International Journal of Digital Earth

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjde20

A lightweight distillation CNN-transformer architecture for remote sensing image super-resolution

Yu Wang, Zhenfeng Shao, Tao Lu, Lifeng Liu, Xiao Huang, Jiaming Wang, Kui Jiang & Kangli Zeng

To cite this article: Yu Wang, Zhenfeng Shao, Tao Lu, Lifeng Liu, Xiao Huang, Jiaming Wang, Kui Jiang & Kangli Zeng (2023) A lightweight distillation CNN-transformer architecture for remote sensing image super-resolution, International Journal of Digital Earth, 16:1, 3560-3579, DOI: 10.1080/17538947.2023.2252393

To link to this article: https://doi.org/10.1080/17538947.2023.2252393

Published online: 01 Sep 2023.

INTERNATIONAL JOURNAL OF DIGITAL EARTH
2023, VOL. 16, NO. 1, 3560–3579
https://doi.org/10.1080/17538947.2023.2252393

A lightweight distillation CNN-transformer architecture for remote sensing image super-resolution

Yu Wang [a,b], Zhenfeng Shao [a], Tao Lu [c], Lifeng Liu[d], Xiao Huang [e], Jiaming Wang [c], Kui Jiang [f] and Kangli Zeng [g]

[a]State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, People's Republic of China; [b]School of General Aviation, Jingchu University of Technology, Jingmen, People's Republic of China; [c]Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan, People's Republic of China; [d]Department of Engineering, University of Durham, Durham, UK; [e]Department of Environmental Sciences, Emory University, Atlanta, GA, USA; [f]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, People's Republic of China; [g]NERCMS, School of Computer Science, Wuhan University, Wuhan, People's Republic of China

ABSTRACT

Remote sensing images exhibit rich texture features and strong autocorrelation. Although the super-resolution (SR) method of remote sensing images based on convolutional neural networks (CNN) can capture rich local information, the limited perceptual field prevents it from establishing long-distance dependence on global information, leading to the low accuracy of remote sensing image reconstruction. Furthermore, it is difficult for existing SR methods to be deployed in mobile devices due to their large network parameters and high computational demand. In this study, we propose a lightweight distillation CNN-Transformer SR architecture, named DCTA, for remote sensing SR, addressing the aforementioned issues. Specifically, the proposed DCTA first extracts the coarse features through the coarse feature extraction layer and then learns the deep features of remote sensing at different scales by fusing the feature distillation extraction module of CNN and Transformer. In addition, we introduce the feature fusion module at the end of the feature distillation extraction module to control the information propagation, aiming to select the informative components for better feature fusion. The extracted low-resolution (LR) feature maps are reorganized through the up-sampling module to obtain high-resolution (HR) feature maps with high accuracy to generate high-quality HR remote sensing images. The experiments comparing different methods demonstrate that the proposed approach performs well on multiple datasets, including NWPU-RESISC45, Draper, and UC Merced. This is achieved by balancing reconstruction performance and network complexity, resulting in both competitive subjective and objective results.

ARTICLE HISTORY
Received 22 May 2023
Accepted 15 August 2023

KEYWORDS
Super-resolution; remote sensing; lightweight network; CNN-Transformer

1. Introduction

Remote sensing images are a valuable data source for obtaining ground information, with applications in both civilian and military fields. These images provide rich texture details that are crucial

CONTACT Zhenfeng Shao shaozhenfeng@whu.edu.cn State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079 People's Republic of China

for various tasks, such as change detection (Bai et al. 2022; Zhu et al. 2022), object identification (Jing Wang et al. 2022), land cover classification (Y. Xu et al. 2022; Xue et al. 2022), hyperspectral image classification (Asker 2023; Firat et al. 2023a; Fırat et al. 2023b; Fırat, Asker, and Hanbay 2022) and other fields (R. Li et al. 2022; Yu et al. 2023; Yuan et al. 2023; S. Zhang et al. 2021). However, the obtained images are often low-resolution (LR) due to limitations in transmission conditions and imaging equipment. Increasing the resolution of remote sensing images using hardware can be both time-consuming and costly. With the proposal of the comprehensive positioning, navigation, and timing system (PNT system) (Prol et al. 2022), remote sensing algorithms based on mobile terminals and edge devices have become a new research direction. Therefore, it is particularly important to design a lightweight super-resolution (SR) reconstruction technology from a software perspective to improve the resolution of remote sensing images.

The SR reconstruction techniques refer to the reconstruction of high-quality HR images from observed low-quality LR images via specific algorithms. Existing SR reconstruction techniques are divided into three major categories, i.e. interpolation (Lei Zhang and Wu 2006), reconstruction (X. Li et al. 2010), and learning-based methods (Zeng et al. 2019). Interpolation-based methods aim to insert new pixel points through the spatial position relationship between sub-pixels and surrounding pixels. Some widely used interpolation methods are nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation. These methods are relatively easy to implement and offer fast performance. Reconstruction-based methods aim to mathematically combine the overall information of an HR image to reconstruct it; commonly used methods include iterative back-projection (Qiu, Cheng, and Wang 2021), projection onto convex sets (B. Wang et al. 2015), and maximization of a posteriori estimation (Jakhetiya et al. 2016). Learning-based methods obtain some prior knowledge to guide image reconstruction by learning the mapping relationship between LR images and HR images. Common learning-based methods include flow learning and sparse coding methods. These traditional SR reconstruction methods mainly rely on the construction of constraint terms and the accuracy of inter-image alignment to achieve the reconstruction effect. However, these methods are not suitable for SR reconstruction with large magnification, and the reconstruction results often own issues such as blurred edge texture and low-quality details.

With the continuous development of deep learning, convolutional neural networks (CNN) have shown great potential in the field of image SR reconstruction thanks to their powerful fitting ability, which has attracted wide attention from scholars. Dong et al. (2014) first applied CNN to an image SR reconstruction task and completed image feature extraction using three-layer convolution, which obtained great reconstruction results compared with traditional SR reconstruction methods. However, the interpolation of the LR images before the image feature extraction operation increases the computational demand, and the errors by the interpolation process introduce uncertainty to the reconstruction results. To address the reconstruction problem caused by interpolation amplification, Dong, Change Loy, and Tang (2016) further proposed the FSRCNN algorithm. The algorithm solves the interpolation problem by adding deconvolution layers to the network, increasing network depth, and adjusting the convolutional kernel size to improve the reconstruction effectiveness and training speed. Kim, Lee, and Lee (2016) first applied a recurrent neural network to SR tasks, which solves the problem of model gradient explosion (or disappearance) using a recursive supervision strategy and the idea of residual learning. Shi et al. (2016) proposed an image SR method based on sub-pixel convolution layers, which computes convolution directly on LR image features to obtain HR images and achieves satisfactory performance in speed and reconstructed results. Ledig et al. (2017) first introduced generative adversarial networks into SR tasks. The model consists of a deep neural network structure with two networks that pit the generative network and the discriminative network against each other. These two networks are trained iteratively on each other, leading to recovered high-frequency information in SR tasks.

The deep learning-based SR algorithms mentioned above are primarily designed for reconstructing natural images. However, remote sensing images present unique challenges due to their high spatial distribution, varied sizes and shapes of ground objects, and the need to extract high-

frequency information for effective SR reconstruction. Therefore, these natural image-based SR algorithms cannot be directly applied to the task of remote sensing image reconstruction. To this end, Lei, Shi, and Zou (2017) proposed a new remote sensing image SR algorithm with a combined local-global network that adopts a multinomial multifork structure to learn the multilevel representation of remote sensing images. W. Xu et al. (2018) designed a remote sensing SR algorithm based on a depth-storage connection network, which enables the reconstruction network that demonstrates a better reconstruction capability by efficiently combining image details with environmental information. Jiang, Wang, Yi, Jiang, Xiao et al. (2018) designed a deep distillation recurrent network, which extracts remote sensing image features by combining ultra-dense residual blocks and multi-scale purification units and further promotes feature representation through a distillation mechanism. Dong et al. (2019) employed a new multi-perceptual attention network and a migration learning strategy to reconstruct remote sensing images by combining enhanced residual blocks and residual channel attention mechanisms. He et al. (2022) proposed a ResNet-based dense spectral Transformer to achieve spectral SR of multispectral remote sensing images. The network combines the Transformer with ResNet to meet the needs of learning long-range relationships for remote sensing images. Zhilin Wang, Shang, and Wang (2022) proposed a remote sensing image SR network based on a swin Transformer fusion attention network. The network uses a swin Transformer module with a fused attention mechanism to extract high-frequency information and uses a gradient method to extract edge details of the image, which effectively enhances the network's ability to reconstruct details.

The above-mentioned remote sensing image SR algorithm based on deep learning mainly extracts deep features of remote sensing images through the CNN network. However, since CNN lacks the ability to model long-range dependencies, it may not handle global features in remote sensing images well. In contrast, Transformer is a neural network suitable for sequence data, which can effectively model dependencies between sequences. In remote sensing image SR, Transformer can help the model learn the global features of the image, so as to better improve the performance of the model. Therefore, combining CNN and Transformer can merge their advantages to enhance the performance of remote sensing image SR models. At the same time, the existing Transformer-based remote sensing SR methods have too many network parameters and a large amount of calculation, which makes it challenging to deploy them on related hardware. To address these challenges, we propose a lightweight distillation CNN-Transformer architecture for remote sensing SR, called DCTA, which mainly solves the problem of insufficient feature information and a large amount of calculation when combining CNN and Transformer for remote sensing image reconstruction. As shown in Figure 1, our proposed DCTA network achieves the highest reconstruction performance on the NWPU-RESISC45 dataset with only 0.53M parameters. Specifically, the DCTA network uses fewer layers and filters, which significantly reduces the number of network parameters. Moreover, considering the particularity of remote sensing images, this network combines CNN and Transformer modules to effectively extract remote sensing image features.

