TTST: A Top-*k* Token Selective Transformer for Remote Sensing Image Super-Resolution

Yi Xiao, *Graduate Student Member, IEEE*, Qiangqiang Yuan, *Member, IEEE*, Kui Jiang, *Member, IEEE*, Jiang He, *Graduate Student Member, IEEE*, Chia-Wen Lin, *Fellow, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— Transformer-based method has demonstrated promising performance in image super-resolution tasks, due to its long-range and global aggregation capability. However, the existing Transformer brings two critical challenges for applying it in large-area earth observation scenes: (1) redundant token representation due to most irrelevant tokens; (2) single-scale representation which ignores scale correlation modeling of similar ground observation targets. To this end, this paper proposes to adaptively eliminate the interference of irreverent tokens for a more compact self-attention calculation. Specifically, we devise a Residual Token Selective Group (RTSG) to grasp the most crucial token by dynamically selecting the top-*k* keys in terms of score ranking for each query. For better feature aggregation, a Multi-scale Feed-forward Layer (MFL) is developed to generate an enriched representation of multi-scale feature mixtures during feed-forward process. Moreover, we also proposed a Global Context Attention (GCA) to fully explore the most informative components, thus introducing more inductive bias to the RTSG for an accurate reconstruction. In particular, multiple cascaded RTSGs form our final Top-*k* Token Selective Transformer (TTST) to achieve progressive representation. Extensive experiments on simulated and real-world remote sensing datasets demonstrate our TTST could perform favorably against state-of-the-art CNN-based and Transformer-based methods, both qualitatively and quantitatively. In brief, TTST outperforms the state-of-the-art approach (HAT-L) in terms of PSNR by 0.14 dB on average, but only accounts for 47.26% and 46.97% of its computational cost and parameters. The code and pre-trained TTST will be available at https://github.com/XY-boy/TTST for validation.**

***Index Terms*— Remote sensing image, super-resolution, sparse transformer, selective attention.**

I. Introduction

With the growing demand for fine-scale remote sensing applications, remote sensing image with **H**igh **S**patial **R**esolution (HSR) is playing an indispensable role in various research scenarios, such as land-cover segmentation [1], classification [2], [3], object detection [4], and change detection [5]. However, limited by the hardware devices, images from aerial sensors can merely characterize partial spatial details [6], [7], [8], resulting in suboptimal scene representation and visual quality. Therefore, improving the spatial resolution of remote sensing imagery is crucial for both human perception and downstream tasks.

In contrast to upgrading hardware, **S**uper-**R**esolution (SR) technologies [9], [10], [11], [12], [13], [14], [15], [16], which recover high-resolution images from low-resolution observations, provide more flexible and economical solutions to tackle this issue. As a classical low-level vision task, super-resolution is highly ill-posed, particularly when the scaling factor is large, *e.g.*, ×8 and ×16. Early works often employ interpolation [17] or various image priors [18], [19], [20] to constrain the infinite solution space. However, they tend to restore unsatisfactory results with severe artifacts and suffer from harsh optimization processes. Subsequently, with the booming of deep learning [21], the deep-learning-based SR approaches have achieved significant performance [22], [23], [24]. Among them, **C**onvolutional **N**eural **N**etworks (CNNs) based models have almost dominated SR for years as they could tame the ill-posedness with local fitting capability. For example, NLSA [25] proposed a sparse non-local attention to alleviate the redundant inference of non-local modeling and yielded impressive performance. However, CNNs are still weak at exploring global dependencies, which are vital for SR.

Until recently, several transformer-based SR methods [26], [27], [28] have emerged and achieved promising performance. The key success of Transformer is the self-attention which is more powerful than CNN in representing the long-range dependencies. More recently, hybrid attention [27] that incorporates the strength of CNN and transformer has been proposed. Nevertheless, most transformer-based SR models mainly engage in natural images [27], [29], [30], [31]. They usually use dense self-attention to aggregate global features, involving all tokens for similarity computation. However, as illustrated in Fig. 1, there is inadequate consideration of the characteristics in large-area earth observation scenes, including the scale diversity and redundancy characteristics.

Manuscript received 26 July 2023; revised 24 November 2023; accepted 26 December 2023. Date of publication 9 January 2024; date of current version 12 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42230108 and Grant 61971319. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nam Ik Cho. *(Corresponding authors: Qiangqiang Yuan; Kui Jiang.)*

Yi Xiao, Qiangqiang Yuan, and Jiang He are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: xiao_yi@whu.edu.cn; yqiang86@gmail.com; hej96.work@gmail.com).

Kui Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: kuijiang_1994@163.com).

Chia-Wen Lin is with the Department of Electrical Engineering and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Liangpei Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

Digital Object Identifier 10.1109/TIP.2023.3349004

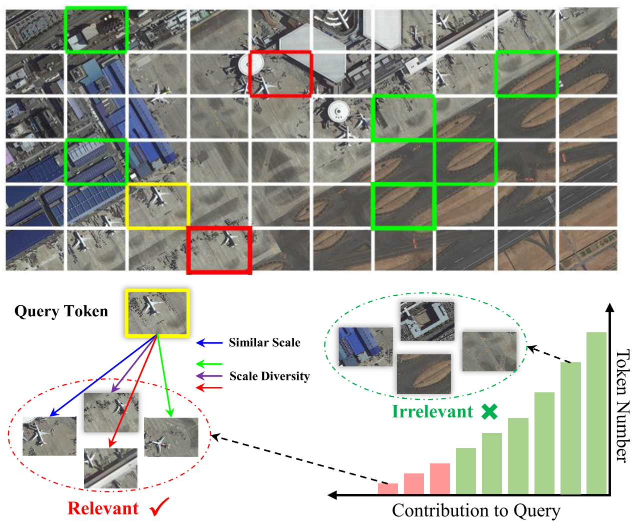


Fig. 1. The overlooked characteristics of remote sensing imagery for previous transformer-based SR models. (1) **Scale diversity**: there exists scale diversity of similar ground observation targets. (2) **Redundant token representation**: there are obvious redundant contents for correlation learning.

More precisely, there exists **scale diversity of similar ground observation targets**, where the latent scale correlations are barely explored by the single-scale representation methods, leading to undesired reconstruction results with artifacts. In addition, due to the large-range imaging, **there are obvious redundant contents for correlation learning, token representation of transformer** in particular. It is completely ignored by the previous transformer-based SR methods, which intuitively makes them more challenging for direct reuse on remote sensing SR tasks.

To this end, a novel **T**op-*k* **T**oken **S**elective **T**ransformer (TTST) is proposed in this study to mitigate the aforementioned issues. Specifically, instead of using all the tokens for *dense* attention matrix calculation, TTST chooses to search a *sparse* mask by selecting only the top-*k* highest attention values (*i.e.,* similarity scoring) of tokens for applying *channel-wise selection*. This enables TTST to capture the most relevant components over the entire HSR images while maintaining a moderate complexity with respect to explicit *spatial-wise selection*. In essence, the learnable mask is a sparse representation of query-key pairs, which aligns with the fact that the informative token is sparsely distributed across remote sensing images. Simultaneously, for better conservation of multi-scale information, a **M**ulti-scale **F**eed-forward **L**ayer (MFL) is developed to explore the latent scale relations of similar objects and enrich the interaction of multi-scale features. Furthermore, based on the observation that valuable prior knowledge exists in large-range areas, we devise a **G**lobal **C**ontext **A**ttention (GCA) module to dynamically adjust the large respective field of CNNs, thus introducing more inductive bias to TTST for better reconstruction.

In brief, the main contribution of this paper is three-fold:

1) A Top-*k* Token Selective Transformer (TTST) is proposed for remote sensing image super-resolution, considering the scale diversity and redundant token representation in challenging remote sensing scenarios.
2) To eliminate the interference of irrelevant tokens, TTST adaptively selects the most critical tokens based on the top-*k* selective mechanism, making the long-range modeling more effective and compact.
3) To explore the latent scale relations, a Multi-scale Feed-forward Layer (MFL) is devised, which helps to aggregate more multi-scale cues into the global representation.

The remainder of this paper is organized as follows. Section II reviews some important works related to our TTST. In Section III, we describe the implementation details of our TTST. Section IV contains extensive experiments on various remote sensing datasets. In Section V we summarize the whole paper.

II. Related Work

In this section, we first comprehensively review remote sensing image super-resolution. Then we introduce some work related to this paper, including Top-*k* selective mechanism and Large Kernel Convolution.

A. Remote Sensing Image Super-Resolution

1) CNN-Based: Drawing inspiration from SRCNN [32], a crowd of CNN-based SR methods have emerged and made remarkable progress on image SR task [33], [34], [35], [36], [37]. Generally, they use well-designed attention to remove the interference of irrelevant information, such as channel attention [38] and holistic attention [39]. To extract global prior knowledge (*e.g.,* self-similarity), some scholars have proposed self-similar attention [26] and non-local sparse attention [25]. Despite encouraging the representation of global context, exhaustive non-local modeling brings huge computationally complex and is less efficient in remote sensing imagery. In addition, their global modeling ability remains weak due to the limited receptive field of CNN.

2) Transformer-Based: Owing to the strong long-range representation capability of self-attention, the transformer has demonstrated comparable and even superior performance against CNN-based methods [40], [41], [42]. Intuitively, global attention is well suited to large-scale remote sensing images. However, it also inevitably involves irrelevant information for self-attention calculation. Recently, Lei et al. [43] proposed a multi-stage enhanced transformer that explores features at different scales with self-attention. However, it fails to conduct a global search at each stage and neglects the token redundancy issue. More recently, Chen et al. [27] proposed to exploit channel attention to introduce more global context into a transformer. Limited by the small respective field, *e.g.,* 3 × 3 convolution, the valuable prior knowledge is not fully explored. Fang et al. [44] incorporated CNN and Transformer for lightweight SR. Chen et al. [45] proposed a spatial-gate feed-forward network for better feature propagation. Zhu et al. [46] aggregated the long-range information in the spatial frequency domain. Nevertheless, they lack explicit consideration of scale variances in remote sensing imagery, thus resulting in suboptimal performance in image restoration tasks [47], [48].

In summary, we urgently need a practical scheme to distill irrelevant information during self-attention and pay more attention to the scale variation in remote sensing scenes.

B. Top-k Selective Mechanism

The top-k selective mechanism allows self-adaptive and efficient modeling of the attention mechanism. Zhao et al. [51] first introduce such selective strategy in NLP tasks. Later, some works [52], [53] proposed to use k-NN attention to select top-k most similar tokens. In particular, they calculate the pixel-wise similarity of query-key pairs and operate spatial-wise selection. In remote sensing areas, a similar patch-wise masking strategy has been investigated for boosting vision transformers [54].

This paper represents the first attempt to introduce a top-k selective mechanism for SR task of remote sensing imagery. Unlike previous works that employ a manual k, we proposed conducting a more flexible channel-wise similarity measurement and applying dynamic selection in the channel dimension. Compared to pixel-wise selection, our method is more computationally efficient as predicting a score for each pixel is naturally a laborious task.

