T^{3}SR: Texture Transfer Transformer for Remote Sensing Image Superresolution

Durong Cai and Peng Zhang

***Abstract*—Remote sensing image superresolution has made significant progress in recent years, aiming to restore natural and realistic high-resolution images from low-resolution images. However, most image superresolution remote sensing methods are improved only by deepening their network and expanding the network size, consuming substantial computing resources and imposing a bottleneck in development. Here, we propose an end-to-end image superresolution network called texture transfer transformer for remote sensing image superresolution (T^{3}SR). For the first time, T^{3}SR introduces image texture transfer into remote sensing, which achieves the most advanced results. Specifically, T^{3}SR divides image superresolution into two stages: texture transfer and feature fusion. First, to solve the problems of missing textures, artifacts, and blurring in a single image superresolution approach, we design a texture transfer module to serve the shallow texture transfer. Second, to further reduce the dependence of the model on the reference image, we propose a U-Transformer-based feature fusion scheme to reduce the dependence on the reference image. Finally, we conduct numerous experiments on standard public datasets to fully evaluate our approach. In addition to verifying the method's superiority based on the reference image paradigm, we also test the performance without the reference image. All results show that our method yields an abundant texture and finish with better visual results. Moreover, the best score is also obtained in the quantitative parameters of PSNR and SSIM. Compared with the best available approach, T^{3}SR has an improved performance by 0.79 dB and 0.33 dB in the datasets of WHU-RS19 and RSSCN7, respectively.**

***Index Terms*—Image superresolution, reference image, self-reference, texture transfer.**

I. Introduction

IMAGE superresolution (SR) refers to the process of obtaining a natural and realistic high-resolution (HR) image from a low-resolution (LR) image. It is a basic task of image processing and computer vision. Likewise, image superresolution is also known by more common names, such as scaling, interpolation, and magnification. More broadly, generating the HR image from an LR image is considered as superresolution as long as technical means are used. In general, HR images have higher pixel density, higher definition, and more detailed textures than LR images. Therefore, HR images have a wider application in various fields than LR images. However, due to the sensor and high-cost constraints, most real-world images are dominated by low resolution. Especially in remote sensing, owing to the equipment limitation, remote sensing images are generically low resolution. If the LR image observed from the satellite can be recovered from the HR image, it is of great significance for various applications such as environmental monitoring, resource exploration, and surveillance. Therefore, superresolution in remote sensing [1], [2], [3] has become one of the most essential scientific areas in recent years.

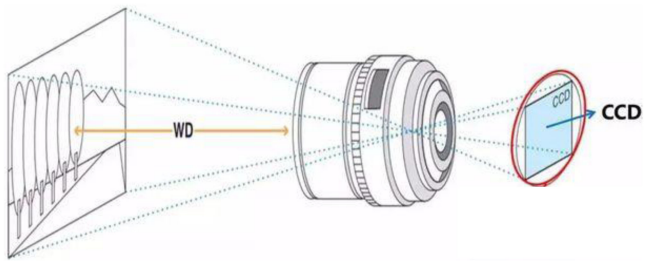

Fig. 1. Architecture of camera imaging.

Image superresolution in remote sensing has attracted more and more attention in the industry, mainly concerning the following aspects: First, the resulting image is mainly low resolution due to the sensor's limited physical parameters. As shown in Fig. 1, the charge coupled device (CCD) generates and stores the corresponding charges according to the difference in the light coming from the lens. However, due to the difficulty of the process and the rising production cost, the size of the CCD is often small. Therefore, when the size of the CCD is constant, the larger the distance from the objective to the lens WD, the lower the imaging resolution of the device. In remote sensing, imaging equipment is generically located on orbiting satellites in units of 10 000 m, and the lower ones are also carried by drone in 100 m. Therefore, the image captured is often lower than 0.5 m resolution in remote sensing, limiting the development of computer vision remote sensing.

Furthermore, as a member of low-level vision tasks, image superresolution determines the image quality and affects other vision tasks. For example, in the object detection task, an HR image has clearer texture details, better representing the target features so that the model can better capture and locate the target. This demonstrates that applying image superresolution to target detection tasks is appropriate for detecting dense small targets

Manuscript received 12 April 2022; revised 20 June 2022 and 30 July 2022; accepted 9 August 2022. Date of publication 16 August 2022; date of current version 9 September 2022. This work was supported in part by the Science and Technology Planning Project of Guangdong Science and Technology Department Guangdong Key Laboratory of Advanced IntelliSense Technology under Grant 2019B121203006, and Shenzhen Science and Technology Program under Grant KQTD20190929172704911. *(Corresponding author: Peng Zhang.)*

The authors are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, China (e-mail: 1750826167@qq.com; zhangpeng5@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3198557

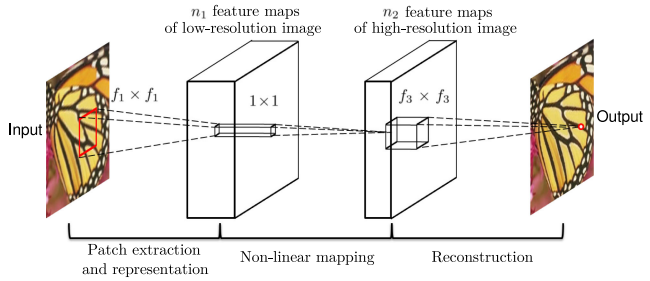

Fig. 2. Overall architecture of SRCNN.

and improving the overall performance. Similarly, combining image superresolution with other high-level vision tasks such as classification, localization, and segmentation can also bring certain benefits to itself [4], [5].

Convolutional neural networks (CNNs) have led to significant improvements in the performance of superresolution. Nevertheless, most methods are developed for the generic scene, while techniques specifically designed for remote sensing scenes are relatively scarce. Unlike the generic scene, remote sensing images have a lower resolution, i.e., one pixel in the picture corresponds to several square meters of ground. A remote sensing image of 5000 $\times$ 5000 pixels cannot be used directly, as it needs to be used after cropping. Due to the properties of a remote sensing scene, the resolution is so low that pixels between neighbors often have the same texture and the image background is also simpler, leading to the image superresolution based on a reference image (RefSR).

The RefSR-based methods transfer the texture from the reference image to the reconstructed image, improving the performance over the past superresolution methods. RefSR-based methods complement the missing details in the LR image by transferring the rich textures of the HR image and generating more detailed and realistic textures with the help of a reference image. The existing RefSR methods are [6], [7], [8] and play an important role in their respective fields. Unfortunately, these methods require high similarity between the reference and LR images. When the correlation is weak or dissimilar, their performance degrades significantly, becoming even worse than single image superresolution methods (SISR).

A classical SISR method is SRCNN, shown in Fig. 2, which does not have a clear definition of the spatial mapping from the LR to the HR image, and obtains a high-dimensional mapping relationship with the help of massive data training. Given that the texture of the HR image is overly destroyed during degradation, the blurring is aggravated with the increase of the magnification factor. Although superresolution methods based on GAN [9] have been proposed to alleviate the above problems, the illusions and artifacts caused by the GAN-based generation further pose a significant challenge to the superresolution task. To overcome the shortcomings of the SISR-based methods, Zhang et al. [7] proposed the superresolution image based on neural texture migration (SRNTT), a superresolution method with a reference image-based paradigm. Unfortunately, the RefSR methods require a high correlation between the reference image and LR image, and their performance degrades significantly when the degree of correlation is low, becoming even inferior to the SISR methods.

In summary, research on image superresolution of remote sensing is of great significance as it can not only solve the problems posed by physical sensors but also bring benefits to other computer vision tasks. Regrettably, the current superresolution methods still face many challenges and shortcomings, summarized in the following problems: First, the problems of missing textures, artifacts, and blurring that exist in SISR methods. Second, the method of RefSR has a strong dependence on the reference image.

