

Article

Enhanced Window-Based Self-Attention with Global and Multi-Scale Representations for Remote Sensing Image Super-Resolution

Yuting Lu ¹ , Shunzhou Wang ², Binglu Wang ^{3,*}, Xin Zhang ³ , Xiaoxu Wang ¹ and Yongqiang Zhao ¹ 

- ¹ School of Automation, Northwestern Polytechnical University, Xi'an 710072, China; lyt1996@mail.nwpu.edu.cn (Y.L.); woyao1982@nwpu.edu.cn (X.W.); zhaoyq@nwpu.edu.cn (Y.Z.)
² Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen 518055, China; shunzhouwang@pku.edu.cn
³ Radar Research Laboratory, School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; xin.zhang@bit.edu.cn
 * Correspondence: wbl@nwpu.edu.cn

Abstract: Transformers have recently gained significant attention in low-level vision tasks, particularly for remote sensing image super-resolution (RSISR). The vanilla vision transformer aims to establish long-range dependencies between image patches. However, its global receptive field leads to a quadratic increase in computational complexity with respect to spatial size, rendering it inefficient for addressing RSISR tasks that involve processing large-sized images. In an effort to mitigate computational costs, recent studies have explored the utilization of local attention mechanisms, inspired by convolutional neural networks (CNNs), focusing on interactions between patches within small windows. Nevertheless, these approaches are naturally influenced by smaller participating receptive fields, and the utilization of fixed window sizes hinders their ability to perceive multi-scale information, consequently limiting model performance. To address these challenges, we propose a hierarchical transformer model named the Multi-Scale and Global Representation Enhancement-based Transformer (MSGFormer). We propose an efficient attention mechanism, Dual Window-based Self-Attention (DWSA), combining distributed and concentrated attention to balance computational complexity and the receptive field range. Additionally, we incorporated the Multi-scale Depth-wise Convolution Attention (MDCA) module, which is effective in capturing multi-scale features through multi-branch convolution. Furthermore, we developed a new Tracing-Back Structure (TBS), offering tracing-back mechanisms for both proposed attention modules to enhance their feature representation capability. Extensive experiments demonstrate that MSGFormer outperforms state-of-the-art methods on multiple public RSISR datasets by up to 0.11–0.55 dB.

Keywords: remote sensing image super-resolution; transformer model; window-based self-attention; global receptive field; multi-scale representation



Citation: Lu, Y.; Wang, S.; Wang, B.; Zhang, X.; Wang, X.; Zhao, Y. Enhanced Window-Based Self-Attention with Global and Multi-Scale Representations for Remote Sensing Image Super-Resolution. *Remote Sens.* **2024**, *16*, 2837. <https://doi.org/10.3390/rs16152837>

Academic Editor: Pedro Melo-Pinto

Received: 29 June 2024

Revised: 27 July 2024

Accepted: 31 July 2024

Published: 2 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The field of remote sensing technology has witnessed remarkable advancements over the past few decades, presenting unprecedented opportunities to comprehend the Earth's surface. The copious amounts of remote sensing image (RSI) data acquired from satellites, aircraft, and remote sensing instruments play a pivotal role in various applications, including ecosystem monitoring [1], object detection [2], and scene understanding [3]. Nevertheless, the quality of remote sensing images is frequently influenced by sensor noise, the movement of imaging platforms, adverse weather conditions, and other factors [4–6], leading to a decline in the spatial resolution of remote sensing images. This reduction hampers the comprehensive analysis and utilization of remote sensing imagery. Consequently, there is significance in the endeavor to reconstruct high-resolution (HR) RSI data by extracting spatial information from low-resolution (LR) RSI data.

Super-resolution (SR) stands as an effective technique for reconstructing the corresponding HR image from a given LR input. With the development of deep learning techniques [7–9], many approaches have emerged in recent years to solve the inherently ill-posed problem of SR. A substantial number of these methods leverage convolutional neural networks [10–15] and have demonstrated remarkable success in the SR domain, benefiting from their robust end-to-end feature representation capabilities. However, the convolutional approach adopts a local mechanism, lacking direct interactions between distant pixels, a crucial aspect for achieving optimal performance [16,17], thus imposing limitations on the model’s overall effectiveness. Recently, transformers, initially proposed in natural language processing (NLP), have exhibited noteworthy performance across various high-level vision tasks [18–22]. At the core of the transformer lies the self-attention (SA) mechanism, which is capable of establishing global dependencies. Recognizing the potential of transformers, some researchers have ventured into applying these models to low-level tasks [23–25], including remote sensing image super-resolution.

While transformers have exhibited notable capabilities in super-resolution tasks, their application to remote sensing images is still limited. Vanilla vision transformers [26,27] incorporating global attention can effectively model dependencies between image patches. However, the expansive participating receptive field of transformers incurs a quadratic computational cost, presenting challenges for remote sensing images that encompass broad spatial coverage. In attempts to mitigate computational demands, some studies [28–31] have introduced inductive biases inspired by CNNs, giving rise to window-based self-attention by constraining attention to localized regions. Nevertheless, remote sensing images often contain a large number of similar natural and artificial landforms, such as hills, grasslands, deserts, buildings, etc., in which the pixels of the same landform have similar distributions but may be far away. Therefore, window-based self-attention naturally has a local receptive field and cannot capture the interaction between distant but similar pixels in remote sensing images. Furthermore, the fixed window size of transformers poses difficulties in explicitly extracting multi-scale features, proving challenging for super-resolving remote sensing images that contain various ground objects at various scales.

To address these challenges, we propose a novel Dual Window-based Self-Attention (DWSA) module, which consists of two parallel operations: distributed attention and concentrated attention. Both of the two attention operations are calculated in windows, but through different window segmentation methods: concentrated attention evenly divides $H \times W \times C$ images into $(\frac{H}{G} \times \frac{W}{G}, G \times G, C)$ with non-overlapping windows of size $G \times G$, while distributed attention divides $H \times W \times C$ images into $(G \times G, \frac{H}{G} \times \frac{W}{G}, C)$ with grid windows of size $G \times G$. Through different window division methods, concentrated attention realizes the capture of local information, while distributed attention realizes the perception of global information. In comparison to full self-attention, DWSA inherently achieves global interactions with linear complexity, providing greater efficiency and rendering it suitable for large-size remote sensing images. In contrast to local attention, DWSA enhances model capacity by introducing a global receptive field, thereby improving super-resolution performance. Additionally, to overcome the limitation of the fixed window size in window-based self-attention in capturing multi-scale information, we propose the Multi-scale Depth-wise Convolution Attention (MDCA) module to be placed before the proposed DWSA module to help it aggregate multi-scale information in remote sensing images. Employing three convolution branches with varying kernel sizes, MDCA facilitates the extraction of multi-scale features, providing a multi-level understanding of remote sensing images. This module comprehensively captures features across scales within the remote sensing image. Despite these enhancements, an analysis of MDCA and DWSA behavior reveals constraints in their feature representation capabilities. To address this, we propose a new Tracing-Back Structure (TBS) for each MDCA and DWSA module to fully improve its feature representation capability. Specifically, we compute the difference between features before and after passing through the proposed attention module and subsequently modulate this difference before fusing it with the output of the attention module

to achieve a more comprehensive feature representation and thus enhance the performance of the model. Collectively, we designate our proposed model as the Multi-Scale and Global Representation Enhancement-based Transformer (MSGFormer). As illustrated in Figure 1, MSGFormer surpasses other state-of-the-art transformers and CNN-based networks on the AID dataset for RSISR, maintaining a balance between performance and model size.

In summary, this paper contributes in three main aspects:

- We present a novel Dual Window-based Self-Attention (DWSA) module, comprising distributed global attention and concentrated local attention, for remote sensing image super-resolution. This innovative approach enables the utilization of a global receptive field while maintaining linear complexity.
- We introduce the Multi-scale Depth-wise Convolution Attention (MDCA) module, which is particularly crucial in addressing the limitation posed by the fixed window size of the transformer, achieved through a multi-branch convolution strategy to enhance model performance.
- We have developed a new Tracing-Back Structure (TBS) to comprehensively enhance the feature representation capabilities of the proposed MDCA and DWSA modules. Accordingly, we introduce the Multi-Scale and Global Representation Enhancement-based Transformer (MSGFormer). The evaluation of various public remote sensing datasets demonstrates that MSGFormer attains state-of-the-art performance.

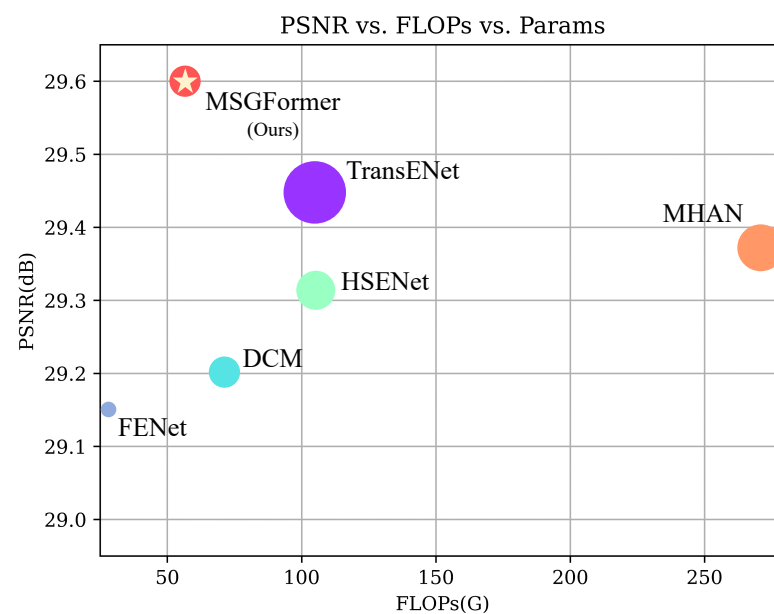


Figure 1. Model complexity and performance comparison between our proposed MSGFormer model and other RSISR methods on the AID dataset [32] for $\times 4$ SR. Circle sizes indicate the number of parameters.