C. Large Kernel Convolution

Recent studies [55], [56], [57] have demonstrated the effectiveness of large kernel convolutions in improving the performance of vision tasks. In light of this, some effort has been paid to increase the respective field of CNNs with larger kernel convolution. Liu et al. [58] modified the standard ResNet with 7 × 7 Depth-Wise Convolutions, resulting in favorable performance improvement in classification tasks. Recently, Ding et al. [59] even introduced 31 × 31 convolutions and achieved competitive performance with a vision transformer. To reduce the complexity of large kernel CNN, Guo et al. [50] proposed a large kernel convolutional decomposition strategy (LKA) without sacrificing the respective field. However, limited by a single large respective field, LKA naturally ignores the different importance of multi-scale features in remote sensing imagery. To pick up the most critical features, an effective selective attention [49] has been proposed. Nevertheless, multi-scale representations are rarely explored with a single-scale convolution.

As large kernel convolution particularly well-suits the large-range context characteristic in remote sensing imagery, we grasp this merits to devise our Global Context Attention. Different from previous works, our GCA can adaptively aggregate the most critical global contextual features with various large respective fields. Fig. 2 illustrates the detailed structure comparisons between our GCA and related works.

III. Methodology

A. Overview of TTST

As illustrated in Fig. 3, our TTST consists of three major components: 1) Feature extraction, which extracts global context feature from I_{LR} with Global Context Attentions (GCAs); 2) Residual Token Selective Groups (RTSGs), where each RSTG contains a Top-k Token Selective Attention (TTSA), a vanilla Window-based Self-Attention (WSA), a Multi-scale Feed-forward Layer (MFL), and an optional GCA module; 3) The Reconstruction part, which aims to restore the super-resolved image I_{SR}. The details of these components are described below.

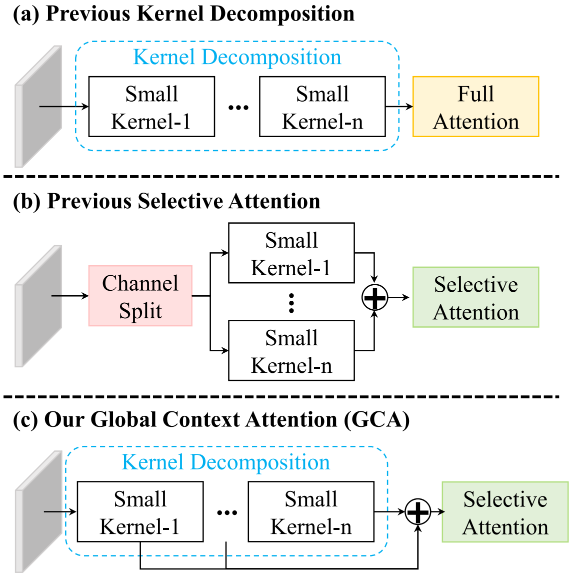


Fig. 2. Architectural comparisons between our Global Context Attention (GCA) and other related works [49], [50].

B. Top-k Token Selective Group

1) Top-k Token Selective Attention: The key motivation of our TTSA is to distill the interference of noisy tokens when calculating self-attention. In particular, TTSA aims to leverage the sparsity by selecting the token with the top-k highest relevance to the query, grasping the most critical information for restoration.

Formally, given a query Q, key K, and value V with the shape of $d \times H \times W$, the dense attention matrix $\mathcal{M} \in \mathbb{R}^{d\times d}$ can be generated by dot-product operation between Q and transposed K across channels. Rather than computing a spatial-wise matrix with the shape of $HW \times HW$, channel-wise similarity measurement helps to reduce memory consumption for efficient inference. Next, an adaptive selection strategy is adopted to mask out the irrelevant elements (*i.e.,* lower attention values) in M. As shown in Fig. 3, k is dynamically set to a sequence of values. Using $k_1 = \frac{1}{2}$ as an example, only the elements with top 50% scores can be reserved for activation, while the remaining 50% elements are masked to 0. Similarly, when $k_4 = \frac{4}{5}$, the sparse rate is 20%. Different from fixed k that lacks flexibility in exploring the latent magnitude of sparsity, the proposed dynamic selection allows the selective process *from sparse to dense* by setting k to multiple values. Specifically, we generate a binary mask matrix to achieve this section operator:

$$[m_k]_{ij} = \begin{cases} 1, & m_{ij} \in index_k \\ 0, & \text{otherwise.} \end{cases} , \quad (1)$$

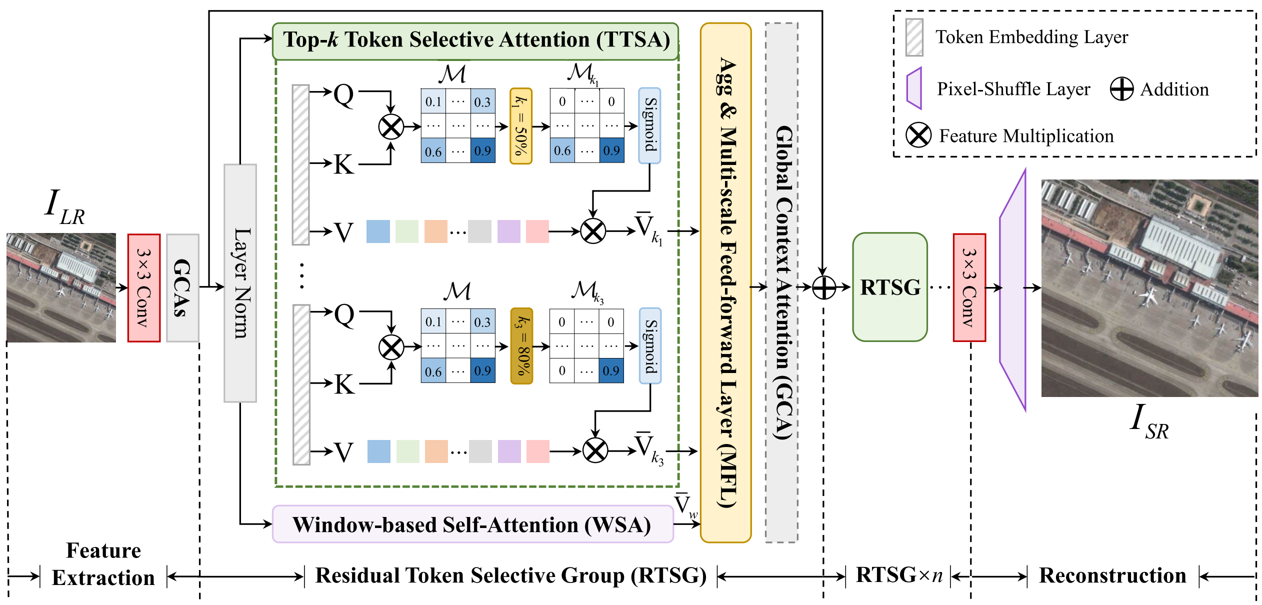


Fig. 3. Overview of the proposed TTST. (1) Feature extraction receives the LR input I_{LR} and introduces inductive bias to transformer with GCA. (2) Residual Token Selective Groups, where each RTSG consists of a TTSA, a WSA, an MFL, and an optional GCA module (marked by a dashed box). (3) Reconstruction part.

where $index_k$ means the coordinates of the top-k highest values. After that, the sparse attention matrix is activated by the softmax function. These processes and be written as follows:

$$\mathcal{M}_k = \text{Softmax}\left(m_k \odot \frac{\text{QK}^{\text{T}}}{\lambda}\right) \tag{2}$$

Here, we denote the activated attention matrix with a dense rate of k as $\mathcal{M}_k = [\text{W}_1, \cdots \text{W}_C]^{\text{T}}$, the i-th feature maps of sparse representation $\bar{\text{V}}_k$ can be formulated by the following:

$$\bar{v}_i = \sum_{i=1}^{d} \omega_i \odot v_i, \tag{3}$$

where ω_i means the i-th attention value of W_i and $\odot$ is channel-wise multiplication. The output of i-th head is the average results of $\bar{\text{V}}_{k_i}$:

$$\bar{\text{V}}_i = \sum_{i=1}^{K} \bar{\text{V}}_{k_i} \bigg/ K, \tag{4}$$

where $K = 4$ means four dynamic value of k. As we adopt the multi-head design, we concatenate all the output in each head and aggregate them with a 1×1 convolution:

$$\bar{\text{V}}_t = \text{Conv}\left(Concat\left(\bar{\text{V}}_i\right)\right). \tag{5}$$

The algorithm of our Top-k Token Selective Attention is summarized in Algorithm. 1.

2) Window-Based Self-Attention: WSA is capable of capturing long-range dependencies, which has become an empirical operation in most existing models [11], [27]. Here, we follow the standard WSA in [27] to generate the long-range representation $\bar{\text{V}}_w$. Finally, we aggregate the output of TTSA and WAS by element-wise addition, *i.e.,* $\text{X} = \bar{\text{V}}_w + \bar{\text{V}}_i$. This integration ensures that our model benefits from both the long-range dependencies captured by WSA and the enhanced locality of TTSA.

Algorithm 1 Top-k Token Selective Attention.

```
Input: LR feature $F_{LR} \in \mathbb{R}^{C\times H\times W}$
Initialization: $e_i \in \mathbb{R}^{d\times H\times W}$ is the feature in $i$th
  head, $d = \frac{C}{6}$, $k_i \in \left[\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}\right]$, $M_{k_i} \in \mathbb{R}^{d\times d}$ is a zero
  matrix, and $\lambda = \sqrt{d}$.
foreach $e_i$ do
    $\text{Q, K, V} \leftarrow Chunk\,(\text{DWConv}\,(e_i))$;     // d×H×W
    $\text{Q, K, V} \leftarrow Reshape\,(\text{Q, K, V})$;          // d×HW
    $M \leftarrow \frac{\text{QK}^\text{T}}{\lambda}$;          // Dense Attention d×d
    foreach $k_i$ do
        $index_{k_i} \leftarrow topk\,(\mathcal{M})$;
        $m_{k_i} \leftarrow mask\,(\mathcal{M}_{k_i}, index_{k_i}, 1)$;
        $\mathcal{M}_{k_i} \leftarrow m_{k_i} \odot \mathcal{M}$;
        $\mathcal{M}_{k_i} \leftarrow sigmoid\,(\mathcal{M}_{k_i})$;          // Sparse
          Attention d×d
        $\bar{\text{V}}_{k_i} \leftarrow \mathcal{M}_{k_i} \otimes \text{V}$.
    end
    $\bar{\text{V}}_i \leftarrow \sum_{i=1}^{K} \bar{\text{V}}_{k_i} \Big/ K$;                       // d×HW
    $\bar{\text{V}}_i \leftarrow Reshape\,(\bar{\text{V}}_t)$;                   // d×H×W
end
$\bar{\text{V}} \leftarrow \text{Conv}\,(Concat\,(\bar{\text{V}}_i))$                       // C×H×W
```

C. Multi-Scale Feed-Forward Layer

The naive MLP layer employs linear projection for feature propagation, which does not explicitly consider multi-scale features. Prior works [45], [47], [60] have modified the MLP layer for better feature propagation. Chen et al. [45] replaced the MLP layer with a spatial-gate network. Wang et al. [47]

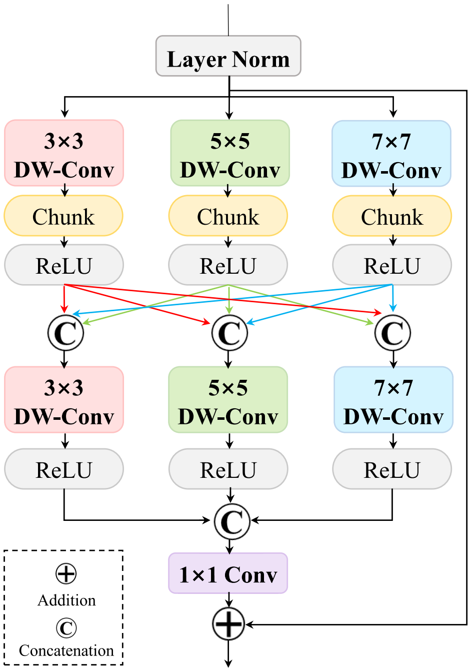


Fig. 4. The conceptual illustration of the proposed Multi-scale Feed-forward Layer (MFL). Here, DW-Conv represents the Depth-wise Convolution, and the chunk is the channel split operation.

added a depth-wise convolutional block to enhance the locality. Nevertheless, limited by the single-scale design, they all neglect to explore multi-scale properties in remote sensing imagery. In fact, boosting the representation of multi-scale objects has fully demonstrated its effectiveness in better remote sensing imagery super-resolution [61]. Therefore, we devise a efficient yet effective multi-scale feed-forward layer to generate an enriched set of features.

