

## Remote sensing image super-resolution via cross-scale hierarchical transformer

Yi Xiao, Qiangqiang Yuan, Jiang He & Liangpei Zhang

**To cite this article:** Yi Xiao, Qiangqiang Yuan, Jiang He & Liangpei Zhang (14 Dec 2023):  
Remote sensing image super-resolution via cross-scale hierarchical transformer, Geo-spatial  
Information Science, DOI: [10.1080/10095020.2023.2288179](https://doi.org/10.1080/10095020.2023.2288179)

**To link to this article:** <https://doi.org/10.1080/10095020.2023.2288179>



© 2023 Wuhan University. Published by  
Informa UK Limited, trading as Taylor &  
Francis Group.



Published online: 14 Dec 2023.



Submit your article to this journal



Article views: 594



View related articles



View Crossmark data

# Remote sensing image super-resolution via cross-scale hierarchical transformer

Yi Xiao  <sup>a</sup>, Qiangqiang Yuan  <sup>a</sup>, Jiang He  <sup>a</sup> and Liangpei Zhang  <sup>b</sup>

<sup>a</sup>School of Geodesy and Geomatics, Wuhan University, Wuhan, China; <sup>b</sup>State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

## ABSTRACT

Global and local modeling is essential for image super-resolution tasks. However, current efforts often lack explicit consideration of the cross-scale knowledge in large-scale earth observation scenarios, resulting in suboptimal single-scale representations in global and local modeling. The key motivation of this work is inspired by two observations: 1) There exists hierarchical features at the local and global regions in remote sensing images, and 2) they exhibit scale variation of similar ground objects (e.g. cross-scale similarity). In light of these, this paper presents an effective method to grasp the global and local image hierarchies by systematically exploring the cross-scale correlation. Specifically, we developed a Cross-scale Self-Attention (CSA) to model the global features, which introduces an auxiliary token space to calculate cross-scale self-attention matrices, thus exploring global dependency from diverse token scales. To extract the cross-scale localities, a Cross-scale Channel Attention (CCA) is devised, where multi-scale features are explored and progressively incorporated into an enriched feature. Moreover, by hierarchically deploying CSA and CCA into transformer groups, the proposed Cross-scale Hierarchical Transformer (CHT) can effectively explore cross-scale representations in remote sensing images, leading to a favorable reconstruction performance. Comprehensive experiments and analysis on four remote sensing datasets have demonstrated the superiority of CHT in both simulated and real-world remote sensing scenes. In particular, our CHT outperforms the state-of-the-art approach (TransENet) in terms of PSNR by 0.11 dB on average, but only accounts for 54.8% of its parameters.

## ARTICLE HISTORY

Received 7 August 2023  
Accepted 20 November 2023

## KEY WORDS

Super-resolution;  
transformer; cross-scale;  
hierarchical attention;  
remote sensing

## 1. Introduction

High spatial resolution remote sensing imagery plays an indispensable role in various earth observation tasks (Deren et al. 2023; Wang et al. 2022b), such as land-cover classification (Li et al. 2023; Xu et al. 2022), object detection (Yu et al. 2021) and change detection (Bai et al. 2022; Chen, Yokoya, and Chini 2023; Chen et al. 2022, 2023a; Han et al. 2023). Recent advancements in aerospace sensors have brought impressive capability in acquiring high-resolution imagery, providing massive high-quality data for the research community and significantly advancing downstream applications. However, limited by the intrinsic resolution and complex imaging environment, remote sensing images captured from the aerial platforms can merely reflect partial spatial details, resulting in a suboptimal scene representation.

A straightforward solution to alleviate this issue is improving the imaging resolution of satellite sensors. Nevertheless, it is obviously time-consuming and uneconomic. Fortunately, Super-Resolution (SR) techniques (He et al. 2023a, 2023b; Liu et al. 2023; Ma et al. 2023; Xiao et al. 2023), which algorithmically recover

High-Resolution (HR) images from Low-Resolution (LR) observations, provide a practical alternative. Given an LR observation, SR aims to restore the latent HR counterpart. As the HR solution space is infinite, SR is inherently an ill-posed issue (Zhang et al. 2022b). Broadly, SR can be categorized into two types: traditional and deep-learning-based approaches. Traditional methods often employ interpolation (Freeman, Jones, and Pasztor 2002; Zhou, Yang, and Liao 2012) interpolation or rely on hand-crafted prior knowledge (Timofte, De Smet, and Van Gool 2013; Zhang et al. 2023) to compact the solution space. However, these methods have limited universality and suffer from high computational complexity. Subsequently, owing to the non-linear representation capability of Convolutional Neural Networks (CNNs) (Fernandez-Beltran, Latorre-Carmona, and Pla 2017; Xiao et al. 2022; Miao et al. 2022a, 2023; Jiang et al. 2021; He et al. 2021; Yu et al. 2023; Zhou, Li, et al. 2023a; Zhou, Liu, et al. 2023b), more latent priors can be explored by CNNs to tame the ill-posedness. Notable efforts include deep (Kim, Kwon Lee, and Mu Lee 2016) and wide (Lim et al. 2017) architectures,

**CONTACT** Qiangqiang Yuan  [yqiang86@gmail.com](mailto:yqiang86@gmail.com)

© 2023 Wuhan University. Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

dense models (Jiang et al. 2020; Zhang et al. 2018b) and attention-based approaches (Mei, Fan, and Zhou 2021; Niu et al. 2020). However, few efforts have been made to enhance the representation of cross-scale correlations in remote sensing scenes, as they are mainly developed for natural images.

Recently, a cross-scale exploration network (Lei and Shi 2022) was proposed to grasp the cross-scale similarity in remote sensing imagery. Despite achieving favorable performance, the CNN-based approaches are still less capable of modeling the long-range properties, thus reaching a plateau of performance. Drawing upon the success of Transformer (Ma et al. 2022), the transformer-based SR methods (Chen et al. 2023b; Lei and Shi 2022; Xiao et al. 2022b) have demonstrated comparable, and even superior performance against CNN-based models. Owing to the global modeling strength of self-attention (Pei et al. 2023; Peng et al. 2023; Xiao et al. 2023c), the transformer naturally excels in exploring global dependencies, which is beneficial for SR tasks. More recently, a multi-stage transformer-enhanced network (Lei, Shi, and Wenjing 2022) was developed for remote sensing image SR. Nevertheless, these methods have overlooked the cross-scale characteristics present in remote sensing images, failing to adequately consider these valuable priors.

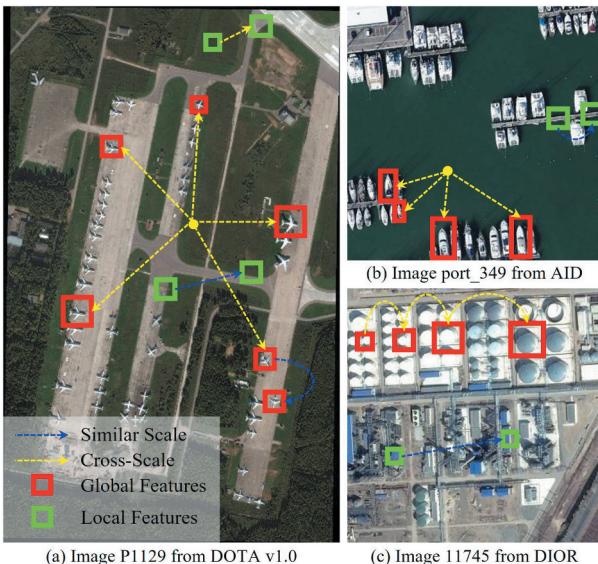
More precisely, compared to natural images, remote sensing images have some unique properties. Firstly, in large-scale observation scenes, hierarchical cross-scale priors exist at both local and global ranges. This can be intuitively illustrated in Figure 1. To capture local features such as edges and colors, convolution layers with single-scale (*e.g.*,  $3 \times 3$ ) designs

are commonly employed. However, this single-scale approach only captures partial representations, leading to suboptimal exploration of cross-scale localities (green boxes in Figure 1). On the other hand, although transformers could extract global dependency with self-attention, they fail to consider the cross-scale global similarity (red boxes in Figure 1). This limitation arises because the scale of the feature space used for self-attention calculations is often fixed. For instance, the self-attention matrix is typically computed by measuring the similarity between query-key pairs at the same scale, lacking flexibility in estimating cross-scale self-attention.

To mitigate the aforementioned challenges, this paper proposes an effective Cross-scale Hierarchical Transformer (CHT) to systematically model the local and global cross-scale features in remote sensing images. Specifically, instead of directly computing the self-attention matrix from query-key pairs, we proposed a Cross-scale Self-Attention (CSA) module to implicitly model the cross-scale global representations in two subspaces. The CSA module incorporates an auxiliary token (denoted as  $S$ ) with diverse feature scales compared to the query ( $Q$ ) and key ( $K$ ), allowing for intermediate self-attention computation between  $Q - S$  and  $K - S$ . This approach enables CHT to measure the cross-scale relevance in different feature spaces with varying scales. Essentially, the auxiliary token decomposes the attention calculation between query-key pairs into two subspaces, making the scale of self-attention more flexible. Furthermore, to better explore cross-scale features at the local range, a Cross-scale Channel Attention (CCA) mechanism is developed to generate a series of multi-scale local features and progressively refine them in a coarse-to-fine manner. By jointly incorporating the strengths of CSA and CCA within transformer groups, our CHT leverages image hierarchies and achieves notable performance improvements for remote sensing image super-resolution tasks.

In short, the main contribution of this paper is listed as follows:

- (1) The proposed CHT hierarchically models both global and local cross-scale representations, enabling a comprehensive understanding of the valuable cross-scale knowledge in remote sensing images.
- (2) To explore the cross-scale global features, a cross-scale self-attention module is proposed, which could implicitly compute cross-scale self-attentions within feature spaces at different scales.
- (3) To model the cross-scale local representations, we devise a cross-scale channel attention model, where cross-scale features are explored



**Figure 1.** Remote sensing images present a hierarchy of features in local and global ranges. The cross-scale local (green boxes), *e.g.*, edges and colors, and global (red boxes) features, *e.g.*, cross-scale self-similarity, cannot be fully explored by single-scale CNNs and transformer.

and mutually enriched with each other in a coarse-to-fine manner.

This paper is organized as follows. Section 2 reviews remote sensing imagery super-resolution. In Section 3, we describe the details of our CHT. Section 4 reports extensive experiments and ablation analysis. Section 5 is the conclusion.

## 2. Related work

### 2.1. Natural image super-resolution

#### 2.1.1. CNN-based methods

CNN-based SR methods have achieved remarkable success in natural images. Inspired by the Super-Resolution Convolution Neural Network (SRCNN) (Dong et al. 2015), the CNN-based SR methods are booming. Kim et al. (2016) promoted the performance of SRCNN by increasing the depth of the network. Lim et al. (2017) choose to expand the width of CNN and achieve favorable performance. However, early methods often employ interpolation to enlarge the spatial size of LR input, which may introduce interference in the pre-processing operation. To this end, Shi et al. (2016) devise an efficient sub-pixel convolution layer to realize the upscale procedure, which has been widely adopted in current SR models. Later, researchers introduced the concepts of residual connections (Wang et al. 2018) and dense connections (Zhang et al. 2018b) to improve the optimization process.

#### 2.1.2. Attention-based methods

To further enhance the representation capability of CNNs, a crowd of attention has been explored to guide the network to pay more attention to the critical features. Zhang et al. (2018a) proposed a residual channel attention network. After that, Niu et al. (2020) put forward to combine channel attention, layer attention, and spatial attention to residual block and achieve impressive performance. Recently, Mei, Fan, and Zhou (2021) proposed sparse non-local attention to increasing the respective of CNNs. Although such non-local operators can achieve decent performance in modeling the global features, the inherent limitation of CNNs still limits the performance.