The primary contributions of this paper can be summarized as follows:

(1) We propose a lightweight remote sensing SR network that mainly extracts deep multi-scale features by combining CNN and Transformer networks to enhance high-frequency feature representation for remote sensing image reconstruction.
(2) We create a novel distillation CNN-Transformer block (DCTB). This module extracts deep features from remote sensing images at various scales, effectively enhancing the network's perceptual field and fully extracting global feature information.
(3) We design a lightweight dual attention distillation block (DADB), featured by its computational efficiency. This module connects enhanced contrast channel residual attention (ECCRA) and spatial residual attention (ESRA) to extract enhanced multi-scale features in each channel direction and spatial scale.

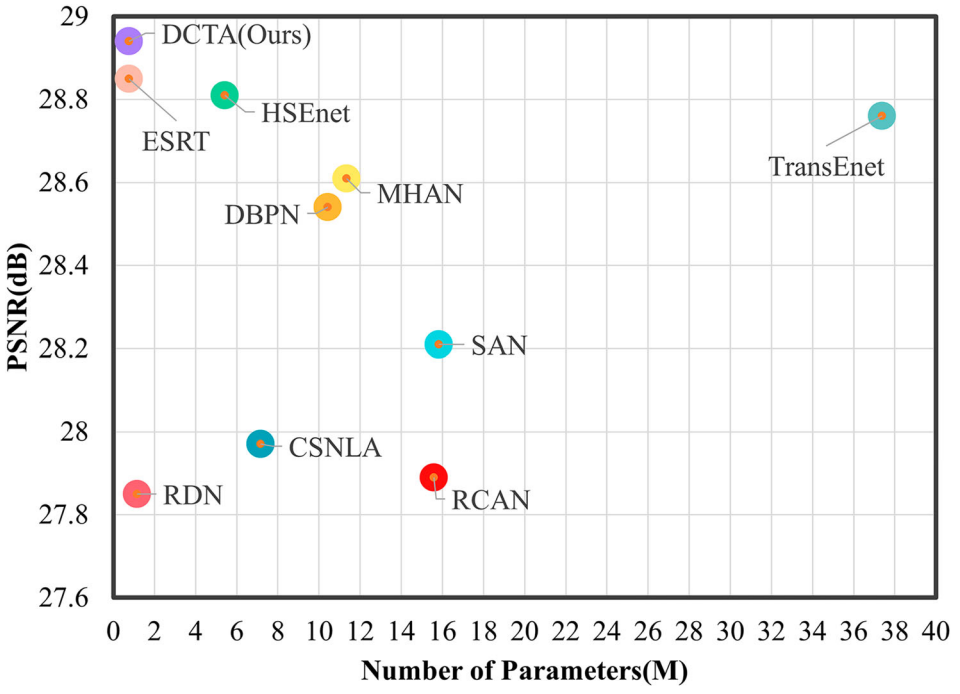


Figure 1. The number of model parameters and PSNR value trade-offs is compared with other SR algorithms on NWPU-RESISC45 datasets for ×4 SR.

(4) While keeping fewer network parameters and floating-point operators, the designed DCTA network achieves state-of-the-art remote sensing image reconstruction performance on three remote sensing datasets. Related ablation experiments also verified the validity of DCTA's structure design.

The rest of this paper is organized as follows: Section 2 provides a summary of related works on CNN-based and Transformer-based SR for remote sensing images. Section 3 elaborates on the detailed structure of the proposed DCTA network. Section 4 covers the experimental design, including a comparison of the subjective and objective experimental results as well as ablation experiments. Finally, Section 5 summarizes the study.

2. Related work

In this section, we provide an overview of the research related to remote sensing image SR using both CNN-based and Transformer-based methods.

2.1. CNN-based remote sensing image SR

The CNN-based algorithms have shown remarkable performance in various remote sensing image tasks, such as change detection (D. Wang et al. 2023; Zhu et al. 2022), pan-sharpening (Jiaming Wang, Shao, Huang, Lu, and Zhang 2022), super-resolution (Jiaming Wang, Shao, Huang, Lu, Zhang, and Li 2022), owing to their advanced feature representation capabilities. As a result, CNN-based SR for remote sensing images has become a mainstream approach. For instance, Lei, Shi, and Zou (2017) introduced a multi-fork network structure that combines local and global information to learn multi-level features for remote sensing image processing. Jiang, Wang, Yi,

and Jiang (2018) presented a remote sensing image SR network that progressively enhances the image details using a dense link network. Pan et al. (2019) proposed a residual dense back-projection network for remote sensing image SR, which utilizes a residual back-projection module to simplify the network and speed up the reconstruction process. Lei, Shi, and Zou (2019) developed a dual-path framework for remote sensing image generation using a coupled discriminative GAN network with coupled discriminative loss. D. Zhang et al. (2020) designed a remote sensing image SR network based on a hybrid higher-order attention network that adopts a higher-order attention model for remote sensing feature detail recovery and connects the feature extraction and feature refinement networks by introducing frequency-aware connections to greatly improve the SR performance. Liu et al. (2022) proposed and demonstrated a pairwise learning-based graph neural network that considers self-similar feature blocks in remote sensing images by aggregating across scales. The successful reconstruction of remote sensing images in SR heavily relies on the extraction of high-frequency information, making it crucial to develop a refined CNN-based feature extraction module.

2.2. Transformer-based remote sensing image SR

The Transformer was originally proposed in 2017 by Vaswani et al. (2017) and has gained significant popularity in the natural language processing (NLP) field. In recent years, there has been growing interest in exploring the potential of Transformers in computer vision applications. Transformers have been adopted to slice and encode images, replacing convolutions, to obtain internal connections through attention models. Recently, the visual Transformer has also been widely used in remote sensing images for tasks such as semantic classification (Strudel et al. 2021), building extraction (Chen et al. 2022), super-resolution (Lei, Shi, and Mo 2021), etc. Cai and Zhang (2022) proposed a texture transfer transformer-based SR network for remote sensing images, which reduces the dependence on reference images through a feature fusion scheme of the U-transformer and produces rich remote sensing texture information. Tu et al. (2022) designed a remote sensing image SR network by combining the Swin Transformer and CNN, which extracts depth features from the residually dense Swin Transformer to produce HR images. An et al. (2022) designed a new end-to-end remote sensing SR framework, which solves the multi-frame remote sensing image SR problem by introducing Transformer. Lei, Shi, and Mo (2021) and S. Wang et al. (2021) have previously proposed Transformer-based remote sensing image SR networks that respectively employ a multi-level enhancement structure and a context transformation layer to extract contextual features and improve SR reconstruction. In contrast, our approach aims to combine both CNN and Transformer frameworks to extract depth features of remote sensing images, resulting in a lightweight yet effective remote sensing image SR framework. This approach enhances the network's perceptual field, improving the high-frequency feature detail representation of remote sensing images.

3. The proposed method

In this section, we describe the overall structure of DCTA and the detailed structure of the DCTB components, i.e. the dual attention distillation block (DADB) and the affine-swin transformer block (ASTB).

3.1. Network architecture

In this work, we design an SR network for remote sensing images based on distillation CNN-Transformer architecture, i.e. DCTA. As shown in Figure 2, the network can be divided into four major components, i.e. the coarse feature extraction layer (CFEL), the feature distillation extraction module (FDEM), the feature fusion module (FFM), and the up-sampling module. Specifically, given an LR remote sensing image I_{LR} with a spatial size $H \times W$ and an HR remote sensing image I_{HR} with a

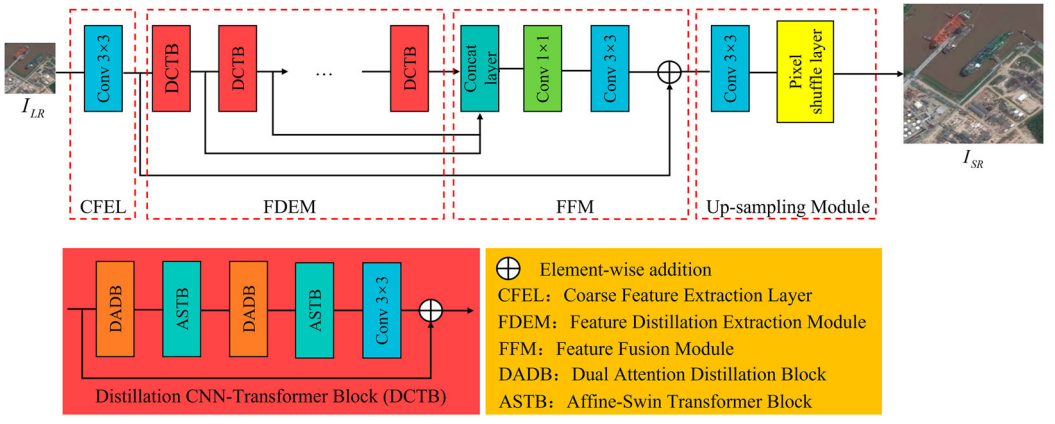


Figure 2. Overall structure of the proposed Distillation CNN-Transformer Architecture (DCTA).

spatial size $tH \times tW$, the proposed DCTA network aims to reconstruct the SR remote sensing image I_{SR}:

$$I_{SR} = \mathcal{F}(I_{LR}), \tag{1}$$

where $\mathcal{F}(.)$ denotes the SR function of the proposed network, and H, W, and t indicate the height, width, and scale factor, respectively. The four-step process follows:

(1) Coarse Feature Extraction Layer. First, the LR remote sensing image I_{LR} is fed into a 3×3 convolutional layer, which extracts the coarse features of the I_{LR}, with the function expression:

$$\mathcal{M}_0 = \mathcal{F}_1(I_{LR}), \tag{2}$$

where $\mathcal{M}_0$ denotes the coarse feature of the I_{LR}, and $\mathcal{F}_1(.)$ denotes the CFEL's 3×3 convolution operation.

(2) Feature Distillation Extraction Module. Second, $\mathcal{M}_0$ is fed to the DCTB, whose function expression is:

$$\mathcal{M}_1 = \mathcal{F}_{DCTB}(M_0), \tag{3}$$

where $\mathcal{M}_1$ denotes the refined distillation features extracted by DCTB. Considering that the FDEM is stacked by multiple DCTBs, the feature expression extracted by each DCTB can be defined as follows:

$$\mathcal{M}_i = \mathcal{F}_{DCTB}(M_{i-1}), \tag{4}$$

where $\mathcal{M}_i$ denotes that FDEM extracts fine features from remote sensing images using i-DCTBs.