The rest of this article is organized as follows. Section 2 reviews related methods. The proposed methodology is reported in Section 3. Section 4 evaluates the experimental results, while Section 5 provides a detailed discussion of the findings. Finally, the conclusion is presented in Section 6.

2. Related Works

2.1. CNN-Based SR

In recent years, CNNs have significantly advanced the progress of image super-resolution. SRCNN [33], a pioneering method in this domain, employs three convolution layers for this mapping. FSRCNN [34] enhances computational efficiency while maintaining quality through a restructured SRCNN architecture. VDSR [35] introduces a unified

framework for multi-scale image handling, incorporating residual learning and additional network layers. Lim et al. [36] achieved outstanding performance through the simplification of the ResNet model architecture. SRDD [10] uses an end-to-end network that learns a high-resolution dictionary while leveraging deep learning advantages.

In the context of remote sensing images, Lei et al. [37] first proposed LGCNet, a CNN-based model for RSISR, aiming to enhance super-resolution performance through the incorporation of both local and global contrast features. Haut et al. [38] coordinated multiple network design improvements for advanced RSISR performance. Dong et al. [39] developed a second-order multi-scale super-resolution network (SMSR) designed to address challenging cases of reconstruction tasks. Zhang et al. [40] employed High-Order Attention modules for hierarchical feature utilization to enhance RSISR performance. Inspired by the lattice structure, FENet [41] utilizes Lightweight Lattice Blocks (LLBS) for improved expressiveness. Despite the impressive results demonstrated by CNN-based methods, CNNs exhibit a restricted receptive field due to their local mechanism. This limitation hampers the ability to capture long-range dependencies between pixels, consequently constraining the overall capability.

2.2. Transformer-Based SR

Transformers, initially devised for natural language tasks [42], have exhibited significant advancements in diverse vision applications, including image classification [43], object detection [44,45], and semantic segmentation [46,47]. A distinguishing attribute of these models lies in their robust ability to capture long-range dependencies among sequences of image patches, demonstrating adaptability to varying input content [48]. Consequently, transformer models have been explored for low-level vision tasks, including super-resolution [17,28,49].

Vanilla vision transformers directly apply self-attention mechanisms to patches extracted from images. For instance, IPT [26] addresses image super-resolution problems using a pre-trained transformer model. Additionally, TransENet [23] introduces a novel super-resolution framework for remote sensing images, enhancing high-dimensional feature representation after the upsampling layers. Nevertheless, the computational complexity of full self-attention in vanilla transformers can increase quadratically with the number of image patches, limiting their applicability to RSISR [50]. Recent super-resolution methods often address this issue by confining self-attention to local image regions to mitigate complexity [17,28–31]. However, this design choice confines context aggregation within local neighborhoods, deviating from the primary motivation of employing self-attention over convolutions, and is consequently less suited for RSISR. In contrast, we present a transformer model capable of learning long-range dependencies while maintaining computational efficiency.

2.3. Hybrid CNN–Transformer Structure

The hybrid CNN–transformer structure has emerged as a popular solution, combining the strengths of both to address super-resolution challenges. Notably, recent studies [51–53] have underscored the efficacy of integrating transformers and convolutions, leveraging the merits of both. The CvT model [53] pioneered the inclusion of depth-wise and point-wise convolutions preceding self-attention. CMT [51] introduced a hybrid network, employing transformers for capturing long-range dependencies and CNNs for modeling local features. MobileViT [54], EdgeNeXt [55], MobileFormer [56], and EfficientFormer [57] reintegrated convolutions into transformers for efficient network design, demonstrating exceptional performance in image classification and downstream applications. In the realm of super-resolution, the hybrid structure of CNNs and transformers continues to shine. Lu et al. [27] utilized CNNs to dynamically adjust the size of the feature map and transformers to capture long-term dependencies. HNCT [58] and CTCNet [59] use a combination of convolutions and transformers to exploit the collaborative utilization of local and global features. DAT [60] and HAT [17] employ a parallel convolution and transformer structure,

enabling the output of the two branches to adapt and fuse, enhancing both local and global coupling. However, existing hybrid networks still lack a comprehensive exploration of the capabilities of convolutions and transformers, posing challenges in enhancing their performance. In this paper, we introduce a novel Tracing-Back Structure to address this limitation and underscore its significance.

3. Methods

In this section, we present our proposed MSGFormer framework. The detailed structure of MSGFormer is elucidated in Section 3.1. Subsequently, Sections 3.2–3.4 provide in-depth discussions on the components of the DWSA, MDCA, and TBS, respectively.

3.1. Overview of MSGFormer

The architecture of our method consists of three primary modules: shallow feature extraction, deep feature extraction, and image reconstruction, as depicted in Figure 2a. Initially, when given an LR input $RSI I_{LR} \in \mathbb{R}^{H \times W \times C_{in}}$, a convolution layer processes it, yielding the shallow feature $F_S \in \mathbb{R}^{H \times W \times C}$. Here, H and W denote the height and width of the input image, while C_{in} and C represent the channel number of input and intermediate features.

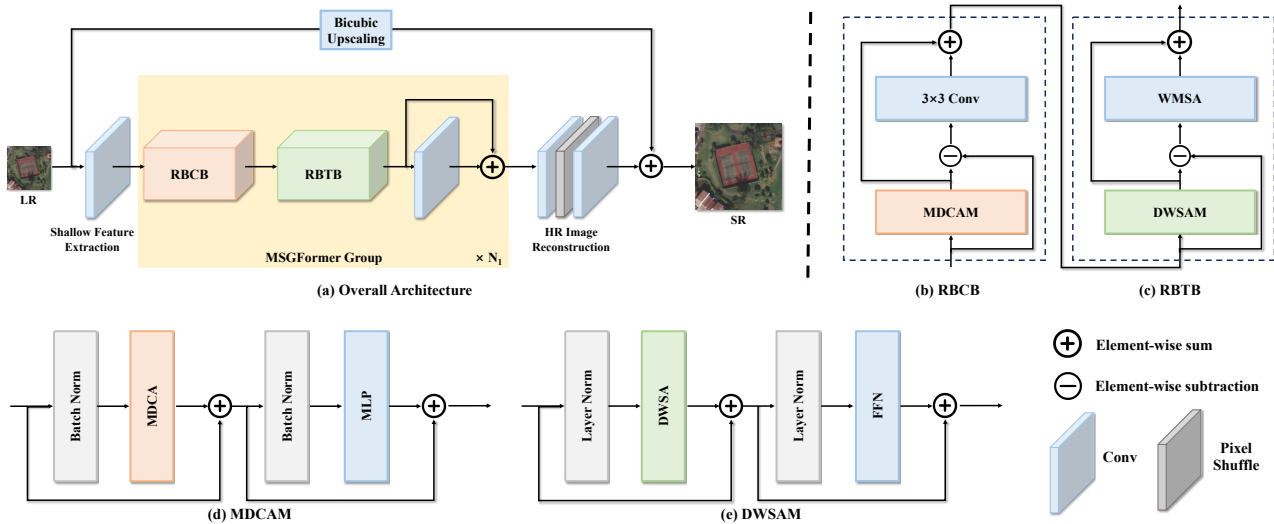


Figure 2. Illustrations of the proposed method. (a) The architecture of MSGFormer. (b) The Tracing-Back Conv Block (TBCB), where WMSA represents vanilla window-based self-attention. (c) The Tracing-Back Transformer Block (TBTB). (d) The Multi-scale Depth-wise Convolution Attention Module (MDCA). (e) The Dual Window-based Self-Attention module (DWSAM).

Subsequently, the shallow feature F_S undergoes processing within the deep feature extraction module, resulting in the deep feature $F_D \in \mathbb{R}^{H \times W \times C}$. This module consists of multiple MSGFormer groups (MSGs), totaling N_1 . Each MSG incorporates a Tracing-Back Conv Block (TBCB) and a Tracing-Back Transformer Block (TBTB) in a sequential manner, as depicted in Figure 2a. The TBCB and TBTB, as detailed in Figure 2b,c, incorporate a Multi-scale Depth-wise Convolution Attention Module (MDCA) and a Dual Window-based Self-Attention module (DWSAM), both augmented by the Tracing-Back Structure (TBS). A 3×3 convolution layer, accompanied by a residual connection, refines features extracted from transformer blocks at the end of each MSG.

Finally, the HR output image $I_{HR} \in \mathbb{R}^{H_{out} \times W_{out} \times 3}$ is reconstructed through the reconstruction module, where H_{out} is the height of the output image, and W_{out} denotes the image width. This module involves upsampling the deep feature F_D using the pixel shuffle method [61]. Moreover, the results of upsampling are incorporated with the inclusion of

a bilinear interpolation of the low-resolution image [62] to enhance the recovery process. Network parameter optimization is performed using the L_1 loss,

$$\mathcal{L} = \|I_{SR} - I_{HR}\|_1, \quad (1)$$

where I_{SR} is obtained by taking I_{LR} as the input of the MSGFormer, and I_{HR} is the corresponding ground-truth HR RSI.