#### 2.1.3. Transformer-based methods

Inspired by the success of transformers in Natural Language Processing (NLP) tasks, transformers have also shown great potential in image restoration, due to the long-range modeling capability of self-attention (Wang et al. 2023). In particular, Yang et al. (2020) learned a texture transformer and obtained superior performance by transferring more global textures for SR. Liang et al. (2021) leveraged the swim-transformer

architecture for image restoration. Chen et al. (2023b) proposed a hybrid attention network to activate more useful pixels in the SR process. Lu et al. (2022) develop an efficient transformer to save memory cost for light-weight SR.

### 2.2. Remote sensing image super-resolution

In remote sensing scenarios, early efforts often employed hand-crafted prior (Pan et al. 2013) to tackle this issue. With the booming of deep learning, researchers began to reuse the model proposed for natural images by simply retraining them in remote sensing images (Jiang et al. 2018, 2019; Lei, Shi, and Zou 2020; Luo et al. 2017). However, these methods usually face challenges in effectively capturing both local and global features using naive convolutions. In addition, they barely consider the unique properties of remote sensing imagery, cross-scale characteristics in particular.

Subsequently, researchers begin to consider the unique characteristics of remote sensing images, and various elaborate structures have been developed to explore these valuable representations. Zhang et al. (2020b) proposed a scene-adaptive strategy to boot the generalization of SR models on various remote sensing scenes. To exploit the multi-scale knowledge in remote sensing images, Dong et al. (2020b) designed a multi-scale second-order attention network. Similarly, Zhang et al. (2020a) put forward a mixed high-order attention network to explore the high-frequency information that is critical in remote sensing scenes. Dong et al. (2020a) proposed a dense-sampling network, which considers the multi-level priors and enhances them with a wide activation mechanism. Recently, Feng et al. (2022) developed a pyramid structure network to model multi-scale features for better SR. Lei and Shi (2022) explicitly considered cross-scale self-similarity with non-local attention. Wang et al. (2022c) introduced a lightweight lattice block for efficient SR of large-scale remote sensing images. Liang et al. (2023) employed the graph neural network for effective SR. Xiao et al. (2023a) introduced the diffusion model to restore realistic results. Although these methods paid more attention to the unique properties of remote sensing data, they still struggled to handle the exploration of long-range features, especially the global cross-scale priors as shown in Figure 1. Many works (Wu, Hong, and Chanussot 2022, 2023) have explored and demonstrated the significance of cross-scale information in remote sensing images, prompting us to place more emphasis on both local and global cross-scale knowledge.

More recently, the transformer has demonstrated the superior capability of modeling long-range representations. Yao et al. (2023) proposed a novel

transformer framework by extending the conventional vision transformer into the multimodal domain. Tu et al. (2022) integrated a swim-transformer block into the Generative Adversarial Network (GAN) to boost the performance with the self-attention mechanism. Lei et al. (2022) employed self-attention to mine more global dependencies and enhance them in multiple stages. However, it is worth noting that remote sensing data often exhibits intricate cross-scale and hierarchical distributions in both local and global features. Relying solely on single-scale characterizations can yield suboptimal results. Therefore, it is imperative to develop a systematic approach, which can comprehensively model both local and global information while fully considering the inherent cross-scale characteristics.

### 2.3. Transformer in remote sensing applications

In addition to super-resolution, transformer has demonstrated impressive performance in various remote sensing applications, such as classification, detection, segmentation, pan-sharpening, and change detection. Roy et al. (2023) incorporated the vision transformer into CNN and introduced a new multimodal fusion transformer network for image classification. Li, Chen, and Zeng (2022) introduced a transfer CNN into a transformer for accurate object detection in large-scale remote sensing images. Wang et al. (2022a) developed a transformer-based network for efficient semantic segmentation. To reduce the spectral redundancy, Zhang et al. (2022a) exploited the channel reduction strategy and proposed a transformer-based network for pan-sharpening. Chen, Qi, and Shi (2021) proposed a bitemporal

transformer to explore the spatial-temporal context for change detection. All of these studies have demonstrated the tremendous potential for the application of transformers in the field of remote sensing.

## 3. Methodology

### 3.1. Overview

As shown in Figure 2, our CHT comprises three main components: 1) Shallow feature extraction, which is responsible for mapping the low-resolution input  $I^{LR}$  to a higher-dimensional feature space. 2) Deep feature extraction, where multiple transformer groups are employed to explore deep feature representations. 3) Reconstruction, which focuses on reconstructing the high-resolution output  $I^{SR}$ .

### 3.2. Shallow and deep feature extraction

#### 3.2.1. Shallow feature extraction

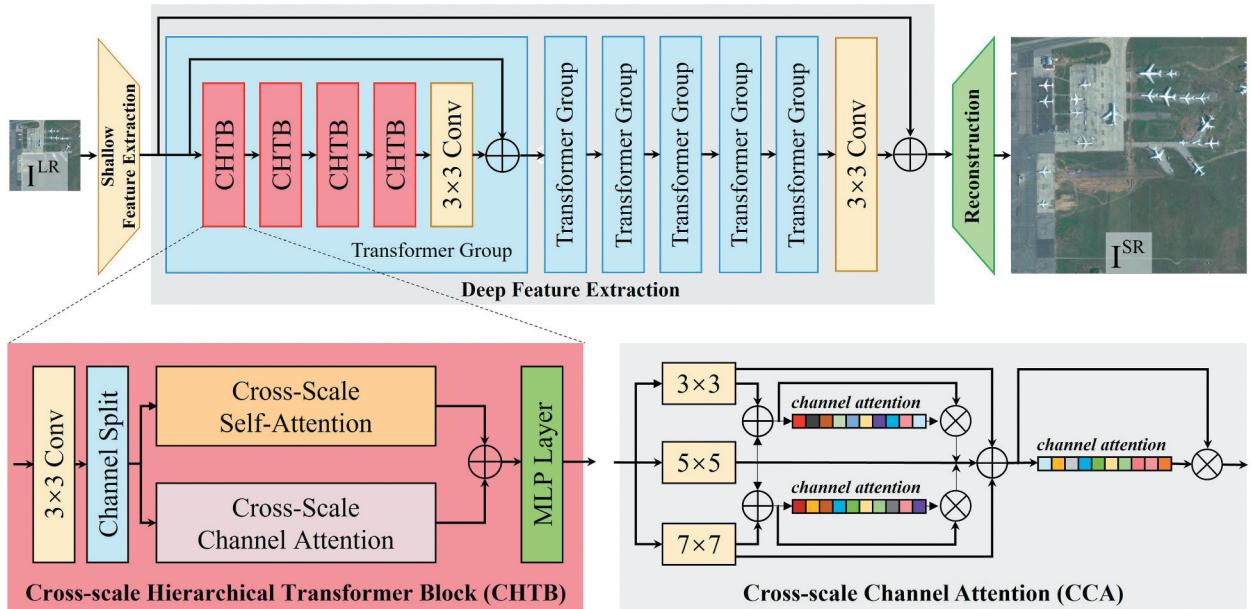
Formally, given a low-resolution input  $I^{LR} \in \mathbb{R}^{H \times W \times 3}$  ( $H$ ,  $W$  and 3 are the height, width and RGB channels number, respectively), we exploit a  $3 \times 3$  convolution  $\text{Conv}(\cdot)$  to extract a shallow feature  $\mathcal{F}_0 \in \mathbb{R}^{H \times W \times D}$ , which means:

$$\mathcal{F}_0 = \text{Conv}(I^{LR}) \quad (1)$$

where  $D$  represents the channel number.

#### 3.2.2. Deep feature extraction

As illustrated in Figure 2, our deep feature extraction includes multiple cascaded transformer groups followed by a  $3 \times 3$  convolutional layer, enabling us to effectively explore deep feature representations at local



**Figure 2.** Overview of the proposed CHT, which consists of three parts: shallow feature extraction, deep feature extraction, and reconstruction.

and global ranges. To ensure stable optimization, we incorporate a global connection. Let us denote the  $i$ -th transformer group as  $g_i$ . The output of  $g_i$  can be expressed as:

$$\mathcal{F}_i = g_i(\mathcal{F}_{i-1}), i = 1, 2, \dots, L \quad (2)$$

where  $\mathcal{F}_{i-1}$  is the output of  $g_{i-1}$  and  $L$  is the number of transformer groups.

Within each transformer group, we incorporate several Cross-scale Hierarchical Transformer Blocks (CHTB) along with a  $3 \times 3$  convolutional layer. The CHTB is a well-designed module that effectively models global and local cross-scale representations through the utilization of hierarchical CSA and CCA. Additionally, inserting a convolutional layer at the end of the CHTB helps to introduce more inductive bias to the transformer for better reconstruction.

### 3.3. Cross-scale hierarchical transformer block

Given an input feature  $\mathcal{F}_{i,0}$  to the 1-st CHTB in the  $i$ -th transformer group, denoted as  $h_{i,1}$ , the output of  $h_{i,j}$  can be formulated as follows:

$$\mathcal{F}_{i,j} = h_{i,j}(\mathcal{F}_{i,j-1}), j = 1, 2, \dots, M \quad (3)$$

where  $M$  is the number of CHTB in each transformer group. As depicted in Figure 2, in each CHTB, we first use a  $3 \times 3$  convolution to extract intermediate features from the input. Next, the intermediate features are split along the channel dimension, and these split feature maps are fed into the hierarchical CSA and CCA modules. The CSA module captures global cross-scale dependencies, allowing the model to leverage prior knowledge about long-range interactions. On the other hand, the CCA module focuses on capturing local cross-scale dependencies, enabling the model to emphasize the importance of channel-wise interactions within each scale. After obtaining the outputs from the CSA and CCA modules, the results are merged using the summation operator to combine the global and local cross-scale information effectively. Finally, a Multi-Layer Perceptron (MLP) network is employed to perform feature feed-forward, further enhancing the learned representations and preparing them for the subsequent CHTB.

#### 3.3.1. Cross-scale self-attention

Taking inspiration from the insights presented in Figure 1, we aim to capture global cross-scale similarity using the proposed CSA. Figure 3 illustrates the structure of CSA. Instead of directly computing self-attention from query-key pairs, we introduce a scale compress as an intermediate operator to implicitly calculate the cross-scale self-similarity. In particular, given an input  $X \in \mathbb{R}^{C \times HW}$ , the standard tokens in

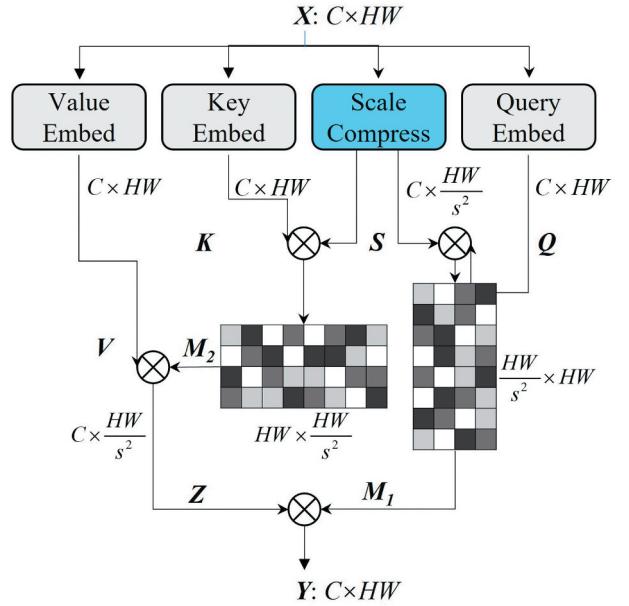


Figure 3. The diagram of the proposed CSA.

