*remote sensing*

MDPI

*Article*

# DESAT: A Distance-Enhanced Strip Attention Transformer for Remote Sensing Image Super-Resolution

**Yujie Mao [1,2], Guojin He [1,2,3,4,*], Guizhou Wang [1,2,3,4], Ranyu Yin [1,3,4], Yan Peng [1,3,4] and Bin Guan [5]**

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; maoyujie24@mails.ucas.ac.cn (Y.M.); pengyan@aircas.ac.cn (Y.P.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Kashgar Aerospace Information Research Institute, Kashgar 844000, China
[4] Key Laboratory of Earth Observation of Hainan Province, Aerospace Information Research Institute, Chinese Academy of Sciences, Sanya 572029, China
[5] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China; guanbin@whu.edu.cn
* Correspondence: hegj@aircas.ac.cn; Tel.: +86-010-82178190

**Abstract:** Transformer-based methods have demonstrated impressive performance in image super-resolution tasks. However, when applied to large-scale Earth observation images, the existing transformers encounter two significant challenges: (1) insufficient consideration of spatial correlation between adjacent ground objects; and (2) performance bottlenecks due to the underutilization of the upsample module. To address these issues, we propose a novel distance-enhanced strip attention transformer (DESAT). The DESAT integrates distance priors, easily obtainable from remote sensing images, into the strip window self-attention mechanism to capture spatial correlations more effectively. To further enhance the transfer of deep features into high-resolution outputs, we designed an attention-enhanced upsample block, which combines the pixel shuffle layer with an attention-based upsample branch implemented through the overlapping window self-attention mechanism. Additionally, to better simulate real-world scenarios, we constructed a new cross-sensor super-resolution dataset using Gaofen-6 satellite imagery. Extensive experiments on both simulated and real-world remote sensing datasets demonstrate that the DESAT outperforms state-of-the-art models by up to 1.17 dB along with superior qualitative results. Furthermore, the DESAT achieves more competitive performance in real-world tasks, effectively balancing spatial detail reconstruction and spectral transform, making it highly suitable for practical remote sensing super-resolution applications.

**Keywords:** remote sensing; image super-resolution; deep learning; transformer; self-attention; Gaofen-6 satellite

**Citation:** Mao, Y.; He, G.; Wang, G.; Yin, R.; Peng, Y.; Guan, B. DESAT: A Distance-enhanced Strip Attention Transformer for Remote Sensing Image Super-Resolution. *Remote Sens.* **2024**, *16*, 4251. https://doi.org/10.3390/rs16224251

Academic Editors: Sidike Paheding and Ashraf Saleem

Received: 28 September 2024
Revised: 9 November 2024
Accepted: 13 November 2024
Published: 14 November 2024



## 1. Introduction

High-resolution (HR) remote sensing images are essential for advanced remote sensing applications, such as precision agriculture [1], land cover mapping [2], forest classification [3], photovoltaic panel extraction [4], and ground object detection [5]. However, the spatial and temporal resolutions of Earth observation satellite images often fall short of the demands of these tasks due to limitations in sensor hardware and imaging techniques. Additionally, the high cost of satellite and sensor technology often restricts the availability of HR remote sensing images, raising the cost of remote sensing applications. Therefore, it is crucial to retrieve at most the already acquired information in order to use the observation at its maximal potential. For this purpose, single image super-resolution (SR) has shown to be a very good solution.

As a classical low-level task in computer vision, single-image SR aims to reconstruct HR images from low-resolution (LR) inputs by enhancing spatial details and texture

information through software algorithms. Single-image SR techniques can be broadly categorized into four types: reconstruction-based, interpolation-based, learning-based, and transformer-based algorithms [6]. Early works utilized interpolation [7] or prior information [8,9] to reconstruct spatial details. However, the complexity of image degradation and the introduction of inaccurate prior knowledge often resulted in noticeable artifacts and unsatisfactory texture reconstruction results.

In recent years, significant advancements in deep learning have driven the development of numerous deep-learning-based SR methods, which have demonstrated remarkable performance [10–16]. Initially, many SR methods based on convolutional neural networks (CNNs) [11–15,17,18] have been proposed and dominated this field due to their ability to model local nonlinear features. For example, the very deep residual channel attention network (RCAN) [17] applies a residual-in-residual design and incorporates a channel attention mechanism to achieve impressive SR performance.

However, CNNs are limited in capturing long-range dependencies in images due to their inherent inductive biases. In contrast, transformers [19], with their multi-head self-attention mechanisms, excel at feature representation and modeling long-range relationships, allowing them to achieve state-of-the-art (SOTA) performance in various high-level computer vision tasks [20–22]. Recently, transformer-based SR methods [16,23–26] have shown superior performance compared to CNN-based methods. For example, SwinIR [16] introduced the Swin Transformer architecture [20], achieving remarkable SR performance, while many subsequent studies [24–27] proposed different attention strategies and model structures to improve it and achieve more competitive performance.

In challenging remote sensing super-resolution tasks, some transformer-based remote sensing image super-resolution methods [28–30] have been designed to address scale diversity, a key challenge in remote sensing. However, most existing methods rely on simulated datasets where LR images are generated using simplified degradation models like bicubic downsampling [31]. However, the significant discrepancies between these assumed models and real-world degradation can lead to unsatisfactory SR performance in real-world scenarios. Additionally, for remote sensing scenes with scale diversity, super-resolution models trained on simulated datasets often struggle to effectively enhance the spatial resolution of real satellite images [32].

To address this issue, ground-truth-based remote sensing super-resolution has been proposed. Numerous researchers have constructed real-world cross-sensor super-resolution datasets using satellite imagery from various sources and have trained SR models based on these datasets, which have been shown to be effective in improving the resolution of real-world satellite images [31–34]. However, the LR-HR image pairs used in these studies are sourced from different satellites with imaging times that are not exactly aligned. Variations in satellite attitude and atmospheric conditions further complicate geometric registration. Additionally, the high costs associated with obtaining HR images constrain dataset sizes, limiting the extent of SR model training and real-world applications.

To mitigate the impact of geometric registration errors, SR models need to avoid overemphasizing local details. Increasing the receptive field in SR networks enables them to learn a more comprehensive spatial understanding, which reduces sensitivity to minor misalignments between the LR-HR pairs. Additionally, tailoring models to the characteristics of remote sensing images is essential for efficient feature extraction, which, in turn, reduces the need for extensive training data. These design strategies, also adopted in this study, help address the inherent variability in remote sensing data, resulting in more robust super-resolution performance. Additionally, we constructed the GF6SRD dataset, a real-world cross-sensor super-resolution dataset, using WFV (16 m) and PMS (8 m) imagery from the same Gaofen-6 satellite. The GF6SRD dataset contains up to 18,306 LR-HR image pairs (512 × 512 pixels for HR). Since both PMS and WFV images are from Gaofen-6, they share nearly identical imaging times, satellite attitudes, and atmospheric conditions, significantly reducing the need for complex geometric registration and enhancing

consistency across LR-HR pairs. The GF6SRD dataset thus offers a more reliable foundation for training super-resolution models in real-world applications.

Early transformer-based SR methods, such as SwinIR, primarily focused on improving feature representation but did not demonstrate clear advantages over CNNs in modeling long-range dependencies compared to CNNs [25]. Later studies, including the HAT [25], RGT [27], and TTST [29], introduced various attention mechanisms to expand the receptive field of SR transformers, thereby improving performance. Furthermore, in Earth observation images, there is a strong spatial correlation between ground objects with close distance, as described by the first law of geography [33]. However, many previous studies have not fully explored these spatial correlations in remote sensing super-resolution tasks.

To address these challenges, a novel distance-enhanced strip attention transformer (DESAT) is proposed in this study. Specifically, to better capture the spatial correlation between adjacent ground objects, we design a distance-enhanced strip attention block (DSAB). This design integrates distance priors into the self-attention mechanism within strip windows which enlarges the receptive field and enhances spatial detail reconstruction, especially for regular textures. Additionally, the DESAT employs a novel attention-enhanced upsample block (AEUB), which combines the pixel shuffle layer with an attention-based upsample branch implemented through distance-enhanced attention mechanism in overlapping windows. This dual-branch upsample block transforms LR deep features into HR outputs from both attention-feature and pixel-wise perspectives. These innovations enable the DESAT to achieve a broader receptive field, more efficient feature extraction, and enhanced spatial detail reconstruction, which is critical for remote sensing super-resolution tasks. Furthermore, these advancements allow the DESAT to partially mitigate challenges such as geometric registration errors and limited dataset sizes, which are commonly encountered in real-world scenarios. To comprehensively assess the DESAT's performance, we conducted experiments using both the simulated AID dataset [34] and our proposed real-world GF6SRD dataset.

In brief, the main contributions of this study are summarized as follows:

1. Considering the spatial correlation between the adjacent ground objects, the proposed novel distance-enhanced strip attention transformer (DESAT) integrates distance priors into the strip window self-attention mechanism, achieving superior spatial detail reconstruction ability in challenging remote sensing image super-resolution tasks.
2. The proposed attention-enhanced upsample block (AEUB) enhances the pixel shuffle layer by introducing an overlapping window attention-based upsample branch, followed by a transformer fusion block for better feature integration and transformation between LR and HR features.
3. We constructed a comprehensive real-world cross-sensor remote sensing super-resolution dataset (GF6SRD) using images from two sensors with different resolutions onboard the Gaofen-6 satellite. The dataset comprises 18,306 LR-HR image pairs (512 × 512 pixels for HR), facilitating robust SR model training, and evaluation across diverse geographic and temporal conditions.

The remainder of this paper is organized as follows: Section 2 further introduces related works of this study; Section 3 details the implementation of our DESAT and the construction of the GF6SRD dataset; Section 4 presents extensive experiments conducted on both the simulated AID dataset and the real-world GF6SRD dataset; Section 5 further discusses the proposed method; and Section 6 summarizes the paper and specifies the future direction of this work.

## 2. Related Works

### *2.1. Natural Image Super-Resolution*

With the advancement of deep learning, various SR networks have emerged and achieved significant performance. The super-resolution convolutional neural network (SRCNN) [11,12] was the first deep learning-based approach to super-resolution. It interpolates LR images to the target resolution and then reconstructs HR images using a three-layer convolutional neural network. To accelerate SRCNN, the fast super-resolution convolutional neural network (FSRCNN) [13] removes the pre-network interpolation step and adds a deconvolutional layer after the convolutional neural network to upsample the LR features to the target resolution. Very deep convolutional networks for super-resolution (VDSR) [14] further extend this approach by using a 20-layer deep convolutional network combined with residual learning, expanding the receptive field and accelerating model convergence. Enhanced deep residual networks for super-resolution (EDSR) [15] introduce modifications to the ResNet architecture [35] by removing the batch normalization layer and relocating the ReLU activation outside each residual block, resulting in substantial performance improvements. The holistic attention network (HAN) [18] adds a layer attention module and a channel spatial attention module, effectively modeling holistic interdependencies to capture more informative features for SR tasks.

Until recently, transformer-based super-resolution (SR) methods [16,23–26] have shown superior performance compared to CNN-based methods due to the excellent long-range relationship modeling capabilities of the self-attention mechanism. For example, the pre-trained image processing transformer (IPT) [23], a pioneering work in applying transformers to image super-resolution, employs a transformer pre-trained on a large dataset and achieves impressive SR performance at a high computational cost. SwinIR [16], built on Swin Transformer architecture [20], enhances computational efficiency by removing the patch merging layer, leading to significant performance improvements. The dual aggregation transformer (DGT) [24] aggregates spatial and channel features alternately to reduce computational complexity. The hybrid attention transformer (HAT) [25] introduces overlapping window cross-attention and hybrid attention blocks to activate more pixels in SR tasks, thus enhancing performance. The introduction of N-Gram context to transformers for super-resolution enlarges the receptive field and significantly improves performance [26]. Recursive-generalization self-attention has been proposed to better aggregate features, followed by cross-attention to extract global information, delivering good SR results at low computational costs [27].

### *2.2. Remote Sensing Image Super-Resolution*

Deep learning SR methods have achieved significant success in natural images. Building on this progress, many researchers have developed specialized SR networks tailored for remote sensing images. For instance, a top-k selective mechanism has been introduced to better handle scale diversity and reduce redundant token representation, achieving excellent remote sensing SR performance [28]. A lightweight feature extraction block and a sequence-based upsample block have been proposed for an efficient hybrid CNN-transformer approach for better SR performance [29]. The hybrid-scale hierarchical transformer network (HSTnet) [30] utilizes a hybrid-scale feature exploitation module to leverage internal recursive information across scales in remote sensing images.

Furthermore, to better simulate the real-world cross-sensor remote sensing super-resolution tasks, ground-truth-based remote sensing super-resolution is proposed. For instance, Galar et al. [36] constructed 21,754 LR and HR image pairs (192 × 192 pixels for HR) from Sentinel-2 and Planet satellites, using an improved EDSR to enhance the resolution of Sentinel-2 images to 2.5 m. Romero et al. [37] created a dataset with 5827 LR-HR image pairs (140 × 140 pixels for HR) from WorldView and Sentinel satellites, applying ESRGAN to improve Sentinel-2 images with a scaling factor of five. Zabalza et al. [32] built a larger dataset of 56,821 LR-HR pairs (192 × 192 pixels for HR) from Sentinel-2 and Planet

satellites, utilizing the spectral attention residual network (SARNet) to enhance Sentinel-2 images from 10 m to 5 m and 2.5 m. Additionally, based on Gaofen-1/6 and Gaofen-2/7 satellites, Zhao et al. [38] constructed 9246 LR and HR image pairs (200 × 200 pixels for HR), applying an improved generative adversarial network with a self-attention module and texture loss to enhance Gaofen-1/6 images from 2 m to 1 m.