(3) Feature Fusion Module. Third, the above-extracted distillation features are input to the FFM to achieve efficient feature fusion without increasing network's parameters, and the process is expressed as follows:

$$\mathcal{M}_{FFM} = \mathcal{F}_{3\times3}(F_{1\times1}(\mathcal{M}_{Concat})) + \mathcal{M}_0, \tag{5}$$

where $\mathcal{F}_{3\times3}$ and $\mathcal{F}_{1\times1}$ represent convolution operations of 3×3 and 1×1, respectively.

(4) Up-sampling Module. Finally, the highly efficient fused remote sensing features are upsampled and reconstructed, outputting the final reconstructed SR remote sensing image:

$$I_{SR} = \mathcal{F}_{up}(F_{3\times3}(\mathcal{M}_{FFM})), \tag{6}$$

where $\mathcal{F}_{up}$ denotes the reconstruction function of the pixel shuffle layer.

The loss function for our DCTA model is formulated as follows:

$$\mathcal{L}_{SR}(\Theta) = \frac{1}{N}\sum_{n=1}^{N}\left\| I_{SR}^{n} - I_{HR}^{n} \right\|_1, \tag{7}$$

where Θ denotes the DCTA parameters, $\|\|_1$ denotes the l_1 norm, and I_{SR}^{n} and I_{HR}^{n} denote the nth reconstructed SR image and corresponding ground truth image, respectively.

3.2. Dual attention distillation block (DADB)

In the following subsection, we provide a detailed description of the DADB's structure, as shown in Figure 3. The proposed enhanced spatial residual attention (ESRA) and enhanced contrast channel residual attention (ECCRA) enable DCTA to extract enhanced remote sensing multi-scale features at each channel direction and spatial scale while maintaining a small number of model parameters. Note that the ESRA and ECCRA are linked together to effectively distill fine remote sensing features while reducing the extracted redundant remote sensing information features, as shown in Figure 4. The detailed structure of DADB is shown in Table 1. The details of ESRA and ECCRA are discussed in the following sections:

(1) Enhanced Spatial Residual Attention (ESRA). To maximize the effectiveness of DCTA and to focus remote sensing image features on spatial scales of critical importance. To this end, we

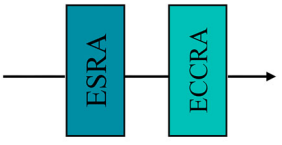
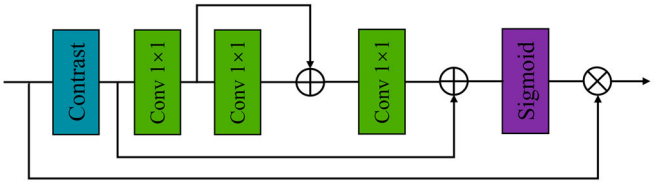


Figure 3. The overall structure of the proposed Dual Attention Distillation Block (DADB).



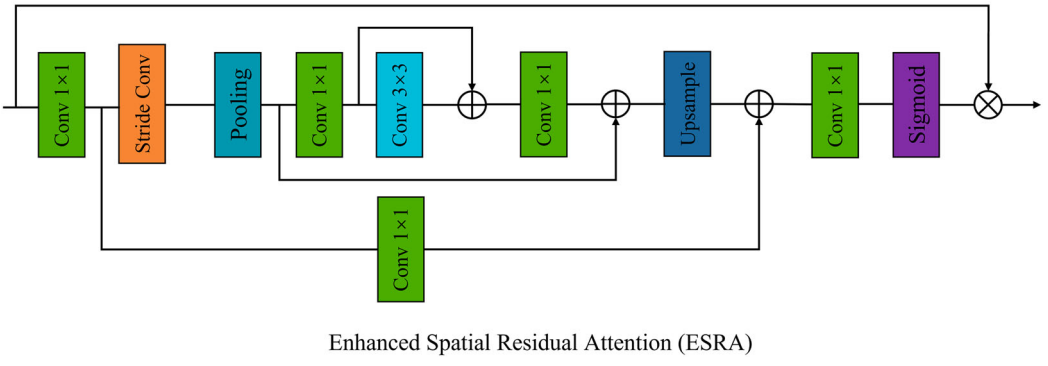


Figure 4. The proposed Enhanced Contrast Channel Residual Attention (ECCRA) and Enhanced Spatial Residual Attention (ESRA).

Table 1. Detailed structure of DADB.

Block name	Layer name	Layer details	Connected to
ESRA	Conv1_a	kernel_size = 1 × 1	CFEL
	Conv_stride	kernel_size = 3 × 3, stride = 2, padding = 0	Conv1_a
	Pooling	kernel_size = 7 × 7, stride = 3	Conv_stride
	Conv1_b	kernel_size = 1 × 1	Pooling
	Conv3	kernel_size = 3 × 3, padding = 1	Conv1_b
	Addition_1	Conv1_b, Conv3	Conv3
	Conv1_c	kernel_size = 1 × 1	Addition_1
	Addition_2	Conv1_c, Pooling	Conv1_c
	Upsample	mode = bilinear	Addition_2
	Conv1_d	kernel_size = 1 × 1	Conv1_a
	Addition_3	Upsample, Conv1_d	Upsample
	Conv1_e	kernel_size = 1 × 1	Addition_3
	Sigmoid	--	Conv1_e
	Multiplication	Sigmoid, CFEL	Sigmoid
ECCRA	Contrast	kernel_size = 1 × 1	ESRA
	Conv1_a	kernel_size = 1 × 1, padding = 0	Contrast
	Conv1_b	kernel_size = 1 × 1, padding = 0	Conv1_a
	Addition_1	Conv1_a, Conv1_b	Conv1_b
	Conv1_c	kernel_size = 1 × 1, padding = 0	Addition_1
	Addition_2	Conv1_c, Contrast	Conv1_c
	Sigmoid	--	Addition_2
	Addition_3	Sigmoid, ESRA	Sigmoid

designed an ESRA block based on enhanced spatial attention (Fang et al. 2022), which enhances more attention to the regions of interest by refining the features. Specifically, the extracted coarse remote sensing image features $\mathcal{M}_0$ are first fed to the 1 × 1 convolution layer to reduce the channel size. Then the maximum pooling layer is passed through a stride convolution kernel to ensure the expansion of the perceptual field with a smaller number of parameters, followed by input to the group convolution layer of the residuals. To recover the spatial dimension and channel size, an upsampling layer and a 1 × 1 convolution layer are attached. Finally, the attention mask is generated through the sigmoid layer, and element multiplication is performed with the rough remote sensing image feature $\mathcal{M}_0$ to realize the extraction of fine remote sensing image features. The expression of the ESRA block can be defined as:

$$\mathcal{M}_{ESRA} = \mathcal{F}_{ESRA}(\mathcal{M}_0), \tag{8}$$

where $\mathcal{F}_{ESRA}$ represents the feature extraction process of the ESRA block.

(2) Enhanced Contrast Channel Residual Attention (ECCRA). To effectively capture the global information of remote sensing images and further improve the detailed structural information, we designed an ECCRA block. The proposed ECCRA differs from the traditional attention block, as it uses the contrast information of the sum of standard deviation and mean to calculate the weight of channel attention and then uses the residual in the residual within the module to solve the problems such as disappearing during the feature information transfer. Let's $Q = [q_1, \ldots, q_c, \ldots, q_C]$ as the input, which C represent feature maps. Therefore, the contrast information value can be calculated by:

$$H(q_c) = \sqrt{\frac{1}{H \times W} \sum_{(i,j) \in q_c} \left(q_c^{i,j} - \frac{1}{H \times W} \sum_{(i,j) \in q_c} q_c^{i,j} \right)^2} + \frac{1}{H \times W} \sum_{(i,j) \in q_c} q_c^{i,j}, \tag{9}$$

where $H(q_c)$ denotes the global contrast information evaluation function.

Specifically, the extracted remote sensing image features $\mathcal{M}_{ESRA}$ are first fed into the contrast layer, then into three 1 × 1 convolutions via residuals. Finally, the calculated feature information is fed into the sigmoid layer, where elemental multiplication is performed with the extracted remote

sensing image features $\mathcal{M}_{ESRA}$. Thus, the expression of ECCRA is defined as:

$$\mathcal{M}_{ECCRA} = \mathcal{F}_{ECCRA}(\mathcal{M}_{ESRA}), \tag{10}$$

$$\mathcal{M}_{DADB} = \mathcal{M}_{ESRA} + \mathcal{M}_{ECCRA}, \tag{11}$$

where $\mathcal{F}_{ECCRA}$ denotes the feature extraction process of the ECCRA block and $\mathcal{M}_{DADB}$ denotes the feature output of the overall DADB.

3.3. Affine-swin transformer block (ASTB)

Swin Transformer (Liang et al. 2021) is a Transformer architecture that employs local attention and offset windows and has achieved great performance in computer vision tasks. For remote sensing image reconstruction tasks, we propose an ASTB based on a Swin Transformer, with which features can be extracted from shallow to deep by varying downsampling scales and gradually expanding the network's perceptual field while using localization to reduce module ground computation complexity. As shown in Figure 5, the remote sensing image features extracted by DADB are passed to the layer normative (LN) layer and the multi-head self-attention (MSA) layer in turn, and the output of the MSA layer and the output of the DADB are operated with element-wise addition. Then the extracted features are passed to the LN layer and the ResMLP layer in turn, and finally, the element-wise addition operation is performed with the output of the MSA to output the final remote sensing image features. Therefore, the overall process of ASTB can be defined as:

$$\mathcal{M}^{'} = \mathcal{F}_{MSA}(\mathcal{F}_{LN}(\mathcal{M}_{DADB})) + \mathcal{M}_{DADB}, \tag{12}$$

$$\mathcal{M}_{ASTB} = \mathcal{F}_{ResMLP}(\mathcal{F}_{LN}(\mathcal{M}^{'})) + \mathcal{M}^{'}, \tag{13}$$

where $\mathcal{F}_{LN}$, $\mathcal{F}_{MSA}$, and $\mathcal{F}_{ResMLP}$ represent the LN, MSA, and ResMLP function operations, respectively.