As illustrated in Fig. 4, after a layer normalization operator $\mathrm{X}_l = \mathrm{LN}(\mathrm{X})$, we feed the normalized feature X_l into three parallel branches to explore multi-scale representations with 3×3, 5×5, and 7×7 DW-Conv, respectively.

$$\mathrm{X}_3 = f^{dwc}_{3\times3}(\mathrm{X}_l), \mathrm{X}_5 = f^{dwc}_{5\times5}(\mathrm{X}_l), \mathrm{X}_7 = f^{dwc}_{7\times7}(\mathrm{X}_l). \quad (6)$$

To enhance the interaction of multi-scale localities, we split the multi-scale representations into three parts along the channel dimension using the chunk operation and concatenated these parts after ReLU activation. This exploration and incorporation process can be formulated as follows:

$$\begin{cases} \bar{\mathrm{X}}_3 = \sigma\left(f^{dwc}_{3\times3}\left[\mathrm{X}^{p_1}_3, \mathrm{X}^{p_1}_5, \mathrm{X}^{p_1}_7\right]\right), \\ \bar{\mathrm{X}}_5 = \sigma\left(f^{dwc}_{5\times5}\left[\mathrm{X}^{p_2}_3, \mathrm{X}^{p_2}_5, \mathrm{X}^{p_2}_7\right]\right), \\ \bar{\mathrm{X}}_7 = \sigma\left(f^{dwc}_{7\times7}\left[\mathrm{X}^{p_3}_3, \mathrm{X}^{p_3}_5, \mathrm{X}^{p_3}_7\right]\right), \\ \bar{\mathrm{X}} = f_{1\times1}\left[\bar{\mathrm{X}}_3, \bar{\mathrm{X}}_5, \bar{\mathrm{X}}_7\right] + \mathrm{X}_l, \end{cases} \quad (7)$$

where $\sigma(\cdot)$ is ReLU activation, $[\cdot]$ represents the channel-wise concatenation, and $f_{1\times1}$ denotes a 1×1 convolution.

D. Global Context Attention

As discussed in the *Introduction*, large-scale remote sensing scenes often exhibit significant redundancy (*e.g.,* self-similarity), which can be treated as valuable prior knowledge for restoration. Moreover, these global contexts can vary in scale. Therefore, we propose to generate multiple global context features yielded from different large respective fields and perform adaptive selection to explore the most useful context. To achieve this, we decompose a large-scale kernel into a sequence of Depth-Wise Convolutions (DW-Conv) with diverse kernels. This decomposition explicitly allows us to obtain a series of global features considering the scale variation.

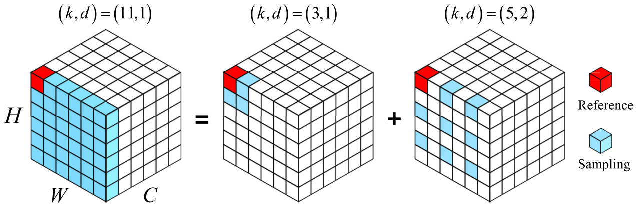


Fig. 5. The diagram of large kernel decomposition. A conventional 11×11 **D**epth-**W**ise **Conv**olution (**DW-Conv**) can be decomposed into two efficient operations: a 3×3 DW-Conv and a 5×5 DW-Conv with dilation rate 2. Here k means the kernel size and d is the dilation rate.

1) Kernel Decomposition: As illustrated in Fig. 5, a large-scale 11×11 DW-Conv can be subdivided into two operations: a 3×3 DW-Conv convolution and a 5×5 DWD-Conv with the dilation rate of 2. Generally, given an input $\bar{\mathrm{X}}$, a series of contextual features with varying respective fields are obtained:

$$U_{i+1} = f^{dwc}_i(U_i), \quad (8)$$

where $U_0 = \bar{\mathrm{X}}$ and $f^{dwc}_i(\cdot)$ is i-th DW-Conv in the decomposed sequence. As illustrated in Fig. 7, we set $i = 1$ for a simple explanation. The contextual features will serve as candidates for further dynamic selection.

There are two **merits** of our kernel-decomposition strategy. (1) It allows us to extract global prior knowledge using convolution with the large respective fields while maintaining a lightweight architecture, compared to simply applying a single larger-kernel convolution. (2) The kernel sequence explicitly produce multiple global representation yields from various respective field, which makes us easier to explore multi-scale prior knowledge and perform succeeding selective attention. As depicted in Fig. 6, with the help of kernel-decomposition strategy, our TTST significantly enlarges the respective field and activates a larger number of pixels for super-resolution restoration compared to standard small kernel convolutions.

2) Context Selective Attention: To grasp the different contributions of global contexts from candidates with different large spatial fields, we introduced a channel-wise selective attention mechanism [49]. Firstly, the candidates are aggregated to obtain a holistic global representation denoted as U. Subsequently, a spatial-wise global pooling operator is conducted to squeeze U to a flattened feature S. After a simple linear projection layer, a compact feature Z is obtained. Mathematically, these processes can be written as follows:

$$S = \mathcal{P}(U), Z = \sigma\left(\mathcal{F}_{fc}(S)\right), \quad (9)$$

where $\mathcal{P}(\cdot)$ means global pooing, $\sigma(\cdot)$ is RuLU activation and $\mathcal{F}_{fc}(\cdot)$ is a fully-connected layer. After that, to generate the channel-wise attention to guide the selection, a softmax is

TABLE I
The Detailed Summary of the Dataset Properties Involved in This Paper, Including AID, DOTA v1.0, DIOR, and NWPU-RESIC45

Properties	Train	Test			
Data Name	AID [63]	AID [63]	DOTA [64]	DIOR [65]	NWPU-RESISC45 [66]
Used (Total) Image Numbers	3000 (10000)	900 (10000)	900 (2806)	1000 (23463)	315 (31500)
Original Size	600 × 600	600 × 600	800 $\sim$ 4000	800 × 800	256 × 256
Spatial Resolution	0.5 $\sim$ 8 m	0.5 $\sim$ 8 m	-	-	0.2 $\sim$ 30 m
Categories	30	30	15	20	45
Task	Scene Classification	Scene Classification	Object Detection	Object Detection	Scene Classification

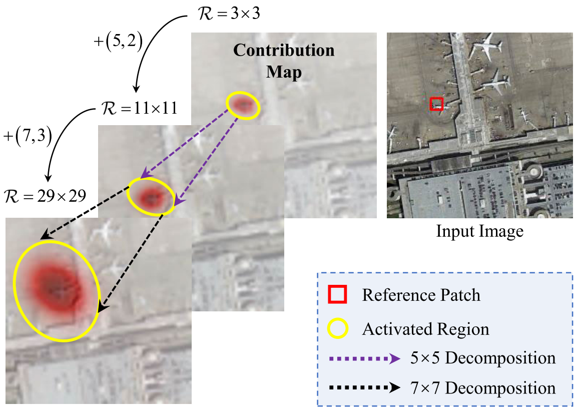


Fig. 6. Local Contribution Map (LAM) [62] of our TTST. With the kernel decomposition strategy, TTST gradually increases the respective field from 3 to 29 and activates more useful pixels in a large range in remote sensing imagery for better restoration.

applied to the channel-wise elements:

$$W_{1k}=\frac{e^{A_k Z}}{e^{A_k Z}+e^{B_k Z}},\ W_{2k}=\frac{e^{B_k Z}}{e^{B_k Z}+e^{A_k Z}},\tag{10}$$

where $A,\ B\in\mathbb{R}^{c\times d}$ are two learnable parameters. A_k means the k-th row of A and W_{1k} represents the k-th attention value in W_1, likewise B_k and W_{2k}. As shown the Fig. 7, when there are two candidates U_1 and U_2, the attention follows $W_{1k}+W_{2k}=1$. The final selected Y is obtained with the channel-wise multiplication:

$$\mathrm{Y}=U_1\cdot W_1+U_2\cdot W_2.\tag{11}$$

E. Resconstruction

To restore the final high-resolution image, a 3 × 3 convolution is inserted before the widely used pixel-shuffle layer [67] to enlarge the channel dimension. Finally, the super-resolved image $I_{SR}\in\mathbb{R}^{Hr\times Wr\times 3}$ can be received. Here, r is the scaling factor, and 3 means the RGB channels.

IV. Experiment

A. Remote Sensing Datasets

In this paper, we report the results of the super-resolution models on four remote sensing datasets, including AID [63], DOTA v1.0 [64], DIOR [65] and NWPU-RESISC45 [66]. The detailed properties of these datasets are summarized in Table I.

B. Implementation Details

1) Model Details: In this study, all the models were implemented for ×4 SR, *i.e.,* $r=4$. For the structure of TTST, the number of GCA in the feature extraction part is set to 5. To ensure deep feature exploration, we stacked 36 RTSG in our final model, consistent with [27]. The channel number in TTST is set to 180. In both TTSA and WSA, the number of multi-head self-attention is 6. The selective rate in TTSA is dynamically set to 1/2, 2/3, 3/4, and 4/5, allowing for a flexible trade-off between sparse and dense selection. Since HAT-L employs one convolution layer at every 6 transformer groups to introduce inductive bias, for fair comparison, we also insert a GCA at every 6 RTSGs in our final TTST.

2) Training Details: To ensure a fair comparison, all the SR methods considered in this study were retrained using the training set mentioned above, without any pre-training and fine-tuning processes. During the training process of our TTST, we randomly select 4 image patches with the size of 64 × 64 in each iteration. The learning rate is initialized to 1×10^{-4} and halved when reaching half of the 500 epochs. We adopt the widely used $\mathcal{L}_1$ loss function to optimize our TTST, *i.e.,* $\mathcal{L}_1=\|I_{SR}-I_{GT}\|_1$, where I_{SR} means the super-resolved image and I_{HR} denotes the ground-truth image. Adam optimizer is used in all models. All experiments involved in this paper were conducted on the same device, *i.e.,* a single NVIDIA RTX 3090 GPU and a 3.40 GHz AMD Ryzen 5700X CPU.