However, these methods use the pixel shuffle layer as their upsample module, which may lead to potential performance bottlenecks. Moreover, as a primary source of Earth observation data, remote sensing images have significant spatial correlations between ground objects. Meanwhile, distance measurement in remote sensing images is relatively easy, facilitating spatial correlation modeling.

In general, it is necessary to consider the improvement of the upsample module, and the modeling of the spatial correlation of ground objects can better extract the features of remote sensing images, which may be beneficial to improving the performance of remote sensing image super-resolution tasks.

## 3. Materials and Methods

In this section, the different sensors onboard the Gaofen-6 satellite and the construction of the real-world super-resolution GF6SRD dataset are first introduced, followed by a presentation of the overall architecture of the DESAT and its detailed implementation. Finally, the proposed distance-enhanced strip attention block (DSAB) and attention-enhanced upsample block (AEUB) are described in detail. The overall methodology of this paper is illustrated in Figure 1.

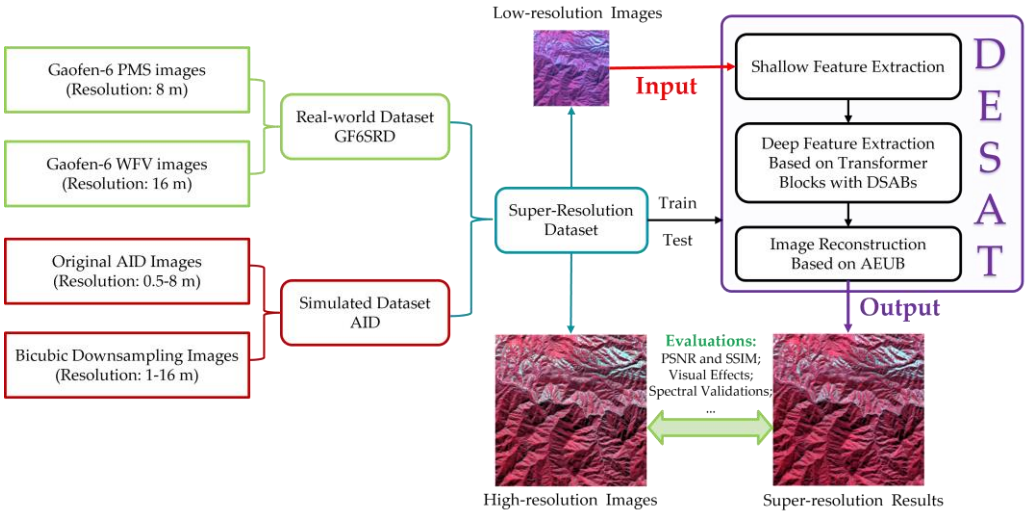


**Figure 1.** The overall methodology of this paper.

### *3.1. Real-World Cross-Sensor Super-Resolution Dataset of Gaofen-6 Satellite (GF6SRD)*

Many super-resolution studies rely on simulated datasets for training and testing SR models. These simulated datasets typically consist solely of HR images, lacking real HR-LR image pairs as references for model training. Most studies generate LR images from HR images based on an assumed degradation model, such as the most widely used bicubic kernel-based downsampling [31]. However, real remote sensing images with different resolutions are often captured by distinct sensors under different imaging conditions, leading to the extremely complicated remote sensing image degradation process, which is difficult to represent via the simplistic degradation model used in the simulated datasets. To address this issue, we construct a real-world cross-sensor super-resolution GF6SRD dataset with different resolutions (16 m WFV and 8 m PMS) onboard the Gaofen-6 satellite. This dataset consists of real HR-LR image pairs, offering a more accurate representation of real-world remote sensing super-resolution tasks. The remainder of Section

3.1 first introduces these two sensors of the Gaofen-6 satellite and then presents the cross-sensor GF6SRD dataset.

3.1.1. Introduction to the Sensors of the Gaofen-6 Satellite

The Gaofen-6 satellite, launched in June 2018, is equipped with two multispectral sensors. Its primary applications include agricultural and forestry surveys, as well as disaster management. The satellite carries two distinct multispectral sensors: a 16 m resolution wide field view (WFV) camera with a swath width exceeding 800 km and a 2/8 m resolution panchromatic/multispectral (PMS) camera with a swath width exceeding 90 km.

As shown in Table 1, the WFV and PMS sensors offer complementary advantages. While the PMS camera provides higher resolution, it has fewer spectral bands, a narrower swath, and a lower revisit frequency than WFV. Additionally, while WFV images are freely accessible, PMS images are not, which restricts the broad application of PMS imagery. To fully leverage the strengths of both sensors, constructing a cross-sensor super-resolution dataset is essential for developing robust SR models that can enhance the resolution of WFV imagery for real-world applications.

**Table 1.** Technical details of different sensors of Gaofen-6 satellite.

| Technical Specifications | WFV | PMS |
|---|---|---|
| Spatial Resolution | 16 m | 2/8 m Pan/MS [1] |
| Image Swath | >800 km | >90 km |
| Revisit Period | 4 days | 41 days |
| Spectral Range | Coastal (B1): 0.40 μm~0.45 μm<br>Blue (B2): 0.45 μm~0.52 μm<br>Green (B3): 0.52 μm~0.59 μm<br>Yellow (B4): 0.59 μm~0.63 μm<br>Red (B5): 0.63 μm~0.69 μm<br>Red Edge1 (B6): 0.69 μm~0.73 μm<br>Red Edge2 (B7): 0.73 μm~0.77 μm<br>NIR (B8): 0.77 μm~0.89 μm | Pan (P): 0.45 μm~0.90 μm<br>Blue (B1): 0.45 μm~0.52 μm<br>Green (B2): 0.52 μm~0.60 μm<br>Red (B3): 0.63 μm~0.69 μm<br>NIR (B4): 0.76 μm~0.90 μm |

[1] Pan/MS represents panchromatic/multispectral.

3.1.2. Construction Details of the Cross-Sensor GF6SRD Dataset

As illustrated in Figure 2, the process of constructing the GF6SRD dataset involves four main steps: Gaofen-6 satellite image downloading, image preprocessing, image screening, and image clipping.

After completing these steps, the resulting GF6SRD dataset consists of 18,306 WFV-PMS (LR-HR) image pairs, where the WFV images are 256 × 256 pixels with a resolution of 16 m, and the PMS images are 512 × 512 pixels with a resolution of 8 m.

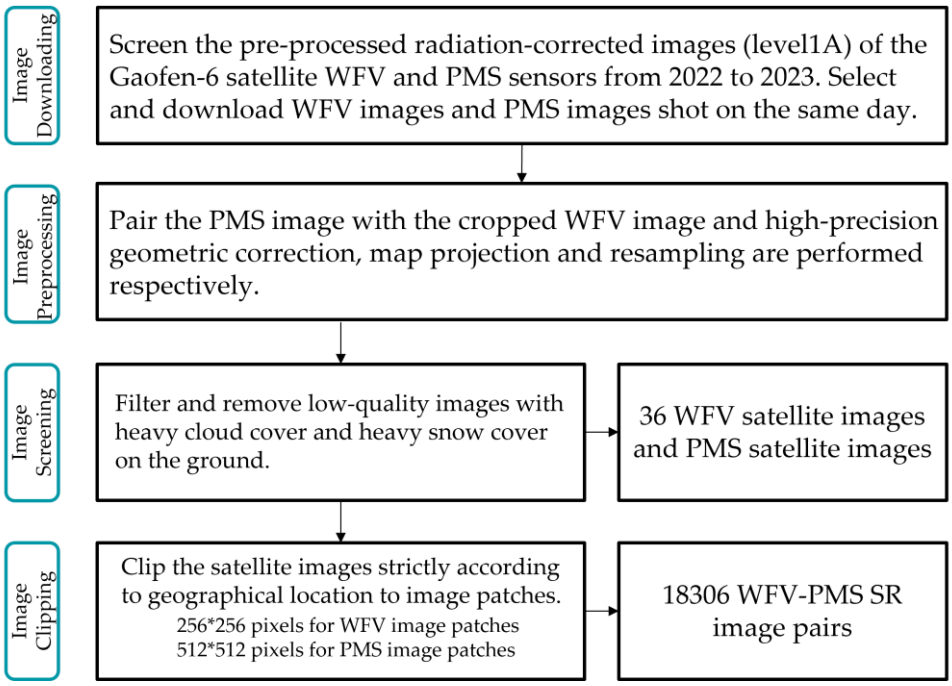


**Figure 2.** Construction flow chart of GF6SRD dataset.

Figure 3 provides examples of the WFV-PMS image pairs, with the same relative stretch applied to the RGB bands for color representation.

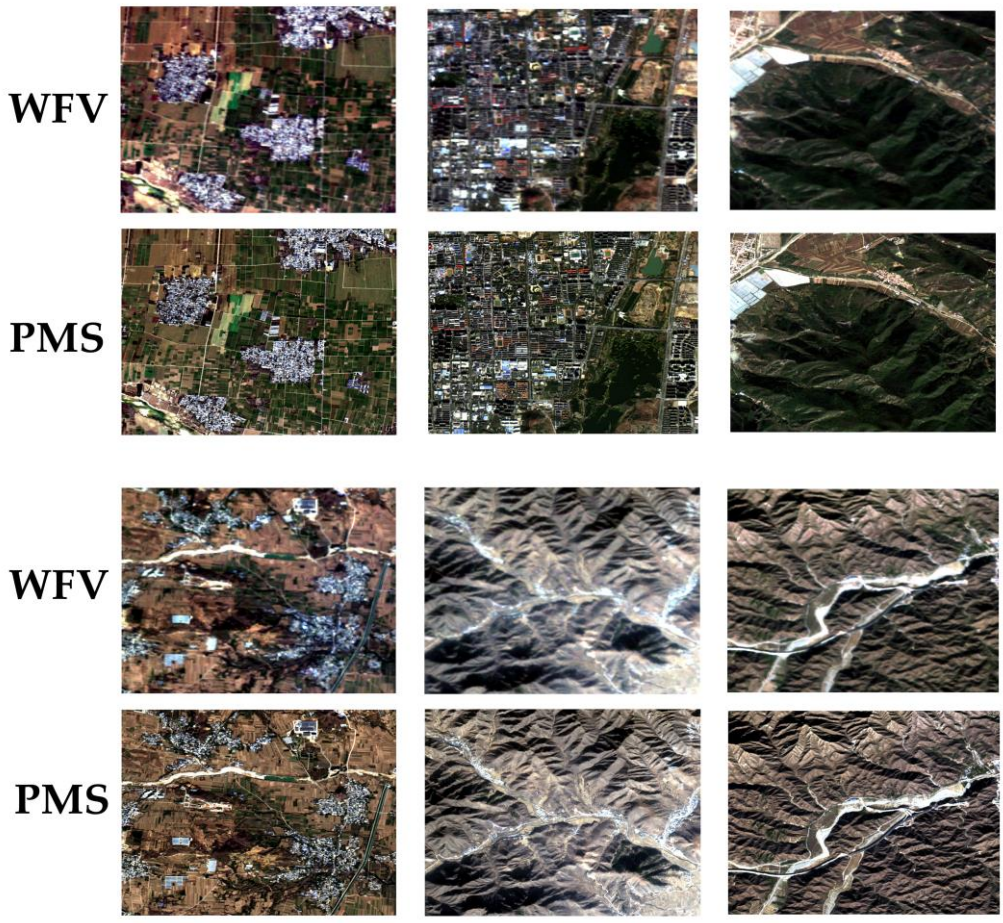


**Figure 3.** Some WFV-PMS image pairs of the proposed GF6SRD dataset.

All the image pairs in the GF6SRD dataset were clipped from 36 pairs of original satellite images, with acquisition dates detailed in Table 2. To better simulate a real-world cross-sensor remote sensing super-resolution task, the dataset was partitioned into training and test sets based on acquisition dates. The test set includes 2710 WFV-PMS image pairs captured on 22 January 2023, 29 September 2023, and 9 November 2023. And the training set contains 15,596 WFV-PMS image pairs captured on other dates.

**Table 2.** Numbers of WFV-PMS original satellite image pairs taken at different times.

| Acquisition Time | Number of WFV-PMS Image Pairs |
|---|---|
| 18 January 2022 | 2 |
| 3 February 2022 | 3 |
| 8 March 2022 | 3 |
| 18 April 2022 | 2 |
| 26 April 2022 | 2 |
| 7 October 2022 | 2 |
| 5 November 2022 | 3 |
| 13 November 2022 | 2 |
| 28 December 2022 | 3 |
| 22 January 2023 | 2 |
| 29 May 2023 | 1 |
| 22 July 2023 | 2 |
| 19 August 2023 | 2 |
| 29 September 2023 | 1 |
| 3 October 2023 | 3 |
| 9 November 2023 | 3 |

### *3.2. Methods*

#### 3.2.1. The Overall Architecture

As illustrated in Figure 4, the overall network of our DESAT comprises three primary components: shallow feature extraction, deep feature extraction, and image reconstruction. This architecture design has been widely adopted in previous studies [16,25,26,28]. Specifically, for a given LR input $I_{LR} \in R^{h \times w \times C_{in}}$, the DESAT first employs a single convolutional layer to extract shallow features $F_S \in R^{h \times w \times C}$, where $C_{in}$ and $C$ denote the number of input and hidden feature channels, respectively. Subsequently, a series of residual distance-enhanced strip attention groups (RDSG), followed by a convolutional layer, are employed to extract deep features $F_D \in R^{h \times w \times C}$ from the shallow features. Finally, the super-resolution output image $I_{SR} \in R^{H \times W \times C_{out}}$ is reconstructed from the deep features $F_D$ and shallow features $F_S$ through a reconstruction block consisting of our AEUB and convolution layers. For simplicity, the DESAT is optimized using the Charbonnier loss [39] defined as $L_C = \sqrt{\|I_{SR} - I_{GT}\|^2 + \epsilon^2}$ with $\epsilon = 1 \times 10^{-3}$, where $I_{SR}$ denotes the SR image l, $I_{GT}$ is the high-resolution ground-truth image, and the $\epsilon$ is a small constant that prevents instability by smoothing the loss function around zero differences.