In the above ASTB, a new residual MLP layer is designed to mitigate the issue of gradient explosion in the feature information transfer process. By introducing the affine transformation layer, the network's training becomes more stable without adding additional training costs. As shown in Figure 6, the transfer process for ResMLP follows: *Affine* $\longrightarrow$ *Linear* $\longrightarrow$ *GeLU* $\longrightarrow$ *Linear* $\longrightarrow$ *Affine*. Benefiting from ResMLP, our proposed ASTB shows advanced reconstruction performance in remote sensing image SR tasks by exploiting cross-window information and then alternating with ResMLP. To strike a balance between network complexity and reconstruction performance, we incorporate two ASTBs within each DCTB.

4. Experimental results

In this section, we present our experimental results from four aspects, i.e. the dataset, implementation details, comparison with the state-of-the-art, and ablation studies.

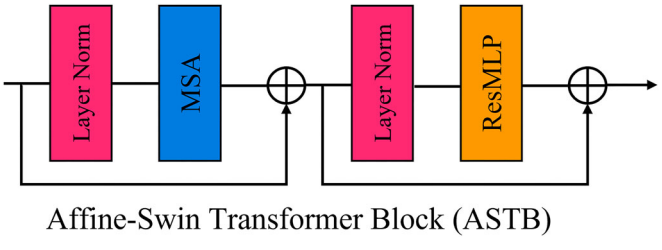


Figure 5. The proposed Affine-Swin Transformer Block (ASTB).

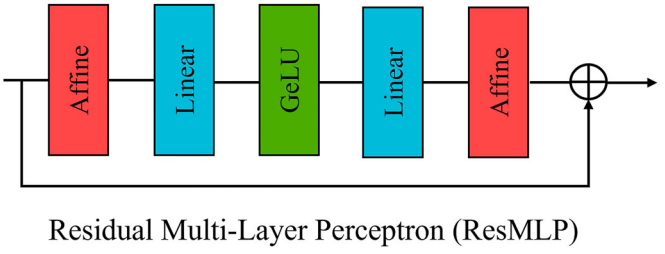


Figure 6. The proposed Residual Multi-Layer Perceptron (ResMLP).

4.1. Datasets

Our DCTA is evaluated through experiments conducted on three publicly available remote sensing datasets: the NWPU-RESISC45 dataset (Cheng, Han, and Lu 2017), the Draper dataset, and the UC Merced (Yang and Newsam 2010) dataset. Figure 7 displays selected samples from these datasets. The NWPU-RESISC45 dataset comprises 45 scene categories, each containing 700 remote sensing images with a size of 256 × 256 pixels. For training and testing, we utilized 2,250 and 90 images, respectively, from this dataset.

The Draper dataset comprises 324 scene categories, each containing 5 images, with an original image size of 3, 099 × 2, 329 pixels. To obtain HR images, we crop the original images to a size of 192 × 192 pixels via Bicubic interpolation. We selected 1,000 remote sensing images for training and 200 images for testing from this dataset.

The UC Merced dataset is composed of 21 categories of remote sensing scenes, each containing 100 images with a size of 256 × 256 pixels. We divided the dataset into two parts, with 1,050 images used for training and the remaining 1,050 images utilized for testing.

4.2. Implementation details

In this study, we focus on the 4× scale factor for remote sensing image reconstruction. The LR images in the training set are all obtained by bicubic downsampling. In the training phase, the images for training are enhanced by random rotation and horizontal flips. Similar to previous work (Y. Wang et al. 2023), We use five evaluation metrics to measure the reconstruction quality of SR images, including PSNR, SSIM (Zhou Wang et al. 2004), FSIM (Lin Zhang et al. 2011), VIF (Sheikh and Bovik 2006), and ERGAS (Ranchin and Wald 2000). We also analyze the number of network parameters (Params) and floating-point operations (FLOPs) in the models. Note that all reconstruction results are evaluated on the Y channel of the YCbCr color space.

We used the Adam optimizer (Kingma and Ba 2014) to train our model with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-8$. The initial learning rate was set to 5e-4 and halved every 200 epochs, for a total of 1,200 epochs. For each training mini-batch, 16 randomly cropped patches of size 48 × 48 were used as input. The specific device parameters are shown in Table 2. The source code is available at https://github.com/Yu-Wang-0801/DCTA.

4.3. Compared with state-of-the-arts

In this subsection, we compare the experimental results of the proposed method with other comparative algorithms in quantitative and qualitative terms, including Bicubic, DBPN (Haris, Shakhnarovich, and Ukita 2019), RDN (Yulun Zhang et al. 2020), RCAN (Yulun Zhang et al. 2018), CSNLA (Mei et al. 2020), SAN (Dai et al. 2019), MHAN (D. Zhang et al. 2020), HSEnet (Lei and Shi 2021), TransEnet (Lei, Shi, and Mo 2021), and ESRT (Lu et al. 2022).

(1) Quantitative Results on NWPU-RESISC45 Dataset: The best results are highlighted in red font in Table 3, which presents the performance comparison of various methods on the NWPU-RESISC45

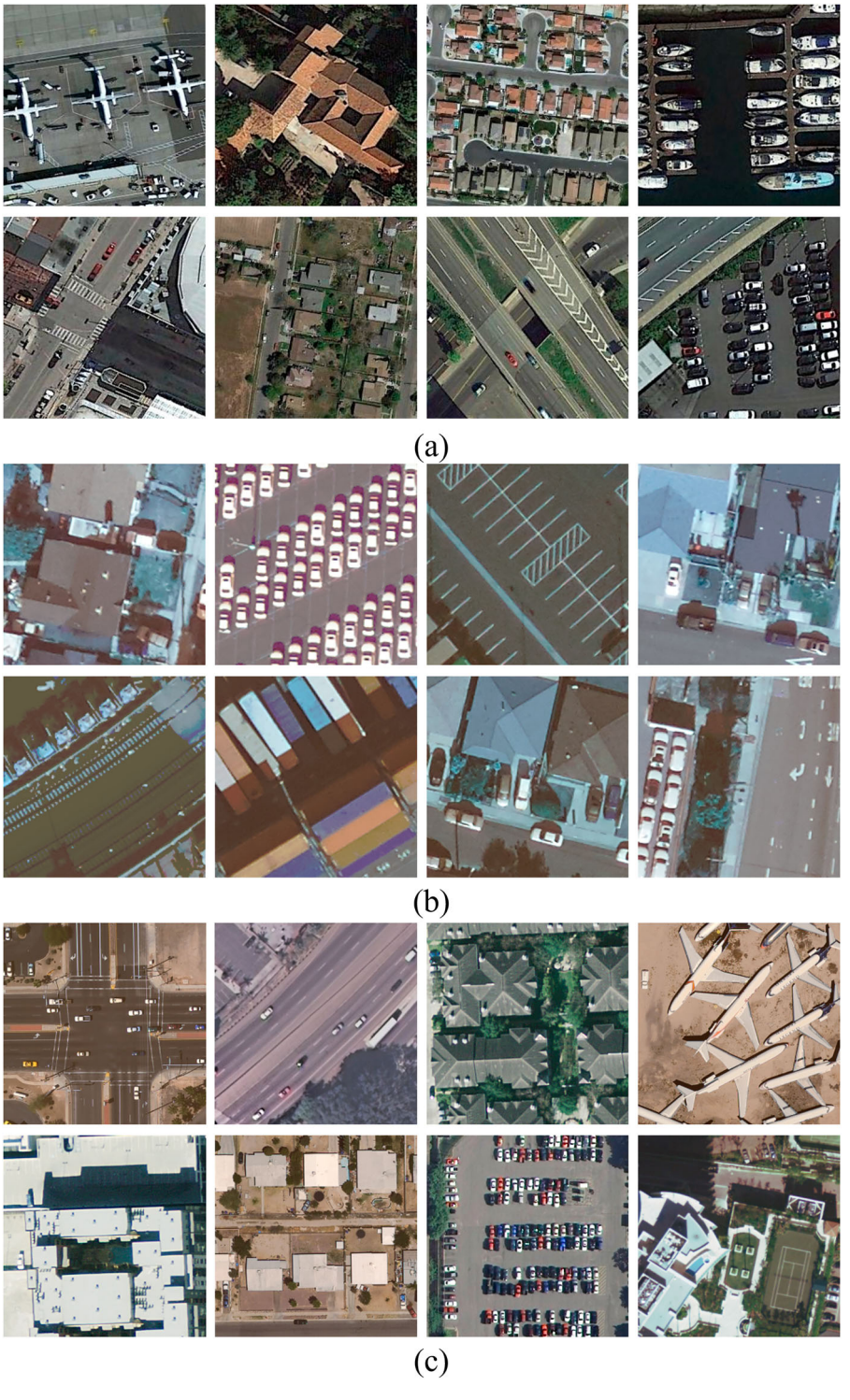


Figure 7. Selected samples from (a) the NWPU-RESISC45 dataset, (b) the Draper dataset, and (c) the UC Merced dataset.

Table 2. Experimental Environment.

Configuration	Parameter
System	Microsoft Windows10 Professional 64bits
Language	Python 3.7
Famework	Pytorch 1.7.0
GPU	NVIDIA GeForce RTX 2080Ti
CPU	Intel(R) Core(TM) i7-7820X CPU @3.60GHz
CUDA	10.2.89
CUDNN	v8.2.1

dataset. The results indicate that our proposed method achieves better objective evaluation metrics than the other compared algorithms, as demonstrated in Table 3. Specifically, while maintaining fewer network parameters and floating-point operators, our method outperforms other algorithms on five commonly used objective evaluation metrics. Figure 8 shows the reconstruction of different comparison algorithms on the NWPU-RESISC45 dataset images, where we label the regions where the reconstructed images are noticeable. The texture detail recovery capability of the proposed method was observed to be superior to that of the other competing methods.