This article focuses on the above challenges and proposes T[3]SR: a texture transfer transformer for superresolution of remote sensing images. The main contributions of this work are summarized as follows:

1) For the imaging characteristics of remote sensing superesolution, we propose T[3]SR, which is the first time introducing texture transfer into the superresolution of remote sensing. More specifically, we divide superresolution into two stages: texture transfer and feature fusion. For shallow semantic information, we transfer the texture directly, while for deep semantic information, we design a feature fusion network, called U-Transformer, to improve the performance and reduce the dependence on the reference image.
2) Addressing the problems of high-frequency detail loss, artifacts, and blurring of the SISR methods. Specifically, we construct an adaptive texture migration module based on the RefSR paradigm. This module can enrich the texture details of superresolution images by adaptively transferring textures according to the texture similarity between the reference and LR images. Among them, the reference attention module is also a variant of a transformer, containing elements, for instance, K, Q, V, and the attention matrix. It differs from the single input and self-attentive mechanism strategy used by a transformer in the past. We take the reference and LR images as input and use the reference attention mechanism. Extensive experiments show that the module can adaptively learn the high-definition texture of the reference image and migrate it to the subsequent superresolution image reconstruction.
3) To address the problem that the RefSR methods require a high correlation of reference images, we design a feature fusion network named U-Transformer based on the structure of the transformer and U-Net. We rely on this module combination because U-Net is designed to compensate for the disadvantage of fewer training samples, and the transformer is designed to enhance image features through a self-attentive mechanism. Therefore, the U-Transformer module effectively reduces the module's dependence on the reference image to achieve weak dependence and even more self-reliance. In the subsequent experiments, we investigate the model's performance in the lack of reference image context, with the results proving that the module can effectively reduce the model's dependence on the reference image.

The rest of this article is organized as follows. Section II will review the related work of the SISR, RefSR, and transformer-SR-based methods, and the remote sensing superresolution methods. Section III details T^3SR, containing a generic exposition and a modular explanation. In Section IV, complete experiments are conducted in both generic and remote sensing scenes to verify our method's effectiveness. Finally, Section V concludes this article.

II. Related Work

In this section, we present the related work regarding the SISR, RefSR, transformer-SR, and remote sensing methods. It also gives a brief analysis of the advantages and disadvantages of the above superresolution paradigms and inspiration for our approach.

A. Single Image Superresolution

In recent years, data-driven deep learning approaches have surpassed traditional superresolution methods. For a more comprehensive understanding of the technical development history of SISR, the reader is referred to [10], [11], [12]. The SISR-based methods aim to train a complex neural network through a large amount of data to establish an image mapping function between the LR and HR images. Next, we present the development history of the SISR methods.

In 2014, the first deep learning-based SISR approach was SRCNN [13] using a three-layered CNN proposed by Dong et al. (see Fig. 2). The arrival of SRCNN brought a new way for superresolution, which laid the foundation for the following methods. Then, in 2015, He et al. proposed Res-Net [14], which solved the network's deep gradient disappearance problem using a residual structure. On this basis, Kim et al. [15] developed VDSR and introduced the residual structure in superresolution for the first time. The problem of inadequate feature extraction was solved by increasing the number of network layers through the residual structure. Subsequently, Lim proposed EDSR [16], which removed some unnecessary modules in the residual structure, making it applicable to low-level vision problems like superresolution. Next, Lim proposed MDSR to address the problems of EDSR [17]. In 2020, Tian et al. [18] proposed CFSRCNN to improve model stability. However, the feature extraction ability of these two networks is insufficient. In 2021, Kong et al. proposed ClassSR [19], the first superresolution method that combines classification and superresolution to accelerate superresolution through data features.

All the above SISR methods ignore texture similarity and improve the performance only by expanding the network size, which makes the superresolution development difficult. Subsequently, the RefSR-based methods were proposed to extend the development of SISR.

B. Reference-Based Image Superresolution

Unlike SISR, RefSR introduces a reference image to assist the image reconstruction process in addition to the LR image. Generally, the reference image is an HR image with a texture or content similar to that of the LR image.

The first RefSR network is CrossNet [6], which contains an image encoder, a cross-scale warping layer, and a fusion decoder. First, the encoder is responsible for the feature extraction of the LR and reference images. Second, the positions of both images are aligned by a cross-scale distortion layer. Finally, the decoder connects the above feature maps to generate an HR image. However, since CrossNet uses spatial alignment, when the similarity between the reference and LR images is low, the restoration effect is inferior to SISR-based methods. Subsequently, methods such as SRNTT [7] and TTSR [8] were proposed. The former was proposed by Zhang et al. and exchanged the most similar features of the reference and LR images through convolutional layers. This method is effective for image detail restoration, but SRNTT ignores the correlation between the reference image and the LR image, which makes all image features equally available to the subsequent network. Therefore, TTSR combines the feature maps with a transformer to design a texture converter that uses an attention method to assign weights to features, informing the network for features to be focused on during learning.

The above-mentioned RefSR methods bring a new paradigm to image superresolution. They obtain texture details by texture migration, which in the past required substantial computational resources. However, these methods focus on the texture migration module and ignore the importance of the subsequent feature fusion. As a result, these methods rely heavily on a reference image, which is undesired.

C. Vision Transformer

In 2017, the transformer was first introduced by Vaswani et al. [20] and was originally designed to process sequences in NLP. In recent years, we have progressively explored transformer architecture in computer vision. A visual transformer slices and encodes an image in the same way NLP processes language sequences and then gets the internal connections through attention. To date, it has been widely used for common vision tasks, like object classification [21], [22], [23], [24], [25], [26], [27], object detection [24], [28], [29], [30], and semantic segmentation [23], [24], [31], [32].

The functions performed in the above tasks can be divided into two areas. The first one is to introduce self-attention in traditional CNN. For example, Fu et al. [33] proposed DANET to extract image information by combining spatial and channel attention. Wang et al. [34] utilized self-attention to improve the model's performance on several advanced visual tasks. An alternative solution is replacing the CNN with self-attentive blocks. For instance, Dosovitskiy [25] used a transformer block for image classification. Transformer models have also been used for low-level vision tasks, for example, superresolution [8], [35], image coloring [36], denoising, [37], [38], and decontamination [38]. The overall architecture of SwinIR [35] is shown in Fig. 3.

D. SR of Remote Sensing

Remote sensing has always been one of the essential scenes, where superresolution is applied. In recent years, the industry's research focus on image superresolution technology for remote sensing scenes is based on deep learning. In 2017, Lei et al. [39]

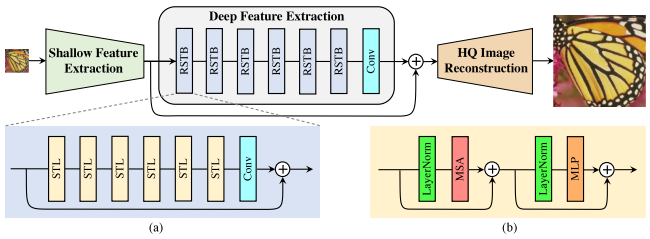

Fig. 3. Overall architecture of SwinIR.(a) Residual Swin Transformer Block (RSTB) (b) Swin Transformer Layer (STL).

proposed a local–global combined network (LGCNet) for remote sensing based on deep CNN. Motivated by the success of the back-projection network [40], Pan et al. [41] proposed a dense residual backprojection block to enhance the feature extraction ability. Since the attention mechanism is well known, Gu et al. [42] proposed to use the attention module to build a remote sensing superresolution network. Jiang et al. [43] also proposed the edge-enhancement network (EEGAN), a remote sensing image superresolution network using an adversarial generative strategy, aiming to improve the recovery performance of image edge information. Similarly, Ma et al. [44] proposed the DRGAN, a GAN-based method for detailed information recovery.

Despite the rapid development of superresolution technology in remote sensing, current methods remain in the SISR paradigm. Most methods improve performance by combining some techniques, such as dense connection, residual connection, attention mechanism, and adversarial loss. These networks are designed based on deep learning and the SISR's paradigm. When there is too much loss of image information or large-scale superresolution, relying solely on the network design cannot bring significant results. Therefore, RefSR adds a new route to image superresolution by reconstructing low-resolution images by introducing additional reference information, which is very consistent with remote sensing image superresolution. However, this paradigm has not been applied to the remote sensing field, which is the focus of this article.

III. Methodology

In this article, we take the image superresolution of remote sensing as our main task. Based on the existing theoretical basis, we propose T^{3}SR, a texture transfer transformer for superresolution of remote sensing images. Due to the data characteristics, a remote sensing image generically has a larger size, lower image resolution, and more similar textures than a generic scene. Therefore, unlike a generic image, a remote-sensing image of 5000*5000 size cannot be trained directly and must be cropped before use. Low-resolution means one pixel in the image corresponds to a ground size of several square meters. In addition, pixels between adjacent remotely sensed images generically have more similar textures, which can be applied to image restoration. The T^{3}SR architecture comprises three components: shallow texture extraction, reference attention, and self-attention.