transformer, *i.e.*,  $Q$  (query),  $K$  (key), and  $V$  (value), can be obtained with linear projection:

$$Q, K, V = \text{Emb}(X) \quad (4)$$

where  $\text{Emb}(\cdot)$  is set to a fully-connected layer. In our work, we introduced an additional token denoted as  $S$ , which enables self-attention operations at different resolutions. This allows us to explore cross-scale global dependencies across different token spaces. To achieve this, we compress the input feature map  $X$  with a scale factor  $s$ , denoted as  $S = \text{Compress}(X, s)$ . The compression operation is performed using a fully connected layer. Later, the cross-scale self-attention is conducted using Equation (5):

$$M_1 = \sigma\left(\frac{Q \cdot S^T}{\lambda}\right), M_2 = \sigma\left(\frac{K \cdot S^T}{\lambda}\right) \quad (5)$$

where  $M_1 \in \mathbb{R}^{HW \times HW}$  and  $M_2 \in \mathbb{R}^{HW \times HW/s^2}$  are two learnable intermediate attention matrix yielded from different scales,  $\sigma$  means softmax function, and  $\lambda = \sqrt{C}$  is a temperature parameter. The attention  $M_2$  is used to distill  $V$  into a intermediate feature  $Z \in \mathbb{R}^{C \times HW/s^2}$ . Finally, the attention  $M_1$  expands the dimension of  $Z$  and generates the output  $Y \in \mathbb{R}^{C \times HW}$ . These processes can be formulated by the following:

$$Y = (V \otimes M_2) \otimes M_1 \quad (6)$$

#### 3.3.2. Cross-scale channel attention

As shown in Figure 2, given an input of  $P \in \mathbb{R}^{C \times H \times W}$ , we firstly set three parallel convolutions with different kernel sizes to explore the multi-scale features:

$$P_1 = f_{3 \times 3}(P), P_2 = f_{5 \times 5}(P), P_3 = f_{7 \times 7}(P) \quad (7)$$

where  $f_{k \times k}$  means a convolution layer with kernel size of  $k$ . Subsequently, channel attention is adopted to explore the cross-scale correlation. Take  $P_{1,2} = P_1 + P_2$  as an example, the attention map  $att_1$  can be generated by the following:

$$att_1 = \sigma\zeta(P_{1,2}) \quad (8)$$

where  $\zeta$  is a squeeze operation achieved by a  $3 \times 3$  convolution and a global pooling layer,  $\sigma$  denotes softmax activation. Similarly, we could obtain the cross-attention  $att_2$  from  $P_{2,3} = P_2 + P_3$ . Subsequently, we enrich the interaction of cross-scale features with Equation (9):

$$P_4 = P_{1,2} \otimes att_1 + P_{2,3} \otimes att_2 + P_1 + P_2 + P_3 \quad (9)$$

In the end, another channel attention is employed to generate the final cross-scale feature  $O$ :

$$O = P_4 \otimes \sigma\zeta(P_4) \quad (10)$$

### 3.4. Reconstruction

In the reconstruction process, the widely used pixel-shuffle layer (Shi et al. 2016) is adopted to upscale the spatial size of deep feature  $\mathcal{F}_{out}$ , it can be formulated by:

$$\mathcal{F}_{out} = \mathcal{F}_0 + \text{Conv}(\mathcal{F}_L) \quad (11)$$

where  $\mathcal{F}_0$  is the shallow feature and  $\mathcal{F}_L$  means the deep feature from the last transformer group. Before pixel-shuffle layer PS( $\cdot$ ), a  $3 \times 3$  convolution is used to increase the channel dimension. In the end, the super-resolved output  $I^{SR} \in \mathbb{R}^{\mathbb{H} \times \mathbb{W} \times \mathbb{C}}$  can be obtained:

$$I^{SR} = \text{PS}(\text{Conv}(\mathcal{F}_{out})) \quad (12)$$

where  $r$  is the scaling factor.

**Table 1.** An introduction of four datasets involved in this paper.

Data name	Experiments	Usage	Pixel size	Image numbers	Resolution	Categories	Tasks
AID	Simulated	train, test	512 × 512	3000, 900	0.8–8m	30	Classification
DOTA v1.0	Simulated	test	512 × 512	900	—	15	Detection
DIOR	Simulated	test	512 × 512	1000	—	20	Detection
NWPU-RESISC45	Real-world	test	128 × 128	315	0.2–30m	45	Classification



**Figure 4.** Some typical samples from (a) AID, (b) DOTA v1.0, (c) DIOR and (d) NWPU-RESISC45 datasets.

## 4. Experiment

### 4.1. Remote sensing datasets

We choose four widely used benchmarks in the remote sensing field to evaluate the SR performance, including the AID (Xia et al. 2017) and NWPU-RESISC45 (Cheng, Han, and Lu 2017) dataset used for scene classification task and DOTA v1.0 (Xia et al. 2018), and DIOR (Li et al. 2020) dataset designed for objection detection tasks. Among them, AID is used to build the training and test sets. The remainder datasets are employed for model tests. Note that we do not perform simulated degradation on the NWPU-RESISC45 dataset and only adopt it to evaluate the models on real-world degradations. Table 1 reports the details of these datasets. Moreover, some typical samples in these datasets are shown in Figure 4.

### 4.2. Implementation details

#### 4.2.1. Model details

In this study, we focus on four times SR for remote sensing imagery, *i.e.*,  $r = 4$ . For the design of our CHT, in the deep feature extraction process, we employed six transformer groups to explore deep feature representations. The number of CHTB within each transformer group was dynamically set as follows: 4, 4, 8, 8, 4, 4. This configuration allowed us to effectively utilize cross-scale knowledge to enhance the performance of the model. The channel dimension of CHT is set to 180. During self-attention calculation, the head number is fixed to three.

#### 4.2.2. Training details

For a fair comparison, all SR methods involved in this paper were retrained from scratch on the AID training

set without any pre-training and fine-tuning process. During the training process of our CHT, we randomly selected four LR image patches in each mini-batch and also augmented them by  $90^\circ$ ,  $180^\circ$ , and  $270^\circ$  rotation and flipping. The patch size is set to  $64 \times 64$ . To optimize our CHT, the widely used  $L_1$  loss is adopted as the objective function, which measures the pixel-wise distance between the super-resolved image  $I^{SR}$  and the ground-truth image  $I^{HR}$ . In particular,  $L_1 = \|I^{HR} - I^{SR}\|_1$ . We used Adam optimizer to train our CHT with a fixed learning rate of  $10^{-4}$ . All the experiments were conducted on a single NVIDIA RTX 3090 GPU with 24 GB memory and a 3.40 GHz AMD Ryzen 5700X CPU.

#### 4.3. Evaluation metrics

In the experiments conducted on simulated data, we utilize two classical full-reference metrics, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM), to quantitatively evaluate the performance of the SR method. These metrics provide an objective assessment of the fidelity of the restored image. It is important to note that for the calculation of PSNR and SSIM, we consider only the luminance channel (Y channel) of the YCbCr color space. Higher PSNR and SSIM values indicate a higher level of fidelity in the restored image. In our experiments with real-world data, we employ two non-reference metrics, as the ground truth image is unavailable. The first indicator is the Average Gradient (AG), which measures the high-frequency details present in an image. A higher AG value indicates better preservation and enhancement of high-frequency components in the restored image. Additionally, we employ the Natural Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2012) to evaluate the perceptual quality of the SR results. Lower NIQE scores indicate better perceptual quality in the restored image.

#### 4.4. Experiments on simulated data

##### 4.4.1. Selected methods for comparison

In this section, we compare our CHT with the following 11 typical SR models, including eight methods used for natural images: Bicubic interpolation, SRCNN (Dong et al. 2015), VDSR (Kim, Kwon Lee, and Mu Lee 2016), SRResNet (Wang et al. 2018), EDSR (Lim et al. 2017), RDN (Zhang et al. 2018b), HAN (Niu et al. 2020), NLSA (Mei, Fan, and Zhou 2021), and three state-of-the-art models developed for remote sensing imagery: RSINet (Feng et al. 2022), HSENet (Lei and Shi 2022), TransENet (Lei, Shi, and Wenjing 2022). In particular, SRCNN and VDSR adopt shallow and deep convolution for

reconstruction. EDSR is a classical method, which expands the width of CNN by a wide activation strategy. Note that EDSR won the first prize in SR tracking in NTIRE 2017. SRResNet and RDN upgrade the CNN with residual and dense connections. SRResNet and RDN improve the non-linear fitting capability of CNN by residual and dense connection, respectively. HAN proposes to use holistic attention for local feature modeling. NLSA devises a non-local sparse attention to exploring the global features. In models specifically designed for remote sensing images, RSINet and HSENet both exploit cross-scale knowledge with pyramid structure and non-local attention. TransENet is a transformer-based approach. Following their official implementation details, we retrained these models from scratch for a fair comparison. Notably, we also perform a self-ensemble strategy in the test phase to further boost the performance of our CHT, and we denote it as CHT+.

##### 4.4.2. Quantitative comparison

Table 2 presents the quantitative comparisons between our CHT and 11 typical models on the AID test set for a scale factor of four. The best and second best PSNR/SSIM performance are highlighted in red and blue, respectively. It is observed that RDN, EDSR, RSINet, HSENet, and TransENet have achieved competitive results on the AID dataset, surpassing methods proposed as early as five years ago, such as SRCNN, VDSR, and SRResNet. These results highlight the importance of modeling local or global features for improved SR. Comparatively, our CHT demonstrates the best performance on almost all 30 scene types, surpassing the second-best model, TransENet, by 0.11 dB in terms of PSNR. This indicates the superiority of our CHT in various remote sensing scenarios.

In Table 3, we further test these models on DOTA v1.0 and DIOR test sets. The PSNR/SSIM results demonstrate that our CHT+ outperforms all the comparative networks. For example, CHT+ achieves a higher PSNR than NLSA by 0.2 dB and surpasses TransENet by 0.14 dB. Even without the self-ensemble strategy, our CHT still achieves the best performance on both the DOTA v1.0 and DIOR datasets. Note that NLSA utilizes non-local attention to explore global features, which are less powerful compared to our cross-scale self-attention mechanism in capturing long-range dependencies. Although TransENet exploits self-attention to enhance performance, CHT can still surpass it by a large margin. The main reasons may lie in 1) the proposed cross-scale self-attention not only grasps the valuable global prior knowledge but also considers the cross-scale properties of similar ground objects, and 2) the hierarchical CSA and CCA in our CHT effectively benefit both local and global hierarchies in remote sensing imagery.

**Table 2.** Quantitative results on AID test set. Here we report the PSNR/SSIM performance of SISR models on 30 classes of scenes. The best and second best metrics are shown in **red bold** and **blue bold**, respectively.