(2) Quantitative Results on Draper Dataset: To further validate the effectiveness of DCTA, we conducted relevant experiments on the Draper dataset. Table 4 shows the comparison of the algorithms on the Draper dataset, where the best results are marked in red font. From Table 4, it can be seen that the proposed method obtains the best reconstruction results in PSNR and ERGAS. Although DCTA is worse than RCAN in three evaluation metrics (i.e. SSIM, FSIM, and VIF), the number of network parameters of DCTA is only 4.9% of RCAN, and the FLOPs are 4.4% of RCAN. Thus, we believe our method presents a better balance between performance and efficiency than other competing algorithms. The reconstructed images obtained by various competing algorithms on the Draper dataset are compared in Figure 9. It can be observed that our proposed method can reconstruct finer feature information on vehicle and aircraft runway lines compared to the other algorithms.

(3) Quantitative Results on UC Merced Dataset: To further validate the generality of DCTA, we tested the proposed method as well as other competing methods on the UC Merced dataset. Table 5 shows the algorithm comparison on the UC Merced dataset, where the best results are marked in red font. From Table 5, it can be seen that DCTA outperforms the other comparison algorithms for all five commonly used objective evaluation metrics with a low number of network parameters and FLOPs. Figure 10 compares the reconstructed images of different competing algorithms on the UC Merced dataset. As shown in Figure 10, our method outperforms other competing algorithms regarding reconstructed details, indicating the superior performance of the proposed method both from quantitative and qualitative perspectives.

Table 3. Comparison of Params, FLOPs, PSNR, SSIM, FSIM, VIF, and ERGAS results of different algorithms on the NWPU-RESISC45 dataset. Parameters (Params) and floating-point operations (FLOPs) are tested on an LR image with 64 × 64 pixels.

Algorithms	#Params/M	FLOPs/G	PSNR/dB ↑	SSIM ↑	FSIM ↑	VIF ↑	ERGAS ↓
Bicubic	–	–	27.25	0.6786	0.7807	0.2900	2.7699
DBPN	10.43	370.70	28.54	0.7476	0.8369	0.3634	2.3916
RDN	1.15	6.62	27.85	0.7419	0.8375	0.3571	2.4698
RCAN	15.59	65.25	27.89	0.7412	0.8413	0.3576	2.5230
CSNLA	7.16	2838.71	27.97	0.7264	0.8331	0.3457	2.4726
SAN	15.82	66.57	28.21	0.7386	0.8371	0.3545	2.4521
MHAN	11.35	50.62	28.61	0.7516	0.8394	0.3684	2.3804
HSEnet	5.43	19.20	28.81	0.7582	0.8444	0.3755	2.3185
TransEnet	37.38	21.40	28.76	0.7574	0.8453	0.3748	2.3328
ESRT	0.75	4.17	28.85	0.7591	0.8459	0.3764	2.3316
Ours	0.77	2.86	28.94	0.7621	0.8463	0.3802	2.2816

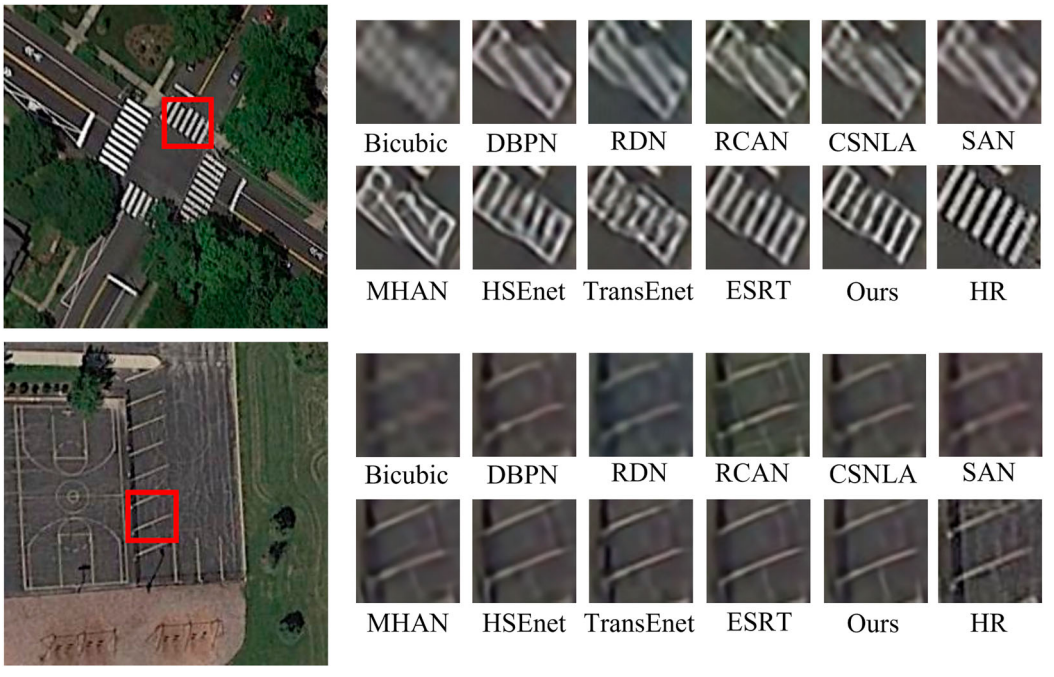


Figure 8. Comparison of subjective results on the NWPU-RESISC45 dataset with other comparison algorithms. Best view via zoomed-in view.

4.4. Ablation studies

In this subsection, we design a series of ablation experiments to evaluate the effectiveness of each component of the DCTA network and the rationality of the number of DCTBs.

(1) Effectiveness of the proposed DADB and ASTB: From Table 6, the reconstruction performance of our network differs by 0.0798 and 0.0158 in terms of ERGAS and spectral angle mapper (SAM) Yifan Zhang, Backer, and Scheunders (2009) metrics when DCTA removes the ASTB component. The results indicate that the ASTB (as explained in Section 3.3) improves the extraction of deep features from remote sensing images and expands the perceptual fields. Additionally, we performed another experiment to evaluate the effectiveness of DADB. In this experiment, we removed DADB from the network and observed a reduction of 0.0216 and 0.0044 in the ERGAS and SAM values of the reconstruction results, respectively. This phenomenon also proves that DADB facilitates the extraction of high-frequency details from remote-sensing images.

Table 4. Comparison of Params, FLOPs, PSNR, SSIM, FSIM, VIF, and ERGAS results of different algorithms on the Draper dataset. Parameters (Params) and floating-point operations (FLOPs) are tested on an LR image with 48 × 48 pixels.

Algorithms	#Params/M	FLOPs/G	PSNR/dB ↑	SSIM ↑	FSIM ↑	VIF ↑	ERGAS ↓
Bicubic	–	–	30.85	0.8217	0.8481	0.3269	1.3906
DBPN	10.43	208.52	33.36	0.8821	0.9042	0.4399	1.0301
RDN	1.15	3.72	33.22	0.8778	0.9003	0.4312	1.0503
RCAN	15.59	36.70	33.62	0.8919	0.9126	0.4594	0.9892
CSNLA	7.16	1596.77	33.38	0.8848	0.9069	0.4444	1.0127
SAN	15.82	37.45	33.44	0.8848	0.9073	0.4445	1.0158
MHAN	11.35	28.48	33.14	0.8819	0.9036	0.4376	1.0506
HSEnet	5.43	10.80	33.64	0.8908	0.9112	0.4551	0.9888
TransEnet	37.38	12.04	33.45	0.8877	0.9123	0.4521	1.0086
ESRT	0.75	2.35	33.49	0.8892	0.9080	0.4500	1.0163
Ours	0.77	1.61	33.73	0.8897	0.9105	0.4536	0.9887

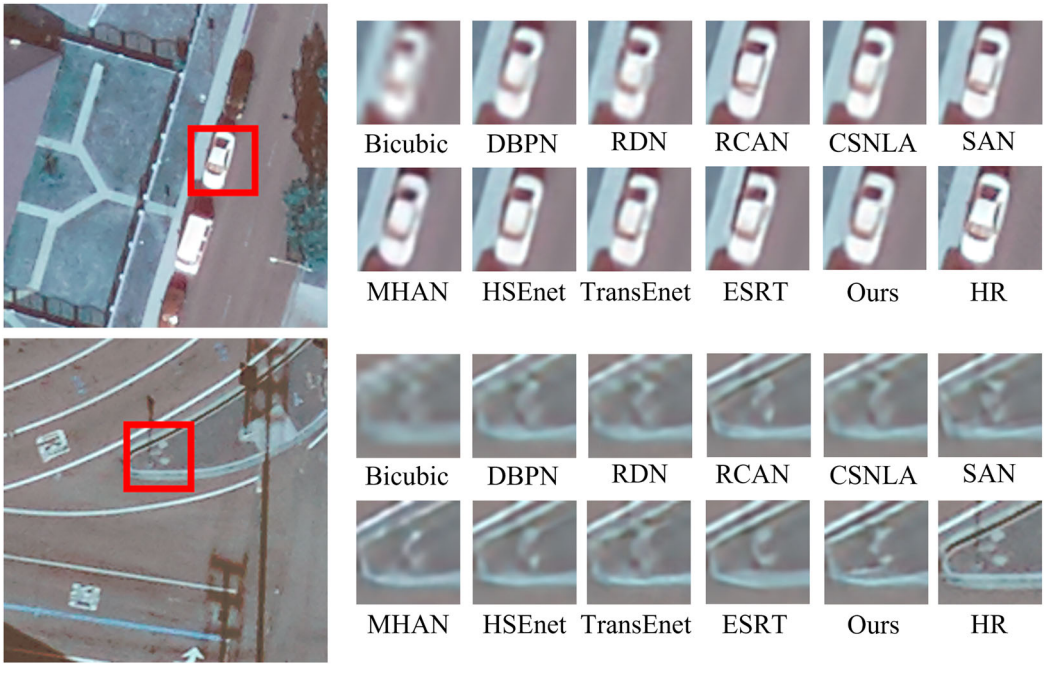


Figure 9. Comparison of subjective results on the Draper dataset with other comparison algorithms. Best view via zoomed-in view.