3.2. Dual Window-Based Self-Attention

The advantage of full self-attention lies in its capacity for global information interaction. Nevertheless, the direct application of attention across the entire space is computationally infeasible, primarily due to the quadratic complexity associated with the attention operator. Inspired by the decomposition of large convolution kernels by [63,64], where large convolution kernels are divided into small convolutional kernels and dilated convolutions, we split the full self-attention module into two parallel attention components, distributed attention and concentrated attention, and accordingly propose a novel attention module termed Dual Window-based Self-Attention (DWSA), as illustrated in Figure 3. DWSA can efficiently capture long-range relationships between pixels, achieve a balance between computational complexity and the participating receptive field, and provide an effective and practical solution for global information interaction. Specifically, assuming that each window contains $M \times M$ pixels, the computational complexity of the full multi-head self-attention (MSA) operation and the Dual Window-based Self-Attention operation on an image of $H \times W \times C$ are

$$\Omega(MSA) = 4HWC^2 + 2(HW)^2C, \quad (2)$$

$$\Omega(DMSA) = 4HWC^2 + 4M^2HWC, \quad (3)$$

where the former is quadratically related to the image spatial size $H \times W$, and the latter is linear when M is fixed (set to 16). The computation of full multi-head self-attention is generally not feasible for large remote sensing images, whereas DWSA offers an acceptable solution.

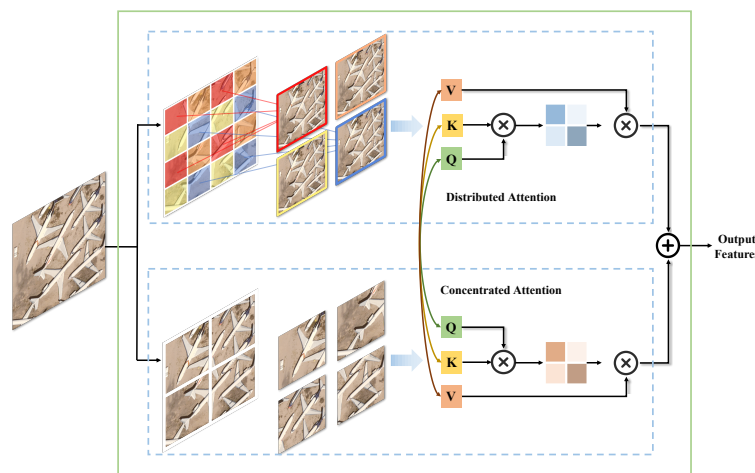


Figure 3. An illustration of Dual Window-based Self-Attention (DWSA). We split the full self-attention module into two parallel attention components: distributed attention and concentrated attention. Both components exhibit linear complexity and achieve global information interaction through distinct window-partitioning methods and shared weights.

3.2.1. Concentrated Attention

Consider an input feature map denoted by $X \in \mathbb{R}^{H \times W \times C}$. Instead of employing self-attention across the entire feature map, we partition the feature map into non-overlapping windows of size $G \times G$, resulting in tensors with dimensions $(\frac{H}{G} \times \frac{W}{G}, G \times G, C)$. Sub-

sequently, self-attention is applied to each partitioned window, allowing information interaction within a localized space. Specifically, for features $X_w \in \mathbb{R}^{G^2 \times C}$ within a window, the corresponding *query*, *key*, and *value* matrices, $Q \in \mathbb{R}^{G^2 \times d}$, $K \in \mathbb{R}^{G^2 \times d}$, and $V \in \mathbb{R}^{G^2 \times C}$, are computed as

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \quad (4)$$

where W_Q , W_K , and W_V are weight matrices. By comparing the similarity between Q and K , we obtain an attention map of size $\mathbb{R}^{G^2 \times G^2}$ and multiply it by V . Overall, the calculation of multi-head self-attention (MSA) can be expressed as

$$MSA(X) = \text{softmax}(QK^T / \sqrt{d})V, \quad (5)$$

where $\sqrt{d}$ is used to control the magnitude of QK^T before applying the softmax function. Similar to the conventional transformer layer [18], the MLP is employed after the MSA module to further transform features. The MLP contains two fully connected layers, and one GELU nonlinearity is applied after the first linear layer.

We employ this concentrated attention with linear complexity in relation to spatial size to enhance interactions among pixels in a local area, as depicted in the lower half of Figure 3.

3.2.2. Distributed Attention

While local self-attention mitigates the computational intensity of full self-attention, it inherently encounters challenges such as limited receptive fields, hindering its ability to model long-range dependencies. Drawing inspiration from concentrated attention, we propose an effective method for achieving global attention, termed distributed attention. We split the feature map into dimensions $(G \times G, \frac{H}{G} \times \frac{W}{G}, C)$ using a uniformly distributed $G \times G$ grid, i.e., sampling the entire feature map at fixed intervals to form a window representing global information, as depicted in the upper portion of Figure 3. Subsequently, we apply self-attention to the partitioned distributed window, facilitating pixel interactions in the global space.

By employing the same concentrated window and distributed window sizes, both of which have only linear complexity concerning spatial size or sequence length, and a shared-weight scheme, we achieve a full balance of computation between local and global operations. This design efficiently mitigates the quadratic complexity of full self-attention by incorporating both local and global considerations, reducing it to linear complexity while retaining non-locality. In contrast to full self-attention, we can capture long-range relationships with minimal computational cost and parameters, thereby enhancing model capacity through the introduction of a global receptive field.

3.3. Multi-Scale Depth-Wise Convolution Attention

The fixed window size of the transformer poses challenges in extracting essential multi-scale features crucial for remote sensing images characterized by significant scale variations. To overcome this constraint, we propose enhancing the transformer's multi-scale capabilities by integrating convolution attention. Specifically, we propose Multi-scale Depth-wise Convolution Attention (MDCA), which incorporates multiple convolutions with varying kernel sizes to capture diverse spatial features across different scales, enabling comprehensive interaction with multi-scale information. As shown in Figure 4, MDCA comprises three components: an initial 3×3 depth-wise convolution for integrating neighborhood features, middle multi-branch depth-wise convolutions for extracting multi-scale information, and a final 1×1 convolution for modeling relationships between different channels. The output of the final 1×1 convolution serves as attention weights,

directly reweighting the input of MDCA. The operational mechanism of MDCA can be expressed mathematically as

$$Att_{end} = Conv_{1 \times 1} \left(\sum_{i=0}^2 mid_conv_i (DW-Conv_{3 \times 3}(F_{input})) + DW-Conv_{3 \times 3}(F_{input}) \right), \quad (6)$$

$$F_{output} = F_{input} \otimes Att_{end}, \quad (7)$$

where $F_{input} \in \mathbb{R}^{H \times W \times C}$ and $F_{output} \in \mathbb{R}^{H \times W \times C}$ are the input and output features of MDCA, and $Att_{end} \in \mathbb{R}^{H \times W \times C}$ is the attention weight. $DW-Conv_{3 \times 3}$ denotes the depth-wise 3×3 convolution and $Conv_{1 \times 1}$ represents the 1×1 convolution. mid_conv_i , $i \in \{0, 1, 2\}$, stands for the convolution of the i th branch in Figure 4, and $\otimes$ denotes element-wise matrix multiplication. This convolution attention enables the capture of key features in the input at different scales and the effective fusion of this information, thereby enhancing the overall information extraction capability.

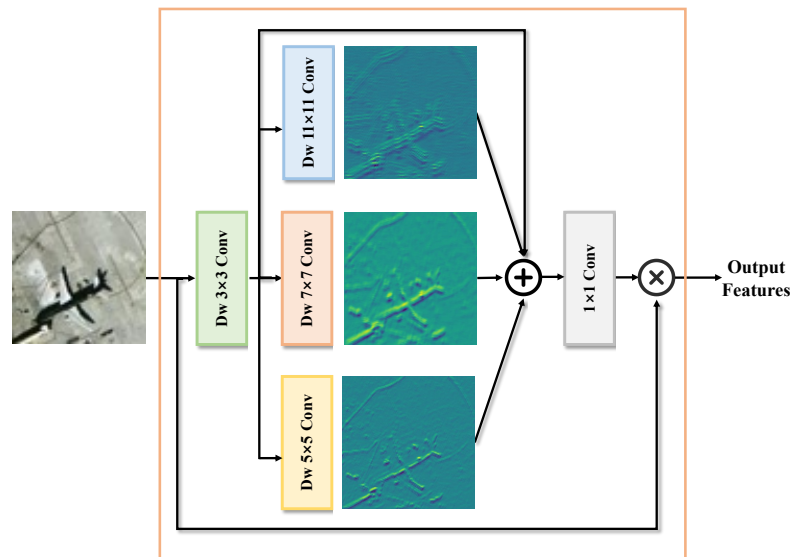


Figure 4. Illustration of Multi-scale Depth-wise Convolution Attention (MDCA). We extract multi-scale features using multi-branch convolutions and then utilize them as attention weights to reweight the input of MDCA.

Multi-scale convolution can effectively capture features at different scales, enhancing the model's sensitivity to multi-scale information. As illustrated in Figure 4, each unique convolution kernel of varying size exhibits an adaptive focus on features across different scales, aligning with expectations. Notably, smaller convolution kernels (i.e., 5×5 and 7×7) excel in capturing local and detailed image information, such as textures and lines, essential for tasks like image super-resolution. Conversely, the larger convolution kernel, with its wider receptive field, comprehensively grasps the global structure and layout of the image, providing valuable contextual information. The amalgamation of convolution kernels across diverse scales facilitates the capture and analysis of information at various scales, leading to a more thorough understanding of the image content.