C. Evaluation Metrics

For simulated experiments, two widely used full-reference indicators are adopted: **P**eak **S**ignal-to-**N**oise **R**atio (PSNR) and **S**tructural **S**imilarity **I**ndex (SSIM) [69]. Note that PSNR and SSIM are calculated on the luminance channel (Y) of YCbCr space. Regarding real-world experiments, two reference-free metrics are used, *i.e.,* the **N**atural **I**mage **Q**uality **E**valuator (NIQE) [70] and the **A**verage **G**radient (AG).

D. Experiments on Simulated Datasets

1) Comparative Methods: To comprehensively evaluate the SR performance of our TTST against state-of-the-art methods on remote sensing imagery, we choose both CNN and Transformer-based models for comparison, including EDSR [68], RCAN [38], HAN [39], NLSA [25], HSENet [26], HAUNet [71], TransENet [43], and HAT-L [27]. Note that TTST+ is the self-ensemble result of our TTST.

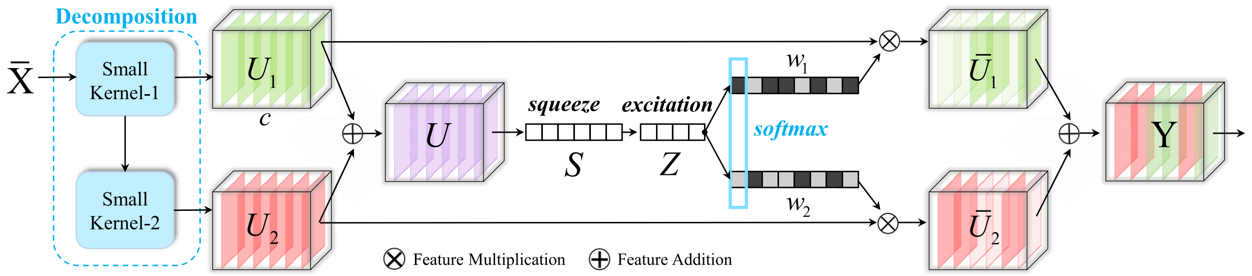


Fig. 7. The overall structure of our Global Context Attention (GCA) module.

TABLE II

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 BLOD** AND **BLUE BLOD**, RESPECTIVELY

Land Cover	Bicubic		EDSR [68]		RCAN [38]		HSENet [26]		NLSA [25]		TransENet [43]		HAT-L [27]		**TTST (Ours)**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airport	27.83	0.7554	29.93	0.8282	30.13	0.8318	30.08	0.8303	**30.16**	**0.8322**	30.15	0.8301	30.15	0.8319	**30.31**	**0.8345**
Bare Land	35.60	0.8564	36.94	0.8837	**36.99**	0.8844	36.79	0.8841	37.00	**0.8845**	36.93	0.8837	36.88	0.8841	**37.07**	**0.8848**
Baseball Field	31.00	0.8305	33.05	0.8765	**33.30**	**0.8789**	33.15	0.8774	33.24	0.8787	33.25	0.8775	33.25	0.8789	**33.41**	**0.8804**
Beach	32.90	0.8446	34.18	0.8727	34.33	0.8751	34.14	0.8746	34.31	0.8749	**34.35**	0.8754	34.34	**0.8756**	**34.47**	**0.8764**
Bridge	30.22	0.8283	32.93	0.8800	**33.13**	**0.8819**	33.06	0.8809	33.12	0.8818	33.08	0.8810	33.04	0.8809	**33.28**	**0.8830**
Center	26.51	0.6944	28.77	0.7921	**28.96**	0.7966	28.83	0.7937	28.95	**0.7971**	28.91	0.7934	28.92	0.7956	**29.16**	**0.8013**
Church	24.29	0.6333	26.30	0.7469	26.54	0.7529	26.47	0.7507	26.51	0.7528	26.52	0.7492	**26.56**	**0.7532**	**26.71**	**0.7574**
Commercial	27.33	0.7174	29.01	0.7940	**29.24**	0.8000	29.21	0.7989	29.21	0.7996	29.21	0.7973	29.21	**0.8007**	**29.40**	**0.8043**
D-Residential	22.93	0.5671	24.38	0.6839	24.63	0.6930	24.60	0.6912	24.60	0.6936	**24.71**	0.6931	24.67	**0.6936**	**24.84**	**0.7011**
Desert	39.26	0.9100	40.20	0.9268	40.24	0.9272	39.57	0.9271	40.27	0.9278	40.29	0.9276	**40.37**	**0.9278**	**40.43**	**0.9283**
Farmland	33.10	0.8226	35.00	0.8683	**35.11**	**0.8701**	35.02	0.8692	35.10	0.8699	34.99	0.8675	35.03	0.8691	**35.15**	**0.8702**
Forest	28.79	0.6605	29.85	0.7315	29.95	0.7345	30.00	0.7363	29.98	0.7369	29.99	**0.7372**	**30.01**	0.7363	**30.06**	**0.7395**
Industrial	26.77	0.6952	28.88	0.7931	29.04	0.7977	28.98	0.7956	**29.04**	**0.7982**	28.98	0.7942	29.04	0.7980	**29.24**	**0.8034**
Meadow	33.86	0.7483	34.63	0.7804	34.65	0.7815	34.55	0.7804	34.69	**0.7821**	34.62	0.7805	**34.70**	0.7815	**34.75**	**0.7823**
M-Residential	26.36	0.6335	28.34	0.7365	**28.52**	0.7415	28.45	0.7390	28.49	**0.7418**	28.48	0.7385	28.46	0.7408	**28.73**	**0.7471**
Mountain	29.51	0.7349	30.63	0.7885	30.72	0.7908	30.72	0.7907	30.74	0.7916	30.76	0.7915	**30.78**	**0.7923**	**30.83**	**0.7939**
Park	29.06	0.7530	30.54	0.8130	30.72	0.8170	30.71	0.8167	30.71	0.8177	30.72	0.8174	**30.71**	**0.8189**	**30.86**	**0.8209**
Parking	24.24	0.7060	27.25	0.8317	27.50	0.8372	27.32	0.8341	27.57	**0.8408**	**27.60**	0.8396	27.56	0.8405	**27.87**	**0.8473**
Playground	32.64	0.8450	35.37	0.8943	**35.61**	0.8964	35.46	0.8952	35.58	**0.8967**	35.53	0.8956	35.49	0.8959	**35.78**	**0.8984**
Pond	30.70	0.8167	32.11	0.8542	32.21	0.8555	32.17	0.8549	**32.22**	**0.8559**	32.19	0.8552	32.18	0.8555	**32.28**	**0.8565**
Port	26.67	0.7986	28.50	0.8596	28.76	0.8635	28.71	0.8623	28.71	0.8631	28.77	0.8626	**28.81**	**0.8638**	**28.98**	**0.8667**
Railway Station	26.78	0.6793	28.72	0.7738	**28.91**	**0.7789**	28.84	0.7762	28.89	0.7783	28.88	0.7756	28.88	0.7780	**29.06**	**0.7824**
Resort	26.79	0.7029	28.52	0.7799	**28.72**	0.7846	28.64	0.7825	28.68	0.7845	28.67	0.7825	28.71	**0.7849**	**28.86**	**0.7883**
River	30.37	0.7402	31.55	0.7891	31.62	0.7906	31.61	0.7904	**31.64**	**0.7914**	31.63	0.7905	31.63	0.7909	**31.70**	**0.7923**
School	27.41	0.7237	29.36	0.8044	29.55	0.8089	29.51	0.8074	**29.55**	0.8097	29.51	0.8074	29.54	**0.8104**	**29.74**	**0.8140**
S-Residential	26.66	0.6006	27.71	0.6728	27.84	0.6759	27.84	0.6754	27.84	**0.6767**	27.85	0.6754	**27.88**	0.6759	**27.95**	**0.6791**
Square	28.55	0.7391	30.84	0.8200	31.03	0.8237	30.94	0.8223	**31.04**	0.8244	30.98	0.8227	31.00	**0.8251**	**31.24**	**0.8279**
Stadium	27.16	0.7547	29.63	0.8387	**29.82**	**0.8425**	29.68	0.8391	29.79	0.8422	29.73	0.8396	29.77	0.8422	**30.03**	**0.8465**
Storage Tanks	25.65	0.6793	27.44	0.7664	27.61	0.7705	27.58	0.7688	**27.61**	**0.7709**	27.58	0.7680	27.60	0.7698	**27.72**	**0.7734**
Viaduct	26.97	0.6755	28.99	0.7757	29.16	0.7805	29.08	0.7772	**29.17**	**0.7813**	29.08	0.7775	29.11	0.7794	**29.32**	**0.7851**
Average	28.86	0.7382	30.65	0.8086	**30.82**	0.8121	30.72	0.8108	30.81	**0.8126**	30.80	0.8109	30.81	0.8124	**30.97**	**0.8156**

2) Quantitative Results: Table II provides quantitative results in terms of PSNR and SSIM on the AID dataset, showcasing the performance of comparative models across 30 scene categories. We observed that the improvement of TransENet over RCAN is marginal. The primary reason for this might be the challenging scene diversity inherent in remote sensing images, posing difficulties for transformer-based SR methods to generalize well across various remote sensing scenes. Notably, our TTST demonstrates superior performance compared to both CNN and Transformer-based methods, achieving a substantial margin of improvement across all land cover types. This indicates the favorable reconstruction performance of TTST in diverse remote sensing scenarios. Specifically, compared to the impressive HAT-L model, our TTST exhibits a remarkable 0.16dB improvement in average PSNR. This illustrates that our TTSA excels at leveraging critical tokens against the channel attention used in HAT-L.

In Table III, we further evaluate the average PSNR and SSIM results on DOTA v1.0 and DIOR datasets. We can see that recent state-of-the-art CNN-based methods have achieved comparable performance on these remote sensing datasets. For instance, NLSA, which explores global prior knowledge using non-local sparse attention, shows a marginal PSNR decrease of 0.02dB compared to RCAN. On the other hand, HAT-L demonstrates notable improvements against CNN-based approaches by leveraging the long-range representation capability of self-attention. In comparison, our TTST achieves a substantial PSNR improvement of 0.14dB compared to HAT-L. This improvement aligns with our motivation to design the token selective attention, which aims to explore the most critical tokens for restoration while eliminating the