The T^{3}SR network is shown in Fig. 4, where the overall structure can be distributed into two modules: the texture transfer module and the U-Transformer fusion module. In the first stage, the LR image and reference image require the following operations: feature extraction, feature correlation, and texture transfer. In the second stage, the U-Transformer module uses the self-attention strategy to remind which information needs to be retained and dropped.

To fully evaluate the performance of T^{3}SR, we chose the standard x4 magnification scale. We perform an $\times 4$ scale upsampling of the LR image for data preparation. This scale makes our method more easily transferable to other image restoration tasks, e.g., deblurring, denoising, dealiasing, and patching. The essence of superresolution is the same as in the above tasks but differs in having an additional zoom operation. Therefore, our method can also be applied to generic image restoration tasks, but the data must be prepared and the model retrained. In this article, we concentrate only on superresolution performance and not on all image restoration tasks.

A. Shallow Texture Extraction

For shallow texture, we use the pretrained weights of VGG-19. The VGG-Net is built by the Visual Geometry Group team of Oxford University. It has five models (A to E), with the E model (VGG-19) used in the ILSVRC 2014 challenge and won first place in ILSVRC positioning and second place in the classification challenge. The overall structure of VGG-19 involves the same convolution kernel size (3×3) and max pooling size (2×2). However, the excessive T^{3}SR parameters make its training difficult if the feature extraction module is retrained. Therefore, instead of retraining the VGG-19 network from zero, we directly adopt the pretrained weights of VGG-19.

First, the upsampling LR image $\mathrm{I}^{LR\uparrow}$, the upsampling after downsampling reference image $\mathrm{I}^{Ref\downarrow\uparrow}$, and the reference image I^{Ref} will pass through the pretrained weights of VGG-19. As shown in (1), (2), and (3), in this way, the feature maps $F_i^{LR\uparrow}$, $F_i^{Ref\downarrow\uparrow}$, and F_i^{Ref}(i means scale) of the corresponding layer are obtained. The overall structure of VGG-19 and some layer outputs are shown in Fig. 5. Here, we only extract feature maps from layers 2, 4, and 8, corresponding to the downsampling scale of $\times1$, $\times2$, and $\times4$

$$F_i^{LR\uparrow} = \varphi_{vgg}(I^{LR\uparrow}) \tag{1}$$

$$F_i^{Ref\downarrow\uparrow} = \varphi_{vgg}(I^{Ref\downarrow\uparrow}) \tag{2}$$

$$F_i^{Ref} = \varphi_{vgg}(I^{Ref}). \tag{3}$$

B. Reference Attention Module

The reference attention module, shown in Fig. 6, is also a variant of the transformer, which contains elements such as K, Q, and V. Different from the previous transformer, we use the transposed matrix multiplication when calculating the correlation between K and Q, which can significantly reduce the memory consumption of GPU. Therefore, the reference attention module is divided into the following four operations to reduce memory consumption.

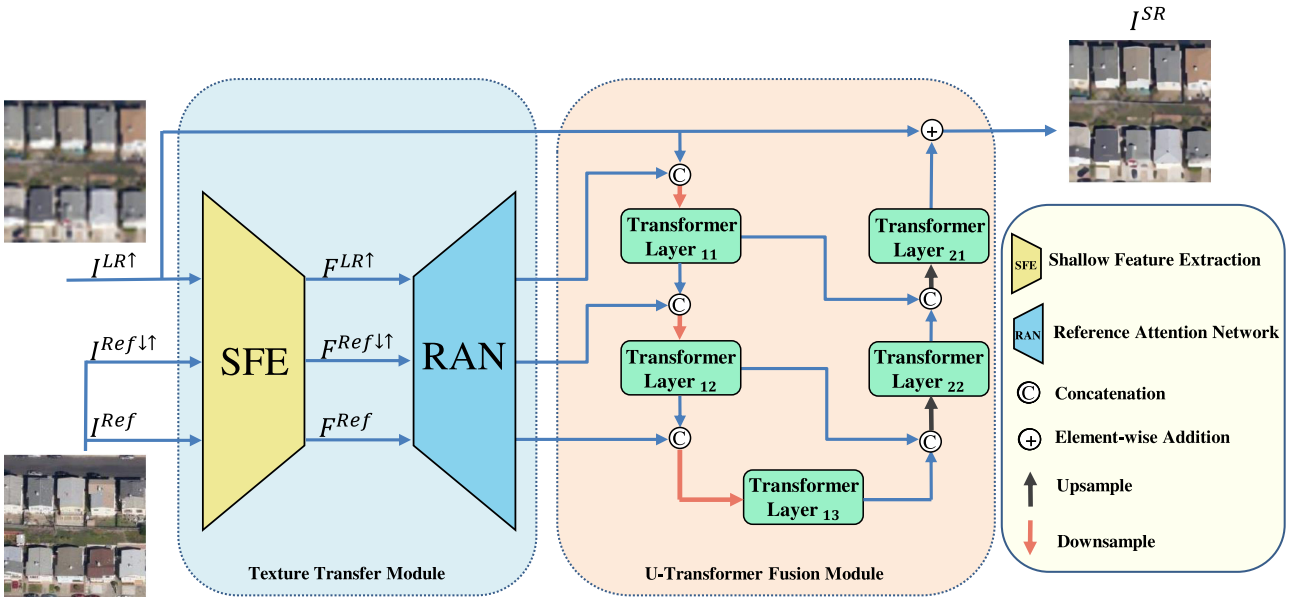

Fig. 4. Overall architecture of Texture Transfer Transformer(T^3SR), which can be divided into texture transfer module and U-Transformer fusion module.

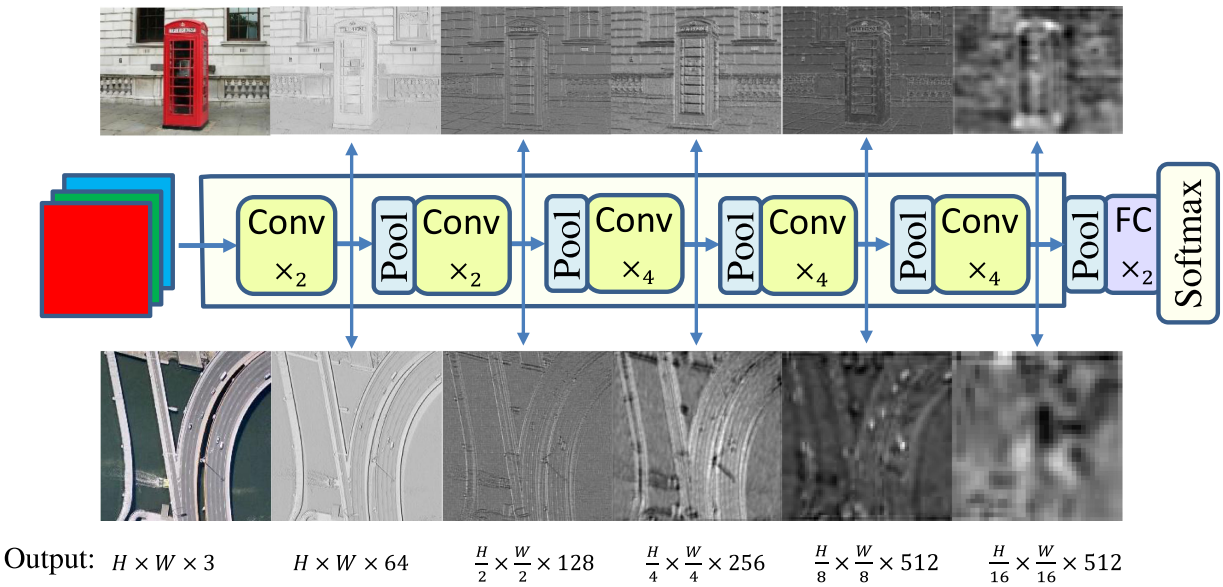

Fig. 5. Overall structure of VGG-19 and visualization of feature maps.

