MULTI-LEVEL ENHANCED TRANSFORMER NETWORK FOR REMOTE SENSING IMAGE SUPER-RESOLUTION

MIN ZHANG[1], XIAO LI[1,*], LI-XIANG XU [2], XIN YUAN[3], YUAN-YAN TANG [4], ZHI-WEI ZHANG [5,*], YUAN-HAO JIN[1], XIANG LI[1]

[1]College of Artificial Intelligence and Big Data, Hefei University, Hefei, 230601, China
[2]College of Artificial Intelligence Engineering, Hefei Institute of Technology, Hefei, 238076, China
[3] School of Electrical and Mechanical Engineering, Adelaide University, Adelaide, SA 5005, Australia
[4]Zhuhai UM Science and Technology Research Institute, FST University of Macau, 999078, Macau
[5]Affiliation: School of Informatics and Engineering, Suzhou University, Suzhou 234000, China
*Corresponding Author's E-mail: zzwloveai@gmail.com

Abstract:

Recently, the development of super-resolution (SR) algorithms has become an important topic in enhancing the resolution of remote sensing images. Current Transformer-based methods often struggle to adequately capture image texture details and edge features. To address this limitation, this paper proposes a novel Transformer-CNN hybrid enhancement network, termed TCENet. First, a hybrid network combining CNN and Transformer is presented to effectively extract multi-scale local features and global features from remote sensing images. Furthermore, a multi-scale and multi-channel attention module is designed, which captures features at different hierarchies by constructing convolutional kernels of varying scales. After each convolutional group, channel attention mechanisms are adopted to extract texture details at multiple scales. Experimental results demonstrate that the proposed method achieves promising performance and exhibits significant application value in the process of remote sensing image super-resolution reconstruction.

Keywords: Remote Sensing Image Super-Resolution; Transformer; Multi-Scale; Attention Mechanism.

1. Introduction

Recently, with the rapid development of remote sensing technology, high-resolution remote sensing images have played an important role in many fields, such as environmental monitoring [1], disaster assessment [2], urban planning [3], and military reconnaissance [4]. However, due to the limitations of the imaging equipment and harsh imaging conditions in space, the image quality is often affected by factors such as motion blur, atmospheric interference, and ultra long-range imaging, resulting in varying degrees of degradation. Therefore, RSI super-resolution (RSISR) reconstruction technology has received much attention in remote sensing applications.

SISR technology mainly includes interpolation-based methods, reconstruction-based methods and deep learning-based methods. Interpolation-based methods mainly include bilinear interpolation and bicubic interpolation [14]. However, interpolation methods cannot generate new high-frequency information, which may lead to over-smoothed edges or blurred regions in output images. These methods sometimes assume prior knowledge about the image, such as smoothness, or are designed to work with sequences of multiple low-resolution images for reconstruction. In recent years, people have focused on the field of deep learning, CNN-based methods perform well in local feature extraction. However, due to the limited receptive field of convolutional operations. Transformer-based methods demonstrate efficient capability in capturing global contextual information through self-attention mechanisms. However, they often struggle to restore local high-frequency details, such as the textures of farmland gullies or building edges, due to the lack of focus on local pixel relationships.

To address the limitations of previous methods, this paper proposes a novel multilevel Transformer-CNN hybrid enhancement network (TCENet). The main contributions of this work can be summarized as follows:

(1) We design a Transformer-CNN hybrid enhancement network (TCENet), which combines the advantages of CNN for reconstructing local features and the Transformer for recovering global features.

(2) We design a multi-level feature enhancement architecture that enables the fusion of low-dimensional and high-dimensional features. Moreover, a multiscale feature extraction module (PFEM) is designed to extract multi-scale features for enhancing local details.

(3) Extensive experimental results on remote sensing image datasets confirm the effectiveness of our method.