To further verify the persuasiveness of the above ablation experiments, we validated the networks with different modes using error comparison plots. The subjective results are shown in Figure 11. From the detail expansion region in Figure 11, it is clear that adding the proposed components leads to better-reconstructed images, proving the effectiveness of our proposed components.

(2) Effectiveness of different numbers of DCTBs: In order to verify the influence of the number of DCTBs on the network, we designed a set of ablation experiments for comparison, and the experimental results are shown in Table 7. It can be seen from Table 7 that with the increase in the number of DCTBs in this network, most of the objective evaluation indicators show an upward trend. We know from experiments that when the number of DCTBs increases to 8, the network parameters will reach 1.01M. Therefore, this network chooses the number of DCTBs to be 6 under careful consideration to ensure optimal reconstruction performance while ensuring a lightweight network.

(3) Effectiveness of the contrast convolution layers: In order to verify the effectiveness of the contrast convolution layer in ECCRA, we conduct related experiments for comparison, and the

Table 5. Comparison of Params, FLOPs, PSNR, SSIM, FSIM, VIF, and ERGAS results of different algorithms on the UC Merced dataset. Parameters (Params) and floating-point operations (FLOPs) are tested on an LR image with 64 × 64 pixels.

Algorithms	#Params/M	FLOPs/G	PSNR/dB ↑	SSIM ↑	FSIM ↑	VIF ↑	ERGAS ↓
Bicubic	–	–	28.63	0.7637	0.8215	0.3536	1.8697
DBPN	10.43	370.70	30.55	0.8328	0.8839	0.4531	1.5031
RDN	1.15	6.62	30.30	0.8170	0.8726	0.4352	1.5471
RCAN	15.59	65.25	30.55	0.8302	0.8832	0.4497	1.5073
CSNLA	7.16	2838.71	30.57	0.8269	0.8812	0.4489	1.4991
SAN	15.82	66.57	30.52	0.8316	0.8835	0.4516	1.5126
MHAN	11.35	50.62	30.34	0.8255	0.8783	0.4418	1.5428
HSEnet	5.43	19.20	30.78	0.8406	0.8886	0.4634	1.4664
TransEnet	37.38	21.40	30.52	0.8257	0.8803	0.4466	1.5066
ESRT	0.75	4.17	30.79	0.8379	0.8881	0.4623	1.4635
Ours	0.77	2.86	31.05	0.8455	0.8927	0.4710	1.4220

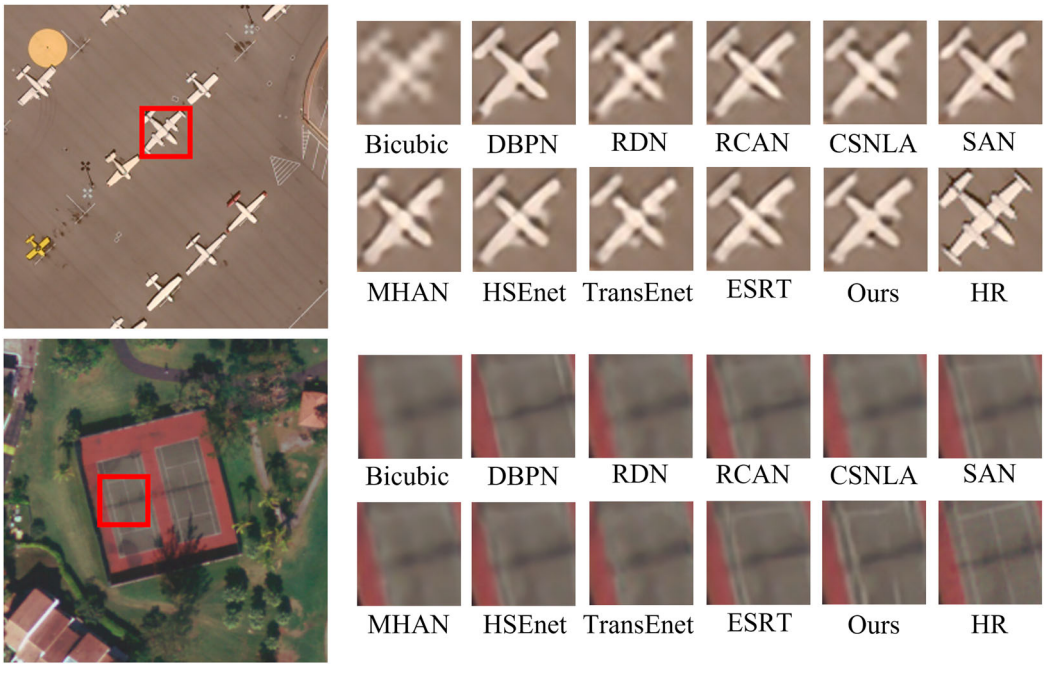


Figure 10. Comparison of subjective results on the UC Merced dataset with other comparison algorithms. Best view via zoomed-in view.

Table 6. The results of the ablation experiments for the NWPU-RESISC45 dataset with scale factor × 4.

Model	DADB	ASTB	ERGAS ↓	SAM ↓
Ours without ASTB			2.3614	0.9488
Ours without DADB			2.3032	0.9374
Ours			2.2816	0.9330

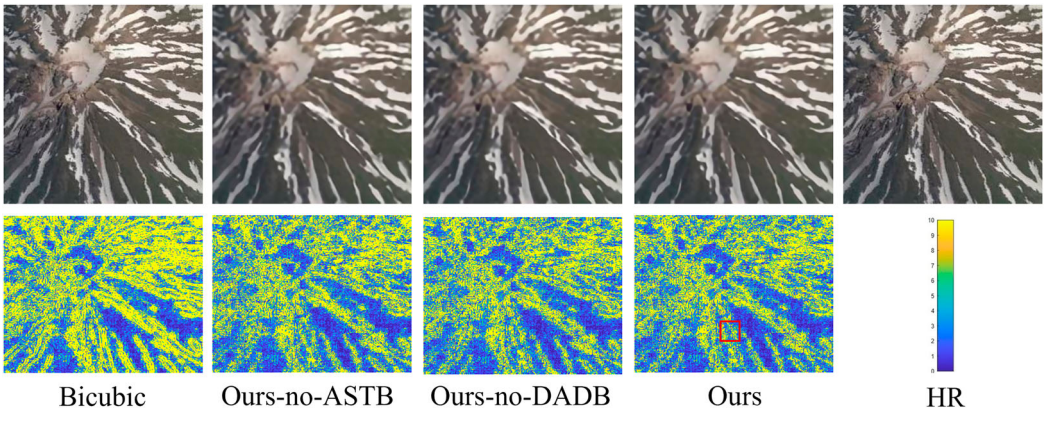


Figure 11. MSE maps between reconstructions and ground truth of different ablation experiments. The error between images increases and decreases with the value corresponding to the color of the MSE map.

experimental results are shown in Table 8. We re-perform related experiments on the NWPU-RESISC45 dataset by removing the contrast convolution layer in ECCRA. From Table 8, we can see that with the assistance of contrastive convolutional layers, our DCTA network can stably improve the reconstruction ability of remote sensing images.

Table 7. Comparison of reconstruction performance of different numbers of DCTBs on NWPU-RESISC45, Draper, and UC Merced datasets.

Datasets	DCTBs	#Params/M	FLOPs/G	PSNR/dB ↑	SSIM ↑	FSIM ↑	VIF ↑	ERGAS ↓
NWPU-RESISC45	2	0.29	1.08	28.81	0.7575	0.8437	0.3753	2.3241
	4	0.53	1.97	28.90	0.7612	0.8458	0.3795	2.2995
	6	0.77	2.86	28.94	0.7621	0.8463	0.3802	2.2816
Draper	2	0.29	0.61	33.65	0.8880	0.9093	0.4495	0.9968
	4	0.53	1.11	33.72	0.8898	0.9102	0.4539	0.9885
	6	0.77	1.61	33.73	0.8897	0.9105	0.4536	0.9887
UC Merced	2	0.29	1.08	30.80	0.8358	0.8871	0.4578	1.4646
	4	0.53	1.97	30.93	0.8409	0.8896	0.4648	1.4418
	6	0.77	2.86	31.05	0.8455	0.8927	0.4710	1.4220

Table 8. Comparison of reconstruction performance of contrast convolution layers on the NWPU-RESISC45 dataset.

Model	PSNR/dB ↑	SSIM ↑	FSIM ↑	VIF ↑	ERGAS ↓
Ours without contrast	28.87	0.7598	0.8452	0.3783	2.3081
Ours	28.94	0.7621	0.8463	0.3802	2.2816

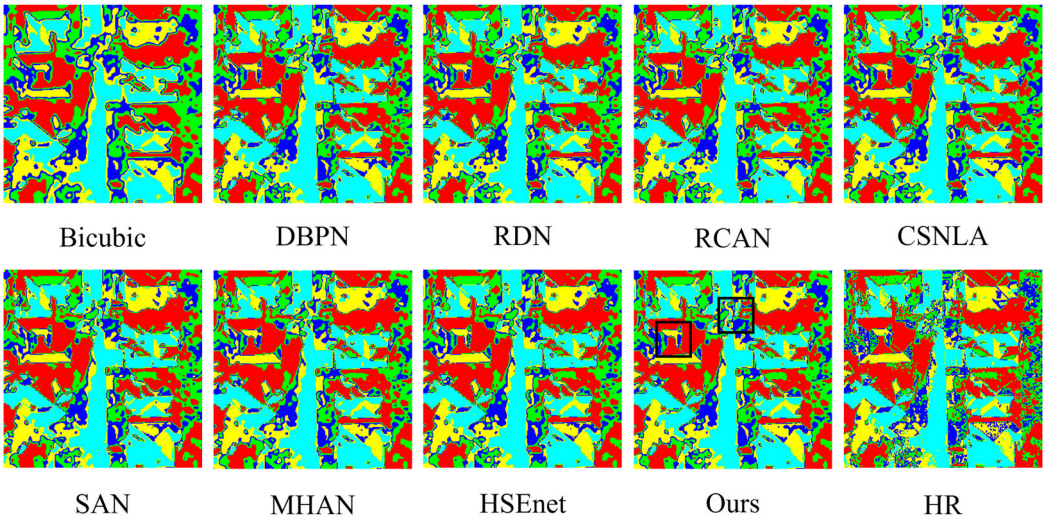


Figure 12. The classification results obtained by different algorithms through the ISODATA classification method are compared. Zoom in for more details.