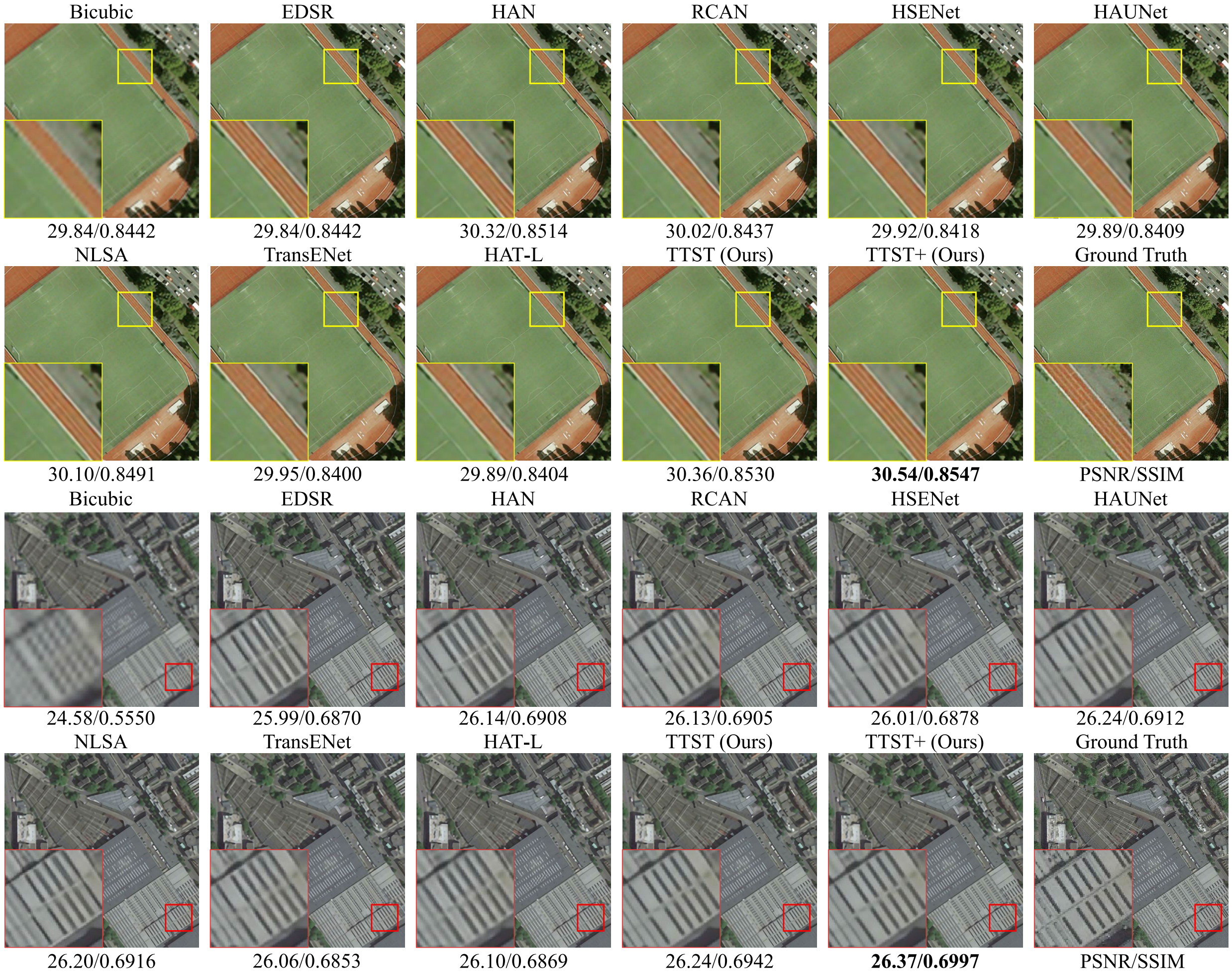


Fig. 8. ×4 visual comparisons on "playground_270" (top) and "railwastation_048" (bottom) samples of AID900. The best PSNR/SSIM is shown in **bold**. Zoom in for better observation.

TABLE III

QUANTITATIVE RESULTS ON AID, DATA V1.0 AND DIOR TEST SET. ALL THESE FLOPS ARE CALCULATED WITH A 3 × 128 × 128 IMAGE INPUT. TTST+ MEANS THE SELF-ENSEMBLE RESULTS OF OUR TTST. THE BEST AND SECOND BEST METRICS ARE SHOWN IN **RED BLOD** AND **BLUE BLOD**, RESPECTIVELY

Type	Methods	#Param.	FLOPs	AID [64]		DOTA [64]		DIOR [65]		Average	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	Bicubic	-	-	28.86	0.7382	31.16	0.7947	28.57	0.7432	29.83	0.7587
CNN-based	EDSR [68]	43.09M	823.34G	30.65	0.8086	33.64	0.8648	30.63	0.8116	31.64	0.8283
	RCAN [38]	15.59M	261.01G	30.82	0.8121	33.86	0.868	30.85	0.8159	31.85	0.8320
	HAN [39]	16.07M	268.89G	30.80	0.8121	33.84	0.8682	30.84	0.8163	31.83	0.8322
	NLSA [25]	44.15M	840.79G	30.81	0.8126	33.86	0.8682	30.82	0.8156	31.83	0.8321
	HSENet [26]	21.70M	306.31G	30.72	0.8108	33.85	0.8667	30.77	0.8143	31.78	0.8306
	HAUNet [71]	9.06M	85.61G	30.88	0.8132	33.94	0.8687	30.87	0.8160	31.90	0.8326
Transformer-based	TransENet [43]	37.46M	87.85G	30.80	0.8109	33.75	0.8675	30.85	0.8148	31.80	0.8311
	HAT-L [27]	40.32M	672.15G	30.81	0.8124	33.99	0.8684	30.87	0.8161	31.90	0.8323
	TTST (Ours)	18.94M	317.68G	**30.97**	**0.8156**	**34.17**	**0.8707**	**30.98**	**0.8178**	**32.04**	**0.8347**
	TTST+ (Ours)	18.94M	317.68G	**31.07**	**0.8174**	**34.31**	**0.8724**	**31.10**	**0.8201**	**32.16**	**0.8367**

interference of irrelevant information in large-scale remote sensing imagery.

3) Qualitative Results: Visual comparisons on AID, DOTA v1.0, and DIOR with scale factor ×4 are displayed in Fig. 8, Fig. 9 and Fig.10, from all of which we can see that our TTST can restore more textures when relevant information (*e.g.,* self-similarity) can be found in these remote sensing scenes. In particular, Transformer-based models without distilling the noisy token cannot recover clean textures accurately. For example, when comparing the visual results of image "playground_270" from the AID dataset in Fig. 8, we observe that TTST produces results that are visually close to the ground truth, while other competitive Transformer-based models without a selective mechanism, such as TransENet and HAT-L, struggle to restore severely degraded details. Moreover, compared to other CNN-based methods like HAN and NLSA, our TTST still maintains favorable visual quality with more high-frequency contextual information.

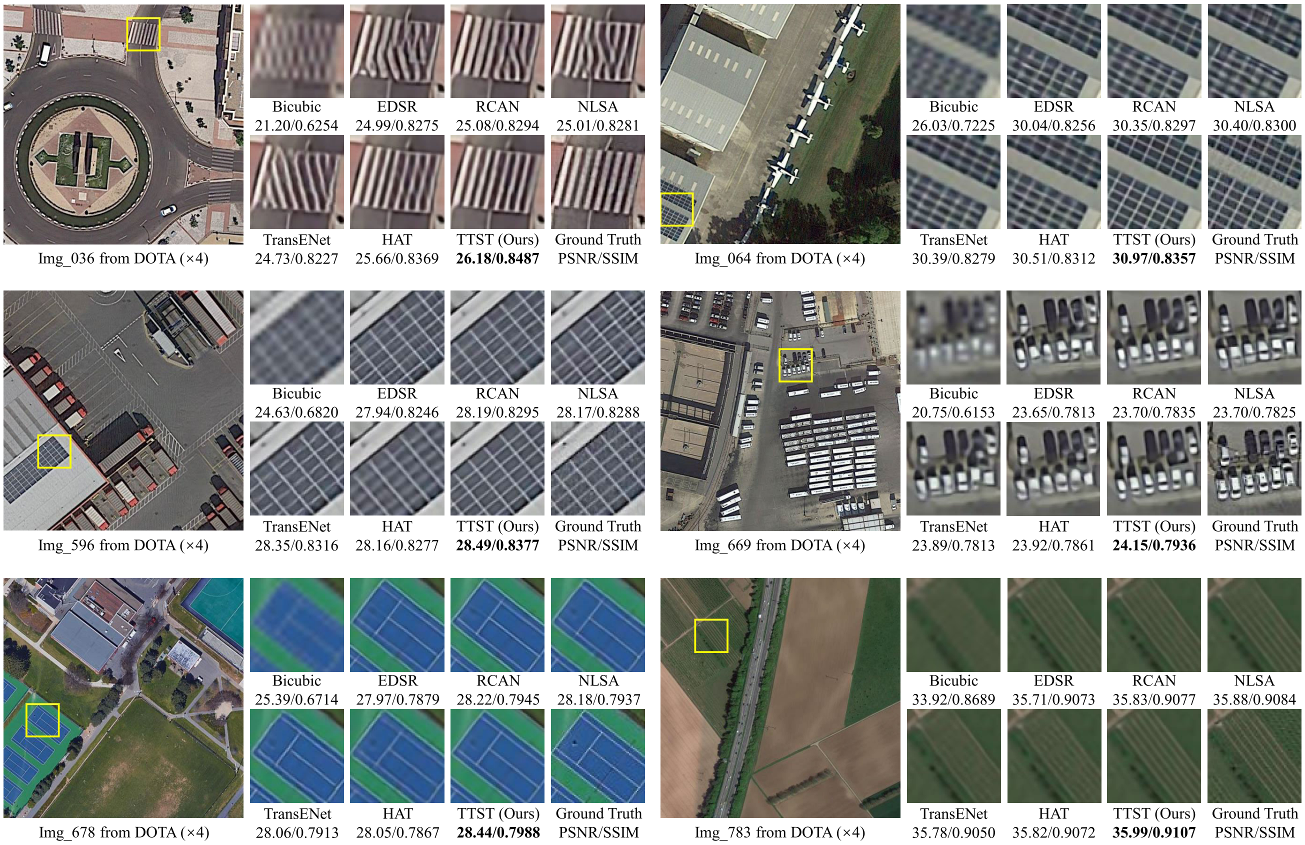


Fig. 9. ×4 visual comparisons on DOTA v1.0. The best PSNR/SSIM is shown in **bold**.

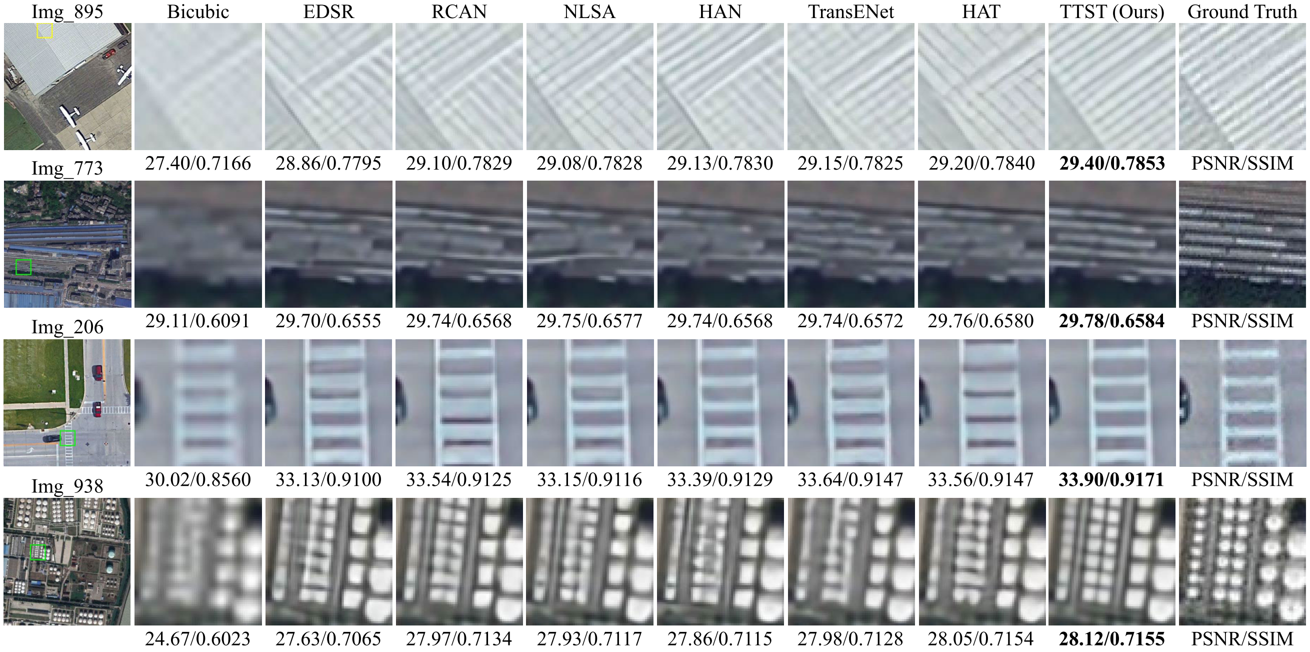


Fig. 10. ×4 visual comparisons on DIOR. The best PSNR/SSIM is shown in **bold**. Zoom in for better comparison.