Land Cover	Bicubic		SRCNN		VDSR		SRResNet		RDN		EDSR		RSINet		HSENet		TransENet		CHT (Ours)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airport	27.83	0.7554	28.80	0.7912	29.74	0.8205	29.79	0.8213	30.10	<b>0.8310</b>	29.93	0.8282	30.07	0.8292	30.08	0.8303	<b>30.15</b>	0.8301	<b>30.28</b>	<b>0.8350</b>
Bare Land	35.60	0.8564	36.33	0.8728	36.74	0.8808	36.80	0.8815	36.80	0.8839	<b>36.94</b>	0.8837	<b>36.96</b>	0.8840	36.79	<b>0.8841</b>	36.93	0.8837	36.84	<b>0.8846</b>
Baseball Field	31.00	0.8305	32.00	0.8550	32.85	0.8719	32.94	0.8730	33.16	<b>0.8778</b>	33.05	0.8765	33.18	0.8771	33.15	0.8774	<b>33.25</b>	0.8775	<b>33.37</b>	<b>0.8805</b>
Beach	32.90	0.8446	33.61	0.8616	34.11	0.8710	34.14	0.8717	34.16	0.8748	34.18	0.8727	<b>34.33</b>	0.8747	34.14	0.8746	<b>34.35</b>	0.8754	34.29	<b>0.8764</b>
Bridge	30.22	0.8283	31.33	0.8522	32.56	0.8733	32.61	0.8736	33.03	<b>0.8814</b>	32.99	0.8800	32.99	0.8793	33.06	0.8809	<b>33.08</b>	0.8810	<b>33.23</b>	<b>0.8836</b>
Center	26.51	0.6944	27.42	0.7358	28.45	0.7783	28.52	0.7790	28.86	<b>0.7945</b>	28.77	0.7921	28.88	0.7932	28.83	0.7937	<b>28.91</b>	0.7934	<b>29.09</b>	<b>0.8015</b>
Church	24.29	0.6333	25.27	0.6899	26.21	0.7353	26.25	0.7361	26.52	0.7506	26.30	0.7469	<b>26.53</b>	0.7506	26.47	<b>0.7507</b>	26.52	0.7492	<b>26.71</b>	<b>0.7576</b>
Commercial	27.33	0.7174	28.22	0.7593	28.91	0.7876	28.98	0.7888	29.21	<b>0.7994</b>	29.01	0.7940	<b>29.22</b>	0.7983	29.21	0.7989	29.21	0.7973	<b>29.35</b>	<b>0.8046</b>
Dense Residential	22.93	0.5671	23.73	0.6305	24.41	0.6755	24.44	0.6754	24.66	0.6921	24.38	0.6839	24.69	0.6921	24.60	0.6912	<b>24.71</b>	0.6931	<b>24.82</b>	<b>0.7009</b>
Desert	39.26	0.9100	39.67	0.9191	40.05	0.9244	40.10	0.9248	39.81	0.9269	40.20	0.9268	<b>40.25</b>	0.9272	39.57	0.9271	<b>40.29</b>	0.9276	39.95	<b>0.9283</b>
Farmland	33.10	0.8226	33.75	0.8407	34.65	0.8610	34.73	0.8623	34.98	0.8691	35.00	0.8683	34.99	0.8681	<b>35.02</b>	0.8692	34.99	0.8675	<b>35.10</b>	<b>0.8707</b>
Forest	28.79	0.6605	29.46	0.7086	29.74	0.7236	29.79	0.7256	29.93	0.7336	29.85	0.7315	29.97	0.7353	<b>30.00</b>	0.7363	29.99	<b>0.7372</b>	<b>30.04</b>	<b>0.7390</b>
Industrial	26.77	0.6952	27.71	0.7417	28.59	0.7800	28.66	0.7819	28.99	<b>0.7966</b>	28.88	0.7931	<b>29.00</b>	0.7950	28.98	0.7956	28.98	0.7942	<b>29.16</b>	<b>0.8028</b>
Meadow	33.86	0.7483	34.20	0.7654	34.42	0.7733	34.46	0.7748	34.53	0.7800	<b>34.63</b>	0.7804	34.61	<b>0.7810</b>	34.55	0.7804	34.62	0.7805	<b>34.72</b>	<b>0.7828</b>
Medium Residential	26.36	0.6335	27.23	0.6842	28.08	0.7244	28.17	0.7253	28.51	0.7405	28.34	0.7365	<b>28.52</b>	0.7396	28.45	0.7390	28.48	0.7385	<b>28.65</b>	<b>0.7462</b>
Mountain	29.51	0.7349	30.25	0.7715	30.62	0.7862	30.65	0.7870	30.71	<b>0.7909</b>	30.63	0.7885	30.73	0.7908	30.72	0.7907	<b>30.76</b>	0.7915	<b>30.80</b>	<b>0.7938</b>
Park	29.06	0.7530	29.89	0.7887	30.46	0.8089	30.52	0.8100	30.68	0.8165	30.54	0.8130	30.70	0.8159	30.71	0.8167	30.72	0.8174	<b>30.81</b>	<b>0.8211</b>
Parking	24.24	0.7060	25.45	0.7586	26.58	0.8086	26.73	0.8119	27.48	0.8363	27.25	0.8317	27.42	0.8344	27.32	0.8341	27.60	0.8396	<b>27.88</b>	<b>0.8470</b>
Playground	32.64	0.8450	33.63	0.8649	34.91	0.8873	35.05	0.8880	35.40	0.8950	35.37	0.8943	35.47	0.8948	35.46	0.8952	35.53	0.8956	<b>35.64</b>	<b>0.8983</b>
Pond	30.70	0.8167	31.39	0.8372	31.97	0.8507	31.99	0.8510	32.14	0.8550	32.11	0.8542	32.16	0.8547	32.17	0.8549	32.19	0.8552	<b>32.26</b>	<b>0.8569</b>
Port	26.67	0.7986	27.64	0.8299	28.40	0.8535	28.47	0.8546	28.75	<b>0.8629</b>	28.50	0.8596	28.77	0.8624	28.71	0.8623	<b>28.77</b>	0.8626	<b>28.97</b>	<b>0.8668</b>
Railway Station	26.78	0.6793	27.69	0.7270	28.44	0.7610	28.51	0.7624	28.86	<b>0.7775</b>	28.72	0.7738	28.82	0.7749	28.84	0.7762	<b>28.88</b>	0.7756	<b>29.03</b>	<b>0.7832</b>
Resort	26.79	0.7029	27.65	0.7429	28.35	0.7718	28.43	0.7735	<b>28.70</b>	<b>0.7840</b>	28.52	0.7799	28.69	0.7824	28.64	0.7825	28.67	0.7825	<b>28.82</b>	<b>0.7881</b>
River	30.37	0.7402	31.00	0.7693	31.44	0.7847	31.48	0.7858	31.60	0.7902	31.55	0.7891	31.60	0.7901	31.61	0.7904	<b>31.63</b>	<b>0.7905</b>	<b>31.68</b>	<b>0.7926</b>
School	27.41	0.7237	28.35	0.7650	29.17	0.7962	29.26	0.7975	29.52	<b>0.8086</b>	29.36	0.8044	<b>29.54</b>	0.8072	29.51	0.8074	29.51	0.8074	<b>29.70</b>	<b>0.8144</b>
Sparse Residential	26.66	0.6006	27.28	0.6462	27.69	0.6666	27.71	0.6674	<b>27.86</b>	<b>0.6757</b>	27.71	0.6728	27.86	0.6755	27.84	0.6754	27.85	0.6754	<b>27.93</b>	<b>0.6790</b>
Square	28.55	0.7391	29.56	0.7780	30.51	0.8107	30.63	0.8120	30.94	<b>0.8230</b>	30.84	0.8200	30.97	0.8214	30.94	0.8223	<b>30.98</b>	0.8227	<b>31.13</b>	<b>0.8282</b>
Stadium	27.16	0.7547	28.11	0.7889	29.25	0.8260	29.33	0.8275	29.73	0.8404	29.63	0.8387	29.70	0.8391	29.68	0.8396	<b>29.73</b>	0.8396	<b>29.98</b>	<b>0.8470</b>
Storage Tanks	25.65	0.6793	26.55	0.7232	27.34	0.7579	27.38	0.7583	27.61	<b>0.7695</b>	27.44	0.7664	<b>27.61</b>	0.7682	27.58	0.7688	27.58	0.7680	<b>27.70</b>	<b>0.7734</b>
Viaduct	26.97	0.6755	27.85	0.7244	28.68	0.7619	28.72	0.7627	<b>29.11</b>	<b>0.7792</b>	28.99	0.7757	29.04	0.7754	29.08	0.7772	29.08	0.7775	<b>29.31</b>	<b>0.7865</b>
Average	28.86	0.7382	29.70	0.7741	30.44	0.8004	30.51	0.8015	30.74	<b>0.8112</b>	30.65	0.8086	30.78	0.8104	30.72	0.8108	<b>30.80</b>	0.8109	<b>30.91</b>	<b>0.8157</b>

**Table 3.** Quantitative comparisons with 11 typical SR models on DOTA v1.0 and DIOR test sets. The best and second best PSNR/SSIM results are highlighted in **red** and **blue**.

Model Type	Method	DOTA v1.0		DIOR		Average	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	Bicubic	31.16	0.7947	28.57	0.7432	29.87	0.7690
	SRCCNN (Dong et al. 2015)	32.10	0.8264	29.49	0.7768	30.80	0.8016
	VDSR (Kim, Lee, and Lee 2016)	33.22	0.8569	30.36	0.8036	31.79	0.8303
	SRResNet Wang et al. (2018)	33.40	0.8585	30.47	0.8053	31.94	0.8319
	RDN (Zhang et al. 2018b)	33.60	0.8670	30.78	0.8147	32.19	0.8409
CNN-based	EDSR Lim et al. (2017)	33.64	0.8648	30.63	0.8116	32.14	0.8382
	HAN (Niu et al. 2020)	33.84	0.8682	30.84	0.8163	32.34	0.8423
	NLSA (Mei, Fan, and Zhou 2021)	33.86	0.8682	30.82	0.8156	32.34	0.8419
	RSINet (Feng et al. 2022)	33.84	0.8665	30.78	0.8136	32.31	0.8401
	HSENet (Lei and Shi 2022)	33.85	0.8667	30.77	0.8143	32.31	0.8405
	TransENet (Lei, Shi, and Mo 2022)	33.75	0.8675	30.85	0.8148	32.30	0.8412
Transformer-based	<b>CHT (Ours)</b>	<b>33.86</b>	<b>0.8711</b>	<b>30.95</b>	<b>0.8187</b>	<b>32.41</b>	<b>0.8449</b>
	<b>CHT+ (Ours)</b>	<b>34.00</b>	<b>0.8730</b>	<b>31.07</b>	<b>0.8211</b>	<b>32.54</b>	<b>0.8471</b>

#### 4.4.3. Qualitative comparison

In Figure 5, we visually exhibit the reconstruction results of these models on the AID test set. Note that we choose two airport scenes with rich cross-scale global features, (e.g., cross-scale similarity of the plane) to better investigate the challenges in remote sensing scenes. From these results, we can find that our CHT can always restore the accurate and sharp details of the planes, while other existing CNN and Transformer-based methods struggle with severe high-frequency loss and artifacts, leading to blurry results and incomplete shapes of the planes. For example, in the “airport\_108” image, existing CNN and Transformer-based methods obtain heavy blurs and fail to recover the entire shape of the planes. Despite HSENet generating some high-frequency details, such as the airfoil, it still can not predict the correct boundary of the plane, and even mix the small plane with the background. In contrast, CHT successfully recovers the complete shape of the plane and distinguishes it from the background without color distortion. These results demonstrate CHT can effectively explore valuable cross-scale local and global priors, which aligns with the motivation behind our CHT.

Furthermore, Figures 6 and 7 illustrate some typical scenes in DOTA v1.0 and DIOR, respectively. From the ground lines in Figure 6, it appears that the comparative methods struggle with severe artifacts, producing incorrect shapes and distorted textures for the ground lines. Only our CHT achieves visually pleasing results, accurately recovering the realistic distribution of the

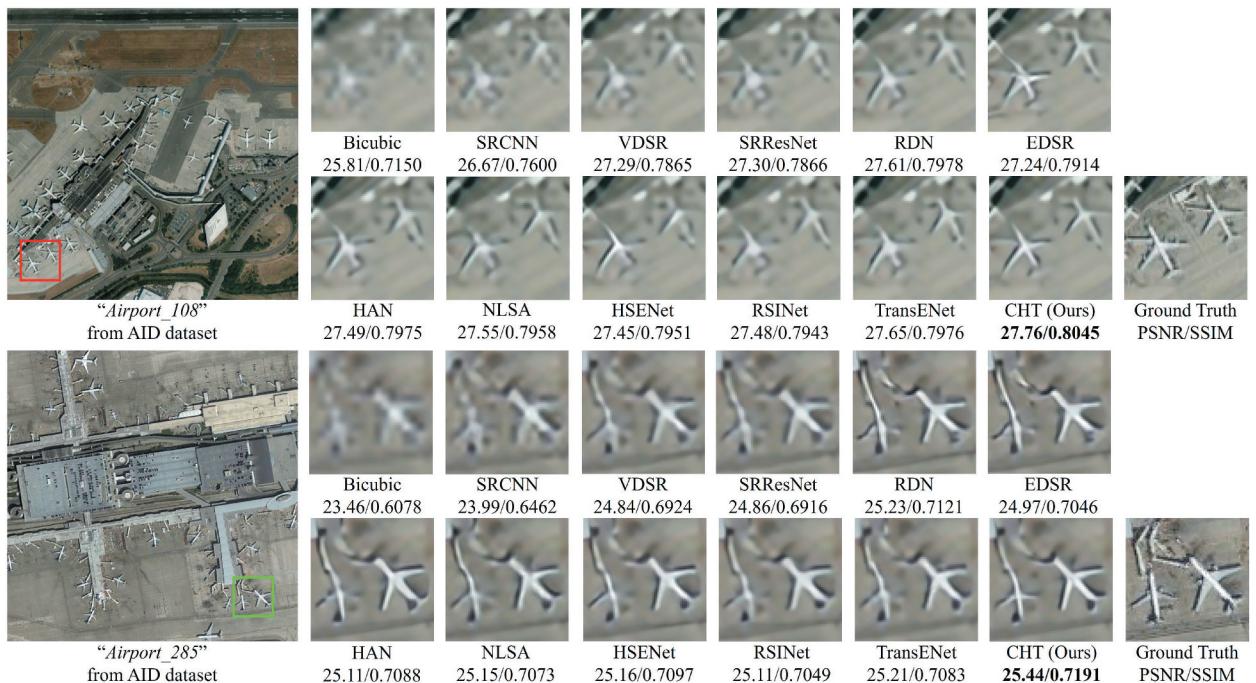
lines. This demonstrates the effectiveness of our proposed cross-scale channel attention in successfully extracting local features, such as edges and outlines. Similarly, in Figure 7, which contains massive cross-scale storage tanks, NLSA and HSENet, despite their ability to explore non-local representations, still struggle with handling such challenging scenes and producing deformed and blurry results. On the other hand, our CHT, benefiting from cross-scale self-attention, can fully explore and exploit valuable priors for better restoration. These visual comparisons provide further evidence of the superiority of our CHT in capturing fine details, preserving realistic distribution, and handling complex scenes with cross-scale features.