2. Related Work

2.1 CNN-Based Super-Resolution

Dong et al. [6] first introduced CNN for the single image super-resolution (SR) task by proposing SRCNN, a simple three-layer convolutional network that learns an end-to-end mapping from low-resolution (LR) to high-resolution (HR) images. Subsequently, Kim et al. [7] developed the Very Deep Super-Resolution Network (VDSR) based on SRCNN. VDSR extends the shallow SRCNN by utilizing a substantially deeper 20-layers architecture, thereby enlarging the receptive field to better exploit contextual information. Ledig et al. [8] subsequently proposed SRResNet by incorporating the residual learning strategy from ResNet [5]. They designed a novel residual block and stacked 16 identical blocks, demonstrating the effectiveness of deep residual architectures in learning complex LR-to-HR mappings. Building upon these frameworks, the Enhanced Deep Super-Resolution Network (EDSR) [9] was proposed. EDSR removes the batch normalization (BN) and rectified linear unit (ReLU) layers from the residual blocks and introduces a residual scaling mechanism. Ledig et al. [10] proposed the Super-Resolution Generative Adversarial Network (SRGAN), which introduced adversarial learning and perceptual loss to generate photo-realistic HR images from LR inputs.

2.2 Transformer -Based Super-Resolution

Since the introduction of the Vision Transformer architecture [11], transformer-based models have been explored and applied to super-resolution (SR) tasks. Liang et al. [12] proposed SwinIR based on the Swin Transformer. By alternately employing regular and shifted window partitioning, its shifted-window attention mechanism enables cross-window interactions, allowing the network to effectively model long-range dependencies. Subsequently, Zhang et al. [13] proposed ELAN, which adopts a grouped multi-scale self-attention (GMSA) mechanism. ELAN partitions feature maps along the channel dimension and assigns different window sizes to each group, thereby capturing dependencies across multiple spatial ranges. Lei et al. proposed a transformer-based enhancement network (TransENet), which fused Transformer-based architecture with a multi-stage enhancement strategy.

3. Proposed Method

In this section, the proposed TCENet is described in detail. First, the overall framework of TCENet is presented. Subsequently, we mainly introduce the Multiscale Feature Extraction Module (PFEM). Finally, encoder module and decoder module are presented.

3.1. Network Architecture

As shown in Figure 1, our proposed TCENet is mainly composed of three components: a shallow feature extraction module, a multilevel feature enhancement module, and a back projection-based reconstruction module. The shallow feature extraction module employs a standard convolutional layer to transform the RGB image into feature maps and extracts low-level image features. The multilevel feature enhancement module consists of an encoder-decoder architecture and a multiscale feature extraction module (PFEM).

Given a LR image $I_{LR} \in R^{H\times W\times 3}$, the RGB image is first transformed into shallow feature maps $F_0 \in R^{H\times W\times C}$ through a 3×3 convolutional layer, where $H\times W$ denotes the spatial dimension and C denotes the number of channels, is defined as follows:

$$F_1 = Conv(I_{LR}) \tag{1}$$

where $Conv(\cdot)$ represents the standard convolutional layer and F_0 denotes the extracted shallow features.

Next, these shallow features pass through multiple PFEM blocks to extract multiscale features. The multi-scale feature extraction is defined as follows:

$$F_n = PFEM_n(PFEM_{n-1}(\cdots PFEM_1(F_0)\cdots)) \tag{2}$$

where $PFEM_n$ denotes the n-th PEEM block.

The feature map F_n is fed into three encoders to extract low-dimensional features. The resulting feature map F_3 is then passed through an upsampling layer to project these features into a higher-dimensional space.

3.2. Multiscale Feature Extraction Module (PFEM)

The multiscale feature extraction module (PFEM) is composed of pyramid dilated convolutions [16], spatial attention, and channel attention. Pyramid dilated convolutions employ different dilation rates to simultaneously capture local details and global contextual information. Spatial attention dynamically adjusts important regions in the spatial dimension of an image, thereby enhancing the model's perception of key targets. Channel attention automatically extracts the importance global contextual information of each channel, enhancing task-related features and suppressing irrelevant or noisy channels. The overall network architecture is illustrated in Figure 2.