Figure 4 shows that each RDSG contains several subgroups utilizing the proposed DSAB, hybrid attention blocks (HAB) [28], an overlap cross-attention block (OCAB) [28], and a 3 × 3 convolutional layer. Building upon the successful residual hybrid attention group [28], our RDSG replaces several HABs with DSABs. This modification incorporates our proposed distance-enhanced strip attention mechanism (DESAM), improving the representation of texture features. Additionally, we added a DESAM with a skip connection to better integrate the horizontal and vertical features before the OCAB. Due to space constraints, readers are referred to previous works for the specific implementations of HAB and OCAB. The details of the proposed DSAB and AEUB are provided in the following sections.

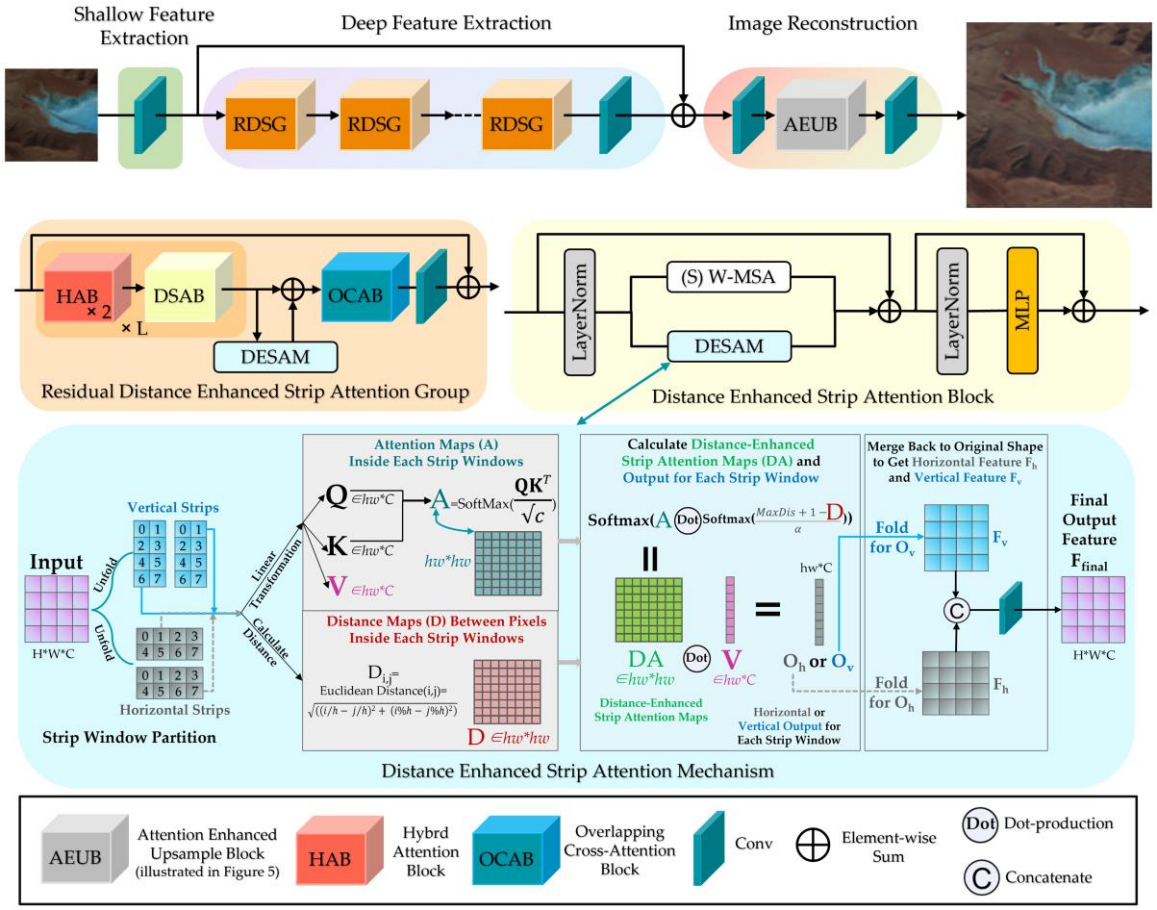


**Figure 4.** The overall architecture of the DESAT and the structure of the DSAB, DESAM, and RDSG.

### 3.2.2. Distance-Enhanced Strip Attention Block (DSAB)

The first law of geography states, "Everything is related to everything else, but near things are more related than distant things." This principle highlights the significant impact of distance on the correlation between different ground objects in remote sensing images. In addition, the high altitude of satellite sensors allows for precise distance measurements between ground objects, enabling the integration of distance priors into the attention mechanism. Moreover, inspired by previous work [40–42], we divide features into horizontally and vertically oriented strip windows to extend the receptive field and improve texture reconstruction while maintaining a low computational cost. Within each strip window, we calculate the Euclidean distance between pixels and incorporate this spatial information into the traditional self-attention mechanism, thereby enhancing feature extraction. This customized mechanism for remote sensing images effectively models spatial correlations between adjacent ground objects, resulting in superior spectral fidelity and spatial detail reconstruction.

As demonstrated in Figure 4, the structure of our DSAB is built by paralleling our proposed distance-enhanced strip attention mechanism (DESAM) over the shifted window multi-head self-attention mechanism (SW-MSA) within the Swin Transformer block [20]. In the following, we focus on the implementation of DESAM. The calculation process of DESAM is depicted in Figure 4. First, our DESAM unfolds the input $F_{input} \in R^{H\times W\times C}$ to $\frac{H}{h}\times\frac{W}{w}$ non-overlapping strip windows $F_{window} \in R^{h\times w\times C}$, oriented both horizontally (with $h < w$) and vertically (with $h > w$). Then, for each window, the Euclidean distance matrix $D \in R^{hw\times hw}$ between pixels is calculated as

$$D_{i,j} = \sqrt{((i/h - j/h)^2 + (i\%h - j\%h)^2)}, \quad (1)$$

where $D_{i,j}$ denotes the distance between the $i$th pixel and the $j$th pixel within the windows. The operations "/" and "%" represent integer division and modulo operations, respectively.

Next, similarly to traditional self-attention mechanism, the query $Q \in R^{hw \times C}$, key $K \in R^{hw \times C}$, and value $V \in R^{hw \times C}$ are generated from the window feature $F_{window} \in R^{h \times w \times C}$ through linear transformation separately as

$$Q = F_{window}W^Q,\ K = F_{window}W^K,\ V = F_{window}W^V, \tag{2}$$

where $W^Q$, $W^K$, and $W^V$ are learnable weight matrices for generating $Q$, $K$, and $V$, respectively. After that, the attention matrix $A \in R^{hw \times hw}$ is computed through the scaled dot-product operation between $Q$ and transposed $K$ across channels as

$$A = {QK^T}/{\sqrt{d_k}}, \tag{3}$$

where $d_k$ the dimensionality of the key, and $\sqrt{d_k}$ is used to control the magnitude of the dot-product operation. Then, to incorporate distance priors, the distance-enhanced attention matrix $DA \in R^{hw \times hw}$ is computed as

$$DA = \mathrm{Softmax}(A \odot \mathrm{Softmax}({MaxDis + 1 - D}/{\alpha})). \tag{4}$$

Where $A$ is the attention matrix, and $D$ is the Euclidean distance matrix. *MaxDis* is the maximum distance between pixels in the strip window, and $\alpha$ is constant to smooth out differences in values within $DA$.

For each window, the horizontal outputs $O_h \in R^{hw \times C}$ and vertical outputs $O_v \in R^{hw \times C}$ are then calculated as

$$O_h = DA_hV_h,\ O_v = DA_vV_v. \tag{5}$$

Finally, these outputs are folded back to their original shape, yielding the horizontal feature $F_H \in R^{H \times W \times C}$ and vertical feature $F_V \in R^{H \times W \times C}$. To integrate these two features, the DESAM concatenates them along the channel dimension and applies a 3 × 3 convolutional layer to produce the final output feature $F_{output} \in R^{H \times W \times C}$, which can be represented by

$$F_{output} = \mathrm{Conv}(\mathrm{Concat}(F_H, F_V)), \tag{6}$$

where Conv(·) is a 3 × 3 convolutional layer with 2C input channels and C output channels.

This process completes the computation of DESAM, yielding the final output $F_{output}$. Similarly to the self-attention mechanism, the time complexity of DESAM is proportional to the number of pixels within the window. In addition, compared to the traditional window self-attention mechanism, our DESAM integrates distance measurements into the attention map, which enables more effective extractions of spatial texture features by differentiating pixels based on their positional relationships within strip windows. Consequently, DESAM enhances DSAB's ability to capture spatial correlations between ground objects, leading to improved spatial detail and texture representation.

#### 3.2.3. Attention-Enhanced Upsample Block (AEUB)

The pixel shuffle layer [43] is the widely used upsample module for super-resolution tasks due to its simplicity and effectiveness. However, much research has focused on optimizing feature extraction modules while often overlooking the potential of improving the upsample process. This has resulted in the upsample module contributing only a small computational load, creating a potential bottleneck in SR models.

To address this, we proposed the attention-enhanced upsample block (AEUB), which integrates the pixel shuffle layer and distance-enhanced attention applied in overlapping windows for better feature transformation. As illustrated in Figure 5, our AEUB consists of two parallel branches: the pixel branch and the attention branch. The pixel branch rearranges pixel-wise spatial details through the pixel shuffle layer, while the attention branch captures broader spatial relationships via distance-enhanced attention. Together, these

two branches ensure more accurate and efficient transformations of LR features into HR outputs.

The pixel branch first utilizes a convolutional layer to increase the number of input channels and then rearranges features from the channel dimension into the spatial dimension through the pixel shuffle layer. In parallel, the attention branch first employs a distance-enhanced attention mechanism in overlapping windows and then merges these window features in a non-overlapping manner to improve spatial resolution through repeated sampling features across spatial dimensions. Subsequently, a novel transformer fusion block combines the outputs of both branches, ensuring better feature integration and HR transformation.

Specifically, for a LR feature input $F_{lr} \in R^{h\times w\times C}$, our AEUB generates two HR features: $F_P \in R^{H\times W\times C}$ from the pixel branch and $F_A \in R^{H\times W\times C}$ from the attention branch. These two features are then fused through the transformer fusion block to produce the final HR feature $F_{hr} \in R^{H\times W\times C}$. The overall operation of AEUB can be represented by

$$F_{hr} = \mathrm{TF}(\mathrm{P}(F_{lr}), \mathrm{A}(F_{lr})), \tag{7}$$

where P (·) and A(·) represent the operations of the pixel branch and attention branch, respectively. And TF(·) denotes the transformer fusion block, which will be illustrated later.

In the pixel branch, the computation is straightforward:

$$F_P = \mathrm{P}(F_{lr}) = \text{Pixel-Shuffle}(\mathrm{Conv}(F_{lr})). \tag{8}$$

In the attention branch, the process involves more complexity. For an upsampling factor of 2, the input LR feature $F_{lr} \in R^{h\times w\times C}$ is unfolded into overlapping windows of size 8 × 8 with an overlapping ratio of 0.5, producing the unfolded features $F_u \in R^{\frac{hw}{16}\times 64\times C}$, which can be represented as.

$$F_u = \mathrm{Unfold}(F_{lr}). \tag{9}$$

For each 8 × 8 window, the distance-enhanced attention mechanism (DEA) is applied to compute the window attention feature $F_w \in R^{8\times 8\times C}$. These window features are then folded back in a non-overlapping manner to produce the attention-based HR feature $F_A \in R^{H\times W\times C}$, expressed as

$$F_A = \mathrm{A}(F_{lr}) = \mathrm{Fold}(\mathrm{DEA}(F_u)), \tag{10}$$

where DEA(·) denotes the distance-enhanced attention mechanism, which is similar to our DESAM, as described in Section 3.2.2, but operates on square windows instead of strip windows. The pseudo-code for the attention branch of our AEUB is shown in Algorithm 1.