3.4. Tracing-Back Structure

By analyzing the behavior of MDCA and DWSA, we observed that they have different preferences in terms of feature representation, aligning with findings in [65]. This suggests that both attention modules have untapped potential. Specifically, the differences (shown in Figure 5c) between the input and output feature maps of MDCA (shown in Figure 5a and Figure 5b, respectively) are higher in the flat region than in the edge part. It shows the proficiency of convolution in extracting local features and texture features. This ability

of MDCA makes it ideal for preserving high-frequency information, which is critical for image sharpness and fine detail. However, an excessive focus on detail may inadvertently neglect the broader context. In contrast, as depicted in Figure 5g, the difference feature map of DWSA is more reflected in the edges and details, indicating that transformers are not good at learning local edges and textures due to the self-attention mechanism. They excel at understanding complex spatial relationships and contextual information in the data, ensuring consistency in super-resolution output across a wider range of scenes, but may inadvertently inhibit the granularity and high-frequency detail that is critical in super-resolution environments.

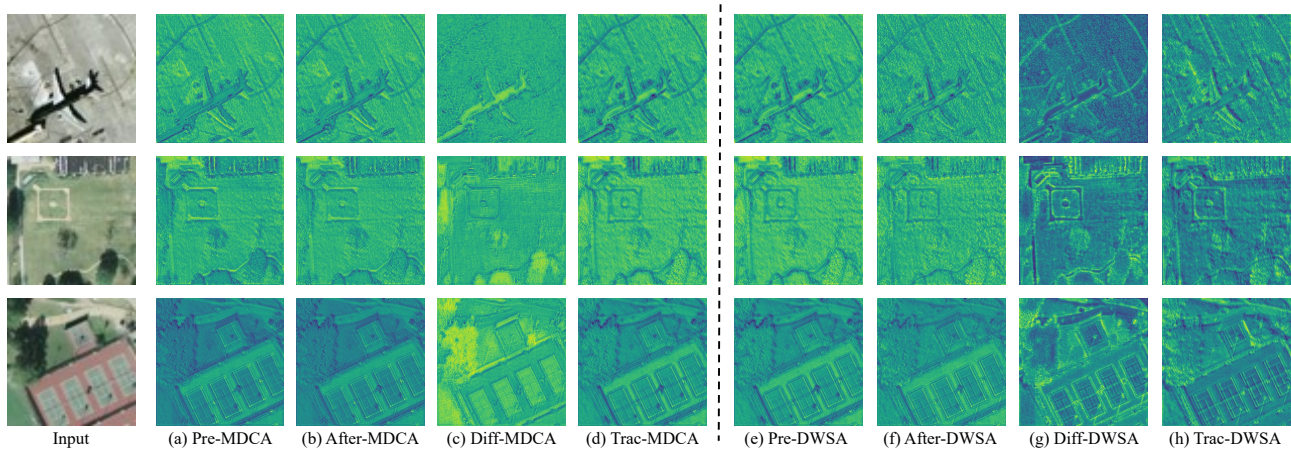


Figure 5. Visualization of feature maps in Tracing-Back Structure. (a) Input feature map of MDCA module. (b) Output feature map of MDCA module. (c) Difference between input and output feature maps of MDCA module. (d) Feature map of MDCA module after Tracing-Back Structure modulation. (e) Input feature map of DWSA module. (f) Output feature map of DWSA module. (g) Difference between input and output feature maps of DWSA module. (h) Feature map of DWSA module after Tracing-Back Structure modulation.

Based on the above analysis, we designed a novel Tracing-Back Structure (TBS) to reconstruct the information flow of the block by accommodating the distinct preferences of MDCA and DWSA modules in feature extraction to enhance the overall feature representation and improve model performance, as shown in Figure 2b,c. Specifically, the TBS first takes the previously calculated features as input and passes them through the attention module to obtain corresponding outputs. Then, the difference between the output and input of the attention module is calculated as follows:

$$F_{dif} = F_{out} - F_{in}, \quad (8)$$

where $F_{in} \in \mathbb{R}^{H \times W \times C}$ and $F_{out} \in \mathbb{R}^{H \times W \times C}$ are the input and output feature maps of the attention module, respectively. $F_{dif} \in \mathbb{R}^{H \times W \times C}$ represents the difference between the input feature map and the output feature map.

Subsequently, modulating this difference and integrating it with the output of the attention module enriches the feature representation, which can be expressed as

$$F_{mod} = Mod(F_{dif}), \quad (9)$$

$$F_{trac} = F_{mod} + F_{out}, \quad (10)$$

where F_{mod} represents the modulated difference features. *Mod* refers to the customized modulation layer, such as the 3×3 convolution layer designed for the MDCA module and the window-based self-attention layer designed for the DWSA module. F_{trac} indicates the output features after the TBS.

As shown in Figure 5d, the MDCA module achieved a greater response in flat areas such as aircraft fuselage and lawns with TBS modulation while still maintaining edge and detail advantages. Similarly, the DWSA module is modulated to achieve greater brightness on the edges of aircraft wings and buildings without losing performance in flat areas, as illustrated in Figure 5h. This demonstrates that the TBS can recover the information lost by the MDCA and DWSA modules due to their respective preferences, thus obtaining a more comprehensive feature representation.

4. Results

4.1. Experimental Setup

4.1.1. Datasets

This study focuses on remote sensing image super-resolution across three magnification factors, $\times 2$, $\times 3$, and $\times 4$, utilizing three publicly accessible remote sensing image datasets: UCMerced [66], RSSCN7 [67], and AID [32].

The UCMerced dataset comprises 21 remote sensing scene classes, each of which contains 100 images, and the size of each image is 256×256 . Following the procedures outlined in [23,68], we partition the UCMerced dataset into two parts for training and testing, with the training set containing 945 remote sensing images and the remaining 1050 for testing. The RSSCN7 dataset contains seven different categories of remote sensing images extracted from Google Earth, with a resolution of 400×400 pixels and a total of 2800 images. We also divided the RSSCN7 dataset into two parts for training and testing, each containing half of the sample, with 20% of the training images reserved for validation. The AID dataset encompasses 10,000 images representing 30 classes of remote sensing scenes, each with a size of 600×600 . Following the methodology articulated in [23], we randomly assigned 80% of the total dataset to training, while the remaining images were designated for testing.

4.1.2. Metrics

We employed the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [69], spatial correlation coefficient (SCC) [70], and spectral angle mapper (SAM) [71] as the evaluation metrics for the RSISR task, and the evaluation of super-resolution results was performed on the RGB channels as follows:

$$PSNR(x, y) = 10 \log_{10} \left(\frac{I_{MAX}^2}{MSE(x, y)} \right), \quad (11)$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2cov(x, y) + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (12)$$

$$SCC(x, y) = \frac{cov(x, y)}{\sigma_x\sigma_y}, \quad (13)$$

$$SAM(x, y) = \arccos \left(\frac{y^T x}{\|x\| \|y\|} \right), \quad (14)$$

where I_{MAX} represents the maximum value in the image, and $MSE(x, y)$ denotes the mean square error between two images, x and y . μ_x and μ_y represent the mean values of images x and y , while σ_x and σ_y represent their variances. The covariance between the two images is denoted by $cov(x, y)$. Constants C_1 and C_2 are introduced to ensure that the denominator is not zero. Higher PSNR, SSIM, SCC, and lower SAM values correspond to superior image reconstruction quality. Furthermore, we also assessed the efficiency of our method by considering key factors, such as the number of parameters and floating-point operations per second (FLOPs).

4.1.3. Implementation Details

We employed the Adam optimizer [72] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\varepsilon = 10^{-8}$. The initial learning rate was set to 4×10^{-4} , with a gradual reduction to 1×10^{-7} at epoch 2000 following the cosine annealing schedule [73]. The training progress employed a mini-batch size of 8 and was conducted on the RTX 4090.

4.1.4. Hyperparameter Details

The MSGFormer model, as introduced, is composed of 10 blocks with an embedding dimension set to 64. In the DWCA module, the convolutional kernel sizes for its three branches are specified as 5, 7, and 11. Furthermore, within the DWSA module of MSGFormer, the size of both the concentrated window and distributed window is configured to be 16, while the number of attention heads is designated as 2.

4.2. Quantitative and Qualitative Comparisons

To assess the efficacy of our proposed approach, we systematically compared it against prominent methods used for natural image super-resolution, such as VDSR [35], SRDD [10], and OmniSR [30]. Additionally, we compared our method with several RSISR techniques, namely, DCM [38], MHAN [40], HSENet [68], TransENet [23], and FENet [41]. For a fair evaluation, we retrained all of the comparative methods using publicly available code provided in the above article on the same dataset.

4.2.1. Quantitative Results

The quantitative results, encompassing various datasets and zooming factors, are detailed in Table 1. As indicated by the summarized statistical outcomes in Table 1, MSGFormer achieves the highest PSNR, SSIM, and SCC values and the lowest SAM values, outperforming state-of-the-art methods across all tested datasets and scaling factors. Specifically, MSGFormer attains the highest PSNR and SSIM values, demonstrating an advantage ranging from +0.55 dB/+0.0126 to +0.11 dB/+0.0034 over the second-ranked HSENet [68] and OmniSR [30]. In comparison with HSENet, which emphasizes mixed-scale feature correlations, and OmniSR, which models pixel interactions from the spatial and channel dimensions, MSGFormer not only effectively perceives multi-scale information but also boasts a broader range of receptive fields, enhancing overall image perception. Notably, larger models like MHAN [40] and TransENet [23] exhibit improved performance when trained on the AID dataset with a larger spatial size but still fall short of outperforming MSGFormer.