#### 4.5. Experiments on real-world data

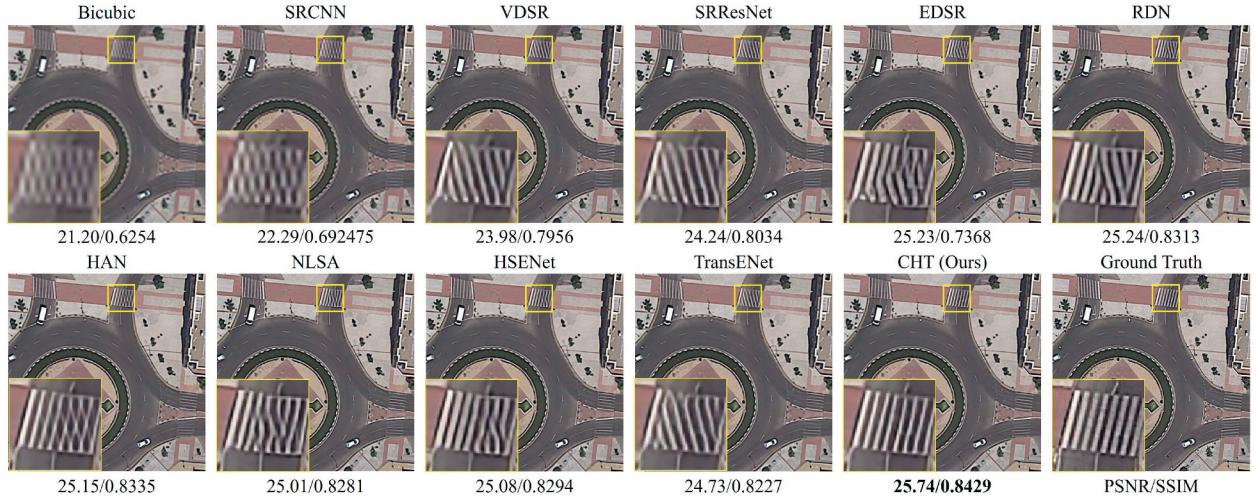
We further test the performance of these methods by applying them to the NWPU-RESISC45 dataset with realistic degradations.

##### 4.5.1. Qualitative comparison

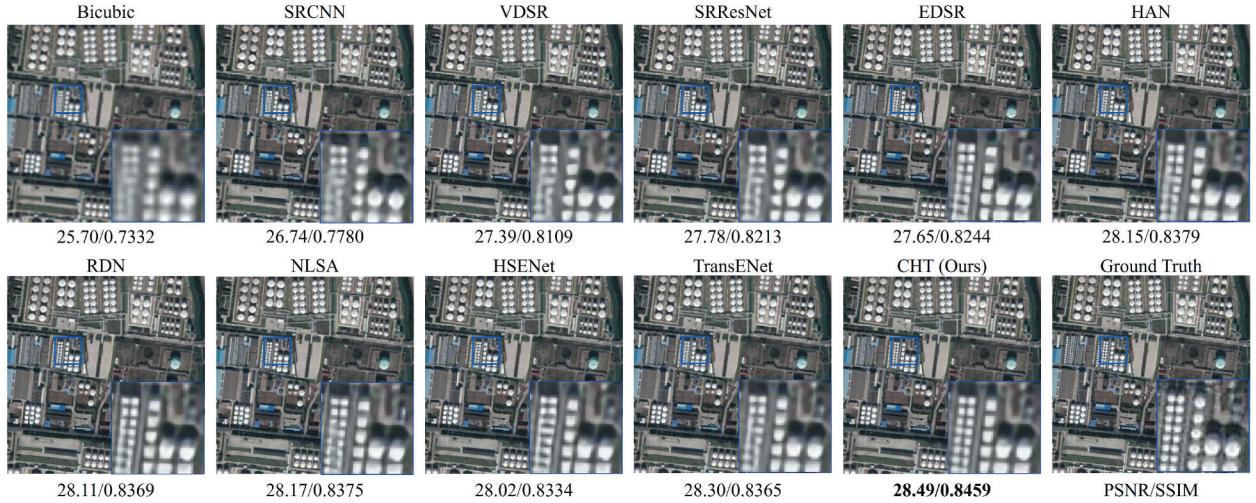
Table 4 presents the AG and NIQE results on NWPU-RESICS45 dataset. Based on the observations from Table 4, our CHT achieves the best performance in AG and secures second place in NIQE. These results demonstrate that our CHT remains competitive in addressing the SR problem in real-world scenarios. Notably, we find that VDSR achieves the best performance in NIQE.



**Figure 5.** Visual comparisons on the AID test set, where the selected scenes contain rich cross-scale similarities. The best PSNR/SSIM is shown in **bold**. Zoom in for better comparison.



**Figure 6.** Visual comparisons on the DOTA v1.0 test set, where the selected scenes contain rich cross-scale local details (e.g., cross-scale ground lines). The best PSNR/SSIM is shown in **bold**. Zoom in for better comparison.



**Figure 7.** Visual comparisons on the DIOR test set, where the selected scenes contain rich cross-scale global priors (e.g., cross-scale storage tanks). The best PSNR/SSIM is shown in **bold**. Zoom in for better comparison.

**Table 4.** Quantitative comparisons with typical SR models on the NWPU-RESICS45 test set. The best and second best AG and NIQE results are highlighted in red and blue.

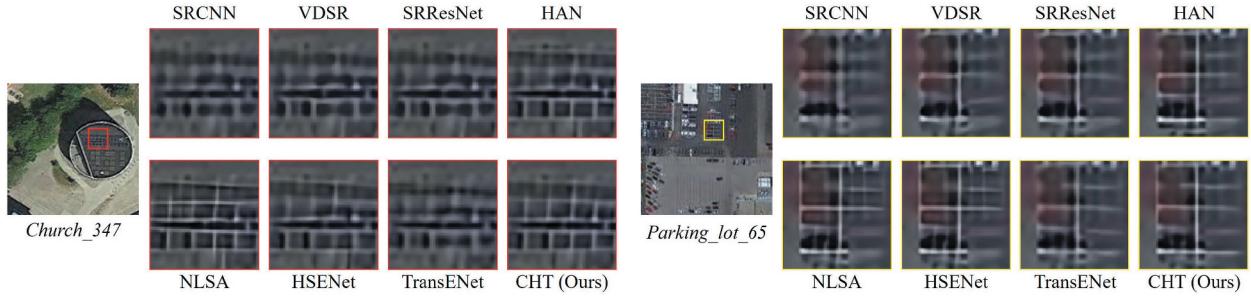
Metrics	SRCCNN	VDSR	SRResNet	HAN	NLSA	HSENet	TransENet	CHT (Ours)
AG ↑	2.8010	2.9264	2.9230	<b>3.0451</b>	3.0343	3.0275	2.9862	<b>3.0681</b>
NIQE ↓	21.032	<b>20.029</b>	20.496	20.428	20.508	20.293	20.572	<b>20.081</b>

This could be attributed to the complex degradations present in realistic scenes, which may interfere with the attention and transformer-based approaches employed by other methods. In this case, our CHT proves its favorable visual perception, showcasing the robust local and global modeling capabilities of the proposed CCA and CSA.

#### 4.5.2. Quantitative comparison

The visual comparisons are shown in Figure 8, where the selected images contain rich global and

local hierarchies. From “church\_347” from the NWPU-RESISC45 dataset, we find that TransENet cannot recover the global similarity, even with the assistance of self-attention. In contrast, our CHT provides more details with rich textures, resulting in a visually appealing restoration. As illustrated in “parking\_lot\_65”, where the marking lines on the ground are present, CNN-based and transformer-based methods fail to restore all of the lines accurately. However, our CHT preserves more sharp information on these lines. These results indicate



**Figure 8.** Visual comparisons on the NWPU-RESICS45 test set with real-world degradations, where the selected scenes contain rich cross-scale global and local similarities. Zoom in for better comparison.

that our CHT maintains its superiority in handling real-world scenes with complex degradations, providing stable and generalized performance for remote sensing image SR.

#### 4.6. Ablation study

To save inference costs, we conduct all these ablation analyses on a small-scale validation set and dub it AID-tiny. It contains 30 images, with each image randomly selected from a category of AID, and does not repeat in the training and test set.

##### 4.6.1. Hierarchical modeling

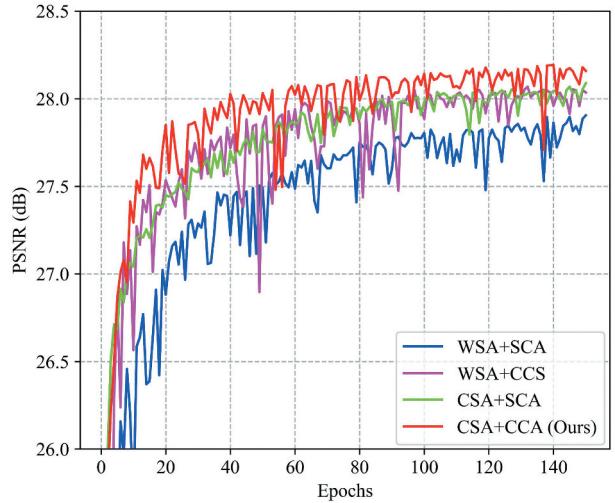
**4.6.1.1. Effect of cross-scale self-attention.** In Table 5, we report the quantitative comparison between the proposed CSA with the initial Window-based Self-Attention (WSA) for four times SR on the AID-tiny dataset. By comparing Model-A with our CHT, we can see that utilizing CSA to capture global features leads to a significant improvement in performance compared to using the single-scale WSA (28.2269 dB vs. 28.1033 dB). This improvement can be attributed to the cross-scale design of CSA, which enables the model to explore more prior knowledge, such as cross-scale self-similarity.

Besides, we visualize the training process of Model-A (pink) and our CHT (red) in Figure 9. It can be found that the PSNR curve of our CHT grows significantly against Model-A, demonstrating the effectiveness of our CSA.

**4.6.1.2. Effect of cross-scale channel attention.** To evaluate the effect of CCA, we compare it with the

**Table 5.** Effect of CSA and CCA. The PSNR performances are calculated on the AID-tiny dataset and the best result is shown in **bold**.