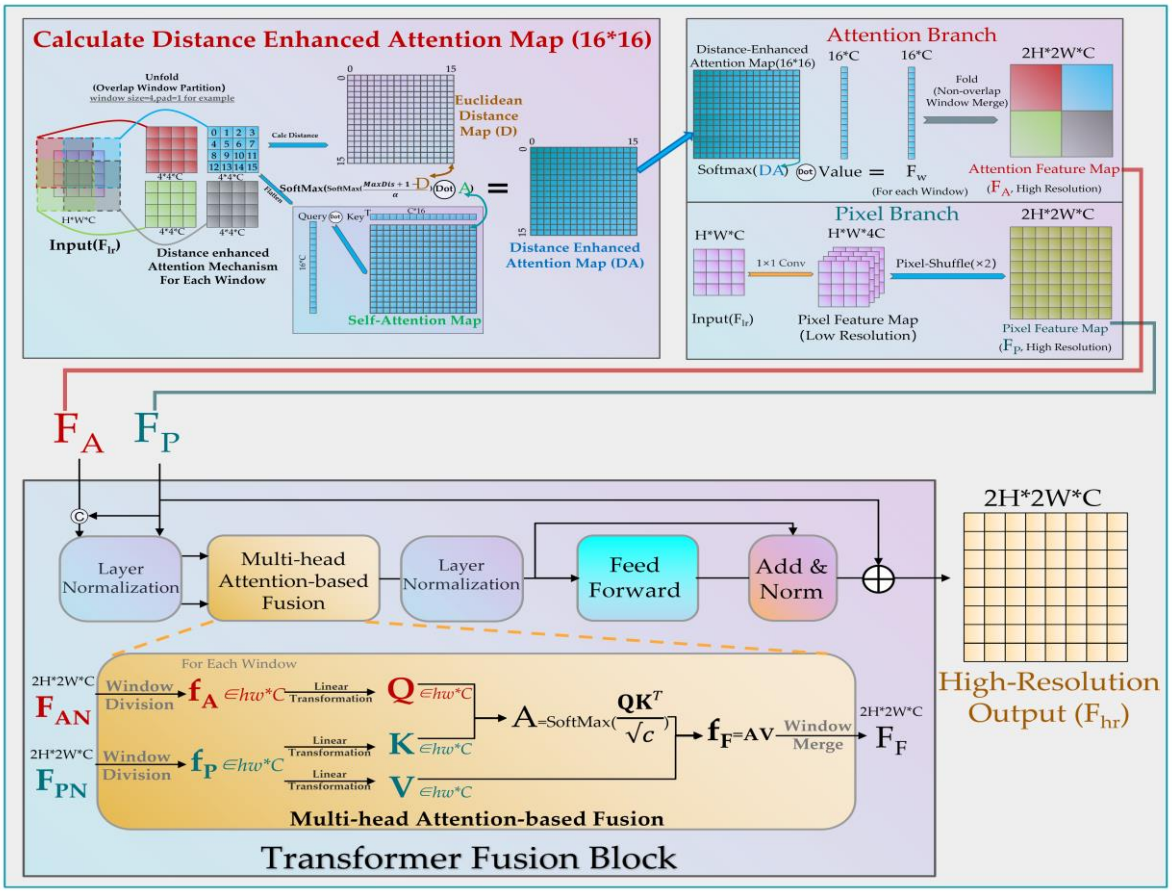


**Figure 5.** Overview of the proposed AEUB module.

**Algorithm 1:** The pseudo-code of our attention branch of AEUB.

**Input:** *LR features* $F_{lr} \in R^{b\times c\times h\times w}$
**Output:** *HR features* $F_A \in R^{b\times c\times 2h\times 2w}$

*Q, K, V* ⇇ *Chunk(Conv(*$F_{lr}$*));* // $b \times c \times h \times w$
*Q, K, V* ⇇ *Reshape(Q, K, V);* // $b \times head \times \frac{c}{head} \times h \times w$
*Q, K, V* ⇇ *Overlapping-Unfold(Q, K, V);* // $b \times head \times \frac{c}{head} \times 64 \times \frac{h}{4} \times \frac{w}{4}$
*Q, K, V* ⇇ *Permute(Q, K, V);* // $b \times head \times \frac{h}{4} \times \frac{w}{4} \times 64 \times \frac{c}{head}$
$A \Leftarrow \frac{QK}{\lambda}$; // $b \times c \times \frac{h}{4} \times \frac{w}{4} \times 64 \times 64$
$DA \Leftarrow A \bullet Dis\_mask$; // $b \times c \times \frac{h}{4} \times \frac{w}{4} \times 64 \times 64$
$F_w$ ⇇ *Softmax(DA)V;* // $b \times head \times \frac{h}{4} \times \frac{w}{4} \times 64 \times \frac{c}{head}$
*Reshaped_*$F_w$ ⇇ *Reshape(Permute(*$F_w$*));* // $b \times c \times 64 \times \frac{h}{4} \times \frac{w}{4}$
$F_A$ ⇇ *Nonoverlapping-Fold (Reshaped_*$F_w$*);* // $b \times c \times 2h \times 2w$

In summary, the attention branch uses an innovative overlapping unfold operation combined with a distance-enhanced attention mechanism to obtain spatially repetitively sampled deep features from the LR input. These features are then efficiently upsampled into HR outputs through a non-overlapping fold operation. Additionally, the upsampling ratio of the attention branch can be adjusted by the window size and overlap rate.

Once the pixel and attention branch features $F_P$ and $F_A$ are obtained, they are fused via the proposed transformer fusion block TF(·), yielding the final HR feature $F_{hr}$. As is depicted in Figure 5, this block generates the query $Q$ from the attention feature $F_A$ and pixel feature $F_P$, while the key $K$ and value $V$ are calculated from pixel feature $F_P$, ensuring optimal integration of both features through the attention mechanism. The detailed computation of the transformer fusion block is as follows:

$$F_{PN} = \mathrm{LN}(F_P),\ F_{AN} = \mathrm{LN}(\mathrm{Concat}(F_A, F_P)),$$
$$F_F = \text{W-MAF}(F_{PN}, F_{AN}), \tag{11}$$
$$F_{hr} = \mathrm{FNN}(\mathrm{LN}(F_F)) + F_{PN,}$$

where LN(·) denotes layer normalization, FNN represents the feedforward neural network, and W-MAF(·) represents our proposed window multi-head attention-based fusion layer, whose specific calculations are illustrated in Figure 5.

## 4. Results

### *4.1. Experimental Setup*

#### 4.1.1. Datasets for Experiments

The results of the experiments are presented to evaluate the super-resolution models on both the simulated AID super-resolution dataset and our proposed real-world GF6SRD dataset.

The AID dataset is a remote sensing classification dataset with resolutions ranging from 0.5 m to 8 m. For simplification, we randomly selected 3500 images from the original 10,000 images in the AID dataset to obtain LR-HR image pairs and form the simulated super-resolution dataset. The original images (600 × 600 pixels) were resampled to 512 × 512 pixels to obtain HR images, which were then downsampled using bicubic interpolation to generate LR images. The simulated AID dataset was split into 3000 training pairs and 500 test pairs.

The GF6SRD dataset, as detailed in Section 3.1, includes 15,596 training image pairs and 2710 test image pairs. Each image pair comprises a 16 m WFV image (LR, 256 × 256 pixels) and an 8 m PMS image (HR, 512 × 512 pixels). Real-world cross-sensor remote sensing super-resolution tasks involve both spatial texture differences and spectral variations between LR and HR images due to differences in the spectral response functions of different sensors. These factors contribute to the instability of SR results. Similarly to the large-input super-resolution task [44], the output image patches from the SR model must be merged to restore the original large-scale remote sensing images (typically exceeding 10,000 × 10,000 pixels). Due to this instability of SR results, noticeable seams often appear at the boundaries of adjacent SR image patches, affecting the overall quality of merged results. To mitigate this issue, we use larger LR image patches of 256 × 256 pixels instead of the more common 64 × 64 pixels, thereby reducing the number of patches from each original remote sensing image and lessening the impact of seams.

#### 4.1.2. Implementation Details

In this study, all models are implemented for ×2 SR. For the structure of our final DESAT, the number of RDSGs is set to four, with each RDSG containing four HABs and two DSABs. The window size for self-attention within each HAB is set to 8, while each DSAB has a strip window length of 16 and a width of 4. The number of channels in the DESAT is set to 96, and the number of attention heads is set to 6. The window size within AEUB is set to 8, and the overlapping ratio is set to 0.5 for ×2 SR. Additionally, we provide a smaller version of the DESAT-S with fewer parameters and computations than the HAT. In the DESAT-S, the number of channels is set to 64, and the number of attention heads is set to 4, while other parameters remain the same as with the DESAT.

All SR models use the same training method to ensure fair comparisons. Specifically, on the AID dataset, total training iterations are set to 600,000, with the learning rate initialized as $5 \times 10^{-5}$ and reduced by half at [150,000, 300,000, and 450,000]. On the GF6SRD dataset, total training iterations are set to 311,920, with the learning rate initialized as $5 \times 10^{-5}$ and reduced by half at [77,980, 155,960, and 233,940]. The batch size is set to one, and models are optimized by the Charbonnier loss [39]. Adam optimizer is employed in all models.

For evaluation, the widely used peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) [45] are adopted. All SR models were trained using the PyTorch framework on an Ubuntu 20.04.2 machine with A100-PCIE-40GB GPU.

### *4.2. Comparison Studies*

#### 4.2.1. Comparison Studies on the Simulated AID Dataset

Quantitative Results: On the simulated AID dataset, Table 3 shows the quantitative comparison of our approaches and the state-of-the-art methods, including CNN-based methods, such as the FSRCNN [13], EDSR [15], and HAN [18], and transformer-based methods, such as the SwinIR [16], SwinIR-NG [26], HAT [25], RGT [27], and TTST [29]. Since the input image resolution is 256 × 256 pixels in this study, which differs from the common 64 × 64 pixels, the original versions of some transformer models mentioned above cannot be directly run due to GPU memory limitations. Meanwhile, to be consistent with our DESAT, the number of channels of all transformer models is set to 96, except for our DESAT-S, and the number of transformer blocks is set to 24. Only our DESAT-S model has the number of channels set to 64 as illustrated in Section 4.1.2.

The quantitative results on the simulated AID dataset show that our DESAT significantly outperforms the other approaches on this dataset. Concretely, the DESAT surpasses the best transformer-based method, the HAT, by 0.09 dB and outperforms the best CNN-based method, the HAN, by 0.05 dB, with much lower parameters and computations. Additionally, our smaller version, the DESAT-S, outperforms the HAT by 0.03 dB with even fewer parameters and computations. Furthermore, the DESAT shows an advantage in perceptual quality, achieving the lowest LPIPS score (0.0906) among all models, with reductions of 0.0005 and 0.0018 relative to the HAN and HAT, respectively. These results underscore the effectiveness of the DSAB and AEUB modules in enhancing perceptual and texture feature representation, particularly highlighting the advantages of our proposed distance-enhanced strip attention mechanism over the channel attention employed in the HAT.

To further evaluate SR performance across different land cover classes on the AID dataset, we analyzed the top six models. Table 4 presents the quantitative results of these models for 30 different land cover classes on the AID dataset. The DESAT achieves the best SR performance in 29 of the 30 land cover classes. Compared with the second-best model, the HAN, the DESAT shows a maximum improvement of 0.14 dB in PSNR and 0.0004 in SSIM. Furthermore, our smaller version, the DESAT-S, also performs comparably to the HAN, while being much more resource efficient.

**Table 3.** Quantitative comparison with state-of-the-art methods on the simulated AID dataset. The top three results are marked in red bold, blue bold and green bold (the "↑" indicates that a larger value for this metric is better, while the "↓" indicates that a smaller value for this metric is better).

| Method | | PSNR (dB)↑ | SSIM↑ | LPIPS↓ | Params | FLOPs | Memory |
|---|---|---|---|---|---|---|---|
| CNN-based | FSRCNN | 38.97 | 0.9712 | 0.1099 | 0.01 M | 0.93 G | 1690 MB |
| | EDSR | 41.09 | 0.9781 | 0.0922 | 40.73 M | 2669.44 G | 8098 MB |
| | HAN | **41.19** | **0.9784** | **0.0911** | 15.92 M | 1035.53 G | 13,054 MB |
| Transformer-based | SwinIR | 41.09 | 0.9780 | 0.0929 | 2.42 M | 160.11 G | 11,558 MB |
| | SwinIR-NG | 41.12 | 0.9781 | 0.0924 | 2.82 M | 149.26 G | 12,270 MB |
| | HAT | 41.15 | 0.9782 | 0.0924 | 4.06 M | 267.45 G | 15,662 MB |
| | RGT | 41.04 | 0.9778 | 0.0930 | 2.14 M | 128.46 G | 15,930 MB |
| | TTST | 40.95 | 0.9775 | 0.0934 | 4.93 M | 323.83 G | 36,629 MB |
| | DESAT-S (ours) | **41.18** | **0.9783** | **0.0916** | 3.03 M | 207.43 G | 15,962 MB |
| | DESAT (ours) | **41.24** | **0.9785** | **0.0906** | 6.39 M | 429.56 G | 21,430 MB |

**Table 4.** Quantitative comparison of 30 land cover classes on the simulated AID dataset. The top two results are marked in red bold and blue bold.