The inputs of the reference attention module are the VGG-19 outputs $F_i^{LR\uparrow}$, $F_i^{Ref\downarrow\uparrow}$ and F_i^{Ref}. To simplify the design of the subsequent convolution, we concatenate the feature map channelwise to obtain $F_i^{(LR\uparrow,Ref\downarrow\uparrow,Ref)} \in R^{H\times W\times 3\,C}$, as shown in (4). Then, we perform convolution and normalization on the Tensor and divide it into three blocks by channel to obtain the Value of Q, K, and V, shown in (5). The expression of normalization is shown in formula (6)

$$F_i^{(LR\uparrow,Ref\downarrow\uparrow,Ref)} = Concat(F_i^{LR\uparrow}, F_i^{Ref\downarrow\uparrow}, F_i^{Ref}) \tag{4}$$

$$Q_i, K_i, V_i = Conv(norm(F_i^{(LR\uparrow,Ref\downarrow\uparrow,Ref)})).Chunk \tag{5}$$

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} \cdot \gamma + \beta. \tag{6}$$

The attention matrix is obtained from (7), representing the weights of the migratable texture. Subsequently, the attention matrix is multiplied with V and passes through a convolution. Finally, the module output is obtained by a residual structure, as shown in (8)

$$Mat_i = K_i \cdot Q_i^T \tag{7}$$

$$F_i^{RA} = Conv(Mat_i \times V_i) + V_i. \tag{8}$$

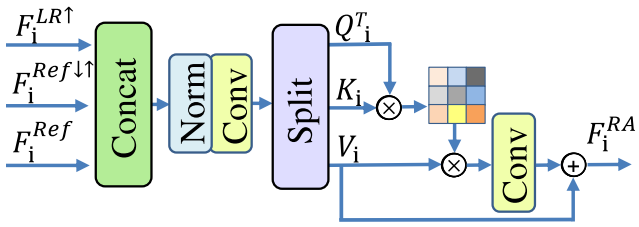

Fig. 6. Overall structure of reference attention module structure.

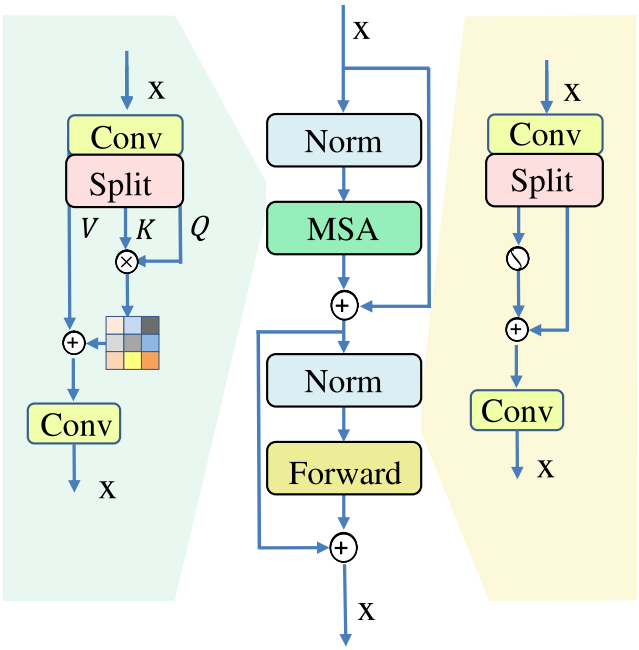

Fig. 7. Overall structure of transformer.

C. U-Transformer

A multihead attention module and a forward feedback network are the two primary components of the transformer suggested in this study. Our method is inspired by SwinIR [35] and Restormer [45], as shown in Fig. 7. The multihead attention module is similar to the reference attention module described before, but the input is different, which is a single Tensor. The data are expanded by convolution and then divided to obtain Q, K, and V. In the forward feedback network, the former multilayer perceptron (MLP) is replaced with an attention network. The input is first passed through a convolutional block equally divided into two branches. Nonlinear activation directly assigns one channel's attention weights to the other channel.

The transformer enhances image features through a self-attention mechanism. Thus, our method works under the dual attention of reference attention and self-attention. It enhances the superresolution performance and discards the strong dependence on the reference image compared to the RefSR methods, suggesting a solution different from previous SISR or RefSR methods.

Fig. 4 depicts the U-Transformer's network structure, which is the same as U-Net overall. We replaced all convolution modules with transformer modules and changed the network depth from four to three layers, corresponding to the downsampling multiples of $\times 1$, $\times 2$, and $\times 4$, respectively. First, $I^{LR\uparrow}$ is passed through a convolution group, followed by concatenation with the output of the reference attention, and then passed through convolution with an input–output channel ratio of 2:1 [see (9)]. At this point, the Tensor completes the first fusion of the scale of $\times 1$ through the transformer module [see (10)]

$$T1\downarrow_{in} = Conv_{2:1}(Concat(Conv_{1:1}(I^{LR\uparrow}), F_1^{RA})) \quad (9)$$

$$T1\downarrow_{out} = Transformer(T1\downarrow_{in}). \quad (10)$$

Then, in the second feature fusion layer, the concatenation object of the reference attention is the fusion output of the previous layer, as shown in the following:

$$T2\downarrow_{in} = Conv_{2:1}(Concat(Conv_{1:1}(T1\downarrow_{out}), F_2^{RA})) \quad (11)$$

$$T2\downarrow_{out} = Transformer(T2\downarrow_{in}). \quad (12)$$

Similarly, the third layer's fusion result is

$$T3\downarrow_{in} = Conv_{2:1}(Concat(Conv_{1:1}(T2\downarrow_{out}), F_3^{RA})) \quad (13)$$

$$T3\downarrow_{out} = Transformer(T3\downarrow_{in}). \quad (14)$$

When the network's upsampling fusion is similar to the downsampling fusion, the difference between them is that the stitched object of the upsampling fusion changes from the reference attention F_i^{RA} to the output $Ti\downarrow_{out}$. Thus, the fusion output of the last layer is presented in the following:

$$T1\uparrow_{in} = Conv_{2:1}(Concat(Conv_{1:1}(T2\uparrow_{out}), T1\downarrow_{out})) \quad (15)$$

$$T1\uparrow_{out} = Transformer(T1\uparrow_{in}). \quad (16)$$

In summary, the mapping of LR to HR image is as follows, where RA is the texture migration module:

$$I^{SR} = Fusion[RA(I^{LR\uparrow}, I^{Ref}), Transformer(I^{LR\uparrow})] + I^{LR\uparrow}. \quad (17)$$

IV. Experiment

Section IV-A introduces seven commonly used datasets, three remote sensing datasets, and four generic scene datasets. In Section IV-B, we introduce the evaluation metrics for image superresolution, while in Sections IV-C and IV-D, we describe the environment and dataset processing for this experiment. Finally, we illustrate our technique's quantitative evaluation and visual comparison with state-of-the-art methodologies.

A. Datasets

For the first time in remote sensing, we offer an image superresolution approach based on a reference image, as the RefSR paradigm cannot be found in remote sensing. Therefore, our experiments include datasets of remote sensing and generic scenes to evaluate our method. The remote sensing datasets

