direction for future research.

In summary, the framework proposed in this article offers a novel approach for SR reconstruction of remote sensing images in resource-constrained scenarios. However, there are also some limitations. First, although the combination of Transformer and CNN can effectively capture image features, the proposed student network is a smaller CNN network in terms of parameters. When dealing with extremely HR images, the limited parameter count of the student network restricts its capability to capture image features, potentially leading to performance bottlenecks. In other words, despite the network requiring more processing time, it may not achieve the desired performance level. Second, the proposed algorithm. Furthermore, experimental results demonstrate that the method achieves excellent performance in SR reconstruction tasks of both simulated and real images. However, even though the processing time for a single image is only 0.03 s, this processing speed is still insufficient for real-time processing of dynamic high-speed video streams. In future work, we will continue to delve into researching and addressing the aforementioned issues to enhance the performance and adaptability of the framework. Our aim is to provide more comprehensive and effective solutions for tasks, such as SR reconstruction in resource-constrained scenarios.

References

[1] B. Zhang et al., "Super-resolution reconstruction of a 3 arc-second global DEM dataset," *Sci. Bull.*, vol. 67, no. 24, pp. 2526–2530, 2022.
[2] Q. Chen, W. Ding, X. Huang, and H. Wang, "Generalized interval type-II fuzzy rough model-based feature discretization for mixed pixels," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 3, pp. 845–859, Mar. 2023.
[3] Q. Chen et al., "Neighborhood rough residual network-based outlier detection method in IoT-enabled maritime transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 11800–11811, Nov. 2023.
[4] C. Lin, C. Qiu, H. Jiang, and L. Zou, "A deep neural network based on prior driven and structural-preserving for SAR image despeckling," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6372–6392, 2023.
[5] B. Tu, Q. Ren, J. Li, Z. Cao, Y. Chen, and A. Plaza, "NCGLF2: Network combining global and local features for fusion of multisource remote sensing data," *Inf. Fusion*, vol. 104, 2024, Art. no. 102192.
[6] X. Du, X. Zheng, X. Lu, and A. A. Doudkin, "Multisource remote sensing data classification with graph fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10062–10072, Dec. 2021.
[7] C. Zhou, Z. He, A. Lou, and A. Plaza, "RGB-to-HSV: A frequency-spectrum unfolding network for spectral super-resolution of RGB videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5609318.
[8] D. Yang, Z. Li, Y. Xia, and Z. Chen, "Remote sensing image super-resolution: Challenges and approaches," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2015, pp. 196–200.
[9] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006.
[10] X. Qifang, Y. Guoqing, and L. Pin, "Super-resolution reconstruction of satellite video images based on interpolation method," *Procedia Comput. Sci.*, vol. 107, pp. 454–459, 2017.
[11] W. Shen, L. Fang, X. Chen, and H. Xu, "Projection onto convex sets method in space-frequency domain for super resolution," J. Comput., vol. 9, no. 8, pp. 1959–1966, 2014.
[12] 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.
[13] K. S. Ni and T. Q. Nguyen, "Image superresolution using support vector regression," *IEEE Trans. Image Process.*, vol. 16, no. 6, pp. 1596–1610, Jun. 2007.

[14] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Variational Bayesian super resolution," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 984–999, Apr. 2011.
[15] X. Liang and Z. Gan, "Improved non-local iterative back-projection method for image super-resolution," in *Proc. 6th Int. Conf. Image Graph.*, 2011, pp. 176–181.
[16] J. Zhang, C. Zhao, R. Xiong, S. Ma, and D. Zhao, "Image super-resolution via dual-dictionary learning and sparse representation," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2012, pp. 1688–1691.
[17] D. Li and S. Simske, "Example based single-frame image super-resolution by support vector regression," *J. Pattern Recognit. Res.*, vol. 1, no. 1, pp. 104–118, 2010.
[18] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016.
[19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 7, pp. 2480–2495, Jul. 2021.
[20] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, "Enriched CNN-transformer feature aggregation networks for super-resolution," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 4956–4965.
[21] L. Zeng, M. Huang, Y. Li, Q. Chen, and H.-N. Dai, "Progressive feature fusion attention dense network for speckle noise removal in OCT images," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2022.3205217.
[22] J. Lin and Q. Chen, "An intelligent sensor data preprocessing method for OCT fundus image watermarking using an RCNN," *Comput. Model. Eng. Sci.*, vol. 138, no. 2, pp. 1549–1561, 2024.
[23] A. Vaswani et al., "Attention is all you need," in Proc. Int. Conf. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 6000–6010.
[24] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using SWIN transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.
[25] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5791–5800.
[26] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 457–466.
[27] M. Zheng, H. Zang, X. Liu, G. Cheng, and S. Zhan, "Efficiently amalgamated CNN-transformer network for image super-resolution reconstruction," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2023, pp. 3–13.
[28] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 1789–1819, 2021.
[29] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 33716–33727.
[30] Z. Hao et al., "Learning efficient vision transformers via fine-grained manifold distillation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 9164–9175.
[31] Z. Hao, Y. Luo, H. Hu, J. An, and Y. Wen, "Data-free ensemble knowledge distillation for privacy-conscious multimedia model compression," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1803–1811.
[32] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv:1803.02155*.
[33] J. Shang, M. Gao, Q. Li, J. Pan, G. Zou, and G. Jeon, "Hybrid-scale hierarchical transformer for remote sensing image super-resolution," *Remote Sens.*, vol. 15, no. 13, 2023, Art. no. 3442.
[34] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5615611.
[35] D. Cai and P. Zhang, "T^3sr: Texture transfer transformer for remote sensing image superresolution," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7346–7358, 2022.
[36] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 2852–2859.
[37] F. Tao et al., "Microbial carbon use efficiency promotes global soil carbon storage," *Nature*, vol. 618, no. 7967, pp. 981–985, 2023.
[38] 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.
[39] 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.
[40] 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.
[41] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.
[42] J. M. Haut, M. E. Paoletti, R. Fernández-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.
[43] 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.
[44] F. A. Kruse et al., "The spectral image processing system (SIPS)–interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2/3, pp. 145–163, 1993.
[45] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

Cong Lin received the B.S. degree in automation and M.S. degree in control theory and control engineering from the Guangxi University of Science and Technology, Liuzhou, China, in 2011 and 2015, respectively. He is currently working toward the Ph.D. degree in information and communication engineering with Hainan University, Haikou, China.

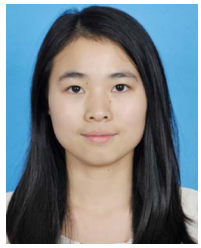
He is currently a Lecturer with the School of Information and Communication Engineering, Guangdong Ocean University, Zhanjiang, China. His research interests include machine learning and image processing.

Xin Mao is currently working toward the bachelor's degree in automation with Guangdong Ocean University, Zhanjiang, China.

His research interests include deep learning and computer vision.

Chenghao Qiu received the B.S. degree in intelligent science and technology from Hainan University, Haikou, China, in 2023. He is currently working toward the M.S. degree in biomedical engineering with the University of Electronic Science and Technology of China, Chengdu, China.

His research interests include computer vision tasks and bio-visual information processing.



Lilan Zou received the M.S. degree in condensed matter physics from Sun Yat-sen University, Guangzhou, China, in 2015. She is currently working toward the Ph.D. degree in materials science and engineering with Hainan University, Haikou, China.

She is currently a Lecturer with the School of Information and Communication Engineering, Guangdong Ocean University, Zhanjiang, China. Her research interests include nonvolatile electronic devices, and their applications in neuromorphic computing and information processing.