| Land Cover | EDSR | | HAN | | SwinIR-NG | | HAT | | DESAT-S | | DESAT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Airport | 41.59 | 0.9793 | **41.69** | **0.9795** | 41.60 | 0.9791 | 41.64 | 0.9792 | 41.66 | 0.9793 | **41.72** | **0.9796** |
| Bare Land | 44.25 | 0.9811 | 44.27 | 0.9812 | 44.27 | 0.9811 | 44.27 | 0.9811 | **44.29** | **0.9812** | **44.31** | **0.9813** |
| Baseball Field | 41.56 | 0.9759 | **41.63** | **0.9761** | 41.57 | 0.9759 | 41.59 | 0.9759 | 41.62 | 0.9760 | **41.66** | **0.9762** |
| Beach | 42.93 | 0.9771 | 43.02 | 0.9773 | 43.02 | 0.9773 | 43.03 | 0.9773 | **43.05** | **0.9774** | **43.08** | **0.9775** |
| Bridge | 42.48 | 0.9784 | **42.56** | **0.9787** | 42.51 | 0.9785 | 42.54 | 0.9785 | 42.56 | 0.9786 | **42.61** | **0.9788** |
| Center | 40.18 | 0.9780 | **40.36** | **0.9784** | 40.27 | 0.9781 | 40.32 | 0.9782 | 40.33 | 0.9783 | **40.43** | **0.9785** |
| Church | 38.69 | 0.9744 | **38.90** | **0.9749** | 38.79 | 0.9745 | 38.85 | 0.9747 | 38.86 | 0.9748 | **38.94** | **0.9751** |
| Commercial | 40.99 | 0.9817 | 41.10 | 0.9820 | 41.07 | 0.9818 | 41.12 | 0.9819 | **41.15** | **0.9820** | **41.22** | **0.9822** |
| D-Residential | 38.78 | 0.9779 | **38.97** | **0.9786** | 38.86 | 0.9781 | 38.92 | 0.9783 | 38.94 | 0.9784 | **39.01** | **0.9787** |
| Desert | 44.69 | 0.9803 | **44.72** | **0.9804** | 44.73 | 0.9804 | 44.74 | 0.9804 | 44.76 | 0.9805 | **44.79** | **0.9806** |
| Farmland | 42.83 | 0.9755 | **42.88** | **0.9758** | 42.83 | 0.9755 | 42.84 | 0.9755 | 42.87 | 0.9756 | **42.91** | **0.9758** |
| Forest | 40.51 | 0.9735 | 40.55 | 0.9737 | 40.55 | 0.9736 | 40.54 | 0.9736 | **40.57** | **0.9737** | **40.61** | **0.9739** |
| Industrial | 39.84 | 0.9796 | **39.98** | **0.9801** | 39.88 | 0.9796 | 39.94 | 0.9798 | 39.96 | 0.9799 | **40.04** | **0.9802** |
| Meadow | 43.10 | 0.9710 | 43.11 | 0.9710 | 43.10 | 0.9709 | 43.11 | 0.9709 | **43.12** | **0.9710** | **43.15** | **0.9712** |
| M-Residential | 38.80 | 0.9711 | **38.97** | **0.9716** | 38.88 | 0.9713 | 38.91 | 0.9713 | 38.93 | 0.9714 | **39.00** | **0.9717** |
| Mountain | 41.68 | 0.9784 | 41.71 | 0.9785 | 41.72 | 0.9785 | 41.72 | 0.9785 | **41.74** | **0.9786** | **41.78** | **0.9787** |
| Park | 40.52 | 0.9779 | 40.59 | 0.9782 | 40.55 | 0.9780 | 40.57 | 0.9780 | **40.61** | **0.9782** | **40.66** | **0.9784** |
| Parking | 39.80 | 0.9833 | **39.97** | **0.9836** | 39.83 | 0.9833 | 39.92 | 0.9835 | 39.92 | 0.9835 | **40.04** | **0.9838** |
| Playground | 41.27 | 0.9758 | **41.36** | **0.9760** | 41.29 | 0.9757 | 41.31 | 0.9757 | 41.34 | 0.9759 | **41.40** | **0.9761** |
| Pond | 42.00 | 0.9795 | 42.06 | **0.9797** | 42.02 | 0.9795 | 42.04 | 0.9795 | **42.06** | 0.9796 | **42.10** | **0.9798** |
| Port | 41.13 | 0.9829 | **41.29** | **0.9832** | 41.19 | 0.9830 | 41.24 | 0.9831 | 41.25 | 0.9831 | **41.31** | **0.9833** |
| Railway | 40.54 | 0.9800 | **40.62** | **0.9803** | 40.51 | 0.9798 | 40.55 | 0.9799 | 40.58 | 0.9801 | **40.64** | **0.9803** |
| Resort | 40.33 | 0.9792 | **40.49** | **0.9798** | 40.42 | 0.9794 | 40.46 | 0.9795 | 40.47 | 0.9796 | **40.52** | **0.9798** |
| River | 41.15 | 0.9730 | **41.20** | **0.9732** | 41.17 | 0.9730 | 41.17 | 0.9730 | 41.20 | 0.9731 | **41.24** | **0.9733** |
| School | 39.39 | 0.9791 | **39.51** | **0.9795** | 39.45 | 0.9792 | 39.48 | 0.9793 | 39.51 | 0.9794 | **39.58** | **0.9797** |
| S-Residential | 39.13 | 0.9686 | **39.24** | **0.9691** | 39.20 | 0.9689 | 39.21 | 0.9688 | 39.24 | 0.9690 | **39.28** | **0.9692** |
| Square | 40.67 | 0.9793 | **40.78** | **0.9796** | 40.75 | 0.9795 | 40.77 | 0.9795 | 40.81 | 0.9797 | **40.93** | **0.9800** |
| Stadium | 40.95 | 0.9811 | **41.11** | **0.9815** | 40.92 | 0.9808 | 40.98 | 0.9809 | 41.00 | 0.9811 | **41.10** | **0.9814** |
| Storage Tanks | 39.72 | 0.9772 | **39.89** | **0.9776** | 39.75 | 0.9771 | 39.80 | 0.9773 | 39.81 | 0.9773 | **39.89** | **0.9776** |
| Viaduct | 40.82 | 0.9785 | **40.89** | **0.9787** | 40.77 | 0.9782 | 40.81 | 0.9784 | 40.84 | 0.9785 | **40.93** | **0.9790** |

Visual Results: Figure 6 provides visual comparisons on the simulated AID dataset. In Figure 6a, for the image "square_229", the DESAT is the only model that correctly reconstructs the path and alignment of the sidewalks. In contrast, other approaches distort the lateral distribution of the sidewalks into an oblique distribution. Figure 6b shows the results for the image "industrial_137". The DESAT accurately recovers the regular texture of densely arranged buildings, while the other approaches exhibit varying levels of artifacts and blurring. In summary, compared with all comparison methods, the DESAT has a more robust texture reconstruction ability, making it the only approach to restore the sidewalk arrangement correctly, and the DESAT also has a clearer reconstruction result for local spatial details. These visual results further validate the DESAT's strong performance on the simulated AID dataset.

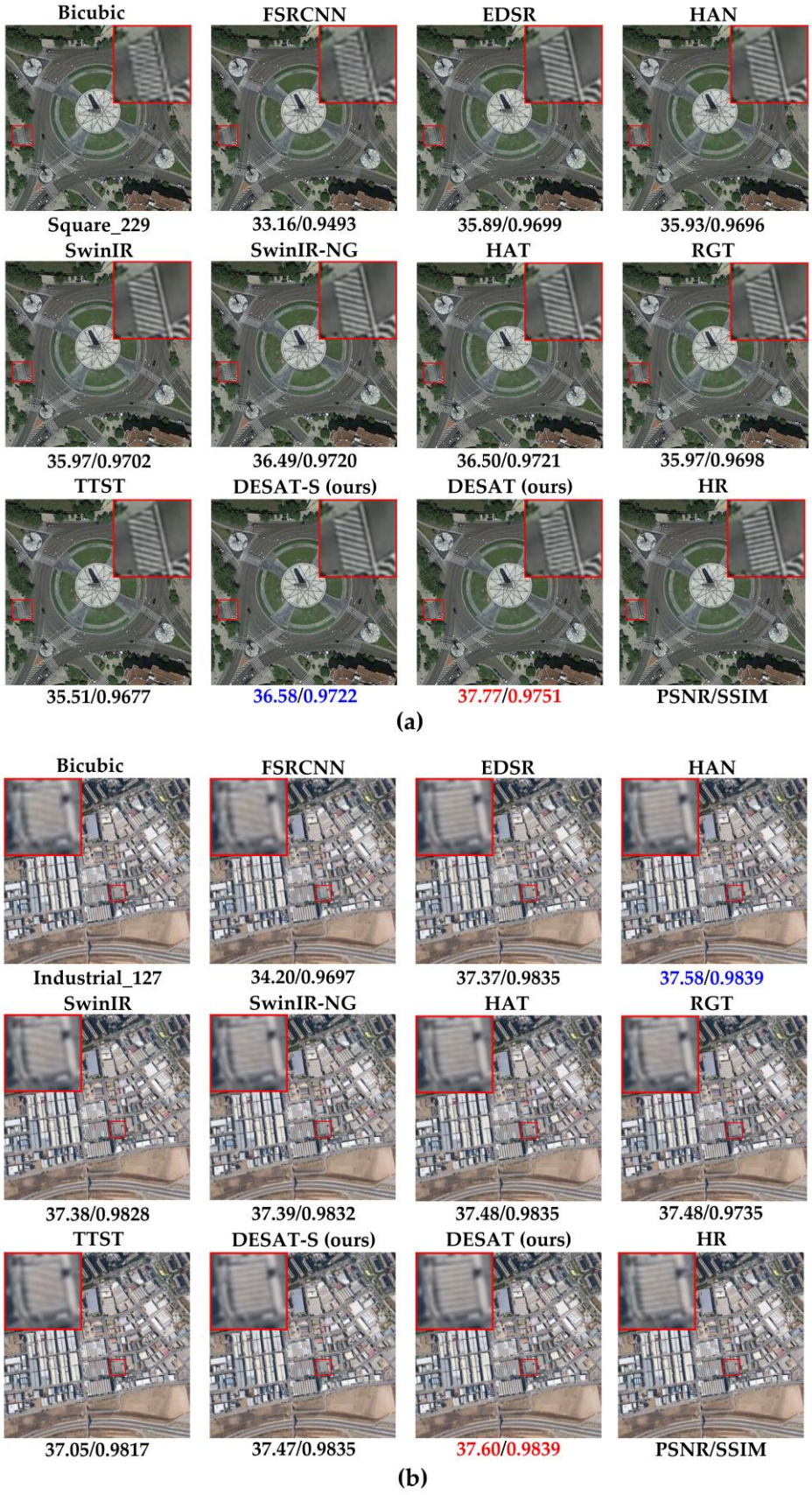


**Figure 6.** Visual comparisons on the simulated AID dataset (the red box represent the zoomed area. And (**a**) represent the visual results of image "square_229", while (**b**) represent the visual results of image "industrial_137").

#### 4.2.2. Comparison Studies on the Proposed Real-World GF6SRD Dataset

Quantitative Results: On our real-world GF6SRD dataset, Table 5 presents the quantitative results of the DESAT compared with the models discussed in Section 4.2.1. The DESAT significantly outperforms all the comparison models on the real-world GF6SRD dataset, showing an improvement of 1.17 dB in PSNR and 0.0090 in SSIM over the best comparison model TTST. Additionally, in perceptual evaluations, the DESAT achieves the lowest LPIPS among all models, with reductions of 0.0153 relative to the best comparison model EDSR. For the smaller version, the DESAT-S, also exceeds all comparison models, outperforming TTST by 0.36 dB in PSNR and 0.0035 in SSIM.

Compared to the SR performance on the simulated AID dataset, the DESAT shows enhanced competitiveness and substantial performance improvement on the real-world GF6SRD dataset. This highlights the model's effectiveness in the more challenging real-world cross-sensor super-resolution tasks, where the introduction of a distance-enhanced strip attention mechanism and attention-enhanced upsample module are particularly beneficial. These results underscore the strong potential of the DESAT for remote sensing engineering applications.

**Table 5.** Quantitative comparison with state-of-the-art methods on the real-world GF6SRD dataset. The top three results are marked in red bold, blue bold and green bold (the "↑" indicates that a larger value for this metric is better, while the "↓" indicates that a smaller value for this metric is better).

| Method | | PSNR (dB)↑ | SSIM↑ | LPIPS↓ | Params | FLOPs | Memory |
|---|---|---|---|---|---|---|---|
| CNN-based | FSRCNN | 30.81 | 0.8370 | 0.4383 | 0.02 M | 1.23 G | 1690 MB |
| | EDSR | 33.36 | **0.9370** | **0.1654** | 40.74 M | 2670.80 G | 8098 MB |
| | HAN | 32.98 | 0.9328 | 0.1844 | 15.93 M | 1035.87 G | 13,054 MB |
| Transformer-based | SwinIR | 33.18 | 0.9326 | 0.1887 | 2.42 M | 160.55 G | 11,296 MB |
| | SwinIR-NG | 33.34 | 0.9302 | 0.1897 | 2.83 M | 149.77 G | 12,270 MB |
| | HAT | 33.35 | 0.9346 | 0.1794 | 4.07 M | 267.88 G | 15,662 MB |
| | RGT | 33.04 | 0.9313 | 0.1831 | 2.15 M | 128.90 G | 15,930 MB |
| | TTST | **33.57** | 0.9341 | 0.1769 | 4.94 M | 324.27 G | 36,667 MB |
| | DESAT-S (ours) | **33.93** | **0.9376** | **0.1621** | 3.04 M | 207.78 G | 15,962 MB |
| | DESAT (ours) | **34.74** | **0.9431** | **0.1501** | 6.42 M | 430.00 G | 21,430 MB |

Visual Results: Figures 7–9 display visual comparisons on the GF6SRD dataset, showing WFV-PMS image pairs alongside SR results from the DESAT and comparison models. All images are displayed using false colors for bands 4, 3, and 2. Due to differences in sensor characteristics, the WFV and PMS images exhibit spectral discrepancies, leading to color variations under the same color stretch.

In Figure 7a, the DESAT reconstructs the arc outline of the outer runway and the square outline of the inner lawn with greater accuracy than other models. It also produces more accurate color predictions for the lawn and runway. Figure 7b demonstrates that the DESAT provides the most precise reconstruction of the small water body in the lower-left corner of the dam, including its boundary.

Figure 8 shows that the DESAT achieves the most accurate color representation and the clearest texture reconstruction of buildings compared to other models. Figure 9 illustrates that the DESAT effectively predicts the internal voids of buildings and the transverse distribution of multiple rectangular structures while providing the most accurate color representation.

These visual results demonstrate that the DESAT significantly outperforms all comparison methods in color accuracy and spatial detail reconstruction. It also proves that the DESAT can produce high-quality super-resolution images with better quantitative metrics and visual effects than other approaches in challenging real-world remote sensing super-resolution tasks.