Similarly, the visual comparisons in Fig. 9 and Fig. 10 provide additional evidence of the superiority of TTST. These comparisons highlight that TTST could restore more realistic distribution than other methods, resulting in clearer and cleaner textures. Specifically, in the case of "Img_036" from the DOTA dataset (Fig. 10), we observe that TransENet and HAT-L fail to accurately restore the shape and boundaries, despite the presence of informative and similar textures in the storage tank of the low-resolution remote sensing image. These visual results further support the effectiveness of TTST in cap-

TABLE IV

Quantitative Results on NWPU-RESISC45 (×4) With Real-World Degradation. The Best and Second Best Metrics Are Shown in **Red Blod** and **Blue Blod**, Respectively

Methods	NWPU-RESISC45 [66]	
	AG ↑	NIQE ↓
Bicubic	2.3556	20.8032
RCAN [38]	3.0296	20.4722
HAN [39]	**3.0451**	20.4276
HSENet [26]	3.0275	**20.2926**
NLSA [25]	3.0343	20.5076
TransENet [43]	2.9862	20.5716
HAT-L [27]	3.0081	20.3153
TTST (Ours)	**3.0393**	**20.2037**

turing global dependencies and enhancing the representation of informative tokens through the proposed top-k token selective attention mechanism.

E. Experiments on Real-World Data

To verify the effectiveness of the proposed TTST in handling real-world degradations, we conduct experiments on the NWPU dataset without applying any simulated degradation.

1) Quantitative Results: Table IV presents the SR performance of various methods in terms of AG and NIQE metrics. The results show that our TTST achieves the best performance in terms of NIQE and secures the second position in AG. This indicates that TTST is robust and competitive in addressing the challenges of SISR for remote sensing imagery with realistic degradations. Notably, while the transformer-based methods no longer outperform the CNN-based approaches as observed in Table IV, our TTST still leads to HAT-L 0.11 in NIQE, demonstrating its superiority to restore realistic results that align with human perception. Moreover, the high AG score indicates that our TTST effectively recovers more high-frequency textures in complex real-world scenes with various degradations.

2) Qualitative Results: The visual comparisons on real-world remote sensing imagery are presented in Fig. 11. From the observations in Fig. 11, it is evident that our TTST produces visually appealing textures with rich high-frequency details. Specifically, when zoomed in on the blue region of interest (ROI), all CNN-based methods exhibit noticeable artifacts and noise, while our TTST generates results with sharper and cleaner edges, providing a visually superior outcome. Moreover, in the red ROI, only our TTST successfully restores realistic details in the outlines. These results affirm the efficacy of our TTST in super-resolving remote sensing images, demonstrating its practical applicability. By leveraging the top-k token selective attention mechanism, our TTST effectively removes noise and blurring artifacts in real-world remote sensing images.

F. Ablation Study

In the ablation section, we conduct extensive discussions of the model design and the key components of our TTST. Notably, we trained these models on AID and observed their PSNR performance on AID-tiny, which contains 30 random images from AID and does not overlap with the training and test set.

TABLE V

The Effect of the Key Components in Our TTST. Model-A Is a Baseline Model, Where We Replace TTSA, MLF, and GCA With a 3 × 3 CNN Layer, a Vanilla MLP Layer, and a 3 × 3 CNN Layer With Similar Parameters to Each Component. The Best PSNR Performance Is Shown in **Blod**

Method	TTSA	MFL	GCA	PSNR (dB)
Model-A	×	×	×	27.944
Model-B	✓	×	×	28.103
Model-C	✓	✓	×	28.147
Model-D (Ours)	✓	✓	✓	**28.201**

TABLE VI

The Generalization of Our TTSA. All These FLOPs Are Calculated With a 3 × 128 × 128 Image Input

Models	#Param.	FLOPs	PSNR (dB)
HAT-L (w/. channel attention)	40.32M	672.15G	28.136
HAT-L (w/. our TTSA)	35.07M	567.83G	28.183 (↑0.047)

1) Top-k Token Selective Attention:

a) Effect of TTSA: The effect of Top-k Token Selective Attention in our TTST on SR performance is listed in Table V. Model-A serves as the baseline, where all the key components, including TTSA, Multi-Scale Feed-forward Layer (MFL), and Global Context Attention (GCA) in TTST, are replaced with CNN layers, a vanilla MLP layer, and CNN layers, respectively.

By comparing the PSNR values in the first and second columns of Table V, we observe that our Top-k Token Selective Attention (TTSA) yields a significant improvement of 0.156dB in PSNR. This improvement is achieved by leveraging the most informative tokens in self-attention, thereby enhancing the reconstruction process. Furthermore, in Fig. 12, we present the average PSNR results (without top-k selection) at the bottom, where our TTSA consistently achieves high-fidelity restoration with superior PSNR performance. For example, in DIOR, DOTA, and AID datasets, our TTSA contributes to a performance improvement of 0.15dB, 0.19dB, and 0.23dB, respectively. To investigate the generalization capability of TTSA, we incorporate the TTSA as a plug-and-play component into HAT-L. The quantitative results are presented in Table VI. TTSA can be comfortably integrated into other transformer-based SR approaches and promote performance (28.183dB vs. 28.136dB), which demonstrates its robustness and favorable generalization capability.

To better understand the effect of the top-k selective mechanism, we further visualize the learned sparse mask $\mathcal{M}_k$ in the first head of the multi-head self-attention. In Fig. 14, we present the sparse mask when the selective rate k is set to 50%. Additionally, we provide the 2nd and 20th feature maps in the key token $\mathbf{K}$. In the 2nd feature map, the airplane with rich texture is poorly distinguished from the background, *i.e.,* it has similar response values across the feature map. In this case, the 2nd feature map is not informative and can be treated as a noisy token. Therefore, the corresponding element in $\mathcal{M}_k$ is zero, which means it will be distilled by our TTSA. In contrast, the 20th feature map exhibits a more informative pattern, where the airplane is well-activated with prominent and clear details. Thus TTSA retains 20th feature maps for self-attention matrix calculation. This visualization demonstrates the effective selection of features.

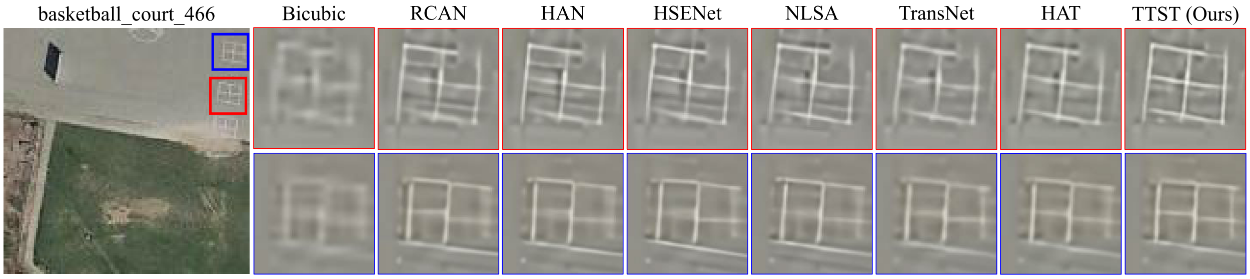


Fig. 11. ×4 visual comparisons on NWPU-RESISC45 with real-world degradation.

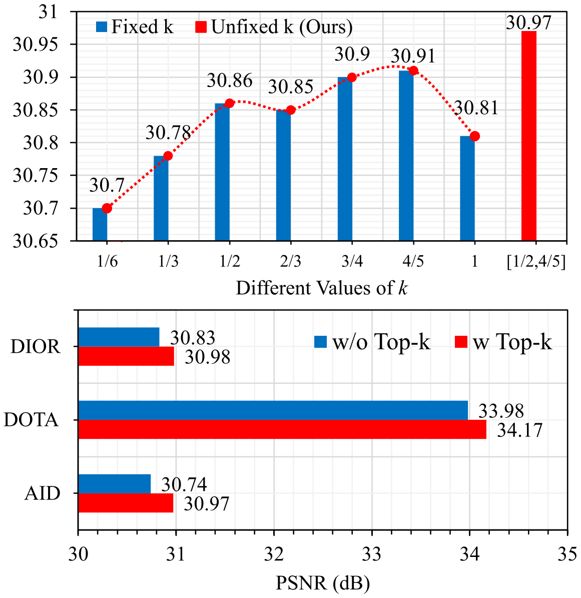


Fig. 12. Ablation analysis of different values of k on AID (top) and the effect of TTSA on all test sets (bottom).

b) Effect of different values of k: The key parameter of our top-k token selective attention is the value of k. As mentioned before, instead of setting k to a single value, we propose to dynamically set k to multiple values, allowing for sparse to dense selection. The PNSR performance of different k is investigated in Fig. 12. We observe that setting k to a small value, such as 1/6, leads to a dramatic drop in performance as there is insufficient long-range information available for restoration. On the other hand, selecting all tokens for self-attention calculation (*i.e.,* k=100%) also degrades the PSNR performance due to the interference of more irrelevant tokens. To strike a favorable balance between sparsity and density, we set k to multiple values with a controllable interval to dynamically capture the most influential tokens. As shown in Fig. 12, a relatively promising performance of 30.97dB is achieved when k is within the range $\left[\frac{1}{2}, \frac{4}{5}\right]$.

c) Effect of channel-wise selection: Formally, the mask size of our TTSA is determined by the channel number in each head, *i.e.,* $\frac{C^2}{num_{head}^2}$, where $C = 180$ and $num_{head} = 6$ in our final TTST. Compared to the spatial-wise selection mechanism, a mask value needs to be predicted for each pixel, resulting in a mask size of $\frac{H^2}{W^2}$, where H and W denote the height and width of the input image, respectively. In the training process, we crop patches with a size of 64 × 64, and the mask size of spatial selection is nearly 4.6 times larger than that of our channel-wise selection. As the image size increases during the inference stage, the computational cost of spatial-wise selection will grow dramatically. Therefore, our channel-wise selection mechanism is memory-efficient and helps reduce inference complexity.