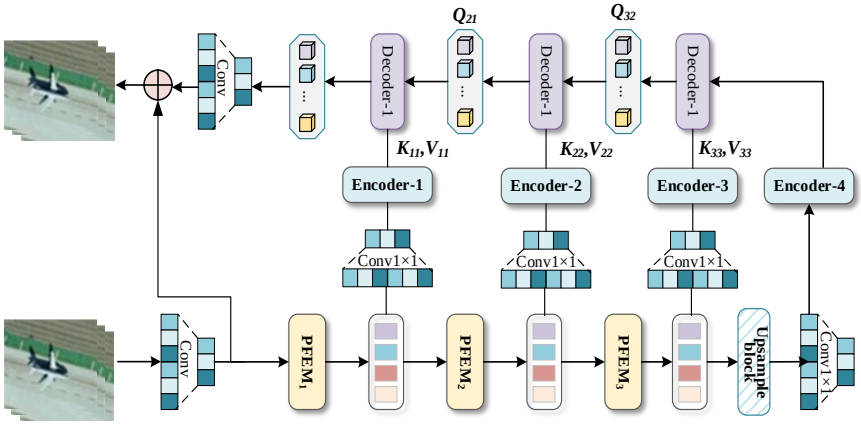

FIGURE 1. Overall architecture of the TCENet framework

The input feature map $X \in R^{H\times W\times C}$ is evenly divided Into k groups in the channel dimension C, where $k = 4$. Each channel subset X_i is fed into an independent processing branch F_i. Within each branch, multiscale feature extraction is performed in the spatial information dimensions (H, W), is defined as follows:

$$F_i = f_i(X_i) \tag{3}$$

where $f_i(\cdot)$ denotes the multiscale feature extraction function of the i-th branch.

Specifically, the branches $f_1(\cdot)$, $f_2(\cdot)$ and $f_3(\cdot)$ employ 1×1, 3×3, 5×5 and 7×7 convolution kernels, respectively. Finally, the output feature maps from the four branches are concatenated along the channel dimension to produce the final fused feature representation:

$$F_{cat} = Concat(F_0, F_1, F_2, F_3) \tag{4}$$

3.3. Encoder Module

The main architecture of the encoder follows the original design proposed in [17], which consists of the multi-head self-attention (MSA) module and multilayer perceptron (MLP) network. A layer normalization (LN) is applied after each module and before the local residual structure. The specific formulation of the encoder is defined as follows:

$$y_i' = MSA\big(LN(y_{i-1})\big) + y_{i-1}, i = 1, \cdots, L_e \tag{5}$$

$$y_i' = MLP(LN(y_i')) + y_i', i = 1, \cdots, L_e \tag{6}$$

$$[f_{E_1}, f_{E_2}, \cdots, f_{E_N}] = y_{L_e} \tag{7}$$

where L_e denotes the number of encoder layers, and f_{E_1} represents the encoder output.

The PFEM-1 module processes F_0 and outputs F_1. Encoder-1 and PFEM-2 module processes

the previous feature map F_1, producing the feature map F_2 and (k_1, v_1). Encoder-2 and the PFEM-3 module then process the feature map F_2, outputting the feature map F_3 and (k_2, v_2). The feature map F_3 is further processed by Encoder-3 to generate (k_3, v_3), while F_3 is simultaneously fed into the upsampling module to produce the feature map F_4. The feature map F_4 then processed by Encoder-3 to output Q, which serves as the query feature for the input of Decoder-3.

3.4. Decoder Module

The decoder architecture consists of the MSA module, the multilayer perceptron (MLP) network, and the cross-attention MSA module, enabling simultaneous processing of the encoder outputs and the decoder input features. The specific formulation of the decoder is defined as follows:

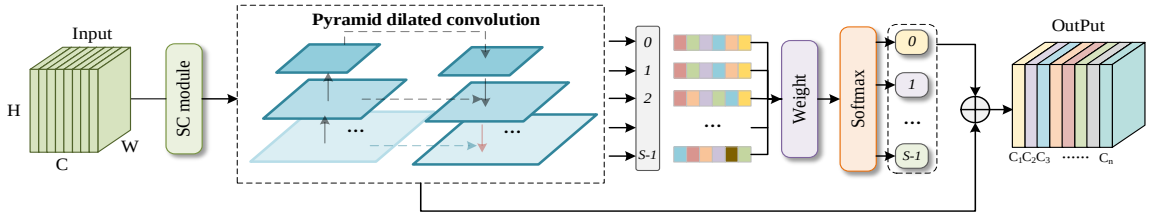

FIGURE 2. Pyramid feature extraction module

$$z_0 = [f_{E_1}, f_{E_2}, \cdots, f_{E_N}] \quad (8)$$

$$z_0' = MSA(LN(z_{i-1})) + Z_{i-1}, i = 1, \cdots, L_d \quad (9)$$

$$z_i'' = MSA\big(LN(z_{i-1}), LN(z_0)\big) + z_{i-1}', i = 1, \cdots, L_d \quad (10)$$