To delve more deeply into the proposed method, we comprehensively discuss the quantitative performance across different classes within the same dataset. We chose the four best-performing methods for intuitive comparison. Utilizing the AID dataset, which contains a diverse range of remote sensing images with relatively large image sizes, we analyzed remote sensing scenes to demonstrate our approach. Table 2 presents the PSNR values for individual classes in the AID dataset with a $\times 4$ upscaling factor. According to Table 2, MSGFormer achieves the highest PSNR value in all categories. Notably, in comparison to TransENet [23], which achieves suboptimal performance in most scenarios due to its full self-attention, MSGFormer, equipped with a global receptive field, multi-scale information perception, and feature enhancement, attains superior performance across all scenarios. Specifically, MSGFormer performs well in various scenarios, including “playground” (+0.37 dB), “baseball field” (+0.32 dB), “bridge” (+0.23 dB), and “parking” (+0.23 dB), emphasizing the efficient use of the global receptive field and multi-scale information, further confirming the effectiveness of our approach. In addition, Table 3 presents a comparative analysis of model parameters, FLOPs, and performance across these methods on the AID dataset with a magnification factor of $\times 4$. It is evident that our MSGFormer achieves the highest PSNR while maintaining fewer parameters and FLOPs. This demonstrates that MSGFormer strikes a balance between model performance and size.

Table 1. The mean PSNR (dB) and SSIM on the UCMerced dataset, RSSCN7 dataset, and AID dataset. The best and second-best results are highlighted in red and blue, respectively.

Method	Scale	UCMerced				RSSCN7				AID			
		PSNR	SSIM	SCC	SAM	PSNR	SSIM	SCC	SAM	PSNR	SSIM	SCC	SAM
VDSR [35]	×2	33.87	0.9280	0.6196	0.0519	30.04	0.8027	0.2967	0.1018	35.11	0.9340	0.6181	0.0544
DCM [38]	×2	33.65	0.9274	0.6291	0.0507	30.03	0.8024	0.2979	0.1019	35.35	0.9366	0.6407	0.0531
MHAN [40]	×2	33.92	0.9283	0.6242	0.0518	30.06	0.8036	0.2996	0.1016	35.56	0.9390	0.6641	0.0520
HSENet [68]	×2	34.22	0.9327	0.6341	0.0500	30.15	0.8070	0.3056	0.1006	35.50	0.9383	0.6626	0.0524
TransENet [23]	×2	34.05	0.9294	0.6275	0.0511	30.08	0.8040	0.2984	0.1013	35.40	0.9372	0.6538	0.0530
SRDD [10]	×2	34.12	0.9303	0.6331	0.0507	30.05	0.8051	0.2993	0.1014	35.33	0.9367	0.6373	0.0531
FENet [41]	×2	33.95	0.9284	0.6243	0.0518	30.05	0.8033	0.2991	0.1016	35.33	0.9364	0.6390	0.0533
OmniSR [30]	×2	34.16	0.9303	0.6326	0.0506	30.11	0.8052	0.3023	0.1010	35.50	0.9383	0.6523	0.0522
MSGFormer (Ours)	×2	34.77	0.9361	0.6560	0.0471	30.30	0.8112	0.3189	0.0993	35.78	0.9411	0.6764	0.0508
VDSR [35]	×3	29.75	0.8346	0.3941	0.0829	27.94	0.7010	0.1495	0.1303	31.17	0.8511	0.3800	0.0836
DCM [38]	×3	29.86	0.8393	0.4025	0.0820	27.96	0.7027	0.1524	0.1301	31.31	0.8561	0.3946	0.0822
MHAN [40]	×3	29.94	0.8391	0.4304	0.0816	28.00	0.7045	0.1551	0.1296	31.55	0.8603	0.4098	0.0801
HSENet [68]	×3	30.04	0.8433	0.4131	0.0806	28.02	0.7067	0.1572	0.1292	31.49	0.8588	0.4053	0.0806
TransENet [23]	×3	29.90	0.8397	0.3988	0.0816	28.02	0.7054	0.1532	0.1292	31.50	0.8588	0.4067	0.0806
SRDD [10]	×3	29.92	0.8411	0.4084	0.0815	27.96	0.7052	0.1552	0.1300	31.38	0.8564	0.3984	0.0817
FENet [41]	×3	29.80	0.8379	0.3941	0.0826	27.97	0.7031	0.1524	0.1300	31.33	0.8550	0.3955	0.0823
OmniSR [30]	×3	29.99	0.8403	0.4073	0.0810	28.04	0.7061	0.1584	0.1290	31.53	0.8596	0.4081	0.0803
MSGFormer (Ours)	×3	30.49	0.8506	0.4370	0.0770	28.20	0.7142	0.1696	0.1267	31.75	0.8646	0.4208	0.0783
VDSR [35]	×4	27.54	0.7522	0.2589	0.1055	26.75	0.6336	0.0825	0.1495	28.99	0.7753	0.2427	0.1055
DCM [38]	×4	27.60	0.7556	0.2610	0.1051	26.79	0.6363	0.0867	0.1490	29.20	0.7826	0.2679	0.1032
MHAN [40]	×4	27.63	0.7581	0.2649	0.1043	26.79	0.6360	0.0850	0.1491	29.39	0.7892	0.2825	0.1008
HSENet [68]	×4	27.75	0.7611	0.2692	0.1034	26.82	0.6378	0.0867	0.1485	29.32	0.7867	0.2765	0.1017
TransENet [23]	×4	27.78	0.7635	0.2701	0.1029	26.81	0.6373	0.0845	0.1485	29.44	0.7912	0.2884	0.1002
SRDD [10]	×4	27.67	0.7609	0.2718	0.1047	26.74	0.6364	0.0842	0.1495	29.21	0.7835	0.2695	0.1030
FENet [41]	×4	27.59	0.7538	0.2568	0.1053	26.80	0.6367	0.0871	0.1487	29.16	0.7812	0.2651	0.1037
OmniSR [30]	×4	27.80	0.7637	0.2779	0.1027	26.85	0.6388	0.0898	0.1480	29.19	0.7829	0.2636	0.1033
MSGFormer (Ours)	×4	28.16	0.7763	0.3029	0.0988	26.96	0.6447	0.0957	0.1467	29.59	0.7960	0.3024	0.0987

Table 2. The mean PSNR (dB) of each class for an upscaling factor of 4 on the AID test dataset. The best and second-best results are highlighted in red and blue, respectively.

Class Name	MHAN	HSENet	TransENet	SRDD	MSGFormer (Ours)
airport	29.20	29.12	29.26	29.05	29.39
bare land	36.36	36.34	36.38	36.34	36.50
baseball field	31.57	31.49	31.63	31.62	31.95
beach	32.62	32.60	32.66	32.62	32.80
bridge	31.65	31.55	31.70	31.54	31.93
center	28.03	27.91	28.09	27.84	28.27
church	24.50	24.43	24.53	24.36	24.65
commercial	27.96	27.90	28.00	27.86	28.12
dense residential	25.08	25.02	25.17	24.92	25.26
desert	39.50	39.47	39.55	39.40	39.60
farmland	34.65	34.59	34.67	34.58	34.84
forest	28.52	28.54	28.59	28.49	28.62
industrial	27.18	27.09	27.24	26.99	27.38
meadow	33.00	32.97	33.00	33.05	33.16
medium residential	28.48	28.41	28.50	28.29	28.60
mountain	29.24	29.22	29.30	29.23	29.34
park	27.97	27.93	28.04	27.96	28.20
parking	26.35	26.16	26.49	25.86	26.72
playground	30.34	30.19	30.38	30.21	30.75
pond	30.53	30.48	30.58	30.69	30.89
port	26.90	26.80	26.95	26.66	27.05

Table 2. Cont.

Class Name	MHAN	HSENet	TransENet	SRDD	MSGFormer (Ours)
railway station	28.59	28.52	28.64	28.42	28.80
resort	28.07	28.00	28.13	27.86	28.17
river	31.00	30.97	31.04	30.99	31.13
school	27.18	27.10	27.25	27.04	27.40
sparse residential	26.62	26.60	26.63	26.55	26.69
square	29.38	29.30	29.46	29.23	29.63
stadium	27.40	27.28	27.48	27.16	27.62
storage tanks	26.20	26.12	26.22	26.03	26.31
viaduct	28.19	28.09	28.24	28.06	28.42
avg	29.39	29.32	29.44	29.28	29.59

Table 3. A quantitative comparison of parameters and flops on the AID dataset for $\times 4$ SR.