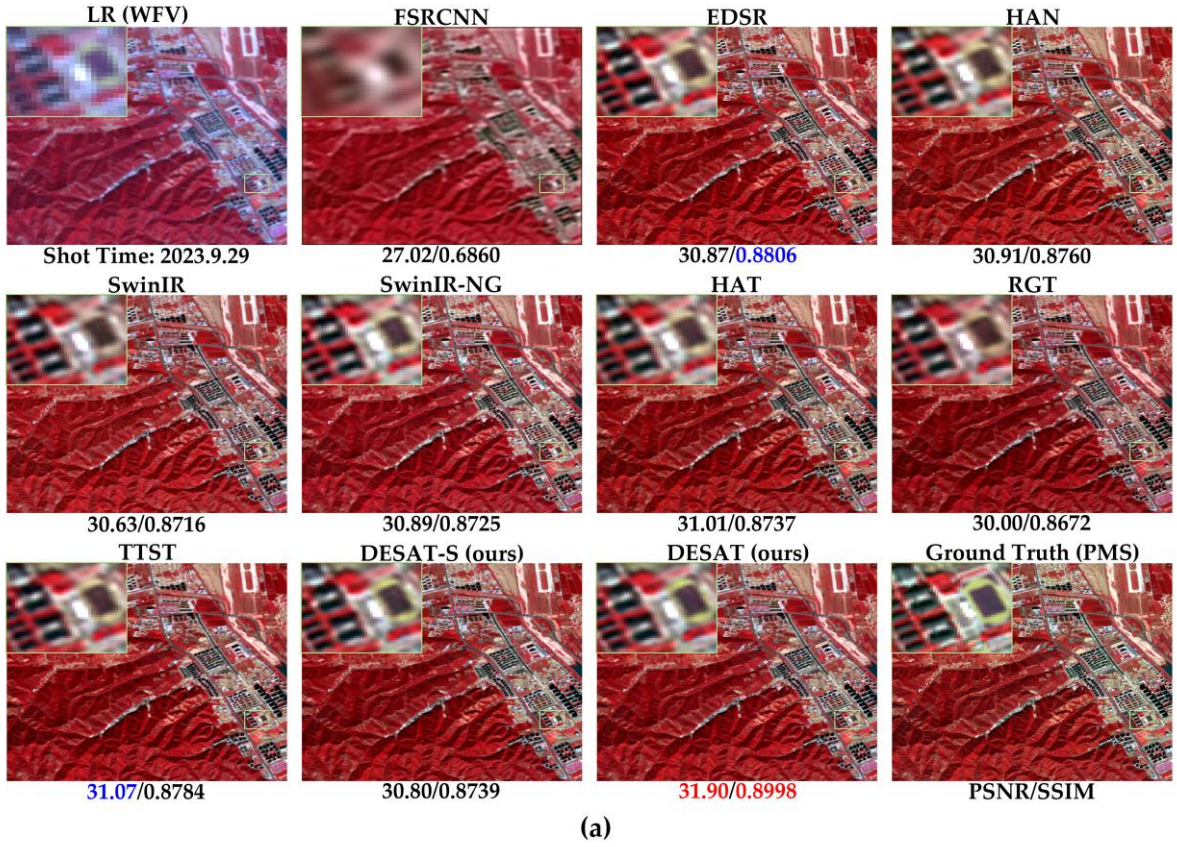


**Figure 7.** Visual comparisons on two WFV-PMS image pairs shot on 29 September 2023 (the green box represent the zoomed area. And (**a**,**b**) represent the visual results of two different areas shot on 29 September 2023, respectively).

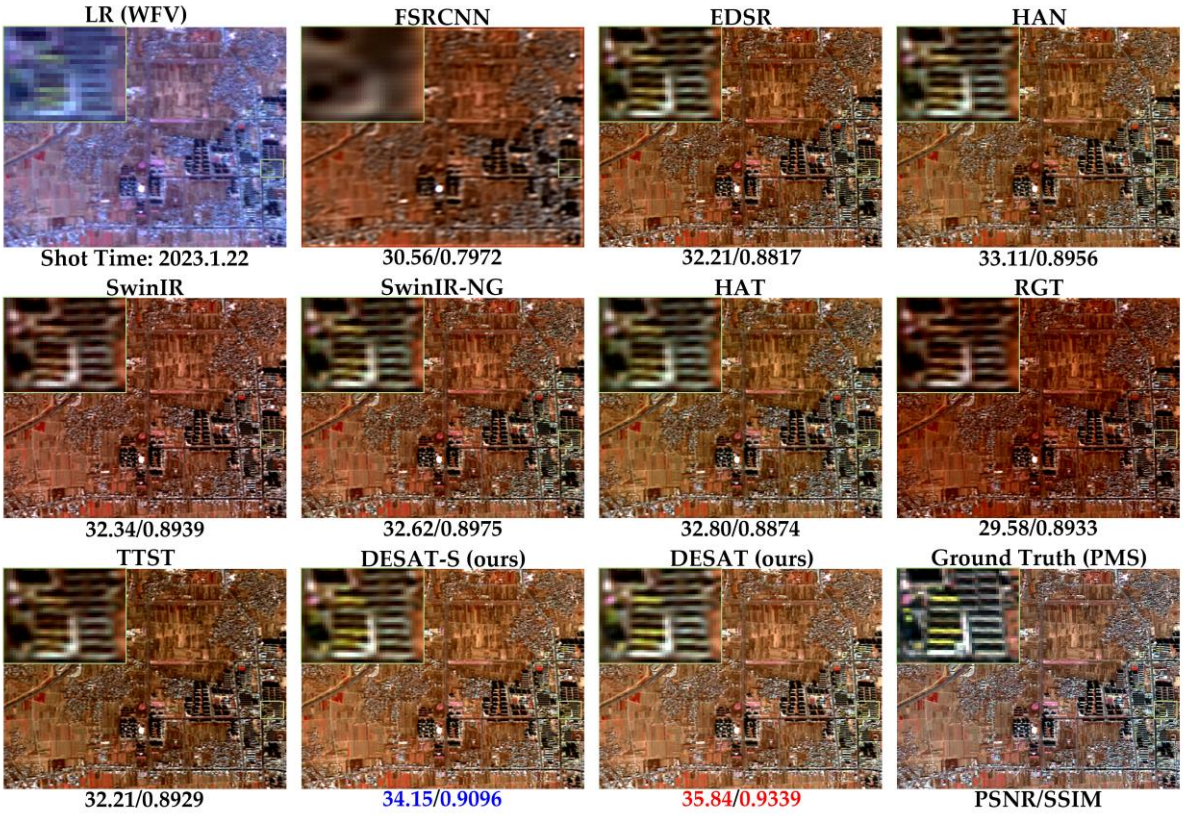


**Figure 8.** Visual comparisons on a WFV-PMS image pair shot on 22 January 2023.

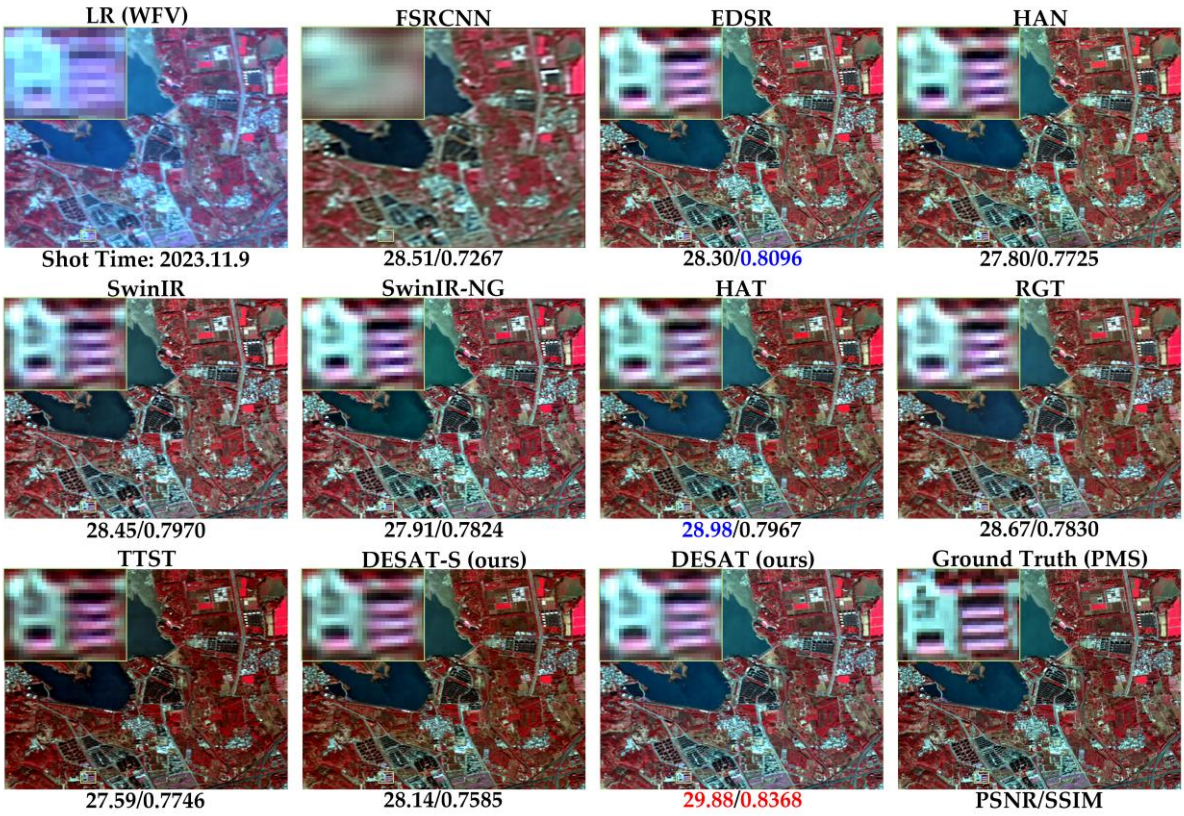


**Figure 9.** Visual comparisons on a WFV-PMS image pair shot on 9 November 2023.

### *4.3. Ablation Experiments*

#### 4.3.1. The Effects of the DSAB and AEUB

To assess the effectiveness of the proposed distance-enhanced strip attention block (DSAB) and attention-enhanced upsample block (AEUB), we conducted a series of ablation experiments. Table 6 presents the quantitative results on both the simulated AID dataset and the real-world GF6SRD dataset. The baseline model is created by replacing DSABs with HABs and using the pixel shuffle layer instead of our AEUB for reconstruction. And the baseline model also includes the OCABs like the DESAT.

On the AID dataset, the inclusion of DSAB results in a performance improvement of 0.06 dB, while adding the AEUB leads to a gain of 0.07 dB. Moreover, incorporating DSAB and AEUB together results in a further performance enhancement of 0.09 dB for the proposed DESAT.

On the GF6SRD dataset, the impact of the DSAB and AEUB is even more pronounced. The addition of DSAB alone raises PSNR by 0.86 dB, while the AEUB contributes a 0.74 dB gain. When combined, the DSAB and AEUB achieve an overall improvement of 1.39 dB in PSNR and 0.0085 increase in SSIM. This substantial enhancement underscores the effectiveness of the DSAB and AEUB in handling the complexities of cross-sensor scenarios in real-world remote sensing super-resolution tasks, where spatial and spectral information must be accurately reconstructed.

These results demonstrate that both the DSAB and the AEUB enhance the DESAT's performance across different datasets, with particularly strong effects observed in challenging real-world datasets like GF6SRD.

**Table 6.** Ablation studies of the proposed DSAB and AEUB on the simulated AID dataset.

| Model | AID | | GF6SRD | |
|---|---|---|---|---|
| | **PSNR** | **SSIM** | **PSNR** | **SSIM** |
| Baseline | 41.15 | 0.9782 | 33.35 | 0.9346 |
| DSAB | 41.21 | 0.9784 | 34.21 | 0.9377 |
| AEUB | 41.22 | 0.9784 | 34.09 | 0.9386 |
| DSAB + AEUB | 41.24 | 0.9785 | 34.74 | 0.9431 |

#### 4.3.2. The Effects of Distance-Enhanced Attention

To evaluate the impact of the distance-enhanced attention mechanism, we compared the DESAT using this mechanism with a version utilizing traditional attention. Table 7 shows the quantitative results on both the AID and GF6SRD datasets. On the AID dataset, the introduction of distance-enhanced attention results in a 0.04 dB increase in PSNR. On the more complex real-world GF6SRD dataset, the distance-enhanced attention mechanism achieves a substantial improvement of up to 0.73 dB over traditional attention.

These results demonstrate that our distance-enhanced attention effectively integrates distance prior to the traditional window self-attention mechanism, leading to significant performance gains, especially in challenging real-world super-resolution tasks.

**Table 7.** Ablation studies of the distance-enhanced attention mechanism on two datasets.

| Module | AID | | GF6SRD | |
|---|---|---|---|---|
| | **PSNR** | **SSIM** | **PSNR** | **SSIM** |
| Traditional Attention | 41.20 | 0.9784 | 34.01 | 0.9389 |
| Distance-enhanced Attention | 41.24 | 0.9785 | 34.74 | 0.9431 |

### *4.4. Interpretations with Local Attribution Maps*

To further interpret the performance improvements of our model, we visualized local attribution maps [46] (LAM) for several state-of-the-art models. LAM utilizes the integrated gradients to interpret SR networks by highlighting contributing pixels and activated regions for reconstruction within the selected local image patch. It provides two

outputs: attribution maps, showing pixels that contribute significantly to local texture reconstruction; and activated regions, reflecting the model's receptive field.

Figure 10 illustrates the LAM results for different models on the AID dataset, revealing the receptive field size as indicated by the activated region. These results indicate that the CNN-based method, the HAN, has a slightly larger activated region compared to EDSR, correlating with the HAN's superior SR performance. Similarly, the transformer-based method, the HAT, shows a larger activated region than SwinIR, aligning with its better performance than SwinIR. Meanwhile, our DESAT exhibits a modestly larger activated region than that of the HAT, reflecting an expanded receptive field due to the DSAB and AEUB. This expanded receptive field improves the DESAT's ability to capture spatial dependencies and reconstruct detailed textures, as shown by its superior quantitative and qualitative performance over the HAT. These results reinforce findings from previous studies [25,28] that correlate receptive field size with super-resolution effectiveness, underscoring the positive impact of DSAB and AEUB on the DESAT's performance.