$$z_i = MLP(MLP(LN(z_i''))) + z_i'', i = 1, \cdots, L_d \quad (11)$$

$$[f_{D_1}, f_{D_2}, \cdots, f_{D_N}] = y_{L_d} \quad (12)$$

where f_{D_i} denotes the output of the decoder, and L_d represents the number of decoder layers. Decoder-3 integrates the Key and Value features (K_3, V_3) from its corresponding encoder (Encoder-3) with the Query feature Q_4 derived from the deeper Encoder-4. Following this principle, high-dimensional features from deeper layers are primarily utilized as the Q components in the decoder, whereas the encoded low-dimensional features are sequentially fed into Decoder-1 through Decoder-3 as K and V pairs to be combined with these high-dimensional queries.

4. Experimental Results and Analysis

4.1. Dataset and Evaluation Metrics

We perform extensive experiments on AID dataset and UCMerced dataset to evaluate the effectiveness of TCENet. The UCMerced dataset contains 2 classes, each consisting of 100 remote sensing scene images of size 256×256 pixels.The AID (Aerial Image Dataset) consists of 10,000 images across 30 categories, with an image size of 600×600.We adopt four metrics to evaluate the reconstruction quality: Peak Signal-to-Noise Ratio PSNR, Structural Similarity (SSIM), Spatial Correlation Coefficient (SCC), and Spectral Angle Mapper (SAM). PSNR is measured in decibels (dB). For the reconstructed images, higher PSNR, SSIM, and SCC values, along with lower SAM values, indicate better quality.

4.2. Experimental Details

All experiments were conducted on a Windows 10 workstation equipped with an NVIDIA GeForce

RTX 3090 GPU. During the training stage, image patches of size 48x48 are randomly cropped from low-resolution remote sensing images. To optimize the model parameters, the Adam optimizer is adopted during training. where $\beta_1 = 0.9$、$\beta_2 = 0.99$ and $\varepsilon = 10^{-8}$. The initial learning rate is set to 10^{-4}, the batch size is set to 8, and the total number of training iterations is 1500. The learning rate is reduced by half after 1000 training iterations.

4.3. Comparative Experiments

Quantitative analysis. To evaluate the performance of the reconstruction algorithm, the proposed method is compared with several state-of-the-art approaches in the field of remote sensing image super-resolution in recent years. These methods include HSENET, TransENet, SRDD, FENet, Omnisr and HAU-Net. The super-resolution reconstruction results of different algorithms at scaling factors of ×3 and ×4 are presented in Tables 1 and 2.

We show the ×2, ×3 and ×4 SR results of all compared methods on the AID dataset in Tables 1 and 2, where the best and second-best results are highlighted in bold and underlined, respectively. It is worth noting that some baseline results are reproduced from our previous work. As shown, the proposed TCENet achieves the highest PSNR and SSIM, along with the lowest SAM across all scaling factors. For instance, at ×2 SR, TCENet outperforms the second-best method significantly in both PSNR and SSIM. The proposed method also shows competitive advantages in the ×3 and ×4 tasks. Overall, these results validate the effectiveness of the proposed method.

TABLE 1. Quantitative evaluation of ×3 SR task on benchmark datasets

Method	AID datasete				UCMerced dataset			
	PSNR	*SSIM*	*SCC*	*SAM*	*PSNR*	*SSIM*	*SCC*	*SAM*
HSENet	31.40	0.8577	0.3988	0.0816	30.09	0.8436	0.4133	0.0801
TransENet	31.47	0.8585	0.4063	0.0812	30.01	0.8440	0.3986	0.0814
SRDD	31.38	0.8562	0.3982	0.0819	29.95	0.8418	0.4087	0.0813
FENet	31.30	0.8550	0.3950	0.0828	29.88	0.8383	0.3940	0.0824
Omnisr	31.58	0.8610	0.3986	0.0818	30.08	0.8421	0.4077	0.0808
HAUNet	<u>31.66</u>	<u>0.8623</u>	<u>0.4146</u>	<u>0.0790</u>	<u>30.21</u>	<u>0.8455</u>	<u>0.4229</u>	<u>0.0786</u>
TCENet(ours)	**31.70**	**0.8629**	**0.4151**	**0.0781**	**30.23**	**0.8478**	**0.4239**	**0.0744**