4.5. Verification of superiority in classification tasks

The remote sensing image SR reconstruction technique is often used as one of the pre-processing steps in computer vision tasks. To better validate the effectiveness of reconstructed remote sensing images for downstream tasks, we used an unsupervised semantic segmentation algorithm (ISODATA) to evaluate reconstructed remote sensing images using various competing algorithms. We conducted the relevant experiments with the software ENVI5.3 platform and set the number of classifications to 5 and the maximum number of iterations to 5. The classification results are shown in Figure 12, where it can be observed that our DCTA-reconstructed images present closer classification results compared to the ones from ground truth images. We also notice that the reconstructed results of other competing algorithms lack relevant texture detail information that facilities correct classification. This experiment further supports the effectiveness of the proposed DCTA model in reconstructing remote sensing images with superior performance.

5. Conclusion

In this paper, we present a novel lightweight super-resolution (SR) architecture, named DCTA, specifically designed for remote sensing applications. Our approach introduces a unique distillation CNN-Transformer block (DCTB) that combines the strengths of CNN and Transformer structures in a lightweight manner. The proposed DCTB enables the extraction of deep features at various scales in remote sensing images, effectively enhancing the network's perceptual field and efficiently utilizing global feature information. To validate the effectiveness of DCTA, we conduct experiments on three datasets: NWPU-RESISC45, Draper, and UC Merced. The results demonstrate that our method achieves an excellent balance between computational cost and reconstruction performance, outperforming existing methods. Furthermore, we verify the efficiency of our design through comprehensive ablation experiments. The deployment of our method on hardware devices is anticipated, as it can enhance the accuracy of downstream tasks related to remote sensing images, such as change detection and building extraction.

Author contributions

Yu Wang: Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Investigation, Validation, Writing–original draft, Writing–review & editing; Zhenfeng Shao: Conceptualization, Resources, Methodology, Writing–review & editing, Supervision; Tao Lu: Methodology, Writing–review & editing, Supervision; Lifeng Liu: Formal analysis, Writing–review & editing; Xiao Huang: Writing–review & editing; Jiaming Wang: Resources, Validation, Writing-review & editing; Kui Jiang: Validation, Writing-review & editing; Kangli Zeng: Software, Writing-review & editing.

Data availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Disclosure statement

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study, in the collection, analyzes, or interpretation of the data, in the writing of the manuscript, or in the decision to publish the results.

Funding

This work was supported by National Natural Science Foundation of China [42090012], Guangxi Science and Technology Plan Project (Guike 2021AB30019), Hubei Province Key R\&D Project (2022BAA048), Sichuan Province Key R\&D Project (2022YFN0031, 2023YFN0022, 2023YFS0381), Zhuhai Industry-University-Research Cooperation Project (ZH22017001210098PWC), Shanxi Provincial Science and Technology Major Special Project (202201150401020), Guangxi Key Laboratory of Spatial Information and Surveying and Mapping Fund Project (21-238-21-01).

ORCID

Yu Wang http://orcid.org/0000-0003-3436-4251
Zhenfeng Shao http://orcid.org/0000-0003-4587-6826
Tao Lu https://orcid.org/0000-0001-8117-2012
Xiao Huang https://orcid.org/0000-0002-4323-382X
Jiaming Wang https://orcid.org/0000-0001-8144-5842
Kui Jiang https://orcid.org/0000-0002-4055-7503
Kangli Zeng https://orcid.org/0000-0002-8592-053X

References

An, Tai, Xin Zhang, Chunlei Huo, Bin Xue, Lingfeng Wang, and Chunhong Pan. 2022. "TR-MISR: Multiimage Super-Resolution Based on Feature Fusion With Transformers." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15: 1373–1388. https://doi.org/10.1109/JSTARS.2022.3143532.

Asker, Mehmet Emin. 2023. "Hyperspectral Image Classification Method Based on Squeeze-and-excitation Networks, Depthwise Separable Convolution and Multibranch Feature Fusion." *Earth Science Informatics* 16 (2): 1427–1448. https://doi.org/10.1007/s12145-023-00982-0.

Bai, Ting, Le Wang, Dameng Yin, Kaimin Sun, Yepei Chen, Wenzhuo Li, and Deren Li. 2022. "Deep Learning for Change Detection in Remote Sensing: A Review." *Geo-Spatial Information Science* 1–27. https://doi.org/10.1080/10095020.2022.2085633.

Cai, Durong, and Peng Zhang. 2022. "Texture Transfer Transformer for Remote Sensing Image Superresolution." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15: 7346–7358. https://doi.org/10.1109/JSTARS.2022.3198557.

Chen, Xin, Chunping Qiu, Wenyue Guo, Anzhu Yu, Xiaochong Tong, and Michael Schmitt. 2022. "Multiscale Feature Learning by Transformer for Building Extraction From Satellite Images." *IEEE Geoscience and Remote Sensing Letters* 19: 1–5.

Cheng, Gong, Junwei Han, and Xiaoqiang Lu. 2017. "Remote Sensing Image Scene Classification: Benchmark and State of the Art." *Proceedings of the IEEE* 105 (10): 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998.

Dai, Tao, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. "Second-Order Attention Network for Single Image Super-Resolution." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11065–11074.

Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. "Learning A Deep Convolutional Network for Image Super-Resolution." In *European Conference on Computer Vision*, 184–199. Springer.

Dong, Chao, Chen Change Loy, and Xiaoou Tang. 2016. "Accelerating the Super-Resolution Convolutional Neural Network." In *European Conference on Computer Vision*, 391–407. Springer.

Dong, Xiaoyu, Zhihong Xi, Xu Sun, and Lianru Gao. 2019. "Transferred Multi-perception Attention Networks for Remote Sensing Image Super-resolution." *Remote Sensing* 11 (23): 2857. https://doi.org/10.3390/rs11232857.

Fang, Jinsheng, Hanjiang Lin, Xinyu Chen, and Kun Zeng. 2022. "A Hybrid Network of CNN and Transformer for Lightweight Image Super-Resolution." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1103–1112.

Firat, Hüseyin, Mehmet Emin Asker, Mehmet Ilyas Bayindir, and Davut Hanbay. 2023a. "3D Residual Spatial–spectral Convolution Network for Hyperspectral Remote Sensing Image Classification." *Neural Computing and Applications* 35 (6): 4479–4497. https://doi.org/10.1007/s00521-022-07933-8.

Fırat, Hüseyin, Mehmet Emin Asker, Mehmet İlyas Bayındır, and Davut Hanbay. 2023b. "Hybrid 3D/2D Complete Inception Module and Convolutional Neural Network for Hyperspectral Remote Sensing Image Classification." *Neural Processing Letters* 55 (2): 1087–1130. https://doi.org/10.1007/s11063-022-10929-z.

Fırat, Hüseyin, Mehmet Emin Asker, and Davut Hanbay. 2022. "Classification of Hyperspectral Remote Sensing Images Using Different Dimension Reduction Methods with 3D/2D CNN." *Remote Sensing Applications: Society and Environment* 25: 100694. https://doi.org/10.1016/j.rsase.2022.100694.

Haris, Muhammad, Greg Shakhnarovich, and Norimichi Ukita. 2019. *Deep Back-Projection Networks for Single Image Super-Resolution*. arXiv preprint arXiv:1904.05677.

He, Jiang, Qiangqiang Yuan, Jie Li, Yi Xiao, Xinxin Liu, and Yun Zou. 2022. "DsTer: A Dense Spectral Transformer for Remote Sensing Spectral Super-resolution." *International Journal of Applied Earth Observation and Geoinformation* 109: 102773. https://doi.org/10.1016/j.jag.2022.102773.

Jakhetiya, Vinit, Weisi Lin, Sunil P. Jaiswal, Sharath Chandra Guntuku, and Oscar C. Au. 2016. "Maximum a Posterior and Perceptually Motivated Reconstruction Algorithm: A Generic Framework." *IEEE Transactions on Multimedia* 19 (1): 93–106. https://doi.org/10.1109/TMM.2016.2609419.

Jiang, Kui, Zhongyuan Wang, Peng Yi, and Junjun Jiang. 2018. "A Progressively Enhanced Network for Video Satellite Imagery Superresolution." *IEEE Signal Processing Letters* 25 (11): 1630–1634. https://doi.org/10.1109/LSP.97.

Jiang, Kui, Zhongyuan Wang, Peng Yi, Junjun Jiang, Jing Xiao, and Yuan Yao. 2018. "Deep Distillation Recursive Network for Remote Sensing Imagery Super-resolution." *Remote Sensing* 10 (11): 1700. https://doi.org/10.3390/rs10111700.

Kim, Jiwon, Jung Kwon Lee, and Kyoung Mu Lee. 2016. "Deeply-Recursive Convolutional Network for Image Super-Resolution." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1637–1645.

Kingma, Diederik P., and Jimmy Ba. 2014. *Adam: A Method for Stochastic Optimization*. arXiv preprint arXiv:1412.6980.

Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, et al. 2017. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4681–4690.

Lei, Sen, and Zhenwei Shi. 2021. "Hybrid-Scale Self-Similarity Exploitation for Remote Sensing Image Super-Resolution." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–10.

Lei, Sen, Zhenwei Shi, and Wenjing Mo. 2021. "Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–11.