Method	Params	FLOPs	PSNR
VDSR [35]	671 K	241.87 G	28.99 dB
DCM [38]	2175 K	71.40 G	29.20 dB
MHAN [40]	11,351 K	278.09 G	29.39 dB
HSENet [68]	5430 K	105.49 G	29.32 dB
TransENet [23]	37,460 K	108.52 G	29.44 dB
SRDD [10]	9337 K	32.73 G	29.21 dB
FENet [41]	351 K	7.54 G	29.16 dB
OmniSR [30]	793 K	17.77 G	29.19 dB
MSGFormer (Ours)	2144 K	61.91 G	29.59 dB

4.2.2. Qualitative Results

In this section, we present qualitative visualization results of different methods to further analyze the model. Figure 6 displays diverse examples of remote sensing image super-resolution results obtained from both the UCMerced and AID datasets, including scenes such as “storage tanks”, “tennis court”, “agricultural”, “center”, “playground”, and “stadium”. Compared to the other methods, the results generated by MSGFormer are very close to real HR images, with clear details. For instance, in scenes demonstrating strong global consistency like “agriculture” and “playground”, only MSGFormer adequately recovers texture information. TransENet [23] and OmniSR [30] capture some details but exhibit distortion, as depicted in Figure 6. Similarly, in scenes with scale changes, such as “center” and “tennis court”, MSGFormer performs effectively, approaching the ground truth. Notably, SRDD [10] achieves better clarity in “tennis court”, and OmniSR [30] appears more realistic in “center”, but both fall short compared to the ground truth. In summary, the proposed MSGFormer visually outperforms other methods.

For a comprehensive understanding of the operational mechanism of the model, we conducted a qualitative comparison of LAM [16] and SR results across various networks, utilizing a magnification factor of $\times 4$ on the UCMerced test dataset. LAM elucidates the significance of each pixel in the input LR image during the reconstruction of a marked patch. As depicted in Figure 7, the red-marked points represent information pixels contributing to the reconstruction. The results indicate that when compared to the CNN-based model FENet [41] and the window-based self-attention model OmniSR [30], MSGFormer exhibits a broader perceived information scope, approaching the perceptual range of the full-attention model TransENet [23]. In contrast to the full self-attention TransENet, MSGFormer not only maintains a similar participating receptive field but also integrates multi-scale information, demonstrating superior performance. Overall, our proposed method consistently delivers competitive results.

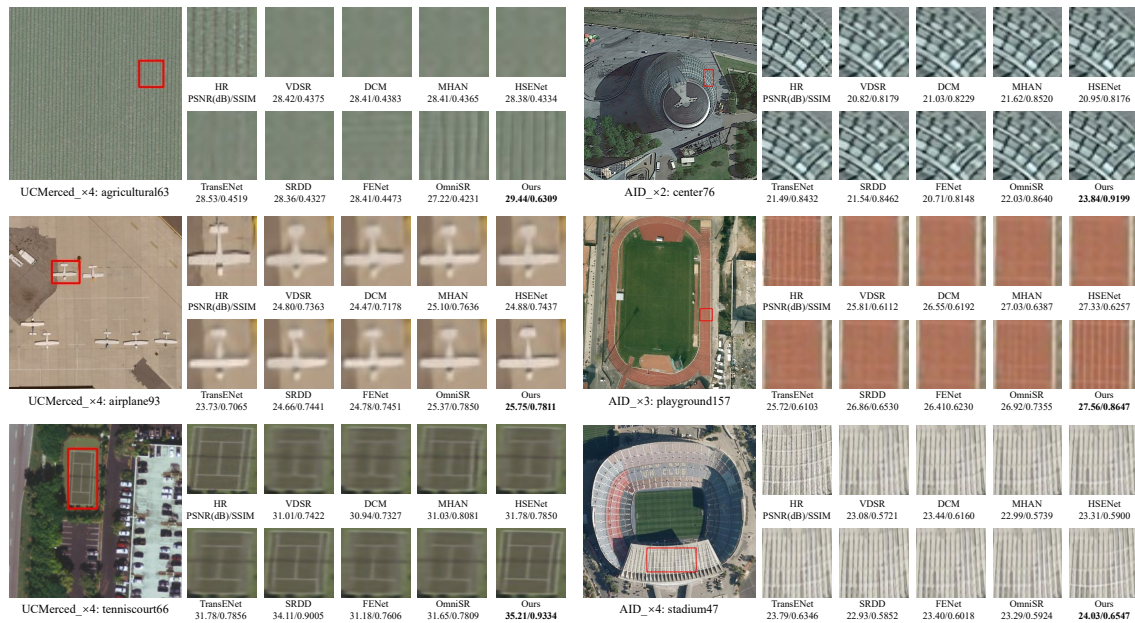


Figure 6. A visual comparison on the UCMerced dataset and AID dataset. The patches for comparison are marked with red boxes in the original images. The PSNR/SSIM is calculated based on the patches to better reflect the performance difference. It is evident that MSGFormer has achieved the most outstanding visual results across various remote sensing scenarios.

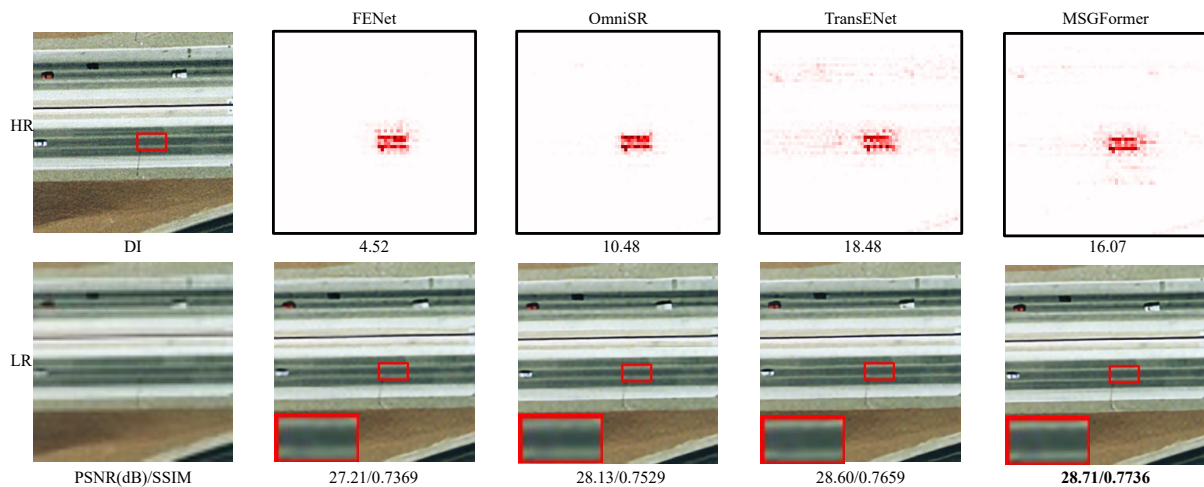


Figure 7. LAM [16] results for different methods. The diffusion index (DI) reflects the range of involved pixels. A higher DI represents a wider range of utilized pixels. The results indicate that MSGFormer utilizes more information than FENet [41] and Omnisr [30] and less information than TransENet [23].

4.3. Ablation Studies

In this section, we rigorously validated the effectiveness of the proposed MSGFormer. Specifically, we conducted an ablation study on MSGFormer to examine the individual contributions of each component. All experiments adhered to a consistent setup, utilizing the UCMerced dataset with a uniform magnification factor of 4. The notable enhancement of MSGFormer is attributed to the integration of DWSA, MDCA, and the TBS. Therefore, we began with a complete model and assessed its impact by selectively removing the corresponding modules. The effects of these proposed components on the model's performance are outlined in Table 4, which can be summarized as follows.

Table 4. Ablation experiments of component analysis by MSGFormer on the UCMerced dataset for $\times 4$ SR.

MDCA	DWSA	TBS	Params	FLOPs	PSNR	SSIM
	✓	✓	1.34 M	9.04 G	28.104 dB	0.7737
✓		✓	2.08 M	11.33 G	28.092 dB	0.7743
✓	✓		1.38 M	8.32 G	28.128 dB	0.7751
✓	✓	✓	2.14 M	12.33 G	28.155 dB	0.7763

4.3.1. Effect of DWSA

To validate the efficacy of DWSA in broadening the involved acceptance field and enhancing the performance of RSISR, we substituted DWSA within the model with self-attention that exclusively exchanges information within a local window for comparative analysis. The outcomes are presented in the second row of Table 4. The model incorporating DWSA demonstrates notable enhancements in the PSNR and SSIM, suggesting that the incorporation of a global receptive field contributes to the enhancement of reconstruction quality. In addition, we further studied the impact of window size on model performance in DWSA, as shown in Table 5. The results indicate that using a larger window enables the model to gather more comprehensive information, leading to improved performance.

To gain a deeper insight into the performance enhancements achieved by DWSA, we employed LAM [16]. As depicted in Figure 8, it is evident that the model employing DWSA encompasses a broader range of utilized pixels, yielding superior reconstruction results. Additionally, the result of our method with DWSA achieves a higher DI, signifying its utilization of the most input pixels and resulting in elevated PSNR and SSIM values. Consequently, the model integrating DWSA expands its coverage over the utilized pixels, benefiting from a more extensive array of pertinent global pixels, further proving the effectiveness of the proposed DWSA.