TABLE 2. Quantitative evaluation of ×4 SR task on benchmark datasets

Method	AID datasete				UCMerced datasete			
	PSNR	*SSIM*	*SCC*	*SAM*	*PSNR*	*SSIM*	*SCC*	*SAM*
HSENet	29.16	0.7811	0.2642	0.1036	27.63	0.7580	0.2689	0.1040
TransENet	29.35	0.7880	0.2821	0.1013	27.79	0.7642	0.2704	0.1023
SRDD	29.19	0.7831	0.2701	0.1030	27.72	0.7612	0.2721	0.1033
FENet	29.10	0.7792	0.2601	0.1043	27.59	0.7538	0.2568	0.1053
Omnisr	29.42	0.7902	0.2889	0.1004	<u>27.85</u>	0.7653	<u>0.2788</u>	<u>0.1010</u>
HAUNet	<u>29.46</u>	<u>0.7914</u>	<u>0.2925</u>	<u>0.0988</u>	27.84	<u>0.7662</u>	0.2784	0.1018
TCENet(ours)	**29.51**	**0.8023**	**0.2938**	**0.0978**	**28.01**	**0.7704**	**0.2910**	**0.1007**

Qualitative comparison. To further evaluate the image reconstruction quality of the proposed method, we conducted a visual comparison with five other algorithms. Figure 3 presents the super-resolution results for a ×4 upscaling factor, highlighting the strengths and characteristics of each method. The results demonstrate that TCENet performs favorably in scenes with complex contours and abundant objects, such as airports, residential areas, and parking lots. However, in regions with relatively simple textures, such as baseball fields, TCENet achieves comparatively moderate performance, where TransENet yields superior results due to its stronger capability in cap-

turing large homogeneous regions. In comparison, the proposed TCENet captures more contextual information while preserving detailed structures, demonstrating strong competitiveness against other methods.

4.4. Ablation Experiments

In this section, the UCMerced dataset is employed to validate the effectiveness of the PFEM module. Table 3 presents the quantitative SR performance on the UCMerced dataset at ×4 upscaling. Compared with the baseline results, PSNR and SSIM are improved by 0.100dB and 0.0039, respectively.