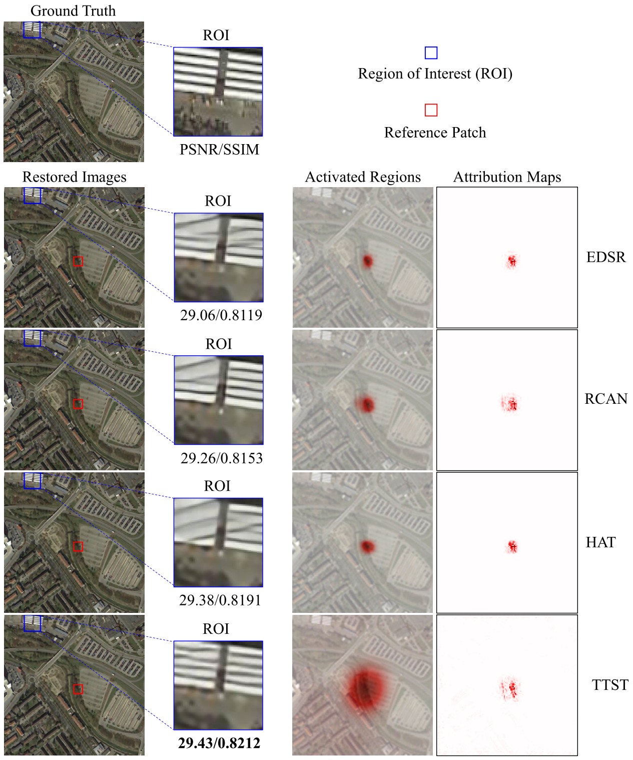


Fig. 13. Local Attribution Maps (LAM) [62] comparisons. We also present the activated area of contribution. The best PSNR performance is shown in **blod**.

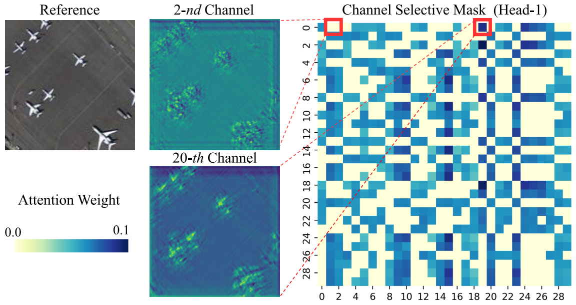


Fig. 14. Visualization of the 2nd and 20th feature maps of K and the mask $\mathcal{M}_k$ with mask rate $k = 50\%$.

To investigate the performance of our channel-wise token selection strategy, we compare it with the k-Nearest Neighbors (k-NN) attention, which performs top-k selection on the spatial dimension with a fixed k value. The results in terms of PSNR,

TABLE VII

Comparison of Our Top-k Token Selective Attention (TTSA) and Other Selective Attention. The FLOPs Are Calculated With a 180 × 64 × 64 Tensor Input. The GPU Memory Is Tested on an NVIDIA RTX 3090 GPU, and the Best PSNR Performance Is Shown in **Blod**

Methods	TTSA (Ours)	k-NN Attention [53]
#Param.	134.46K	129.88K
FLOPs	0.5507G	2.1237G
GPU Memory	832M	1753M
PSNR (dB)	**28.201**	28.107

TABLE VIII

The Effect of Multi-Scale Feed-Forward Layer (MFL). For a Fair Comparison, the Embedding Dimension in the MLP Layer Is Set to 360, Making the Parameter Comparable to Our MFL. The FLOPs Are Calculated With a 180 × 128 × 128 Tensor Input, and the Best PSNR Performance Is Shown in **Blod**

Methods	MLP-based	CNN-based			
		3×3	5×5	7×7	**MFL (Ours)**
#Param.	133.7K	139.3K	156.6K	182.5K	159.5K
FLOPs	2.176G	2.282G	2.566G	2.990G	2.613G
PSNR (dB)	28.014	27.912	28.029	28.103	**28.201**

parameters, Floating Point Operations (FLOPs), and GPU memory are presented in Table VII. It can be observed that our TTSA significantly reduces the computational complexity compared to the k-NN attention. For instance, TTSA reduces the FLOPs by 74% (0.5507G vs. 2.1237G) and memory cost by 53%, while achieving the best PSNR performance (28.201dB vs. 28.107dB).

2) Multi-Scale Feed-Forward Layer: **a) Effect of performing MFL**. The effect of MFL in our TTST on SR performance is listed in Table V. By comparing Model-B and Model-C, we observe that MSF brings 0.044dB improvement. To further assess the effectiveness of our MFL, we conduct a comparison with the standard MLP layer, which is widely used in Transformers [27], [30], [43]. The PSNR performance and model efficiency analysis on the AID-tiny dataset is reported in Table VIII. It is observed that CNN, which retains the critical locality for restoration tasks, outperforms the fully-connected MLP layer. **b) Effect of Multi-Scale Design**. To investigate the effectiveness of multi-scale design, we individually adopt single-scale DW-Conv (*i.e.,* 3×3, 5×5, and 7×7) in MFL for comparison. Although various single-scale DW-Convs are exploited in the feed-forward procedure, they fail to simultaneously leverage multi-scale knowledge. In contrast, our MFL explores and incorporates multi-scale cues during the feed-forward process, resulting in notable performance improvements. Specifically, our MFL achieves a substantial PSNR gain of 0.187dB over the naive MLP layer.

3) Global Context Attention:

a) Effect of GCA: The impact of GCA on SR performance is reported in Table V. A comparative analysis between Model-C and Model-D reveals a notable improvement of 0.054dB in PSNR, highlighting the efficacy of GCA in extracting global context by dynamically adjusting the large respective field.

More intuitively, we visualize the **L**ocal **A**ttribution **M**aps (LAM) of several state-of-the-art models. LAM adopts the integrated gradients method, which interprets an SR network by attributing the existence of certain features of local patches in the output image. The output of LAM represents the pixel-wise importance of the input LR image to restore a certain patch in the SR image. As shown in Fig. 13, darker red pixels indicate a higher contribution to the restoration process. Our TTST exhibits a larger activation of pixels in large-scale remote sensing imagery, indicating that our GCA significantly expands the respective field. Furthermore, more pixels with high contributions are involved in the reconstruction process of our TTST, suggesting that our GCA effectively grasps useful global prior knowledge for improved SR performance.

b) Effect of Kernel decomposition: As reported in Table IX, we provide a comprehensive evaluation of different kernel decomposition sequences (k, d) across various respective fields R. The model parameters, FLOPs, and PSNR performance are also presented for comparison. The results reveal that large-kernel decomposition can significantly reduce the computational complexity compared to applying a single large-kernel convolution. For example, when $\mathcal{R} = 11$, kernel decomposition substantially reduces the computation cost in terms of parameters (120.803K vs. 3920.58K) and FLOPs (1.958G vs. 64.23G). Moreover, models with kernel composition can surpass the single-scale design under the same respective field. This aligns with the motivation of our GCA. While single-scale large-kernel convolution increases the respective field, it fails to consider the scale variation of prior knowledge, which is essential for global contextual exploration. In contrast, our kernel decomposition strategy can effectively characterize multi-scale contexts without compromising the receptive field, resulting in superior SR performance. After extensive analysis, we can find that $\mathcal{R} = 23$ is determined to be the most effective as it offers the best performance in PSNR.

4) Model Complexity: To comprehensively assess the trade-off between performance and complexity, we systematically adjust the model settings of TransENet, HAT-L, and our TTST to create variants with different parameters and FLOPs. The quantitative results are reported in Table X. Initially, we reduce the parameters of TransENet to align its size closely with our TTST (19.44M vs. 18.94M) by modifying the embedding dimension (Dim) from the default 512 to 256. Subsequently, we increase the channel number (C) of feature extraction in TransENet from the default 64 to 180, making its FLOPs similar to our TTST (307.87G vs. 314.68G). Notably, TTST exhibits superior PSNR performance when the model complexity is comparable to TransENet. Similarly, we investigate the model complexity of HAT-L by changing the embedding dimension (Dim) from the default 180 to 128 and the number of **R**esidual **H**ybrid **A**ttention **B**locks (RHAB) from default 12 to 6. These results demonstrate that TTST outperforms both TransENet and HAT-L under similar model complexity.

Furthermore, we explore various model settings of our TTST. When the channel number (C) is set to 64, the complexity of TTST reduces significantly in parameters (5.82M vs. 18.94M) and FLOPs (98.47G vs. 314.68G). However, the PSNR performance diminishes (28.074dB vs. 28.201dB) due to the limited representation capability.

TABLE IX

The Effect of Kernel Decomposition. The Notation (k, d) Indicates That We Decompose the Large Kernel Convolution Into a Sequence of Small Kernel Depth-Wise Convolutions (DW-Convs) With a Kernel Size of k and a Distillation Rate of d. All the FLOPs Are Calculated Using a $180 \times 128 \times 128$ Tensor Input, and the Best PSNR Performance Is Highlighted in **Bold**

Kernel Type	$\mathcal{R}$	(k, d) sequence	#Param.	FLOPs	PSNR (dB)
Small Conv	3	-	291.777K	4.776G	28.093
	5	-	810.176K	13.27G	28.127
	7	-	1587.78K	26.01G	28.102
	9	-	2624.58K	42.99G	28.139
Large Conv & *Small* DW-Convs	11	-	3920.58K	64.23G	28.010
		(3,1)⟶(5,2)	120.803K	1.958G	28.114
	23	-	17139.8K	280.8G	28.129
		(5,1)⟶(7,3)	127.998K	2.080G	**28.201**
	29	-	27248.6K	446.4G	28.107
		(5,1)⟶(7,4)	127.998K	2.080G	28.189

TABLE X

Model Complexity Analysis. The Model Settings of TransENet, HAT-L, and Our TTST Are Modified to Develop Some Variants With Different Model Complexities in Terms of Parameters and FLOPs. All the FLOPs Are Calculated Using a $3 \times 128 \times 128$ Tensor Input, and the Best PSNR Performance Is Highlighted in **Bold**

Models	Setting	#Param.	FLOPs	PSNR (dB)
TransENet [43]	Default	37.46M	87.85G	28.103
	Dim=256	19.44M	58.51G	29.884
	C=180	51.94M	307.87G	28.158
HAT-L [27]	Default	40.32M	672.15G	28.136
	Dim=120	18.15M	310.73G	28.044
	RHAG=6	20.51M	345.63G	28.107
TTST (Ours)	C=64	5.52M	98.47G	28.074
	C=180	18.94M	314.68G	28.201
	RTSG=6	18.94M	314.68G	28.201
	RTSG=12	35.91M	595.63G	**28.255**

Additionally, we increase the number of **R**esidual **T**oken **S**elective **G**roups (RTSG), resulting in further performance improvement (28.255dB vs. 28.201dB). To strike a favorable trade-off between performance and complexity, we set RTSG to 6 as it delivers a satisfactory performance with lower complexity.