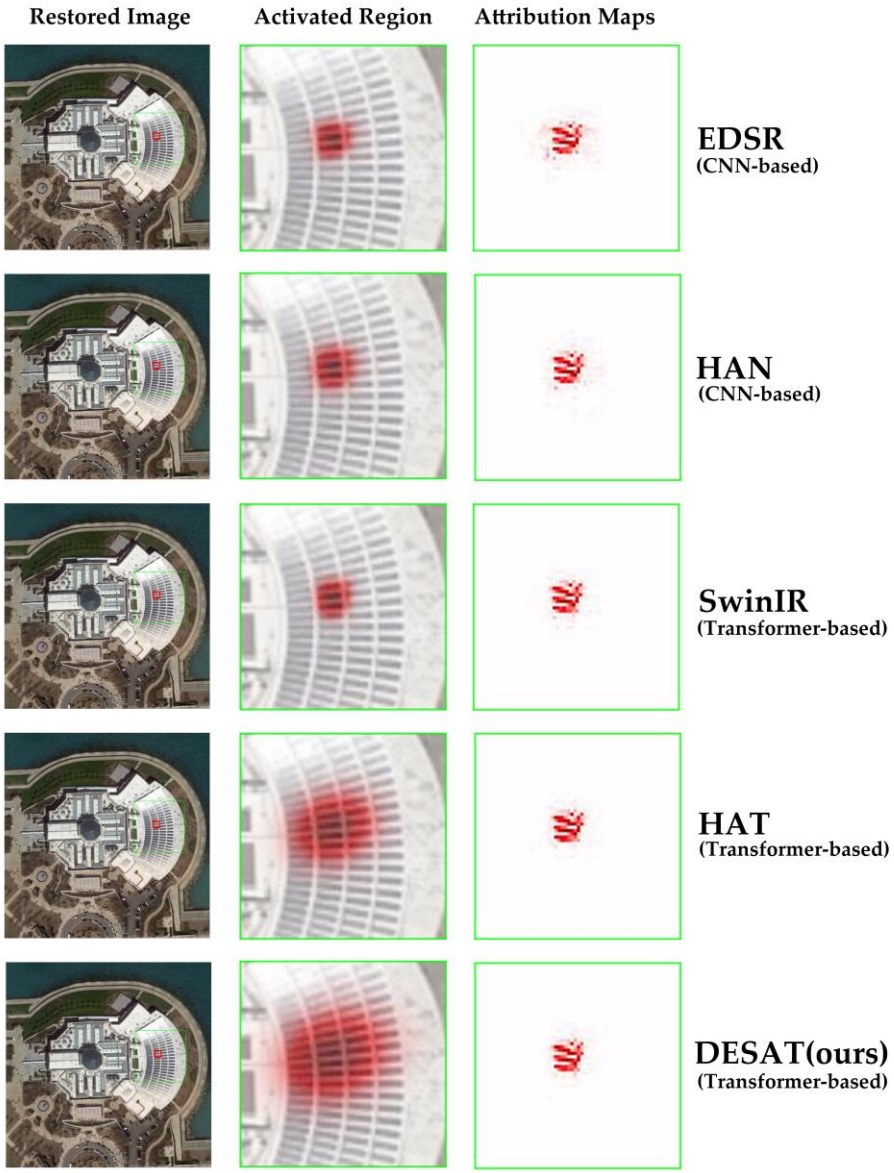


**Figure 10.** Local attribution maps (LAM) results for different models in the AID dataset (the red box represents the selected local image patch, and the green box represents the zoomed area).

### *4.5. Spectral Validations*

Accurate reconstruction of spectral characteristics is crucial for super-resolution remote sensing images as it directly impacts their effectiveness in various applications. Beyond capturing spatial textures, a model's ability to learn and accurately transform spectral information across different sensors is essential, especially in real-world cross-sensor super-resolution tasks. On our GF6SRD dataset, the SR model must address differences in spectral characteristics and color variations between WFV and PMS images due to their distinct imaging processes.

To evaluate the DESAT's ability to transform spectral information, we compared the spectral values of target PMS images with the SR images generated by the DESAT. Specifically, we analyzed all test WFV-PMS image pairs captured on 29 September 2023, performing a detailed comparison of spectral values across all bands. R-squared coefficients were employed to quantify the spectral similarity between the SR and PMS images.

Figure 11 shows a scatter plot of the spectral values from 200,000 randomly selected pixels in these WFV-PMS image pairs. The R-squared coefficients for each band are above 0.95, indicating that the spectral characteristics of the SR images closely match those of the target PMS images.

Additionally, Figure 12 provides visual examples of spectral performance across various land cover categories, including vegetation, urban, water, and bare areas. For each land cover type, we selected representative points to plot spectral curves for the WFV image, the PMS image, and the DESAT-generated SR image. These visual examples highlight the DESAT's ability to accurately transform spectral information across diverse land cover categories.

Overall, these results confirm that the DESAT effectively learns and transforms spectral characteristics in real-world cross-sensor super-resolution tasks, maintaining consistency in spectral information across different sensors.

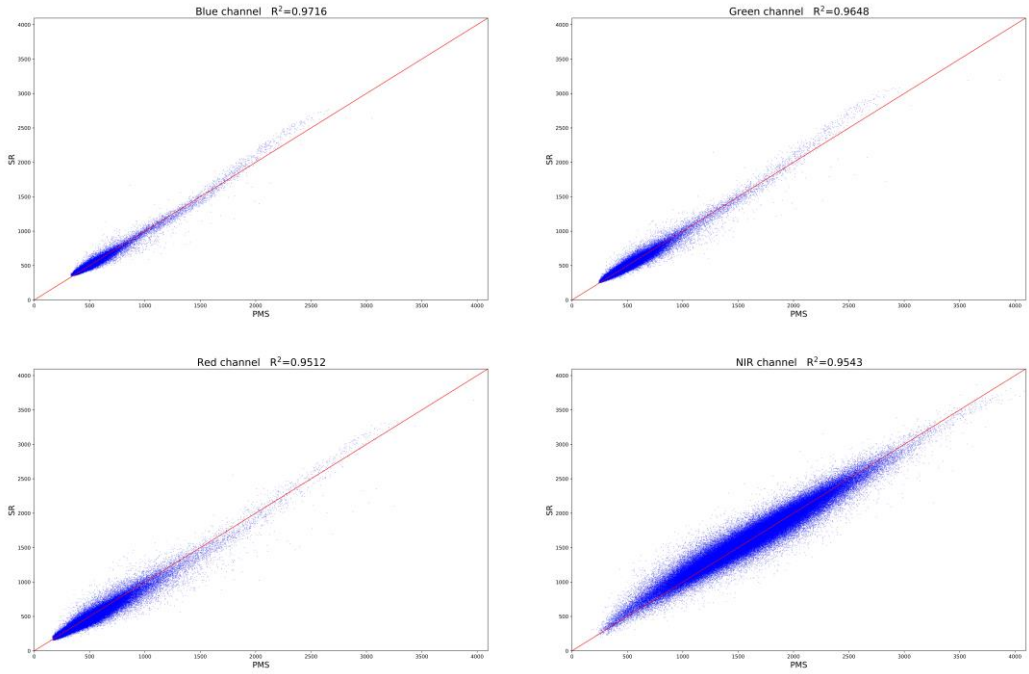


**Figure 11.** Spectral value comparisons in each band between the SR images and the PMS images (the x-axis represents the PMS values, and the y-axis represents the SR values).

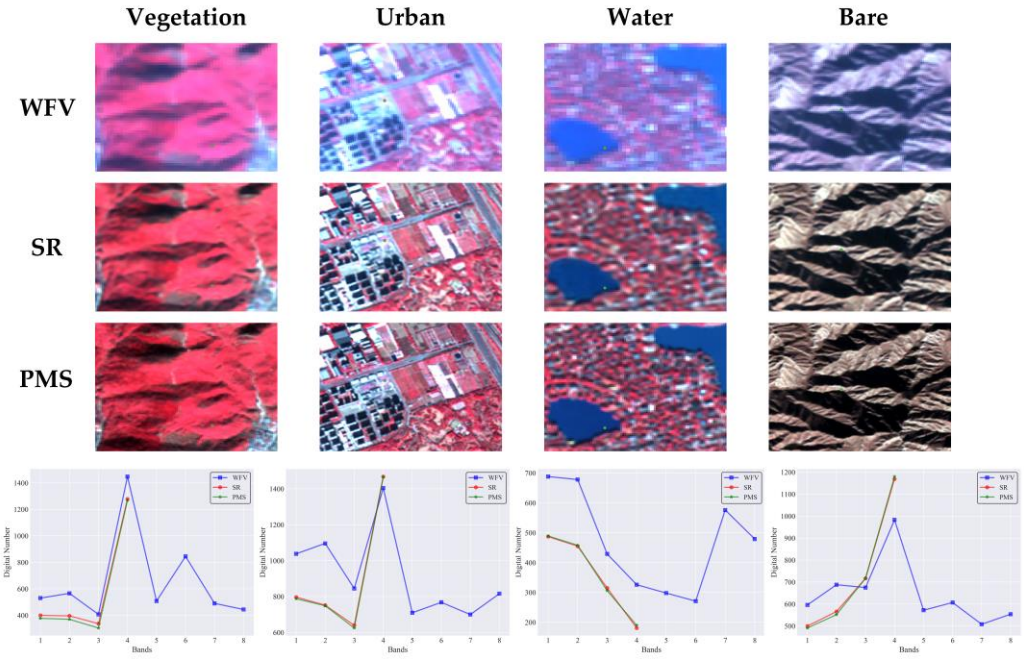


**Figure 12.** Visual examples of different land cover types in WFV images, SR images, and PMS images, along with the representative spectral curves (zoom in for a better view).

### *4.6. Migration Experiments*

To evaluate the generalization capability of the DESAT, we conducted migration experiments by applying the DESAT, trained on the GF6SRD dataset, to WFV images captured by the Gaofen-6 satellite in regions and periods not included in the original dataset. The WFV images were divided into image patches of 256 × 256 pixels, which the DESAT processed to generate SR patches. These patches were then merged to create a wide-field SR image with the same resolution and spectral characteristics as PMS images.

Figure 13 presents the visual results from these migration tests. As illustrated, the SR images generated by the DESAT display sharper and more detailed local textures than the original WFV images, indicating the DESAT's strong generalization across regions and times.

To further quantify migration effectiveness, we employed two no-reference metrics: average gradient (AG) and spatial frequency (SF). With no ground-truth PMS images available for migration experiments, AG and SF provide insights by quantifying spatial detail and sharpness retained in SR images. These metrics are essential in assessing the SR results' practical value in remote sensing, where higher values generally indicate better spatial information and texture fidelity.

Table 8 presents the AG and SF metrics for DESAT-generated SR results versus bicubic-interpolated WFV images. The DESAT shows an AG increase of 12.55 and an SF improvement of 27.04 over bicubic interpolation. This indicates the DESAT's ability to enhance and preserve spatial detail in migration scenarios, confirming its applicability in real-world SR applications across diverse conditions.

**Table 8.** Quantitative evaluation of migration effectiveness: AG and SF metrics for DESAT vs. bicubic interpolation on WFV images.

| Method | AG↑ | SF↑ |
|---|---|---|
| Bicubic Interpolation | 17.53 | 34.19 |
| SR Results from DESAT | 30.08 | 61.23 |

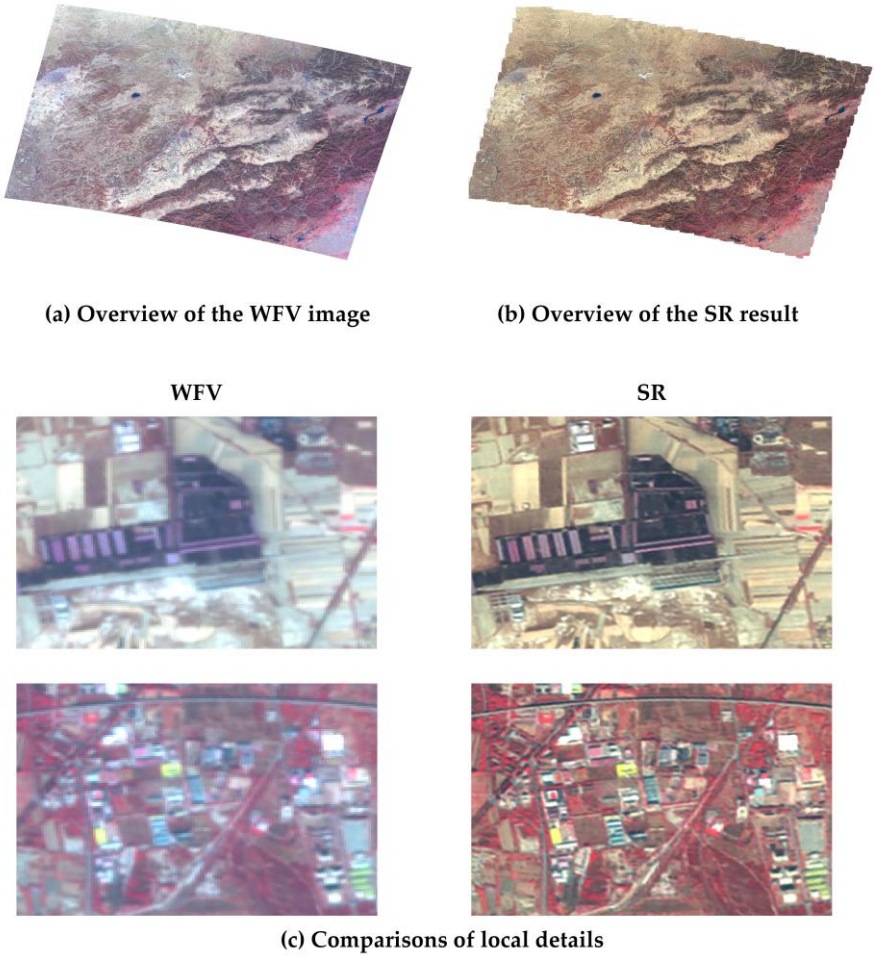


**Figure 13.** The overviews and local details of the WFV image and the super-resolution results.

## 5. Discussion

In this study, we construct a real-world cross-sensor super-resolution dataset, GF6SRD, and propose the DESAT to address the limitations of existing SR methods. Experimental results demonstrate that the DESAT shows competitive performance in both simulated and real-world super-resolution tasks. The discussion integrating these results with theoretical analysis is as follows.