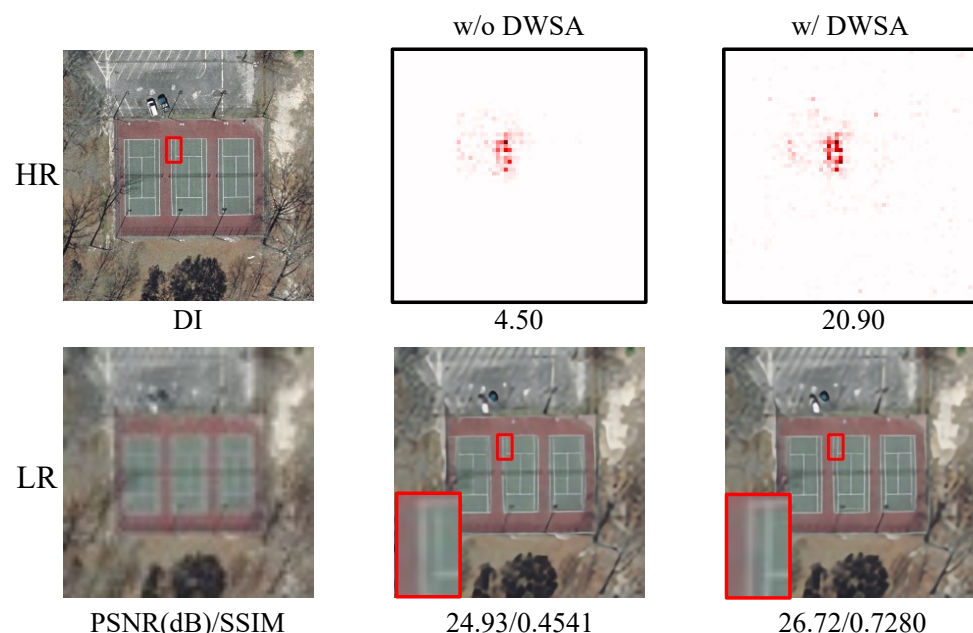
**Figure 8.** LAM result for ablation study on the proposed DWSA. It indicates that the model with DWSA has a larger scope of utilized pixels and generates better reconstruction results.

Table 5. Ablation experiments on the design of DWSA on the UCMerced dataset for $\times 4$ SR.

Window Size	FLOPs	PSNR	SSIM	SCC	SAM
4	12.22 G	28.034 dB	0.7737	0.2968	0.1001
8	12.25 G	28.079 dB	0.7743	0.2990	0.0996
16	12.33 G	28.155 dB	0.7763	0.3029	0.0988

4.3.2. Effect of MDCA

To substantiate the effectiveness of MDCA, we devised a model variant by excluding MDCA from all blocks. The outcomes are presented in the initial row of Table 4. By integrating multi-scale information, the performance of the model is improved by 0.051 dB. Additionally, for a more comprehensive understanding of the performance enhancement achieved by MDCA, we conducted an in-depth analysis of the distinct contributions made by each component within MDCA, as shown in Table 6. $K \times K$ stands for a depth-wise $K \times K$ convolution. Our findings reveal that each component significantly contributes to the overall performance. The integration of multi-scale convolution operations into the model enhances its ability to perceive information across scales, resulting in improved overall performance.

Table 6. Ablation experiments on the design of MDCA on the UCMerced dataset for $\times 4$ SR.

5×5	7×7	11×11	Params	FLOPs	PSNR	SSIM
	✓	✓	2.13 M	12.26 G	28.150 dB	0.7762
✓		✓	2.11 M	12.20 G	28.131 dB	0.7754
✓	✓		2.07 M	12.01 G	28.140 dB	0.7760
✓	✓	✓	2.14 M	12.33 G	28.155 dB	0.7763

4.3.3. Effect of TBS

The primary motivation behind the design of the TBS is to further exploit the potential of MDCA and DWSA by analyzing their behaviors, aiming to achieve a more comprehensive feature representation. To demonstrate the effectiveness of the TBS, we excluded it from all blocks, with results reported in the third row of Table 4. In contrast to the case where MDCA and DWSA information flows directly in series, the TBS utilizes an information tracing-back mechanism to retrieve the features that are ignored due to the inherent behaviors of both attention modules, significantly improving PSNR and SSIM values, providing compelling evidence for the efficacy of the TBS. In addition, we further investigated the impact of feature backtracking for each module on model performance, as shown in Table 7. The results show that the feature backtracking of MDCA and DWSA modules can improve the feature representation ability of the model.

Table 7. Ablation experiments on the design of the TBS on the UCMerced dataset for $\times 4$ SR.

TBS-MDCA	TBS-DWSA	PSNR	SSIM	SCC	SAM
✓		28.082 dB	0.7738	0.2978	0.0995
	✓	28.062 dB	0.7728	0.2975	0.0995
✓	✓	28.155 dB	0.7763	0.3029	0.0988

5. Discussion

In this section, we commence with a detailed analysis of the results within the framework of prior studies and working hypotheses, aiming to underscore the significance of the outcomes. Subsequently, we delve into the strengths of this study, followed by an examination of the limitations and sources of error associated with the proposed method. Finally, we identify potential avenues for future research to guide the trajectory of further investigation.

Prior CNN-based RSISR approaches have been limited by the local processing principle of convolutional kernels, which has prevented direct interactions between distant pixels. This can be clearly seen in Figure 7, where the receptive fields of the CNN-based FENet [41] and OmniSR [30] are primarily focused on the local region. As a result, this limitation has led to reduced image quality, with images exhibiting blur and artifacts, as demonstrated in Figure 6. In contrast, transformer-based methods provide a broad receptive field, primarily due to the effectiveness of the self-attention mechanism. As depicted in Figure 7, the transformer-based TransENet [23] demonstrates a global receptive field. Nevertheless, the significant expansion of their receptive field has resulted in quadratic computing costs, presenting difficulties for remote sensing images that encompass wide spatial coverage, as illustrated in Table 3.

By incorporating Dual Window-based Self-Attention (DWSA) to facilitate the effective interaction of local and global information, MSGFormer is capable of achieving an expansive receptive field similar to transformers while upholding linear computational complexity. As depicted in Figure 7 and Table 3, MSGFormer showcases a receptive field that closely resembles that of the transformer-based approach while maintaining a relatively low computational burden. Moreover, we have developed Multi-scale Depth-wise Convolution Attention (MDCA) to overcome the constraint of fixed window size in capturing multi-scale information, which is a limitation in window-based transformer models, as depicted in Figure 4. Furthermore, we propose a new Tracing-Back Structure for each MDCA and DWSA module to fully exploit the potential of its feature representation, as shown in Figure 5. Through these enhancements, MSGFormer demonstrates superior performance compared to other methods, as demonstrated in Tables 1 and 3.

Despite the promising performance highlighted above, there are still limitations to the proposed method in certain complex scenarios. A typical instance of failure is illustrated in Figure 9, where our method has difficulties in recovering small targets in a homogeneous background. While the hand-crafted sparse dual attention patterns prove effective, they are inherently data-agnostic and may not be optimal. It is possible that relevant keys and values are overlooked while less significant ones are retained. In future research, we intend to investigate the design of more adaptive candidate keys and value sets to cater to individual inputs, thereby addressing the issue of sparse, hand-crafted attention patterns. It is also important to note that MSGFormer is currently tailored for satellite optical remote sensing images and has not been optimized for other types of remote sensing data that necessitate the modeling of channel-to-channel relationships, such as hyperspectral images or synthetic aperture radar (SAR) images. Moving forward, we aspire to adapt our model to diverse applications by leveraging the unique characteristics of various data.

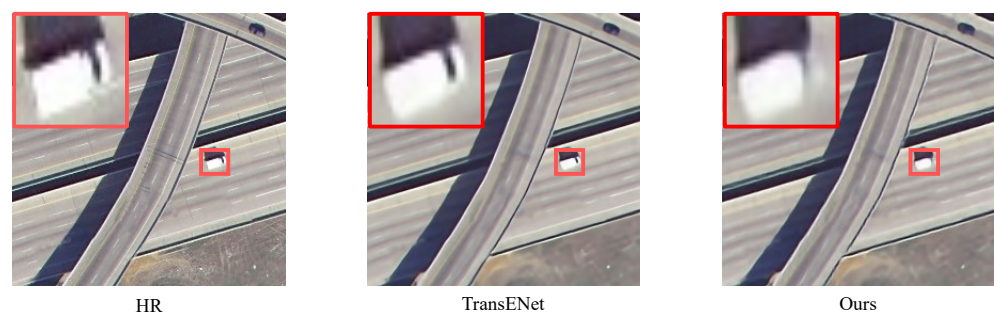


Figure 9. A failure case on the UCMerced dataset. Due to the lack of background information available for small objects in this figure, it is difficult for our framework to achieve a good reconstruction effect.

6. Conclusions

This study introduces MSGFormer, an efficient transformer model for super-resolving remote sensing images, which achieves a desirable balance between model performance and size. MSGFormer aims to fully use the global and multi-scale information of remote sensing images and fully exploit the feature representation capability of the model. MSGFormer

consists of three core parts: Dual Window-based Self-Attention (DWSA), Multi-scale Depth-wise Convolution Attention (MDCA), and a Tracing-Back Structure (TBS). DWSA consists of two parallel attention modules: distributed attention and concentrated attention. This design aims to expand the perceptual range of the model with minimal computational burden by considering both local and global aspects. Additionally, MDCA in MSGFormer introduces multiple convolution kernels of varying scales to assist transformers with a fixed window size in perceiving multi-scale features. To further exploit the potential of MDCA and DWSA, we introduce the TBS based on the distinct behavior of both attention modules to provide powerful feature representation. Ablation studies confirm the effectiveness of the proposed components. Experimental results on three public datasets demonstrate that our method outperforms the state of the art, yielding superior super-resolved results.