V. Conclusion

In this study, we propose a **T**op-*k* **T**oken **S**elective **T**ransformer (TTST) for remote sensing imagery super-resolution (SR), aiming to overcome some critical limitations inherent in the existing transformer-based SR methods. To address the issue of token redundancy, we design a flexible **T**op-*k* **T**oken **S**elective **A**ttention (TTSA) that enables our TTST to prioritize the more valuable tokens with the top-*k* highest similarity to the query, thus filtering out irrelevant information in self-attention. To consider the scale diversity of similar ground objects in remote sensing imagery, we devise a **M**ulti-scale **F**eed-forward **L**ayer (MFL) to generate a set of enriched multi-scale features. Additionally, a **G**lobal **C**ontext **A**ttention (GCA) module is equipped for our TTST to enhance contextual awareness and leverage global prior knowledge in large-scale scenes. TTST outperforms state-of-the-art SR models, both in simulated and real-world remote sensing images. Despite achieving decent performance, remote sensing imagery always suffers from various degradations, making TTST collapse in real-world scenes. Recently, Huang et al. [72] proposed an interpretable transitional learning for degradation representation, achieving a favorable generalization of blind super-resolution. This motivates us to develop an interpretable degradation scheme for remote sensing scenarios, thus promoting the practical application capability of TTST.

References

[1] L. Wu, L. Fang, J. Yue, B. Zhang, P. Ghamisi, and M. He, "Deep bilateral filtering network for point-supervised semantic segmentation in remote sensing images," *IEEE Trans. Image Process.*, vol. 31, pp. 7419–7434, 2022.

[2] J. Yang, B. Du, Y. Xu, and L. Zhang, "Can spectral information work while extracting spatial distribution? An online spectral information compensation network for HSI classification," *IEEE Trans. Image Process.*, vol. 32, pp. 2360–2373, 2023.

[3] Y. Yang, X. Tang, Y.-M. Cheung, X. Zhang, and L. Jiao, "SAGN: Semantic-aware graph network for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1011–1025, 2023.

[4] B. Liu, C. Xu, Z. Cui, and J. Yang, "Progressive context-dependent inference for object detection in remote sensing imagery," *IEEE Trans. Image Process.*, vol. 32, pp. 580–590, 2023.

[5] M. Lin, G. Yang, and H. Zhang, "Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images," *IEEE Trans. Image Process.*, vol. 32, pp. 57–71, 2023.

[6] Q. Zhang, Q. Yuan, M. Song, H. Yu, and L. Zhang, "Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising," *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, 2022.

[7] Q. Zhang, Y. Zheng, Q. Yuan, M. Song, H. Yu, and Y. Xiao, "Hyperspectral image denoising: From model-driven, data-driven, to model-data-driven," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, pp. 1–21, Jun. 2023.

[8] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J. Huang, J. Chanussot, and G. Vivone, "LRTCFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Trans. Image Process.*, vol. 32, pp. 1640–1655, 2023.

[9] Y. Huang, J. Li, X. Gao, Y. Hu, and W. Lu, "Interpretable detail-fidelity attention network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 2325–2339, 2021.

[10] B. Xia, Y. Tian, Y. Zhang, Y. Hang, W. Yang, and Q. Liao, "Meta-learning-based degradation representation for blind super-resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 3383–3396, 2023.

[11] Q. Cai et al., "HIPA: Hierarchical patch transformer for single image super resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 3226–3237, 2023.

[12] P. Wei, Z. Xie, G. Li, and L. Lin, "Taylor neural network for real-world image super-resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 1942–1951, 2023.

[13] J. Song, K. Liu, A. Sowmya, and C. Sun, "Super-resolution phase retrieval network for single-pattern structured light 3D imaging," *IEEE Trans. Image Process.*, vol. 32, pp. 537–549, 2023.

[14] J. He et al., "Spectral super-resolution meets deep learning: Achievements and challenges," *Inf. Fusion*, vol. 97, Sep. 2023, Art. no. 101812.

[15] R. Ran, L.-J. Deng, T.-X. Jiang, J.-F. Hu, J. Chanussot, and G. Vivone, "GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution," *IEEE Trans. Cybern.*, vol. 53, no. 7, pp. 4148–4161, Jul. 2023.

[16] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.

[17] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.*, vol. 22, no. 2, pp. 56–65, Mar. 2002.

[18] K. I. Kim and Y. Kwon, “Single-image super-resolution using sparse regression and natural image prior,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1127–1133, Jun. 2010.

[19] J. Sun, J. Sun, Z. Xu, and H.-Y. Shum, “Gradient profile prior and its applications in image super-resolution and enhancement,” *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1529–1542, Jun. 2011.

[20] R. Timofte, V. De, and L. V. Gool, “Anchored neighborhood regression for fast example-based super-resolution,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[21] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, “Deep learning for single image super-resolution: A brief review,” *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.

[22] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, “Single-frame super-resolution in remote sensing: A practical overview,” *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, Jan. 2017.

[23] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, “Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration,” in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 669–687.

[24] Y. Xiao, Q. Yuan, Q. Zhang, and L. Zhang, “Deep blind super-resolution for satellite video,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5516316.

[25] Y. Mei, Y. Fan, and Y. Zhou, “Image super-resolution with non-local sparse attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3516–3525.

[26] S. Lei and Z. Shi, “Hybrid-scale self-similarity exploitation for remote sensing image super-resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401410.

[27] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, “Activating more pixels in image super-resolution transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22367–22377.

[28] Y. Xiao et al., “Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial–temporal transformer,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, Apr. 2022, Art. no. 102731.

[29] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[30] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, “Transformer for single image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 456–465.

[31] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, “SRFormer: Permuted self-attention for single image super-resolution,” 2023, *arXiv:2303.09735*.

[32] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[33] S. Lei, Z. Shi, and Z. Zou, “Super-resolution for remote sensing images via local–global combined network,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.

[34] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, “Deep distillation recursive network for remote sensing imagery super-resolution,” *Remote Sens.*, vol. 10, no. 11, p. 1700, Oct. 2018.

[35] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, “Edge-enhanced GAN for remote sensing image superresolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[36] K. Jiang, Z. Wang, P. Yi, and J. Jiang, “Hierarchical dense recursive network for image super-resolution,” *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107475.

[37] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, “From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution,” *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.

[38] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[39] B. Niu et al., “Single image super-resolution via a holistic attention network,” in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 191–207.

[40] B. Kim, J. Kim, and J. C. Ye, “Task-agnostic vision transformer for distributed learning of image processing,” *IEEE Trans. Image Process.*, vol. 32, pp. 203–218, 2023.

[41] J. He, Q. Yuan, J. Li, Y. Xiao, X. Liu, and Y. Zou, “DsTer: A dense spectral transformer for remote sensing spectral super-resolution,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, May 2022, Art. no. 102773.

[42] Y. Xiao et al., “Local–global temporal difference learning for satellite video super-resolution,” *IEEE Trans. Circuits Syst. Video Technol.*, early access, pp. 1–14, Sep. 2023, doi: 10.1109/TCSVT.2023.3312321.

[43] S. Lei, Z. Shi, and W. Mo, “Transformer-based multistage enhancement for remote sensing image super-resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611.

[44] J. Fang, H. Lin, X. Chen, and K. Zeng, “A hybrid network of CNN and transformer for lightweight image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1102–1111.

[45] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yang, and F. Yu, “Dual aggregation transformer for image super-resolution,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 12312–12321.

[46] Q. Zhu, P. Li, and Q. Li, “Attention retractable frequency fusion transformer for image super resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1756–1763.

[47] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, “UFormer: A general U-shaped transformer for image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

[48] J. Xiao, X. Fu, A. Liu, F. Wu, and Z.-J. Zha, “Image de-raining transformer,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2022.

[49] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[50] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, “Visual attention network,” 2022, *arXiv:2202.09741*.

[51] G. Zhao, J. Lin, Z. Zhang, X. Ren, Q. Su, and X. Sun, “Explicit sparse transformer: Concentrated attention through explicit selection,” 2019, *arXiv:1912.11637*.

[52] H. Wang, J. Shen, Y. Liu, Y. Gao, and E. Gavves, “NFormer: Robust person re-identification with neighbor transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7287–7297.

[53] P. Wang et al., “KVT: *k*-NN attention for boosting vision transformers,” in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 285–302.

[54] Q. He et al., “AST: Adaptive self-supervised transformer for optical remote sensing representation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 200, pp. 41–54, Jun. 2023.

[55] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[56] H. Yan, Z. Li, W. Li, C. Wang, M. Wu, and C. Zhang, “ConTNet: Why not use convolution and transformer at the same time?” 2021, *arXiv:2104.13497*.

[57] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12159–12168.

[58] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[59] X. Ding, X. Zhang, J. Han, and G. Ding, “Scaling up your kernels to 31 × 31: Revisiting large kernel design in CNNs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11963–11975.

[60] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, “Restormer: Efficient transformer for high-resolution image restoration,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[61] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, “Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610819.

[62] J. Gu and C. Dong, “Interpreting super-resolution networks with local attribution maps,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9195–9204.

[63] G.-S. Xia et al., “AID: A benchmark data set for performance evaluation of aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[64] G.-S. Xia et al., “DOTA: A large-scale dataset for object detection in aerial images,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[65] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[66] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[67] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

[68] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[69] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[70] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[71] J. Wang, B. Wang, X. Wang, Y. Zhao, and T. Long, "Hybrid attention-based U-shaped network for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612515.

[72] Y. Huang, J. Li, Y. Hu, X. Gao, and H. Huang, "Transitional learning: Exploring the transition states of degradation for blind super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6495–6510, May 2023.

Yi Xiao (Graduate Student Member, IEEE) received the B.S. degree from the School of Mathematics and Physics, China University of Geosciences, Wuhan, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan.

His major research interests include remote sensing image/video processing and computer vision. For more information visit the link: https://xy-boy.github.io/.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

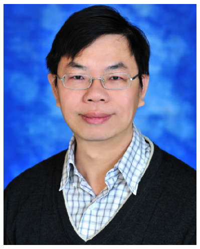
In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He has published more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTION ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the recognition of Best Reviewers of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS 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 he has frequently served as a referee for more than 40 international journals for remote sensing and image processing.

Kui Jiang (Member, IEEE) received the M.E. and Ph.D. degrees from the School of Computer Science, Wuhan University, Wuhan, China, in 2019 and 2022, respectively. Before July 2023, he was a Research Scientist with the Cloud BU, Huawei. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He received the 2022 ACM Wuhan Doctoral Dissertation Award. His research interests include image/video processing and computer vision.

Jiang He (Graduate Student Member, IEEE) received the B.S. degree in remote sensing science and technology from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include hyperspectral super-resolution, image fusion, quality improvement, remote sensing image processing, and deep learning.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU, where he is also the Deputy Director of the AI Research Center. He was with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, from 2000 to 2007. Prior to joining academia, he was with the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, from 1992 to 2000. His research interests include image and video processing, computer vision, and video networking.

Dr. Lin served as a Distinguished Lecturer of IEEE Circuits and Systems Society from 2018 to 2019, a Steering Committee Member of IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2015, and the Chair of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society from 2013 to 2015. His articles received the Best Paper Award of IEEE VCIP 2015 and the Young Investigator Award of VCIP 2005. He received the Outstanding Electrical Professor Award presented by the Chinese Institute of Electrical Engineering in 2019 and the Young Investigator Award presented by the Ministry of Science and Technology, Taiwan, in 2006. He is also the Chair of the Steering Committee of IEEE ICME. He has served as the Technical Program Co-Chair for IEEE ICME 2010, the General Co-Chair for IEEE VCIP 2018, and the Technical Program Co-Chair for IEEE ICIP 2019. He has served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and *Journal of Visual Communication and Image Representation*.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. 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 to lead the Remote Sensing Program in China. 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 of the IEEE 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 also the Founding Chair of the IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He also serves as an associate editor or an editor for more than ten international journals. He is also serving as an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.