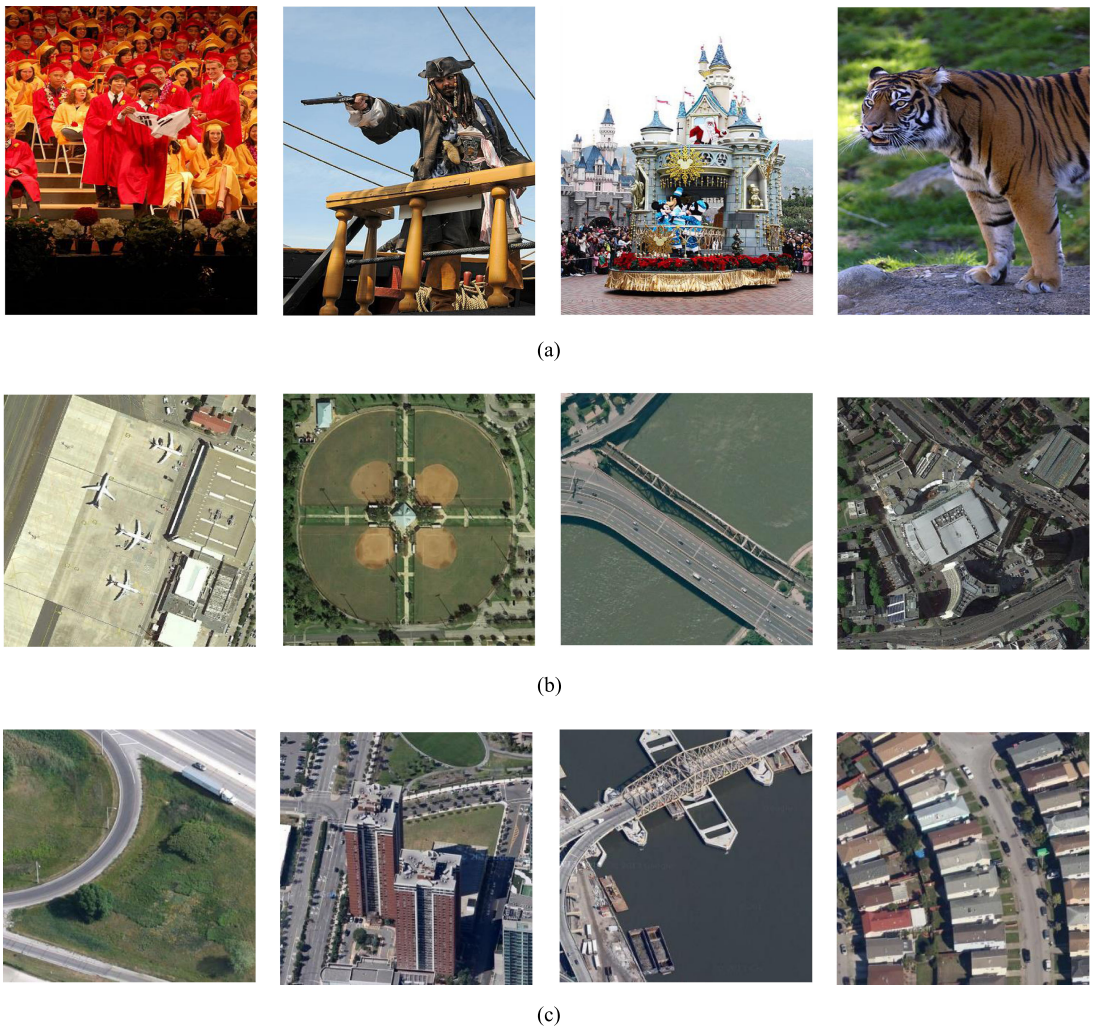

Fig. 8. Example image of the datasets, where group (a) is selected from CUFED, (b) from AID, and (c) from RSSCN7.

TABLE I
SOME ATTRIBUTES OF DATASETS USED IN OUR EXPERIMENTS

Scenes	Use	Dataset	Number	Class	Size
RS	Train	AID	10000	30	600×600
	Test	WHU-RS19	1001	19	600×600
		RSSCN7	2800	7	400×400
Com	Train	CUFED	11871	-	160×160
	Test	Manga109	109	-	-
		Sun80	80	-	-
		Urban100	100	-	-

contain AID [46], WHU-RS19, and RSSCN7, where AID is used for training and WHU-RS19 and RSSCN7 for testing. The datasets in the generic scene are CUFED, Manga109, Sun80, and Urban100, where CUFED is used for training and CUFED5, Manga109, Sun80, and Urban100 for testing. Fig. 8 depicts an example picture of the datasets, and Table I shows the details of the experimental datasets.

AID: This large-scale aerial photography collection is compiled from Google Earth sample photos, with 30 categories and 10 000 images. Airport, Bare-Field, Baseball-Field, Beach, Bridge, Center, Church, Commercial, Dense-Residential, Desert, Farmland, Forest, Industry, Grassland, Medium-Residential, Mountain, Park, Parking-Lot, Playgrounds, Ponds, Ports, Railway Stations, Resorts, Rivers, Schools, Sparse Dwellings, Plazas, Stadiums, Storage Tanks, and Viaducts are among the 30 aerial scene types available. It is worth noting that, despite the Google Earth photographs being postprocessed using RGB reconstructions of the original optical aerial image, there is no discernible difference between the Google Earth image and the genuine optical aerial image. As a result, Google Earth image may be used to test the efficacy of picture superresolution techniques.

RSSCN7: It comprises 2800 photos of remote sensing scenes, each of 400 x 400 pixels size. The images are from Grassland, Forest, Farmland, Parking Lot, Residential Area, Industrial Area, River, and Lake, to mention a few. Because this dataset was acquired over several seasons and weather conditions, it represents real-world remote sensing photography under various settings.

WHU-RS19: It is a collection of satellite imagery exported from Google Earth at a resolution of 0.5 m. It contains satellite images of 19 categories of scenes, including Airports, Beaches, Bridges, Businesses, Deserts, Farmlands, Football Fields, Forests, Industries, Grass, Mountains, Parks,

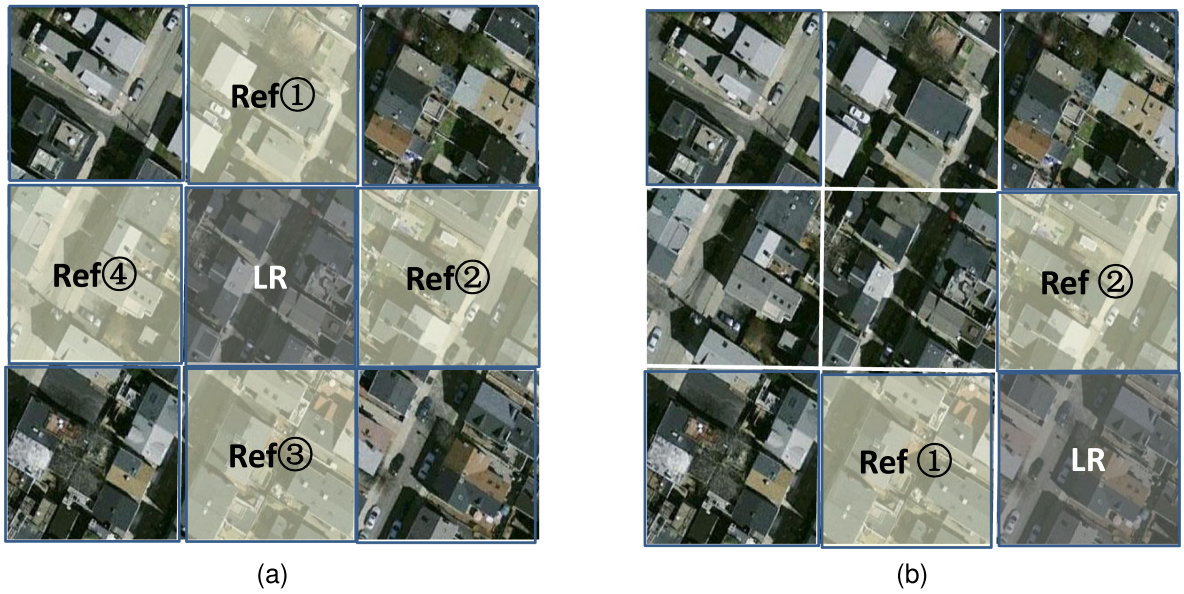

Fig. 9. Reference image generation method, where (a) represents the generic case and (b) the special case where the LR image is located at the edge position.

TABLE II
EXPERIMENTAL ENVIRONMENT

Configure	Parameter
System	Ubuntu 20.04.5 LTS
Language	Python 3.7
Framework	Pytorch 1.10
GPU	Nvidai GeForce RTX 3090
CPU	Inter(R) Core(TM) i7-11700K @3.60GHz
CUDA	11.3.58
CUDNN	v8.2.1.32

TABLE III
OVERALL PSNR (DB) AND SSIM VALUES OF T^{3}SR IN WHU-RS19 AND RSSCN7

Datasets	**RSSCN7**	**WHU-RS19**
Methods	PSNR / SSIM	PSNR / SSIM
Bicubic	27.87 / .679	29.29 / .747
SRCNN	28.41 / .710	30.06 / .783
VDSR	28.85 / .731	30.74 / .807
RDN	29.10 / .741	31.20 / .825
D-DBPN	29.23 / .747	31.36 / .826
RCAN	29.25 / .747	31.32 / .822
SRFBN	29.22 / .745	31.35 / .824
SAN	29.26 / .749	31.38 / .827
MHAN	29.28 / .750	31.40 / .828
T^{3}SR	29.61 / .754	32.19 / .832

Parking Lots, Ponds, Ports, Train Stations, Houses, Rivers, and Viaduct.