Lei, Sen, Zhenwei Shi, and Zhengxia Zou. 2017. "Super-Resolution for Remote Sensing Images Via Local–Global Combined Network." *IEEE Geoscience and Remote Sensing Letters* 14 (8): 1243–1247. https://doi.org/10.1109/LGRS.2017.2704122.

Lei, Sen, Zhenwei Shi, and Zhengxia Zou. 2019. "Coupled Adversarial Training for Remote Sensing Image Super-resolution." *IEEE Transactions on Geoscience and Remote Sensing* 58 (5): 3633–3643. https://doi.org/10.1109/TGRS.36.

Li, Xuelong, Yanting Hu, Xinbo Gao, Dacheng Tao, and Beijia Ning. 2010. "A Multi-frame Image Super-resolution Method." *Signal Processing* 90 (2): 405–414. https://doi.org/10.1016/j.sigpro.2009.05.028.

Li, Rui, Shunyi Zheng, Chenxi Duan, Libo Wang, and Ce Zhang. 2022. "Land Cover Classification From Remote Sensing Images Based on Multi-scale Fully Convolutional Network." *Geo-Spatial Information Science* 1–12. https://doi.org/10.1080/10095020.2022.2053303.

Liang, Jingyun, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. "Swinir: Image Restoration Using Swin Transformer." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liu, Ziyu, Ruyi Feng, Lizhe Wang, Wei Han, and Tieyong Zeng. 2022. "Dual Learning-Based Graph Neural Network for Remote Sensing Image Super-Resolution." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–14.

Lu, Zhisheng, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. 2022. "Transformer for Single Image Super-Resolution." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 457–466.

Mei, Yiqun, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S. Huang, and Honghui Shi. 2020. "Image Super-Resolution with Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5690–5699.

Pan, Zongxu, Wen Ma, Jiayi Guo, and Bin Lei. 2019. "Super-resolution of Single Remote Sensing Image Based on Residual Dense Backprojection Networks." *IEEE Transactions on Geoscience and Remote Sensing* 57 (10): 7918–7933. https://doi.org/10.1109/TGRS.36.

Prol, Fabricio S, R Morales Ferre, Zainab Saleem, Petri Välisuo, Cristina Pinell, Elena-Simona Lohan, Mahmoud Elsanhoury, et al. 2022. "Position, Navigation, and Timing (PNT) Through Low Earth Orbit (LEO) Satellites: A Survey on Current Status, Challenges, and Opportunities." *IEEE Access* (10): 83971–84002.

Qiu, Defu, Yuhu Cheng, and Xuesong Wang. 2021. "Gradual Back-Projection Residual Attention Network for Magnetic Resonance Image Super-Resolution." *Computer Methods and Programs in Biomedicine* 208: 106252. https://doi.org/10.1016/j.cmpb.2021.106252.

Ranchin, Thierry, and Lucien Wald. 2000. "Fusion of High Spatial and Spectral Resolution Images: The ARSIS Concept and Its Implementation." *Photogrammetric Engineering and Remote Sensing* 66 (1): 49–61.

Sheikh, Hamid R., and Alan C. Bovik. 2006. "Image Information and Visual Quality." *IEEE Transactions on Image Processing* 15 (2): 430–444. https://doi.org/10.1109/TIP.2005.859378.

Shi, Wenzhe, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883.

Strudel, Robin, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. 2021. "Segmenter: Transformer for Semantic Segmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7262–7272.

Tu, Jingzhi, Gang Mei, Zhengjing Ma, and Francesco Piccialli. 2022. "SWCGAN: Generative Adversarial Network Combining Swin Transformer and CNN for Remote Sensing Image Super-Resolution." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15: 5662–5673. https://doi.org/10.1109/JSTARS.2022.3190322.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All You Need." *Advances in Neural Information Processing Systems* 30.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing* 13 (4): 600–612. https://doi.org/10.1109/TIP.2003.819861.

Wang, Decheng, Xiangning Chen, Ningbo Guo, Hui Yi, and Yinan Li. 2023. "STCD: Efficient Siamese Transformers-Based Change Detection Method for Remote Sensing Images." *Geo-Spatial Information Science* 1–20. https://doi.org/10.1080/10095020.2022.2157762.

Wang, Benfeng, Xiaohong Chen, Jingye Li, and Jingjie Cao. 2015. "An Improved Weighted Projection Onto Convex Sets Method for Seismic Data Interpolation and Denoising." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9 (1): 228–235. https://doi.org/10.1109/JSTARS.4609443.

Wang, Jing, Hanchi Liu, Peng Jiang, Zhengfang Wang, Qingmei Sui, and Fengkai Zhang. 2022. "GPRI2Net: A Deep-Neural-Network-Based Ground Penetrating Radar Data Inversion and Object Identification Framework for Consecutive and Long Survey Lines." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–20.

Wang, Zhilin, Haixing Shang, and Shuang Wang. 2022. "Super-Resolution Reconstruction of Remote Sensing Images Based on Swin Transformer Fusion Attention Network." In *Second International Conference on Optics and Communication Technology (ICOCT 2022)*, Vol. 12473, 173–181. SPIE.

Wang, Jiaming, Zhenfeng Shao, Xiao Huang, Tao Lu, and Ruiqian Zhang. 2022. "A Dual-Path Fusion Network for Pan-Sharpening." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–14.

Wang, Jiaming, Zhenfeng Shao, Xiao Huang, Tao Lu, Ruiqian Zhang, and Yong Li. 2022. "From Artifact Removal to Super-Resolution." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–15.

Wang, Yu, Zhenfeng Shao, Tao Lu, Changzhi Wu, and Jiaming Wang. 2023. "Remote Sensing Image Super-Resolution Via Multiscale Enhancement Network." *IEEE Geoscience and Remote Sensing Letters* 20: 1–5.

Wang, Shunzhou, Tianfei Zhou, Yao Lu, and Huijun Di. 2021. "Contextual Transformation Network for Lightweight Remote-Sensing Image Super-Resolution." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–13.

Xu, Yue, Jianya Gong, Xin Huang, Xiangyun Hu, Jiayi Li, Qiang Li, and Min Peng. 2022. "Luojia-HSSR: A High Spatial-spectral Resolution Remote Sensing Dataset for Land-cover Classification with a New 3D-HRNet." *Geo-Spatial Information Science* 1–13. https://doi.org/10.1080/10095020.2022.2070555.

Xu, Wenjia, X. U. Guangluan, Yang Wang, Xian Sun, Daoyu Lin, and W. U. Yirong. 2018. "High Quality Remote Sensing Image Super-Resolution Using Deep Memory Connected Network." In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, 8889–8892. IEEE.

Xue, Zhixiang, Xuchu Yu, Anzhu Yu, Bing Liu, Pengqiang Zhang, and Shentong Wu. 2022. "Self-Supervised Feature Learning for Multimodal Remote Sensing Image Land Cover Classification." *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–15.

Yang, Yi, and Shawn Newsam. 2010. "Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification." In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 270–279.

Yu, Xiaoyu, Jun Pan, Mi Wang, and Jiangong Xu. 2023. "A Curvature-Driven Cloud Removal Method for Remote Sensing Images." *Geo-Spatial Information Science* 1–22. https://doi.org/10.1080/10095020.2023.2189462.

Yuan, Jianye, Xin Ma, Zhentong Zhang, Qiang Xu, Ge Han, Song Li, Wei Gong, Fangyuan Liu, and Xin Cai. 2023. "EFFC-Net: Lightweight Fully Convolutional Neural Networks in Remote Sensing Disaster Images." *Geo-Spatial Information Science* 1–12. https://doi.org/10.1080/10095020.2023.2183145.

Zeng, Kangli, Tao Lu, Xuefeng Liang, Kai Liu, Hui Chen, and Yanduo Zhang. 2019. "Face Super-Resolution Via Bilayer Contextual Representation." *Signal Processing: Image Communication* 75: 147–157.

Zhang, Yifan, Steve De Backer, and Paul Scheunders. 2009. "Noise-Resistant Wavelet-Based Bayesian Fusion of Multispectral and Hyperspectral Images." *IEEE Transactions on Geoscience and Remote Sensing* 47 (11): 3834–3843. https://doi.org/10.1109/TGRS.2009.2017737.

Zhang, Yulun, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. "Image Super-Resolution Using Very Deep Residual Channel Attention Networks." In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhang, Sihang, Zhenfeng Shao, Xiao Huang, Linze Bai, and Jiaming Wang. 2021. "An Internal-External Optimized Convolutional Neural Network for Arbitrary Orientated Object Detection From Optical Remote Sensing Images." *Geo-Spatial Information Science* 24 (4): 654–665. https://doi.org/10.1080/10095020.2021.1972772.

Zhang, Dongyang, Jie Shao, Xinyao Li, and Heng Tao Shen. 2020. "Remote Sensing Image Super-resolution Via Mixed High-order Attention Network." *IEEE Transactions on Geoscience and Remote Sensing* 59 (6): 5183–5196. https://doi.org/10.1109/TGRS.2020.3009918.

Zhang, Yulun, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2020. "Residual Dense Network for Image Restoration." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (7): 2480–2495.

Zhang, Lei, and Xiaolin Wu. 2006. "An Edge-Guided Image Interpolation Algorithm Via Directional Filtering and Data Fusion." *IEEE Transactions on Image Processing* 15 (8): 2226–2238. https://doi.org/10.1109/TIP.2006.877407.

Zhang, Lin, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. "FSIM: A Feature Similarity Index for Image Quality Assessment." *IEEE Transactions on Image Processing* 20 (8): 2378–2386. https://doi.org/10.1109/TIP.2011.2109730.

Zhu, Qiqi, Xi Guo, Weihuan Deng, Qingfeng Guan, Yanfei Zhong, Liangpei Zhang, and Deren Li. 2022. "Land-Use/Land-Cover Change Detection Based on a Siamese Global Learning Framework for High Spatial Resolution Remote Sensing Imagery." *ISPRS Journal of Photogrammetry and Remote Sensing* 184: 63–78. https://doi.org/10.1016/j.isprsjprs.2021.12.005.