TABLE 3. Ablation experiments for the blocks on UCMerced-×4

PFEM	PSNR	SSIM
×	27.91	0.7665
√	28.01	0.7704

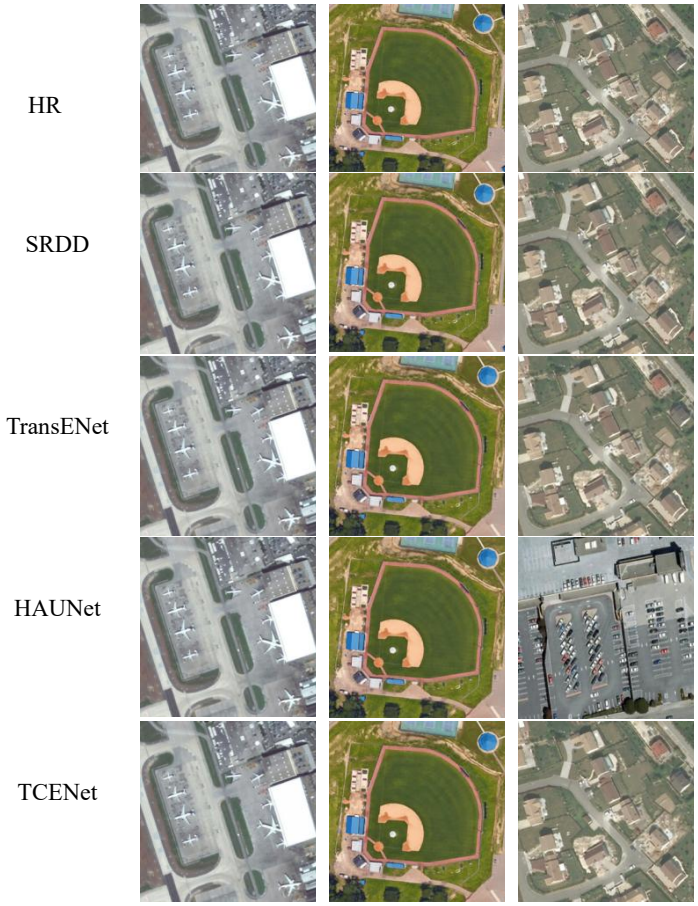

FIGURE 3. Comparison results with other method

4. Conclusions

In this paper, we propose a novel Transformer-CNN hybrid enhancement network (TCENet) for remote sensing image super-resolution, which combines the strengths of both CNNs and Transformers. TCENet integrates multi-scale local features and global features in remote sensing images. TCENet aims at the comprehensive utilization of high- and low-dimensional features. Moreover, a multiscale feature extraction module (PFEM) is introduced to capture multi-scale features and suppress reconstruction artifacts. Extensive experiments on the AID and UCMerced datasets demonstrate the effectiveness and reliability of our method.

Acknowledgements

This paper is supported by the Anhui Provincial University Research Project 2025AHGXZK20110, Institute of Complex Networks and Computational Intelligence, Suzhou University Research Plat for (2021XJPT50), Suzhou Municipal Scientific Research Platform (2022SJPT05), Anhui Province Overseas Visiting Scholar Program for Young Backbone Teachers under the Young and Middle-Aged Teacher Development Initiative gxgwfx2020063, The University Synergy Innovation Program of Anhui Province GXXT-2022-047.

References

[1] Rau J Y, Jhan J P, Hsu Y C, "Analysis of oblique aerial images for land cover and point cloud classification in an urban environment," *IEEE Transactions on Geoscience and Remote Sensing*, 2014, pp. 1304-1319.

[2] Ghaffarian S, Kerle N, Filatova T, "Remote sensing-based proxies for urban disaster risk management and resilience: A review," *Remote sensing*, Vol. 10, No. 11, 2018.

[3] Hu T, Yang J, Li X, et al, "Mapping urban land use by using landsat images and open social data," *Remote sensing*, Vol. 8, No. 2, pp. 151, 2016.

[4] Chen C, Gong W, Chen Y, et al, "Object detection in remote sensing images based on a scene-contextual feature pyramid network," *Remote Sensing*, Vol. 11, No. 3, pp. 339, 2019.

[5] He K, Zhang X, Ren S, et al, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition. 2016*, pp. 770-778.
[6] Dong C, Loy C C, He K, et al, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, Vol. 38, No. 2, 2015, pp. 295-307.
[7] Kim J, Lee J K, Lee K M, “Accurate image super-resolution using very deep convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646-1654.
[8] Ledig C, Theis L, Huszár F, et al, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.
[9] Lim B, Son S, Kim H, et al, “Enhanced deep res idual networks for single image super-resolution,” in *Proceedings of the IEEE conference on comput er vision and pattern recognition workshops*, 2017, pp. 136-144.
[10] Ledig C, Theis L, Huszár F, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681-4690.
[11] Dosovitskiy A, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
[12] Liang J, Cao J, Sun G, “Image restoration using swin transformer,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833-1844.
[13] Zhang X, Zeng H, Guo S, “Efficient long-range a ttention network for image super-resolution,” in *E uropean conference on computer vision.* Cham: *Sp ringer Nature Switzerland*, 2022, pp. 649-667.
[14] Yuan S, Abe M, Taguchi A, “High accuracy bicubic interpolation using image local features,” *IEICE Transactions on Fundamentals of Electronics*, *Communications and Computer Sciences*, Vol.90, No.8, pp. 1611-1615, 2007.
[15] Lei S, Shi Z, Mo W, “Transformer-based multista ge enhancement for remote sensing image super-re solution,” *IEEE Transactions on Geoscience and R emote Sensing*, 2021, pp. 1-11.
[16] Chen L C, Papandreou G, Kokkinos I, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, Vol. 40, No. 4, pp. 834-848, 2017.
[17] Vaswani A, Shazeer N, Parmar N, “Attention is all you need,” *Advances in neural information processing systems*, 2017.