Method	Global Modeling		Local Modeling		PSNR (dB)
	WSA	CSA	SCA	CCA	
Baseline	✓		✓		28.0994
Model-A	✓			✓	28.1033
Model-B		✓	✓		28.1267
CHT (Ours)	✓	✓	✓	✓	<b>28.2269</b>

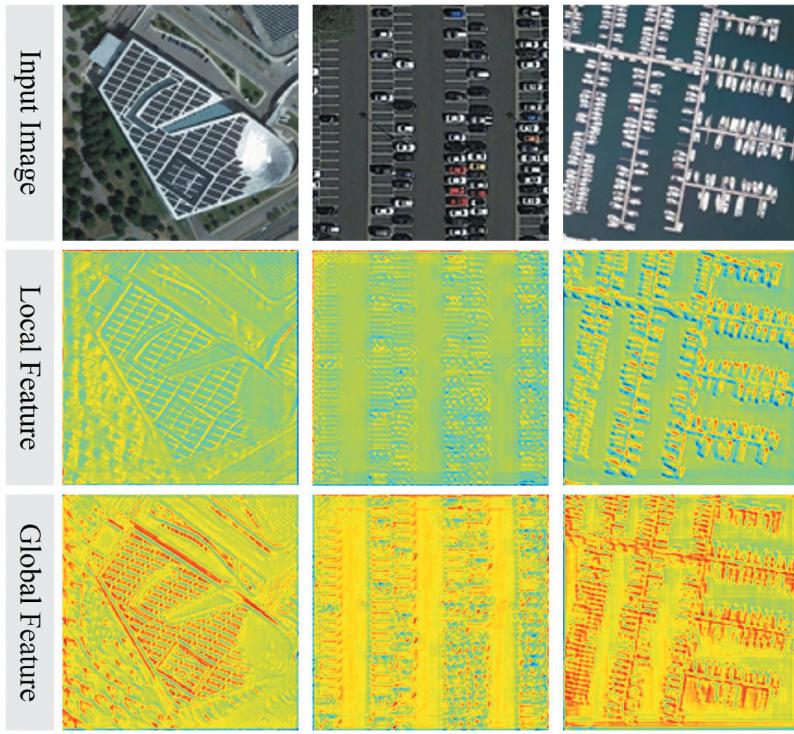


**Figure 9.** The training process of the effect of CSA and CCA. The models are validated on the AID-tiny dataset.

naive Single-scale Channel Attention (SCA) block. The quantitative results are listed in Table 5. The results show that incorporating CCA in our CHT leads to a noticeable improvement in performance compared to SCA (0.1002 dB higher in PSNR). This improvement can be attributed to the fact that while channel attention is capable of modeling local features, it neglects the cross-scale properties present in remote sensing imagery. Consequently, relying solely on SCA results in suboptimal performance.

Furthermore, our CHT outperforms the baseline model that also adopts hierarchical modeling. This result suggests that our CSA and CCA can mutually promote each other, leading to improved modeling of image hierarchies.

**4.6.1.3. Feature visualization.** More intuitively, we visualize the intermediate feature maps from local and global branches in our CHTB. For better visualization, three typical scenarios with rich global and local cross-scale knowledge are sent to our CHT. The results are shown in Figure 10. From this figure, we can find that local features are prominent in localities, such as edge, boundary, and shape information. On the other hand, the features yielded from global cross-scale attention tend to activate more cross-scale



**Figure 10.** Feature visualization of the hierarchical modeling.

regions in the global context. This result aligns with our motivation, which demonstrates that the proposed CHT can systematically explore both local and global cross-scale knowledge.

#### 4.6.2. Cross-scale modeling

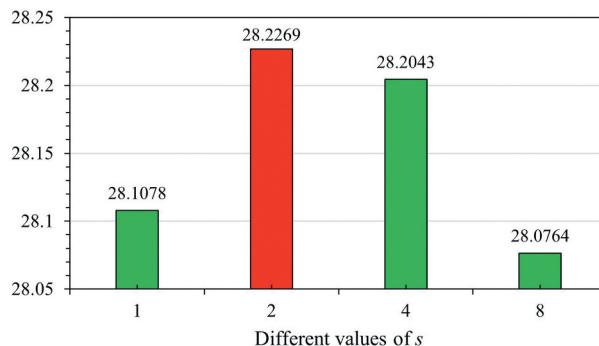
**4.6.2.1. Effect of different scale  $s$  in CSA.** The key hyper-parameter of our CSA is the scale  $s$  of the addition token  $S$ . To investigate the impact of different scale values, we empirically set  $s$  to 1, 2, 4, and 8 and evaluate the performance on the PSNR metric, as depicted in Figure 11. The results indicate that setting  $s = 2$  achieves the best PSNR performance. When  $s = 1$ , the scale of  $S$  aligns with that of  $Q$  and  $K$ , resulting in self-attention calculations within two single-scale feature spaces. This limitation fails to explore the cross-scale global dependency effectively. Once the scale is large, e.g.,  $s = 8$ , we observed that the

performance dropped dramatically. This is mainly attributed to the loss of global information when compressing the feature space too much. Based on these empirical findings, we set  $s = 2$  in our final CHT model as it offers the best performance.

**4.6.2.2. Effect of difference scale in CCA.** To evaluate the effectiveness of the multi-scale design, we conducted an experiment by replacing the multi-scale convolution in CCA with three single-scale convolutions:  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$ . The results in Table 6 demonstrate that the single-scale convolutions can only explore limited representations. In comparison, our CHT outperforms Model-C, Model-D, and Model-E by 0.1002 dB, 0.0614 dB, and 0.0921 dB in PSNR, respectively. This indicates that the cross-scale design of our CHT effectively extracts enriched local details, leading to superior reconstruction.

#### 4.6.3. Model efficiency

We first investigate the different number of transformer groups  $N_{TG}$  and cross-scale hierarchical transformer blocks  $N_{CHTB}$  used in our CHT. The computational cost and PSNR results are reported in



**Figure 11.** The training process of the effect of different values of  $s$  in CSA. The models are validated on the AID-tiny dataset.

**Table 6.** Effect of cross-scale in CCA. The results are tested on AID-tiny and the best PSNR is shown in **bold**.

Scale	Model	$3 \times 3$	$5 \times 5$	$7 \times 7$	PSNR (dB)
Single	Model-C	✓			28.1267
	Model-D		✓		28.1655
	Model-E			✓	28.1348
Cross	CHT (Ours)	✓	✓	✓	<b>28.2269</b>

**Table 7.** The ablation of different numbers of transformer groups. The best PSNR performance is shown in **bold**.

$N_{TG}$	2	4	6	8
FLOPs (G)	144.7	230.5	322.3	412.2
PSNR (dB)	30.63	30.79	<b>30.91</b>	30.90

**Table 8.** The ablation of different numbers of CHTB. The best PSNR performance is shown in **bold**.

$N_{CHTB}$	2	4	6	8
FLOPs (G)	180.6	322.3	463.1	603.8
PSNR (dB)	30.78	30.91	30.87	<b>30.94</b>

**Table 9.** Model efficiency analysis. FLOPs are calculated with an input image of size  $128 \times 128 \times 3$ .  $1\text{M} = 10^6$  and  $1\text{G} = 10^9$ . Note that the PSNR is tested on the AID test set and the best performance is shown in **bold**.

Method	Parameters	FLOPs	PSNR (dB)
Bicubic	–	–	28.86
SRCCNN	0.021 M	0.328 G	29.70
VDSR	0.667 M	10.93 G	30.44
SRResNet	0.734 M	23.13 G	30.51
EDSR	43.09 M	823.3 G	30.65
RDN	22.27 M	372.4 G	30.74
HAN	16.07 M	268.0 G	30.80
NLSA	44.15 M	840.8 G	30.81
RSINet	14.61 M	388.0 G	30.78
HSENet	21.70 M	306.3 G	30.72
TransENet	37.46 M	87.85 G	30.80
CHT (Ours)	20.53 M	322.3 G	<b>30.91</b>

Table 7 and 8. We observed that  $N_{TG} = 6$  yields the best PSNR performance. As  $N_{CHTB}$  increases, the capability of local and global cross-scale modeling also improves. However, the computational consumption also increases dramatically. Although the best PSNR result is achieved with  $N_{CHTB} = 8$ , we choose  $N_{CHTB} = 4$  as the final setting because it provides competitive performance while consuming less computation.

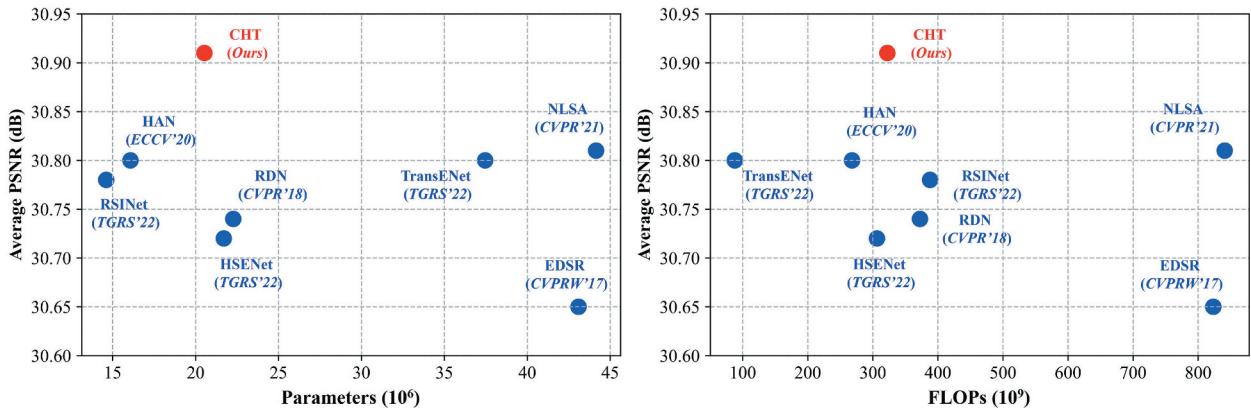
Table 9 provides the number of parameters, the FLOating Point operations (FLOPs), and the PSNR performance of each model on the AID test set. Intuitively, the relationships between PSNR performance and computational cost are shown in Figure 12. The observations reveal that earlier

methods, which did not consider local and global feature modeling and only employed naive CNNs, have fewer parameters and limited performance. However, with the introduction of attention mechanisms, such as holistic attention and non-local attention, both local and global information are fully exploited, resulting in significant performance improvement. Nevertheless, the number of parameters of these models also increases by a large margin. In comparison to the NLSA approach, which utilizes non-local attention for global modeling, our CSA is superior (30.91 dB vs. 30.81 dB) without excessively increasing the computational complexity (20.53 M vs. 44.15 M). This analysis highlights that our CHT strikes a good balance between local-global modeling and efficiency.

## 5. Conclusions

In this paper, we propose an effective transformer-based SR model to mitigate the shortcomings of existing methods when applied to remote sensing imagery. Our CHT aims to systematically model both local and global representations by designing a hierarchical transformer group that incorporates CSA and CCA. To facilitate the representation of cross-scale global knowledge in remote sensing scenes, we introduce an additional token space for implicit self-attention calculation. This auxiliary token allows for the indirect computation of self-attention in two sub-spaces with different feature scales, thereby enabling self-attention to be aware of cross-scale dependencies. Moreover, we devise a cross-scale channel attention mechanism that explores multi-scale features and enriches them with channel attention in a coarse-to-fine manner. The extensive experiments on four remote sensing datasets demonstrate the superiority of our CHT in grasping the local and global features, both in simulated and real-world data.

Despite achieving decent performance on simulation experiments, remote sensing imagery usually suffers from various degradations (Hong et al. 2018),



**Figure 12.** Comparison of computational complexity.

which may lead to several performance drops in our CHT. As investigated in previous research (Liu et al. 2022; Xiao et al. 2023d), when handling multiple degradations, the model trained on simulated data often lacks universal and generalizability. This is because there is a domain gap between single degradation (*e.g.*, bicubic downsampling) and real-world degradations, such as blur and noise. In future work, more efforts should be paid to develop a practical scheme for real-world remote sensing image SR, thus mitigating the performance drop when handling multiple degradations, such as blur and noise. Furthermore, lightweight SR may be another hotspot, especially in the context of real-time inference on large-scale remote sensing images.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported in part by the National Natural Science Foundation of China [grant numbers 42230108, and 61971319].

## Notes on contributors

**Yi Xiao** is currently pursuing the PhD degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan. He is responsible for writing the manuscript and implementing the algorithm. His major research interests include remote sensing image/video processing and computer vision. More details can be found at <https://xy-boy.github.io>.