B. Evaluation Indicators

The peak signal-to-noise ratio (PSNR) is the ratio of signal power to noise power. It is frequently used as an objective metric for picture quality evaluation, such as image compression and recovery, defined as

$$PNSR = 10\log_{10}\frac{(2^N-1)^2}{MSE} \tag{18}$$

$$MSE = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}[X(i,j)-Y(i,j)]^2. \tag{19}$$

MSE represents the mean square error of pixel values between the true and reconstructed images. The height and width of the image are represented by H and W, respectively. N is the number of bits per pixel, and the default value for RGB image is 8. PSNR is calculated in decibels (dB), and the larger the value, the smaller the distortion.

Structural similarity (SSIM), a measure of the similarity of two images, was proposed by the Laboratory for Image and Video Engineering at the University of Texas at Austin. SSIM evaluates the similarity of two images in terms of brightness, contrast, and color structure. The formula is shown below

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2+\mu_y^2+c_1} \tag{20}$$

$$c(x,y) = \frac{2\sigma_{xy} + c_2}{\sigma_x^2+\sigma_y^2+c_2} \tag{21}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y+c_3} \tag{22}$$

$$SSIM = l(x,y)^2\cdot c(x,y)^2\cdot s(x,y)^2. \tag{23}$$

The brightness comparison is $l(x,y)$, the contrast comparison is $c(x,y)$, the structural comparison is $s(x,y)$ and c_1, c_2, c_3 are constants.

C. Implementation Details

The experiment was conducted using a GeForce RTX 3090 GPU with 24 GB of memory. The computer operating system is Ubuntu 20.04.5 LTS and the programming environment is a

TABLE IV
AVERAGE PSNR (DB) AND SSIM VALUES FOR EACH CLASS IN RSSCN7

Methods / Class	SRCNN	VDSR	RDN	D-DBPN	RCAN	SRFBN	SAN	MHAN	T^{3}SR
	RSSCN7(PSNR / SSIM)								
Grass	33.49/.815	34.06/.827	34.34/.833	34.31/.833	34.24/.831	34.32/.833	34.34/.834	34.34/.834	34.57/.835
Filed	32.62/.739	33.09/.751	33.38/.759	33.36/.759	33.32/.757	33.36/.758	33.38/.759	33.39/.759	33.72/.761
Industry	24.86/.669	25.75/.715	26.41/.745	26.39/.744	26.14/.732	26.36/.743	26.42/.746	26.43/.745	26.75/.753
River Lake	30.33/.804	30.92/.822	31.23/.831	31.22/.831	31.11/.828	31.21/.830	31.24/.832	31.24/.832	32.25/.830
Forest	28.05/.609	28.28/.629	28.39/.639	28.38/.638	28.35/.636	28.38/.637	28.41/.640	28.40/.640	28.47/.645
Resident	24.14/.649	24.89/.694	25.37/.720	25.36/.719	25.15/.707	25.33/.717	25.39/.721	25.39/.721	25.60/.732
Parking	24.33/.633	25.11/.679	25.59/.708	25.58/.707	25.36/.694	25.55/.705	25.46/.706	25.61/.709	25.94/.719

TABLE V
AVERAGE PSNR (DB) AND SSIM VALUES FOR EACH CLASS IN WHU-RS19

Methods / Class	SRCNN	VDSR	RDN	D-DBPN	RCAN	SRFBN	SAN	MHAN	T^{3}SR
	WHU-RS19(PSNR / SSIM)								
Airport	27.27/.766	28.28/.807	28.98/.830	29.00/.829	28.77/.821	28.99/.828	29.01/.831	29.02/.830	30.27/.836
Beach	44.42/.977	46.90/.980	47.25/.981	47.37/.981	47.32/.981	47.35/.981	47.39/.981	47.42/.981	49.69/.982
Bridge	33.18/.893	34.34/.908	35.44/.918	35.56/.918	35.25/.916	35.58/.918	35.54/.919	35.56/.919	38.56/.922
Commercial	24.33/.686	25.13/.733	25.69/.761	25.71/.760	25.55/.749	23.70/.758	25.71/.762	25.74/.762	26.01/.768
Desert	40.16/.929	40.65/.933	40.64/.936	40.75/.936	40.83/.936	40.76/.936	40.77/.936	40.77/.937	41.77/.937
Farmland	36.43/.869	37.03/.881	37.63/.893	37.67/.892	37.55/.889	37.67/.891	37.31/.894	37.30/.894	37.92/.895
Football Field	27.62/.778	28.86/.823	29.97/.854	30.20/.853	29.62/.842	29.92/.850	30.02/.856	30.03/.855	30.95/.860
Forest	28.15/.654	28.54/.686	28.74/.696	28.31/.695	28.67/.691	28.69/.693	28.71/.698	28.31/.698	28.87/.702
Industrial	26.24/.731	27.38/.780	28.35/.812	28.32/.811	28.01/.799	28.16/.808	28.29/.813	28.30/.813	28.65/.820
Meadow	37.17/.865	37.57/.872	37.82/.875	37.73/.875	37.77/.874	37.80/.874	37.82/.875	37.81/.875	37.93/.876
Mountain	24.93/.592	25.31/.625	25.49/.635	25.48/.635	25.43/.631	25.45/.633	25.50/.637	25.51/.637	25.72/.646
Park	27.84/.720	28.45/.750	28.90/.767	28.94/.766	28.78/.761	28.92/.764	29.92/.763	28.93/.763	29.26/.774
Parking	26.29/.798	27.34/.841	29.24/.884	29.33/.882	31.52/.866	29.11/.878	29.21/.887	29.23/.887	30.43/.898
Pond	31.78/.864	32.42/.875	32.84/.882	32.81/.881	32.30/.879	32.81/.881	32.81/.882	32.84/.882	34.30/.880
Port	26.60/.811	27.58/.844	28.61/.865	28.59/.865	28.19/.856	28.49/.863	28.60/.867	28.61/.866	30.20/.869
Railway Station	25.69/.669	26.58/.723	27.61/.762	27.61/.761	27.29/.747	27.60/.759	27.63/.764	27.63/.763	28.07/.773
Residential	24.10/.710	25.36/.772	26.14/.804	26.31/.803	25.96/.790	26.17/.800	26.19/.803	26.21/.803	26.67/.816
River	28.60/.733	29.11/.759	29.39/.768	29.38/.768	29.31/.765	29.37/.767	29 37/.768	29.38/.768	29.73/.771
Viaduct	25.19/.692	26.38/.749	27.26/.790	27.28/.79	27.11/.773	27.30/.788	27.37/.790	27.30/.790	27.82/.800

TABLE VI
AVERAGE PSNR (DB) AND SSIM VALUES IN GENERIC SCENE DATASETS

Datasets / Methods	CUFED5	Urban100	Manga1090	Sun80
	PSNR / SSIM			
CrossNet	25.48 / .764	25.11 / .764	23.36 / .741	28.52 / .793
SRNTT-rec	26.24 / .784	25.50 / .783	28.95 / .885	28.54 / .793
SRNTT	25.61 / .764	25.09 / .774	27.54 / .862	27.59 / .756
TTSR-rec	27.09 / .804	25.87 / .784	30.09 / .907	30.02 / .814
TTSR	25.53 / .765	24.62 / .747	28.70 / .886	28.59 / .774
T^{3}SR-self	25.79 / .740	26.45 / .733	29.14 / .866	31.80 / .799
T^{3}SR	27.13 / .805	28.60 / .803	31.82 / .912	32.54 / .814

deep learning framework based on Python 3.7, Pytorch 1.10. The detailed experimental environment is reported in Table II.

Before the model training, we process the datasets, including partitioning the train and test data and selecting LR images corresponding to the reference image. Yang et al. proposed CUFDE, where each pair is generated by SIFT feature matching to obtain different levels of similarity. Each image pair has overlapping regions, which is difficult to achieve in reality. In order to reduce the influence of the matching degree, the model has a stronger generalization ability. Our selection of reference images is shown in Fig. 9.

A 5000×5000 remote sensing image obviously cannot be input directly to the network because a too large size will cause the memory to overfill and stop working. Therefore, first, we need to slide the image to get a subimage of 200×200 and generate an LR image through bicubic interpolation from the HR image. During the sliding process, we randomly select the subimage adjacent in the top, bottom, left, and right directions as the reference image [see Fig. 9(a)]. Of course, if the subimage is located at the edge of the image, we select the direction where the subimage exists as the reference image [see Fig. 9(b)]. The reference image produced by our method makes the matched pairs have some similarity in texture and no overlapping regions. Therefore, our matching pairs are closer to real-life practical applications, making the model focus more on texture similarity than the same texture.