Author Contributions: Conceptualization, Y.L. and S.W.; methodology, Y.L. and S.W.; software, Y.L.; validation, Y.L.; formal analysis, Y.L. and S.W.; investigation, Y.L. and S.W.; resources, Y.L. and S.W.; data curation, Y.L.; writing—original draft preparation, Y.L.; writing—review and editing, S.W. and X.Z.; visualization, Y.L.; supervision, B.W. and X.W.; project administration, Y.Z.; funding acquisition, B.W. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Postdoctoral Science Foundation of China (2022M710393, 2022TQ0035), the Shaanxi Science Fund for Distinguished Young Scholars (2022JC-49) and the Basic and Applied Basic Research Foundation of Guangdong Province (2024A1515012388).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Liu, K.; Zhang, B.; Lu, J.; Yan, H. Towards Integrity and Detail with Ensemble Learning for Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5606813. [\[CrossRef\]](#)
3. Sambandham, V.T.; Kirchheim, K.; Ortmeier, F.; Mukhopadhyaya, S. Deep learning-based harmonization and super-resolution of Landsat-8 and Sentinel-2 images. *ISPRS J. Photogramm. Remote Sens.* **2024**, *212*, 274–288. [\[CrossRef\]](#)
4. Wang, Y.; Yuan, W.; Xie, F.; Lin, B. ESatSR: Enhancing Super-Resolution for Satellite Remote Sensing Images with State Space Model and Spatial Context. *Remote Sens.* **2024**, *16*, 1956. [\[CrossRef\]](#)
5. Wu, J.; Xia, L.; Chan, T.O.; Awange, J.; Yuan, P.; Zhong, B.; Li, Q. A novel fusion framework embedded with zero-shot super-resolution and multivariate autoregression for precipitable water vapor across the continental Europe. *Remote Sens. Environ.* **2023**, *297*, 113783. [\[CrossRef\]](#)
6. Wang, J.; Lu, Y.; Wang, S.; Wang, B.; Wang, X.; Long, T. Two-stage Spatial-Frequency Joint Learning for Large-Factor Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5606813. [\[CrossRef\]](#)
7. Zheng, Q.; Tian, X.; Yu, Z.; Ding, Y.; Elhanashi, A.; Saponara, S.; Kpalma, K. MobileRaT: A Lightweight Radio Transformer Method for Automatic Modulation Classification in Drone Communication Systems. *Drones* **2023**, *7*, 596. [\[CrossRef\]](#)
8. Zhao, Y.; Yin, Y.; Gui, G. Lightweight deep learning based intelligent edge surveillance techniques. *IEEE Trans. Cogn. Commun.* **2020**, *6*, 1146–1154. [\[CrossRef\]](#)
9. Zheng, Q.; Tian, X.; Yang, M.; Wu, Y.; Su, H. PAC-Bayesian framework based drop-path method for 2D discriminative convolutional network pruning. *Multidimens. Syst. Signal Process.* **2020**, *31*, 793–827. [\[CrossRef\]](#)
10. Maeda, S. Image super-resolution with deep dictionary. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 464–480.
11. Ran, R.; Deng, L.J.; Jiang, T.X.; Hu, J.F.; Chanussot, J.; Vivone, G. GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution. *IEEE Trans. Cybern.* **2023**, *53*, 4148–4161. [\[CrossRef\]](#)
12. Wang, J.; Wang, B.; Wang, X.; Zhao, Y.; Long, T. Hybrid attention based u-shaped network for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5612515. [\[CrossRef\]](#)
13. Hu, W.; Ju, L.; Du, Y.; Li, Y. A Super-Resolution Reconstruction Model for Remote Sensing Image Based on Generative Adversarial Networks. *Remote Sens.* **2024**, *16*, 1460. [\[CrossRef\]](#)
14. Yao, S.; Cheng, Y.; Yang, F.; Mozerov, M.G. A continuous digital elevation representation model for DEM super-resolution. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 1–13. [\[CrossRef\]](#)
15. Mardieva, S.; Ahmad, S.; Umirzakova, S.; Rasool, M.A.; Whangbo, T.K. Lightweight image super-resolution for IoT devices using deep residual feature distillation network. *Knowl.-Based Syst.* **2024**, *285*, 111343. [\[CrossRef\]](#)

16. Gu, J.; Dong, C. Interpreting super-resolution networks with local attribution maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9199–9208.
17. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 22367–22377.
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
19. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
21. Ding, M.; Xiao, B.; Codella, N.; Luo, P.; Wang, J.; Yuan, L. Davit: Dual attention vision transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 74–92.
22. Chen, Q.; Wu, Q.; Wang, J.; Hu, Q.; Hu, T.; Ding, E.; Cheng, J.; Wang, J. Mixformer: Mixing features across windows and dimensions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5249–5259.
23. Lei, S.; Shi, Z.; Mo, W. Transformer-based multistage enhancement for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5615611. [[CrossRef](#)]
24. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739.
25. Chen, Z.; Zhang, Y.; Gu, J.; Zhang, Y.; Kong, L.; Yuan, X. Cross Aggregation Transformer for Image Restoration. In Proceedings of the Advances in neural information processing systems, New Orleans, CA, USA, 28 November–9 December 2022; pp. 25478–25490.
26. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12299–12310.
27. Lu, Z.; Li, J.; Liu, H.; Huang, C.; Zhang, L.; Zeng, T. Transformer for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 457–466.
28. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 1833–1844.
29. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17683–17693.
30. Wang, H.; Chen, X.; Ni, B.; Liu, Y.; Liu, J. Omni Aggregation Networks for Lightweight Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 22378–22387.
31. Choi, H.; Lee, J.; Yang, J. N-gram in swin transformers for efficient lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 2071–2081.
32. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [[CrossRef](#)]
33. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
34. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 391–407.
35. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
36. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
37. Lei, S.; Shi, Z.; Zou, Z. Super-resolution for remote sensing images via local-global combined network. *IEEE Trans. Geosci. Remote Sens. Lett.* **2017**, *14*, 1243–1247. [[CrossRef](#)]
38. Haut, J.M.; Paoletti, M.E.; Fernández-Beltrán, R.; Plaza, J.; Plaza, A.; Li, J. Remote sensing single-image superresolution based on a deep compendium model. *IEEE Trans. Geosci. Remote Sens. Lett.* **2019**, *16*, 1432–1436. [[CrossRef](#)]

39. Dong, X.; Wang, L.; Sun, X.; Jia, X.; Gao, L.; Zhang, B. Remote sensing image super-resolution using second-order multi-scale networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 3473–3485. [\[CrossRef\]](#)
40. Zhang, D.; Shao, J.; Li, X.; Shen, H.T. Remote sensing image super-resolution via mixed high-order attention network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5183–5196. [\[CrossRef\]](#)
41. Wang, Z.; Li, L.; Xue, Y.; Jiang, C.; Wang, J.; Sun, K.; Ma, H. FeNet: Feature enhancement network for lightweight remote-sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622112. [\[CrossRef\]](#)
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Washington, DC, USA, 11–15 November 2017; Volume 30.
43. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12581–12600. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
45. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood attention transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 6185–6194.
46. Zhou, L.; Gong, C.; Liu, Z.; Fu, K. SAL: Selection and attention losses for weakly supervised semantic segmentation. *IEEE Trans. Multimed.* **2020**, *23*, 1035–1048. [\[CrossRef\]](#)
47. Lee, Y.; Kim, J.; Willette, J.; Hwang, S.J. Mpvit: Multi-path vision transformer for dense prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 7287–7296.
48. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2022**, *54*, 200. [\[CrossRef\]](#)
49. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning texture transformer network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5791–5800.
50. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [\[CrossRef\]](#)
51. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.
52. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
53. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
54. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
55. Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S.W.; Anwer, R.M.; Shahbaz Khan, F. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 3–20.
56. Chen, Y.; Dai, X.; Chen, D.; Liu, M.; Dong, X.; Yuan, L.; Liu, Z. Mobile-former: Bridging mobilenet and transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5270–5279.
57. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, CA, USA, 28 November–9 December 2022; pp. 12934–12949.
58. Fang, J.; Lin, H.; Chen, X.; Zeng, K. A hybrid network of cnn and transformer for lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1103–1112.
59. Gao, G.; Xu, Z.; Li, J.; Yang, J.; Zeng, T.; Qi, G.J. Ctnet: A cnn-transformer cooperation network for face image super-resolution. *IEEE Trans. Image Process.* **2023**, *32*, 1978–1991. [\[CrossRef\]](#)
60. Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; Yu, F. Dual Aggregation Transformer for Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 12312–12321.
61. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
62. Liu, J.; Chen, C.; Tang, J.; Wu, G. From coarse to fine: Hierarchical pixel integration for lightweight image super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 1666–1674.

63. Guo, M.H.; Lu, C.Z.; Liu, Z.N.; Cheng, M.M.; Hu, S.M. Visual attention network. *Comput. Vis. Media* **2023**, *9*, 733–752. [[CrossRef](#)]
64. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
65. Park, N.; Kim, S. How do vision transformers work? *arXiv* **2022**, arXiv:2202.06709.
66. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
67. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [[CrossRef](#)]
68. Lei, S.; Shi, Z. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5401410. [[CrossRef](#)]
69. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
70. Zhou, J.; Civco, D.L.; Silander, J.A. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. *Int. J. Remote Sens.* **1998**, *19*, 743–757. [[CrossRef](#)]
71. Yuhas, R.H.; Goetz, A.F.; Boardman, J.W. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In Proceedings of the Third Annual JPL Airborne Geoscience Workshop, Pasadena, CA, USA, 1–5 June 1992; Volume 1.
72. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
73. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.