**Qiangqiang Yuan** is a professor at the School of Geodesy and Geomatics, Wuhan University, Wuhan. His research interests include image reconstruction, remote sensing image processing and application, and data fusion. Dr. Yuan was a recipient of the Top-Ten Academic Star of Wuhan University in 2011, the Youth Talent Support Program of China in 2019, and the recognition of Best Reviewers of the IEEE GRSL in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is an associate editor of five international journals and has frequently served as a referee for more than 40 international journals for remote sensing and image processing.

**Jiang He** is currently pursuing the PhD degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan. His major research interests include remote sensing image fusion and pan-sharpening.

**Liangpei Zhang** received the BS degree in physics from Hunan Normal University, Changsha, China, in 1982, the MS degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the PhD degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in

1998. From 2011 to 2016, he was a Principal Scientist for the China State Key Basic Research Project appointed by the Ministry of National Science and Technology of China, Beijing, China, to lead the Remote Sensing Program in China. He is currently a “Chang-Jiang Scholar” Chair Professor appointed by the Ministry of Education of China, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He has published more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He holds 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence. Dr. Zhang is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes in the IEEE Geoscience and Remote Sensing Society (GRSS) 2014 Data Fusion Contest. His students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is the Founding Chair of the IEEE GRSS Wuhan Chapter. He serves as an associate editor or editor for more than ten international journals. He is also serving as an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing.

## ORCID

Yi Xiao  <http://orcid.org/0000-0001-9533-8917>  
 Qiangqiang Yuan  <http://orcid.org/0000-0001-7140-2224>  
 Jiang He  <http://orcid.org/0000-0002-0296-8395>  
 Liangpei Zhang  <http://orcid.org/0000-0001-6890-3650>

## Data availability statement

The AID dataset can be found at <https://captain-whu.github.io/AID/>. The DOTA v1.0 dataset is available at <https://captain-whu.github.io/DOTA/dataset.html>. The DIOR and NWPU-RESISC45 datasets can be downloaded from <https://gcheng-nwpu.github.io/>.

## References

- Bai, T., L. Wang, D. Yin, K. Sun, Y. Chen, L. Wenzhuo, and L. Deren. 2022. “Deep Learning for Change Detection in Remote Sensing: A Review.” *Geo-Spatial Information Science* 26 (3): 262–288. [10.1080/10095020.2022.2085633](https://doi.org/10.1080/10095020.2022.2085633).
- Chen, H., Z. Qi, and Z. Shi. 2021. “Remote Sensing Image Change Detection with Transformers.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–14. <https://doi.org/10.1109/TGRS.2020.3034752>.
- Chen, H., J. Song, W. Chen, B. Du, and N. Yokoya. 2023a. “Exchange Means Change: An Unsupervised Single-Temporal Change Detection Framework Based on Intra- and Inter-Image Patch Exchange.” *ISPRS Journal of Photogrammetry & Remote Sensing* 206:87–105. <https://doi.org/10.1016/j.isprsjprs.2023.11.004>.
- Chen, H., N. Yokoya, and M. Chini. 2023. “Fourier Domain Structural Relationship Analysis for Unsupervised

- Multimodal Change Detection.” *ISPRS Journal of Photogrammetry & Remote Sensing* 198:99–114. <https://doi.org/10.1016/j.isprsjprs.2023.03.004>.
- Chen, H., N. Yokoya, C. Wu, and B. Du. 2022. “Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–18. <https://doi.org/10.1109/TGRS.2022.3229027>.
- Chen, X., X. Wang, J. Zhou, Y. Qiao, and C. Dong. 2023b. “Activating More Pixels in Image Super-Resolution Transformer.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 22367–22377. Vancouver, Canada.
- Cheng, G., J. Han, and X. Lu. 2017. “Remote Sensing Image Scene Classification: Benchmark and State of the Art.” *Proceedings of the IEEE* 105 (10): 1865–1883. [10.1109/JPROC.2017.2675998](https://doi.org/10.1109/JPROC.2017.2675998).
- Deren, L., F. Yang, M. Wang, and R. Dai. 2023. “Internet intelligent remote sensing scientific experimental satellite LuoJia3-01.” *Geo-Spatial Information Science* 26 (3): 257–261. [10.1080/10095020.2023.2208472](https://doi.org/10.1080/10095020.2023.2208472).
- Dong, C., C. Change Loy, K. He, and X. Tang. 2015. “Image Super-Resolution Using Deep Convolutional Networks.” *IEEE Transactions on Pattern Analysis & Machine Intelligence* 38 (2): 295–307. <https://doi.org/10.1109/TPAMI.2015.2439281>.
- Dong, X., X. Sun, X. Jia, X. Zhihong, L. Gao, and B. Zhang. 2020a. “Remote Sensing Image Super-Resolution Using Novel Dense-Sampling Networks.” *IEEE Transactions on Geoscience & Remote Sensing* 59 (2): 1618–1633. <https://doi.org/10.1109/TGRS.2020.2994253>.
- Dong, X., L. Wang, X. Sun, X. Jia, L. Gao, and B. Zhang. 2020b. “Remote Sensing Image Super-Resolution Using Second-Order Multi-Scale Networks.” *IEEE Transactions on Geoscience & Remote Sensing* 59 (4): 3473–3485. <https://doi.org/10.1109/TGRS.2020.3019660>.
- Feng, J., Q. Jiang, C.-H. Tseng, X. Jin, L. Liu, W. Zhou, and S. Yao. 2022. “A Deep Multitask Convolutional Neural Network for Remote Sensing Image Super-Resolution and Colorization.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–15. <https://doi.org/10.1109/TGRS.2022.3154435>.
- Fernandez-Beltran, R., P. Latorre-Carmona, and F. Pla. 2017. “Single-frame super-resolution in remote sensing: A practical overview.” *International Journal of Remote Sensing* 38 (1): 314–354. <https://doi.org/10.1080/01431161.2016.1264027>.
- Freeman, W. T., T. R. Jones, and E. C. Pasztor. 2002. “Example-based super-resolution.” *IEEE Computer Graphics and Applications* 22 (2): 56–65. <https://doi.org/10.1109/38.988747>.
- Han, C., C. Wu, H. Guo, M. Hu, and H. Chen. 2023. “HANet: A Hierarchical Attention Network for Change Detection with Bitemporal Very-High-Resolution Remote Sensing Images.” *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 16:3867–3878. <https://doi.org/10.1109/JSTARS.2023.3264802>.
- He, J., J. Li, Q. Yuan, H. Shen, and L. Zhang. 2021. “Spectral Response Function-Guided Deep Optimization-Driven Network for Spectral Super-Resolution.” *IEEE Transactions on Neural Networks and Learning Systems* 33 (9): 4213–4227. <https://doi.org/10.1109/TNNLS.2021.3056181>.
- He, J., Q. Yuan, J. Li, Y. Xiao, D. Liu, H. Shen, and L. Zhang. 2023a. “Spectral Super-Resolution Meets Deep Learning: Achievements and Challenges.” *Information Fusion* 97:101812. <https://doi.org/10.1016/j.inffus.2023.101812>.
- He, J., Q. Yuan, J. Li, Y. Xiao, and L. Zhang. 2023b. “A Self-Supervised Remote Sensing Image Fusion Framework with Dual-Stage Self-Learning and Spectral Super-Resolution Injection.” *ISPRS Journal of Photogrammetry & Remote Sensing* 204:131–144. <https://doi.org/10.1016/j.isprsjprs.2023.09.003>.
- Hong, D., N. Yokoya, J. Chanussot, and X. Xiang Zhu. 2018. “An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing.” *IEEE Transactions on Image Processing* 28 (4): 1923–1938. <https://doi.org/10.1109/TIP.2018.2878958>.
- Jiang, K., Z. Wang, Y. Peng, C. Chen, Z. Wang, X. Wang, J. Jiang, and C.-W. Lin. 2021. “Rain-Free and Residue Hand-in-Hand: A Progressive Coupled Network for Real-Time Image Deraining.” *IEEE Transactions on Image Processing* 30:7404–7418. <https://doi.org/10.1109/TIP.2021.3102504>.
- Jiang, K., Z. Wang, Y. Peng, and J. Jiang. 2020. “Hierarchical Dense Recursive Network for Image Super-Resolution.” *Pattern Recognition* 107:107475. <https://doi.org/10.1016/j.patcog.2020.107475>.
- Jiang, K., Z. Wang, Y. Peng, J. Jiang, J. Xiao, and Y. Yao. 2018. “Deep Distillation Recursive Network for Remote Sensing Imagery Super-Resolution.” *Remote Sensing* 10 (11): 1700. <https://doi.org/10.3390/rs10111700>.
- Jiang, K., Z. Wang, Y. Peng, G. Wang, L. Tao, and J. Jiang. 2019. “Edge-Enhanced GAN for Remote Sensing Image Superresolution.” *IEEE Transactions on Geoscience & Remote Sensing* 57 (8): 5799–5812. <https://doi.org/10.1109/TGRS.2019.2902431>.
- Kim, J., J. Kwon Lee, and K. Mu Lee. 2016. “Accurate Image Super-Resolution Using Very Deep Convolutional Networks.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654. Las Vegas, NV, USA.
- Lei, S., and Z. Shi. 2022. “Hybrid-Scale Self-Similarity Exploitation for Remote Sensing Image Super-Resolution.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–10. <https://doi.org/10.1109/TGRS.2021.3069889>.
- Lei, S., Z. Shi, and M. Wenjing. 2022. “Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–11. <https://doi.org/10.1109/TGRS.2021.3136190>.
- Lei, S., Z. Shi, and Z. Zou. 2020. “Coupled Adversarial Training for Remote Sensing Image Super-Resolution.” *IEEE Transactions on Geoscience & Remote Sensing* 58 (5): 3633–3643. <https://doi.org/10.1109/TGRS.2019.2959020>.
- Li, C., B. Zhang, D. Hong, J. Yao, and J. Chanussot. 2023. “LRR-Net: An Interpretable Deep Unfolding Network for Hyperspectral Anomaly Detection.” *IEEE Transactions on Geoscience & Remote Sensing* 61:1–12. <https://doi.org/10.1109/TGRS.2023.3279834>.
- Li, K., G. Wan, G. Cheng, L. Meng, and J. Han. 2020. “Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark.” *ISPRS Journal of Photogrammetry & Remote Sensing* 159:296–307. <https://doi.org/10.1016/j.isprsjprs.2019.11.023>.
- Li, Q., Y. Chen, and Y. Zeng. 2022. “Transformer with Transfer CNN for Remote-Sensing-Image Object Detection.” *Remote Sensing* 14 (4): 984. <https://doi.org/10.3390/rs14040984>.
- Liang, G., U. KinTak, H. Yin, J. Liu, and H. Luo. 2023. “Multi-scale hybrid attention graph convolution neural network for remote sensing images super-resolution.”