D. Result

This section verifies the superiority and effectiveness of the proposed method in a comparative experiment and an ablation study. First, we compare T^{3}SR with other outstanding methods under public remote sensing datasets in the comparative experiments. Next, in the ablation study, we verify the effectiveness of each network's module by comparing the methods based on RefSR on the generic scenes because there is no RefSR paradigm method for remote sensing.

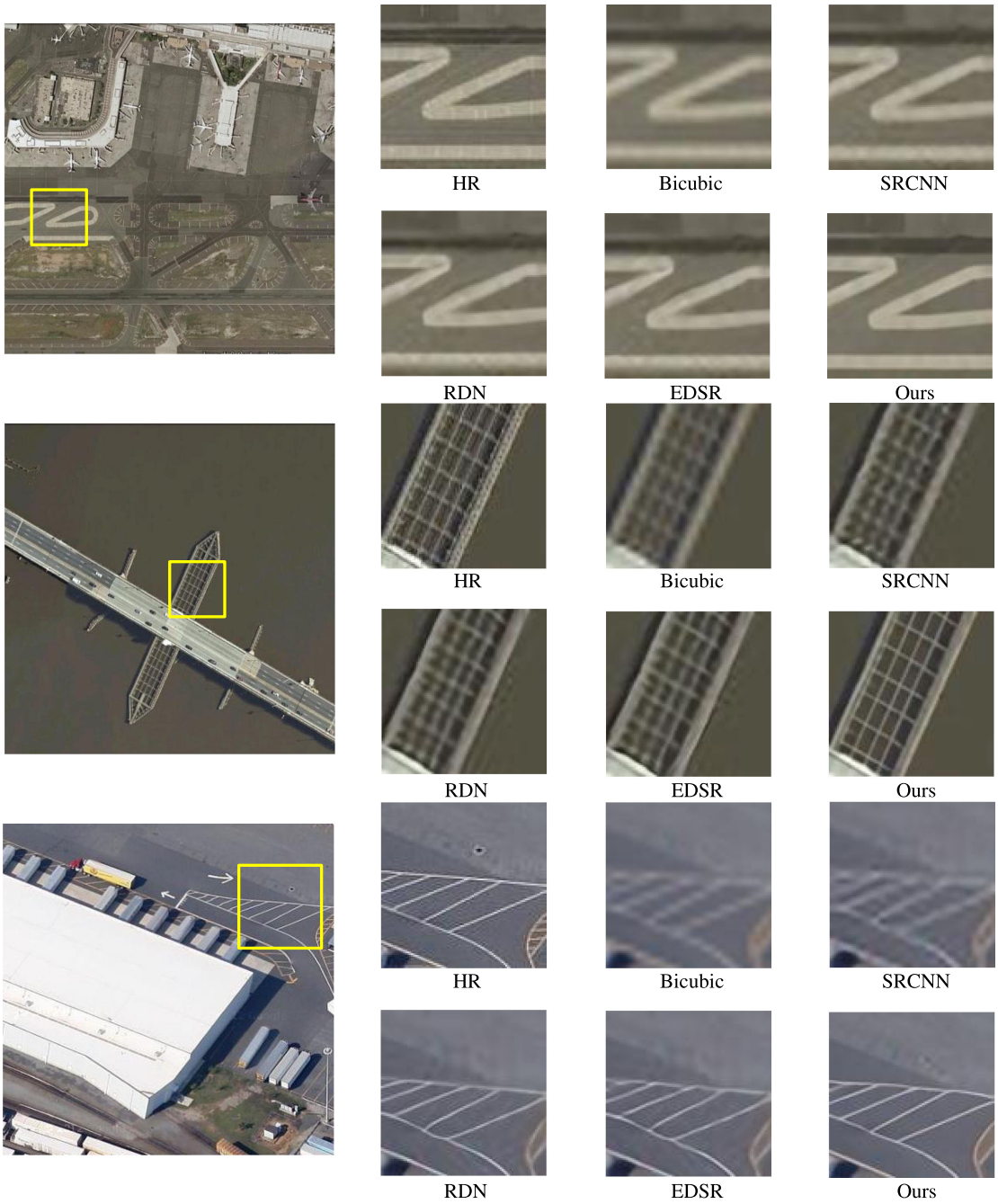

Fig. 10. Visualization of T[3]SR in RSSCN7 dataset.

1) Comparative Experiment: In this section, we conduct experiments in the remote sensing scene as a benchmark. Among them, AID is the training data, and the training datasets contain 30 categories, with a total of 10 000 images of 600 × 600. WHU-RS19 and RSSCN7 are test dataset. Among them, the test dataset WHU-RS19 contains 19 categories and 1013 images of 600×600 size. The test dataset RSSCN7 contains seven categories and 2800 images of 400 × 400. Before training, the selection of reference images has been described in detail in Section IV-C, as illustrated in Fig. 9.

Quantitative evaluation: We compare T[3]SR with classic and as state-of-the-art remote sensing image super-resolution methods, including the SRCNN [13], VDSR [15], RDN [47], RCAN [48], SRFBN [49], SAN [50], D-DBPN [40], and MHAN [51]. All experiments are performed with an image superresolution scale factor of ×4. In order to make the method argument more abundant, first, the test datasets WHU-RS19 and RSSCN7 were used to obtain the overall performance. Next, we evaluate each category on both datasets. Both jointly prove this method's effectiveness and advancement in image superresolution in remote sensing.

The quantitative evaluation results in remote sensing are shown in Tables III–V. Table III compares the overall PSNR and SSIM of T[3]SR in the WHU-RS19 and RSSCN7 remote

Fig. 11. Visualization of T^{3}SR in WHU-RS19 dataset.

sensing datasets. Table IV shows the PSNR and SSIM evaluation results under different categories in the RSSCN7 test dataset. Similarly, Table V shows the evaluation results of this method in WHU-RS19. The red number represents the highest score, and the blue number represents the second highest score. As shown in the comparison results in Table III, T^{3}SR improves by 0.79 dB and 0.33 dB, respectively, compared with MHAN on RSSCN7 and WHU-RS19.

The experimental results show that T^{3}SR achieves the best results in almost all categories in Tables IV and V. Compared with the score of second place, the most significant improvement of PSNR is Bridge, which is 3 dB, and the greatest improvement of SSIM is Residential, which is 1.3%. The Beach's highest PSNR and SSIM scores are 49.69 dB and 98.2%, respectively. Because the texture of Beach is smoother and simpler, T^{3}SR easily obtains the highest PSNR and SSIM scores. The reader is referred to Tables IV and V for detailed data.

Qualitative evaluation: In order to more fully demonstrate the effectiveness of the proposed method, a visual comparison is also carried out in the remote sensing scene, as shown in Figs. 10 and 11. The visual comparison results show that compared with various advanced image superresolution methods, the proposed method restores the result with clearer detailed information and a richer texture outline. For a more precise comparison, a single yellow rectangle is used to mark the enlarged subplots, which is done for the results of all methods. As shown in the display results, the results obtained by T^{3}SR are more realistic than other methods, with clearer lines and better visual effects.

As shown in the figure above, the bicubic upsampling strategy leads to texture loss and structural blurring, which is more pronounced for remote sensing images with weak edge details. For early image superresolution algorithms such as SRCNN, because the network depth is not enough, it often leads to poor recovery, and the shape of some small objects cannot be recovered. Other methods can lead to hallucinations and artifacts, which are also unacceptable, producing erroneous structural and texture information, failing to recover more details, and ultimately resulting in poor image quality for image superresolution. In contrast, the developed method recovers more details and textures, especially in remote sensing, where it achieves the visual effect of the GT image. Even some GT image defects due to imaging conditions show that our method can be optimized to achieve better visual effects.

2) Ablation Study: For the ablation study, we conduct experiments from two perspectives. To verify the effectiveness of T^{3}SR in the paradigm of the RefSR methods, we conduct comparative experiments based on a generic scene, as this article introduces

TABLE VII
CONTRIBUTION OF EACH MODULE

Module \ Method	TTM	TTM+U-Net	TTM+U-Trans
TTM	√	√	√
U-net		√	
U-Trans			√
PSNR	31.80	31.83	32.19
SSIM	0.821	0.823	0.832

the superresolution method of RefSR to remote sensing scenes for the first time. Moreover, we use each module as a variable to further explore the effectiveness of each constituent module. Fig. 9 illustrates the reference image generation.