- Signal Processing* 207:108954. <https://doi.org/10.1016/j.sigpro.2023.108954>.
- Liang, J., J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. 2021. "Swinir: Image Restoration Using Swin Transformer." In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844. Venice, Italy.
- Lim, B., S. Son, H. Kim, S. Nah, and K. Mu Lee. 2017. "Enhanced Deep Residual Networks for Single Image Super-Resolution." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144 Honolulu, HI, USA.
- Liu, A., Y. Liu, G. Jinjin, Y. Qiao, and C. Dong. 2022. "Blind Image Super-Resolution: A Survey and Beyond." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 45 (5): 5461–5480. <https://doi.org/10.1109/TPAMI.2022.3203009>.
- Liu, D., J. Li, Q. Yuan, L. Zheng, H. Jiang, S. Zhao, and Y. Xiao. 2023. "An Efficient Unfolding Network with Disentangled Spatial-Spectral Representation for Hyperspectral Image Super-Resolution." *Information Fusion* 94:92–111. <https://doi.org/10.1016/j.inffus.2023.01.018>.
- Lu, Z., J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng. 2022. "Transformer for Single Image Super-Resolution." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June, 457–466. New Orleans, LA, USA.
- Luo, Y., L. Zhou, S. Wang, and Z. Wang. 2017. "Video satellite imagery super resolution via convolutional neural networks." *IEEE Geoscience & Remote Sensing Letters* 14 (12): 2398–2402. <https://doi.org/10.1109/LGRS.2017.2766204>.
- Ma, J., G. Wang, L. Zhang, and Q. Zhang. 2023. "Restoration and Enhancement on Low Exposure Raw Images by Joint Demosaicing and Denoising." *Neural Networks* 162:557–570. <https://doi.org/10.1016/j.neunet.2023.03.018>.
- Ma, J., S. Yan, L. Zhang, G. Wang, and Q. Zhang. 2022. "ELMformer: Efficient Raw Image Restoration with a Locally Multiplicative Transformer." In *Proceedings of the 30th ACM International Conference on Multimedia*, 5842–5852. Lisbon, Portugal.
- Mei, Y., Y. Fan, and Y. Zhou. 2021. "Image Super-Resolution with Non-Local Sparse Attention." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 3517–3526.
- Miao, Y.-C., X.-L. Zhao, F. Xiao, J.-L. Wang, and Y.-B. Zheng. 2022. "Hyperspectral Denoising Using Unsupervised Disentangled Spatiospectral Deep Priors." *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 60:1–16. <https://doi.org/10.1109/TGRS.2022.3228927>.
- Miao, Y., L. Zhang, L. Zhang, and D. Tao. 2023. "DDS2M: Self-Supervised Denoising Diffusion Spatio-Spectral Model for Hyperspectral Image Restoration." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12086–12096. Paris, France.
- Mittal, A., R. Soundararajan, and A. C. Bovik. 2012. "Making a "Completely blind" Image Quality Analyzer." *IEEE Signal Processing Letters* 20 (3): 209–212. <https://doi.org/10.1109/LSP.2012.2227726>.
- Niu, B., W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen. 2020. "Single Image Super-Resolution via a Holistic Attention Network." In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII* 16, 191–207. Springer.
- Pan, Z., J. Yu, H. Huang, H. Shaoxing, A. Zhang, M. Hongbing, and W. Sun. 2013. "Super-Resolution Based on Compressive Sensing and Structural Self-Similarity for Remote Sensing Images." *IEEE Transactions on Geoscience & Remote Sensing* 51 (9): 4864–4876. <https://doi.org/10.1109/TGRS.2012.2230270>.
- Pei, Y., R. Chen, L. Deren, X. Xiao, and X. Zheng. 2023. "FCN-Attention: A Deep Learning UWB NLOS/LOS Classification Algorithm Using Fully Convolution Neural Network with Self-Attention Mechanism." *Geo-Spatial Information Science* 1–20. <https://doi.org/10.1080/10095020.2023.2178334>.
- Peng, Y., J. He, Q. Yuan, S. Wang, X. Chu, and L. Zhang. 2023. "Automated Glacier Extraction Using a Transformer Based Deep Learning Approach from Multi-Sensor Remote Sensing Imagery." *ISPRS Journal of Photogrammetry & Remote Sensing* 202:303–313. <https://doi.org/10.1016/j.isprsjprs.2023.06.015>.
- Roy, S. K., A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot. 2023. "Multimodal fusion transformer for remote sensing image classification." *IEEE Transactions on Geoscience & Remote Sensing* 64:1–20. <https://doi.org/10.1109/TGRS.2023.3286826>.
- Shi, W., J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. 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. Las Vegas, NV, USA.
- Timofte, R., V. De Smet, and L. Van Gool. 2013. "Anchored Neighborhood Regression for Fast Example-Based Super-Resolution." In *Proceedings of the IEEE International Conference on Computer Vision*, 1920–1927. Sydney, Australia.
- Tu, J., G. Mei, M. Zhengjing, and F. 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 & Remote Sensing* 15:5662–5673. <https://doi.org/10.1109/JSTARS.2022.3190322>.
- Wang, D., X. Chen, N. Guo, Y. Hui, and L. Yinan. 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, L., L. Rui, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson. 2022a. "UNetFormer: A UNet-Like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery." *ISPRS Journal of Photogrammetry & Remote Sensing* 190:196–214. <https://doi.org/10.1016/j.isprsjprs.2022.06.008>.
- Wang, X., K. Yu, S. Wu, G. Jinjin, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. 2018. "Esrgan: Enhanced super-resolution generative adversarial networks." In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 1–10. Munich, Germany.
- Wang, Y., Q. Yuan, S. Zhou, and L. Zhang. 2022. "Global Spatiotemporal Completion of Daily High-Resolution TCCO from TROPOMI Over Land Using a Swath-Based Local Ensemble Learning Method." *ISPRS Journal of Photogrammetry & Remote Sensing* 194:167–180. <https://doi.org/10.1016/j.isprsjprs.2022.10.012>.
- Wang, Z., L. Li, Y. Xue, C. Jiang, J. Wang, K. Sun, and M. Hongbing. 2022. "FeNet: Feature Enhancement Network for Lightweight Remote-Sensing Image Super-Resolution." *IEEE Transactions on Geoscience &*

- Remote Sensing* 60:1–12. <https://doi.org/10.1109/TGRS.2022.3168787>.
- Wu, X., D. Hong, and J. Chanussot. 2022. “Convolutional Neural Networks for Multimodal Remote Sensing Data Classification.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–10. <https://doi.org/10.1109/TGRS.2022.3210948>.
- Wu, X., D. Hong, and J. Chanussot. 2023. “UIU-Net: U-Net in U-Net for Infrared Small Object Detection.” *IEEE Transactions on Image Processing* 32:364–376. <https://doi.org/10.1109/TIP.2022.3228497>.
- Xia, G.-S., X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. 2018. “DOTA: A large-scale dataset for object detection in aerial images.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3974–3983. Salt Lake City, UT, USA.
- Xia, G.-S., J. Hu, H. Fan, B. Shi, X. Bai, Y. Zhong, L. Zhang, and L. Xiaoqiang. 2017. “AID: A benchmark data set for performance evaluation of aerial scene classification.” *IEEE Transactions on Geoscience & Remote Sensing* 55 (7): 3965–3981. <https://doi.org/10.1109/TGRS.2017.2685945>.
- Xiao, Y., Q. Yuan, H. Jiang, Q. Zhang, J. Sun, S. Xin, W. Jialian, and L. Zhang. 2022b. “Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer.” *International Journal of Applied Earth Observation and Geoinformation* 108:102731. <https://doi.org/10.1016/j.jag.2022.102731>.
- Xiao, Y., Q. Yuan, K. Jiang, H. Jiang, X. Jin, and L. Zhang. 2023a. “EDiffsR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution.” *arXiv preprint arXiv:2310.19288*. <https://doi.org/10.48550/arXiv.2310.19288>.
- Xiao, Y., Q. Yuan, K. Jiang, H. Jiang, Y. Wang, and L. Zhang. 2023b. “From Degrade to Upgrade: Learning a Self-Supervised Degradation Guided Adaptive Network for Blind Remote Sensing Image Super-Resolution.” *Information Fusion* 96:297–311. <https://doi.org/10.1016/j.inffus.2023.03.021>.
- Xiao, Y., Q. Yuan, K. Jiang, X. Jin, H. Jiang, L. Zhang, and C.-W. Lin. 2023c. “Local-Global Temporal Difference Learning for Satellite Video Super-Resolution.” *IEEE Transactions on Circuits and Systems for Video Technology* 1–16. <https://doi.org/10.1109/TCSVT.2023.33123110.1109/TCSVT.2023.3312321>.
- Xiao, Y., Q. Yuan, Q. Zhang, and L. Zhang. 2023d. “Deep Blind Super-Resolution for Satellite Video.” *IEEE Transactions on Geoscience & Remote Sensing* 61:1–16. <https://doi.org/10.1109/TGRS.2023.3291822>.
- Xiao, Y., X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang. 2022a. “Satellite Video Super-Resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection.” *IEEE Transactions on Geoscience & Remote Sensing* 60:1–19. <https://doi.org/10.1109/TGRS.2022.3157870>.
- Xu, Y., J. Gong, X. Huang, H. Xiangyun, L. Jiayi, L. Qiang, and M. 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* 26 (3): 289–301. <https://doi.org/10.1080/10095020.2022.2070555>.
- Yang, F., H. Yang, F. Jianlong, L. Hongtao, and B. Guo. 2020. “Learning Texture Transformer Network for Image Super-Resolution.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5791–5800.
- Yao, J., B. Zhang, L. Chenyu, D. Hong, and J. Chanussot. 2023. “Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework.” *IEEE Transactions on Geoscience & Remote Sensing* 61:1–15. <https://doi.org/10.1109/TGRS.2023.3284671>.
- Yu, Y., J. Zhao, Q. Gong, C. Huang, G. Zheng, and M. Jinye. 2021. “Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5.” *Remote Sensing* 13 (18): 3555. <https://doi.org/10.3390/rs13183555>.
- Yu, Y., J. Zhao, C. Huang, and X. Zhao. 2023. “Treat Noise as Domain Shift: Noise Feature Disentanglement for Underwater Perception and Maritime Surveys in Side-Scan Sonar Images.” *IEEE Transactions on Geoscience & Remote Sensing* 61:1–15. <https://doi.org/10.1109/TGRS.2023.3322787>.
- Zhang, D., J. Shao, X. Li, and H. Tao Shen. 2020a. “Remote Sensing Image Super-Resolution via Mixed High-Order Attention Network.” *IEEE Transactions on Geoscience & Remote Sensing* 59 (6): 5183–5196. <https://doi.org/10.1109/TGRS.2020.3009918>.
- Zhang, K., Z. Li, F. Zhang, W. Wan, and J. Sun. 2022a. “Pan-Sharpening Based on Transformer with Redundancy Reduction.” *IEEE Geoscience & Remote Sensing Letters* 19:1–5. <https://doi.org/10.1109/LGRS.2022.3186985>.
- Zhang, Q., Q. Yuan, M. Song, Y. Haoyang, and L. Zhang. 2022b. “Cooperated Spectral Low-Rankness Prior and Deep Spatial Prior for HSI Unsupervised Denoising.” *IEEE Transactions on Image Processing* 31:6356–6368. <https://doi.org/10.1109/TIP.2022.3211471>.
- Zhang, Q., Y. Zheng, Q. Yuan, M. Song, Y. Haoyang, and Y. Xiao. 2023. “Hyperspectral Image Denoising: From Model-Driven, Data-Driven, to Model-Data-Driven.” *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. [https://doi.org/10.1109/TNNLS.2023.3278866](https://doi.org/10.1109/TNNLS.2023.32788610.1109/TNNLS.2023.3278866).
- Zhang, S., Q. Yuan, L. Jie, J. Sun, and X. Zhang. 2020. “Scene-adaptive remote sensing image super-resolution using a multiscale attention network.” *IEEE Transactions on Geoscience & Remote Sensing* 58 (7): 4764–4779. <https://doi.org/10.1109/TGRS.2020.2966805>.
- Zhang, Y., K. Li, L. Kai, L. Wang, B. Zhong, and F. Yun. 2018a. “Image Super-Resolution Using Very Deep Residual Channel Attention Networks.” In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301. Munich, Germany.
- Zhang, Y., Y. Tian, Y. Kong, B. Zhong, and F. Yun. 2018b. “Residual Dense Network for Image Super-Resolution.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2472–2481.
- Zhou, F., W. Yang, and Q. Liao. 2012. “Interpolation-Based Image Super-Resolution Using Multisurface Fitting.” *IEEE Transactions on Image Processing* 21 (7): 3312–3318. <https://doi.org/10.1109/TIP.2012.2189576>.
- Zhou, J., B. Li, D. Zhang, J. Yuan, W. Zhang, Z. Cai, and J. Shi. 2023a. “UGIF-Net: An Efficient Fully Guided Information Flow Network for Underwater Image Enhancement.” *IEEE Transactions on Geoscience & Remote Sensing* 61:1–17. <https://doi.org/10.1109/TGRS.2023.3293912>.
- Zhou, J., Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang. 2023b. “Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction.” *International Journal of Computer Vision* 1–19. <https://doi.org/10.1007/s11263-023-01853-3>.