In the first stage, we adopt the CrossNet, TTSR, and SRNTT approaches as benchmarks to prove the effectiveness of our RefSR approach. CUFED is selected as training data, and the training set contains 11871 reference matching pairs. CUFED5, Manga109, Sun80, and Urban100 are test data. Among them, the CUFED5 dataset contains 126 test images. The Urban100 dataset contains 100 images of urban scenes. The Sun80 dataset has 80 natural images, and the Manga109 contains 109 manga volumes.

As shown in Table VI, our method is compared with the most advanced RefSR methods, including CrossNet [6], SRNTT [7], and TTSR [8]. To demonstrate that our method can reduce the dependence on reference images, we implement a self-reference test (T^{3}SR-self) where the reference image is the upsampled image of the LR image.

All experiments are performed on a scale of x4. Compared to the state-of-the-art, T^{3}SR improves by 0.04 dB, 2.73 dB, 1.73 dB, and 2.52 dB for the CUFED5, Manga109, Sun80, and Urban100, respectively. The red number represents the highest score, while the blue represents second place (see Table VI). T^{3}SR surpasses the current state-of-the-art superresolution methods in the RefSR paradigm. Notably, the T^{3}SR-self result indicates that our approach performs well without a reference image. The quantitative comparison shows that T^{3}SR has state-of-the-art performance and can reduce the severe dependence on the reference image.

In the second stage, we verify the effectiveness of each module. We disassemble T^{3}SR into Texture Transfer Module (TTM) and U-Transformer Fusion Module (U-Trans). In addition, since U-Trans is an improvement of U-Net, it is combined by self-attention rather than using the original CNN. Therefore, this method also sets U-Net variables to verify the optimal solution of each module. All experiments were performed at a magnification factor of $\times 4$ under the dataset WHU-RS19.

As shown in Table VII, each component has its contribution to the overall performance improvement. It can be seen that the combination of TTM and U-Net can bring a specific improvement compared to TTM, but it does not bring significant progress. However, compared to TTM, the combination of TTM and U-Trans improves PSNR and SSIM by 0.39 dB and 1.1%, respectively. Finally, U-Trans can also bring better performance to the network than U-Net, showing that self-attention is better than traditional CNN.

V. CONCLUSION

In this article, we propose T^{3}SR, an end-to-end superresolution network called texture transfer transformer for remote sensing image superresolution. T^{3}SR pioneered the introduction of image texture transfer into remote sensing and achieved the current state-of-the-art results. Specifically, T^{3}SR divides superresolution into two stages: texture migration and deep fusion. First, we design a texture transfer module to serve the transfer of shallow semantic information. Next, we propose a U-Transformer, which serves the feature fusion of deep networks and reduces the model's dependence on the reference image. Finally, we conduct numerous experiments on standard public datasets to fully evaluate our approach. The experimental results show that T^{3}SR can improve the texture details of reconstructed pictures and the visual effects. Moreover, the scores of the quantization parameters PSNR and SSIM are also better than the state-of-the-art methods. In addition, we add self-referencing experiments, demonstrating that our method still achieves excellent results when there is no reference image or the correlation between the reference image, and the reconstructed image is weak.

ACKNOWLEDGMENT

The authors are very grateful for the help or encouragement of their colleagues, the extraordinary work of the technical staff, and the organization's financial support.

REFERENCES

[1] X. Yue et al., "Super-resolution network for remote sensing images via preclassification and deep–shallow features fusion," *Remote Sens.*, vol. 14, no. 4, 2022, Art. no. 925.
[2] M. Zhao, J. Ning, J. Hu, and T. Li, "Hyperspectral image super-resolution under the guidance of deep gradient information," *Remote Sens.*, vol. 13, no. 12, 2021, Art. no. 2382.
[3] H. Huan et al., "End-to-end super-resolution for remote-sensing images using an improved multi-scale residual network," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 666.
[4] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1432.
[5] E. Koester and C. S. Sahin, "A comparison of super-resolution and nearest neighbors interpolation applied to object detection on satellite data," 2019, *arXiv:1907.05283*. [Online]. Available: https://arxiv.org/abs/1907.05283
[6] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 88–104.
[7] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7982–7991.
[8] 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.
[9] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.
[10] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 372–386.
[11] K. Nasrollahi and T. B. Moeslund, "Super-resolution: A comprehensive survey," *Mach. Vis. Appl.*, vol. 25, no. 6, pp. 1423–1468, 2014.

[12] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learning-based single image super-resolution," *PeerJ Comput. Sci.*, vol. 7, 2021, Art. no. e621.
[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[15] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.
[16] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.
[17] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 517–532.
[18] C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C.-W. Lin, "Coarse-to-fine CNN for image super-resolution," *IEEE Trans. Multimedia*, vol. 23, pp. 1489–1502, 2020.
[19] X. Kong, H. Zhao, Y. Qiao, and C. Dong, "Classsr: A general framework to accelerate super-resolution networks by data characteristic," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12016–12025.
[20] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.
[21] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Inc. Advances in Neural Information Processing Systems*, Curran Associates, vol. 32, 2019.
[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*. [Online]. Available: https://arxiv.org/abs/2010.11929
[23] B. Wu et al., "Visual transformers: Token-based image representation and processing for computer vision," 2020, *arXiv:2006.03677*. [Online]. Available: https://arxiv.org/abs/2006.03677
[24] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[25] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*. [Online]. Available: https://arxiv.org/abs/2104.05707
[26] Y. Liu, G. Sun, Y. Qiu, L. Zhang, A. Chhatkuli, and L. Van Gool, "Transformer in convolutional neural networks," 2021, *arXiv:2106.03180*. [Online]. Available: https://arxiv.org/abs/2106.03180
[27] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12894–12904.
[28] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
[29] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.
[30] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[31] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[32] H. Cao et al., "Swin-unet: Unet-like pure transformer for medical image segmentation," 2021, *arXiv:2105.05537*. [Online]. Available: https://arxiv.org/abs/2105.05537
[33] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[34] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[35] 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.
[36] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021, *arXiv:2102.04432*. [Online]. Available: https://arxiv.org/abs/2102.04432
[37] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310.
[38] Z. Wang, X. Cun, J. Bao, and J. Liu, "Uformer: A general u-shaped transformer for image restoration," 2021, *arXiv:2106.03106*. [Online]. Available: https://arxiv.org/abs/2106.03106
[39] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.
[40] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1664–1673.
[41] Z. Pan, W. Ma, J. Guo, and B. Lei, "Super-resolution of single remote sensing image based on residual dense backprojection networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7918–7933, Oct. 2019.
[42] J. Gu, X. Sun, Y. Zhang, K. Fu, and L. Wang, "Deep residual squeeze and excitation network for remote sensing image super-resolution," *Remote Sens.*, vol. 11, no. 15, 2019, Art. no. 1817.
[43] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.
[44] W. Ma, Z. Pan, F. Yuan, and B. Lei, "Super-resolution of remote sensing images via a dense residual generative adversarial network," *Remote Sens.*, vol. 11, no. 21, 2019, Art. no. 2578.
[45] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," 2021, *arXiv:2111.09881*. [Online]. Available: https://arxiv.org/abs/2111.09881
[46] 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.
[47] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[48] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[49] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3867–3876.
[50] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 11065–11074.
[51] D. Zhang, J. Shao, X. Li, and H. T. Shen, "Remote sensing image super-resolution via mixed high-order attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5183–5196, Jun. 2021.

Durong Cai received the B.S. degree in electronic information science and technology from the School of electronic information and electrical engineering, Huizhou University, Guangdong, China, in 2020 and the M.S. degree in electronic information with the School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China, in 2022.

His research interests include deep learning, object detection, and super-resolution.

Peng Zhang received the B.S. degree in microwave engineering, the M.S. degree in information and communication engineering, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology, Changsha, China, in 1997, 2000, and 2004, respectively.

He is currently an Associate Professor with the School of Electronics and Communication Engineering, Shenzhen Campus of Sun Yat-Sen University, Shenzhen, China. His research interests include high resolution satellite remote sensing image processing and application, target detection and recognition in optical and SAR images, data fusion and analysis based on RS, GIS and GPS information, and medical image processing, analysis, understanding, and application.