### *5.1. Impact of DSAB and AEUB*

Ablation studies (Section 4.3.1) show that incorporating the DSAB and AEUB effectively enhances SR performance. Theoretically, the strip window attention mechanism in the DESAT extends the receptive field while maintaining computational efficiency. Compared to the pixel shuffle layer, the larger window size in AEUB (set to 8) further extends the receptive field. This is supported by the LAM results (Section 4.4), which show that the DESAT has a bigger activated region and achieves a larger receptive field. Furthermore, the distance-enhanced attention mechanism successfully integrates distance priors into traditional attention (Section 4.3.2), leading to performance improvements, particularly in real-world tasks.

### *5.2. Comparison with Other Models*

The DESAT outperforms CNN-based models like the EDSR and HAN in SR performance while offering lower computational costs. Compared to transformer-based models such as SwinIR, SwinIR-NG, and RGT, the DESAT provides significant performance improvements. Additionally, our smaller version, the DESAT-S, also surpasses transformer-

based models like the HAT and TTST with fewer parameters and lower computational demands. Notably, compared to its performance on simulated datasets, the DESAT shows greater advancements on more challenging cross-sensor super-resolution tasks, where it demonstrates superior spectral fidelity and spatial detail reconstruction. The significant spectral and spatial differences between PMS and WFV images in the cross-sensor GF6SRD dataset pose a challenge to traditional SR models, which the DESAT addresses effectively by leveraging distance priors through its enhanced attention mechanism.

### *5.3. Performance on Super-Resolution Tasks with Higher Magnification Rates (×4)*

To further assess the DESAT's capabilities, we conducted additional4× super-resolution experiments on the simulated AID dataset, comparing the DESAT with other strong models, including the HAN, SwinIR-NG, and HAT. As illustrated in Table 9, the DESAT consistently demonstrates competitive performance, outperforming all comparison models at the 4× magnification rate. Concretely, the DESAT surpasses the transformer-based method, the HAT, by 0.06 dB in PSNR and 0.0012 in SSIM, while the DESAT also outperforms the CNN-based method, the HAN, by 0.06 dB in PSNR and 0.0008 in SSIM with much lower parameters and computations. Additionally, the DESAT also shows an advantage in perceptual quality, achieving the lowest LPIPS (0.2658) among all models, with reductions of 0.0025 and 0.0010 relative to the HAT and HAN, respectively. These results support the DESAT's adaptability and effectiveness at higher magnification rates, demonstrating its great potential for remote sensing super-resolution applications.

**Table 9.** Comparison of 4× super-resolution results on the simulated AID dataset.

| Method | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| HAN | 31.91 | 0.8771 | 0.2668 |
| SwinIR-NG | 31.89 | 0.8758 | 0.2684 |
| HAT | 31.91 | 0.8767 | 0.2685 |
| DESAT (ours) | 31.97 | 0.8779 | 0.2658 |

However, our real-world GF6SRD dataset was constructed specifically for 2× super-resolution tasks to accurately simulate the complex degradation in real-world remote sensing imagery using native-resolution images from different sensors. Thus, this dataset is only of practical significance and application value in the 2× super-resolution tasks. For future work, we will construct real-world cross-sensor datasets that support higher magnification rates, allowing for a further evaluation of the DESAT's performance in real-world high-magnification super-resolution tasks.

### *5.4. Potential of the AEUB on Non-Integer Scale Super-Resolution Tasks*

Most SR models focus on integer scaling factors, which limits flexibility in real-world applications that often require non-integer scaling. For example, in a real-world remote sensing super-resolution task, the multispectral image of the Gaofen-7 satellite with a spatial resolution of 3.2 m is used as the HR image, and the PMS image of the Gaofen-6 satellite with a spatial resolution of 8 m is used as the LR image. This super-resolution task requires a 2.5× upsampling factor. However, traditional models, including those in Section 4, are constrained to integer scales due to their reliance on the pixel shuffle layer.

The attention branch of AEUB has the potential to overcome this by supporting non-integer scaling through an innovative spatially repeated sampling technique. In our approach, the desired upsampling ratio, $\alpha$, can be adjusted by setting the window overlap ratio, $\beta$ = 1-1/$\alpha$. Following the process described in Section 3.2.3, our approach divides input features into overlapping windows, applies distance-enhanced attention to each window, and then merges the results without overlap to create high-resolution outputs at the desired scales.

This design demonstrates the potential of AEUB for non-integer scale super-resolution, suggesting its applicability to real-world remote sensing image resolution tasks requiring various scaling factors. While the pixel branch in AEUB, which uses the pixel shuffle layer, is optimized for integer scaling, future work will explore ways to further enhance its adaptability for non-integer scales, broadening AEUB's applicability across a wider range of upscaling needs.

### *5.5. Limitations of the Study*

First, although the GF6SRD dataset reduces geometric errors by using two sensors from a single satellite, some WFV-PMS image pairs still exhibit geometric errors within two pixels. This may slightly affect the dataset's effectiveness for remote sensing SR applications. Second, while the DSAB significantly enhances performance, it also increases computational complexity. To address this, DSABs and HABs are used in combination within the DESAT to balance performance and efficiency. Future work will focus on developing a lightweight DSAB that can maintain performance while reducing computational demands.

## 6. Conclusions

In this study, we propose a distance-enhanced strip attention transformer for remote sensing super-resolution to address critical limitations of the existing SR methods. To better capture spatial correlations between ground objects, we designed a distance-enhanced strip attention block (DSAB), which incorporates the first law of geography by integrating distance prior into a strip window attention mechanism. This approach allows the DESAT to more accurately reconstruct spatial textures with spatial correlations while expanding its receptive field. Additionally, to address bottlenecks caused by traditional upsample modules, we proposed an attention enhance upsample block (AEUB), which combines the pixel shuffle layer with overlapping window distance-enhanced attention, effectively transferring deep features into HR outputs. We further introduced a real-world cross-sensor super-resolution dataset, GF6SRD, constructed with PMS and WFV imagery from the Gaofen-6 satellite. This high-quality dataset provides a valuable resource for remote sensing applications.

Comparison experiments on both the simulated AID dataset and the real-world GF6SRD dataset show that the DESAT achieves competitive results, with substantial improvements on the GF6SRD dataset, outperforming state-of-the-art SR methods in both quantitative and qualitative evaluations. This demonstrates the DESAT's effectiveness, particularly in challenging real-world scenarios where traditional methods often encounter limitations. Spectral validation experiments further highlight the DESAT's strong capability in learning and transforming spectral information, while migration studies confirm its robust generalization across different geographic regions and time periods. Future research will focus on developing lightweight SR models and constructing hybrid SR datasets incorporating imagery from various satellites, further enhancing the applicability of super-resolution techniques across a broader range of remote sensing tasks.

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

**Funding:** This work is supported by the program of the National Natural Science Foundation of China (grant number: 62101531, 61731022), the Second Tibetan Plateau Scientific Expedition and Research Program (grant number: 2019QZKK030701), and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant number: XDA19090300).

**Data Availability Statement:** The data of experimental images used to support the findings of this research are available from the corresponding author upon reasonable request. The code and part of our GF6SRD dataset is available at https://doi.org/10.5281/zenodo.14058724.

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

## References

1. Sishodia, R.P.; Ray, R.L.; Singh, S.K. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sens.* **2020**, *12*, 3136. https://doi.org/10.3390/rs12193136.
2. Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-Cover Classification with High-Resolution Remote Sensing Images Using Transferable Deep Models. *Remote Sens. Environ.* **2020**, *237*, 111322. https://doi.org/10.1016/j.rse.2019.111322.
3. Peng, X.; He, G.; Wang, G.; Yin, R.; Wang, J. A Weakly Supervised Semantic Segmentation Framework for Medium-Resolution Forest Classification with Noisy Labels and GF-1 WFV Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4412419. https://doi.org/10.1109/TGRS.2024.3404953.
4. Yang, R.; He, G.; Yin, R.; Wang, G.; Zhang, Z.; Long, T.; Peng, Y.; Wang, J. A Novel Weakly-Supervised Method Based on the Segment Anything Model for Seamless Transition from Classification to Segmentation: A Case Study in Segmenting Latent Photovoltaic Locations. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103929. https://doi.org/10.1016/j.jag.2024.103929.
5. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. https://doi.org/10.1016/j.isprsjprs.2019.11.023.
6. Lepcha, D.C.; Goyal, B.; Dogra, A.; Goyal, V. Image Super-Resolution: A Comprehensive Review, Recent Trends, Challenges and Applications. *Inf. Fusion* **2023**, *91*, 230–260. https://doi.org/10.1016/j.inffus.2022.10.007.
7. Freeman, W.T.; Jones, T.R.; Pasztor, E.C. Example-Based Super-Resolution. *IEEE Comput. Graph. Appl.* **2002**, *22*, 56–65. https://doi.org/10.1109/38.988747.
8. Sun, J.; Zhu, J.; Tappen, M.F. Context-Constrained Hallucination for Image Super-Resolution. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 231–238.
9. Kim, K.I.; Kwon, Y. Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1127–1133. https://doi.org/10.1109/TPAMI.2010.25.
10. Yang, W.; Zhang, X.; Tian, Y.; Wang, W.; Xue, J.-H.; Liao, Q. Deep Learning for Single Image Super-Resolution: A Brief Review. *IEEE Trans. Multimed.* **2019**, *21*, 3106–3121. https://doi.org/10.1109/TMM.2019.2919431.
11. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
12. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. https://doi.org/10.1109/TPAMI.2015.2439281.
13. 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.
14. 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.
15. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21-26 July 2017; pp. 1132–1140.
16. 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.
17. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 294–310.
18. Niu, B.; Wen, W.; Ren, W.; Zhang, Xiangde; Yang, L.; Wang, S.; Zhang, K.; Cao, Xiaochun; Shen, H. Single Image Super-Resolution via a Holistic Attention Network. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 191–207.
19. 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, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
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. 9992–10002.
21. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
22. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 7242–7252.

23. 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. 12294–12305.
24. 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 October 2023; pp. 12278–12287.
25. 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.
26. 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.
27. Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X. Recursive Generalization Transformer for Image Super-Resolution. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024.
28. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Lin, C.-W.; Zhang, L. TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-Resolution. *IEEE Trans. Image Process.* **2024**, *33*, 738–752. https://doi.org/10.1109/TIP.2023.3349004.
29. Zhang, W.; Tan, Z.; Lv, Q.; Li, J.; Zhu, B.; Liu, Y. An Efficient Hybrid CNN-Transformer Approach for Remote Sensing Super-Resolution. *Remote Sens.* **2024**, *16*, 880. https://doi.org/10.3390/rs16050880.
30. Shang, J.; Gao, M.; Li, Q.; Pan, Jinfeng; Zou, G.; Jeon, G. Hybrid-Scale Hierarchical Transformer for Remote Sensing Image Super-Resolution. *Remote Sens.* **2023**, *15*, 3442. https://doi.org/10.3390/rs15133442.
31. Chen, H.; He, X.; Qing, L.; Wu, Y.; Ren, C.; Sheriff, R.E.; Zhu, C. Real-World Single Image Super-Resolution: A Brief Review. *Inf. Fusion* **2022**, *79*, 124–145. https://doi.org/10.1016/j.inffus.2021.09.005.
32. Zabalza, M.; Bernardini, A. Super-Resolution of Sentinel-2 Images Using a Spectral Attention Mechanism. *Remote Sens.* **2022**, *14*, 2890. https://doi.org/10.3390/rs14122890.
33. Tobler, W. On the First Law of Geography: A Reply. *Ann. Assoc. Am. Geogr.* **2004**, *94*, 304–310. https://doi.org/10.1111/j.1467-8306.2004.09402009.x.
34. 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. https://doi.org/10.1109/TGRS.2017.2685945.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Galar, M.; Sesma, R.; Ayala, C.; Albizua, L.; Aranda, C. Super-Resolution of Sentinel-2 Images Using Convolutional Neural Networks and Real Ground Truth Data. *Remote Sens.* **2020**, *12*, 2941. https://doi.org/10.3390/rs12182941.
37. Salgueiro Romero, L.; Marcello, J.; Vilaplana, V. Super-Resolution of Sentinel-2 Imagery Using Generative Adversarial Networks. *Remote Sens.* **2020**, *12*, 2424. https://doi.org/10.3390/rs12152424.
38. Zhao, J.; Ma, Y.; Chen, F.; Shang, E.; Yao, W.; Zhang, S.; Yang, J. SA-GAN: A Second Order Attention Generator Adversarial Network with Region Aware Strategy for Real Satellite Images Super Resolution Reconstruction. *Remote Sens.* **2023**, *15*, 1391. https://doi.org/10.3390/rs15051391.
39. Lai, W.-S.; Huang, J.-B.; Ahuja, N.; Yang, M.-H. Fast and Accurate Image Super-Resolution with Deep Laplacian Pyramid Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2599–2613. https://doi.org/10.1109/TPAMI.2018.2865304.
40. Cui, Y.; Knoll, A. Dual-Domain Strip Attention for Image Restoration. *Neural Netw.* **2024**, *171*, 429–439. https://doi.org/10.1016/j.neunet.2023.12.003.
41. Tsai, F.-J.; Peng, Y.-T.; Lin, Y.-Y.; Tsai, C.-C.; Lin, C.-W. Stripformer: Strip Transformer for Fast Image Deblurring. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 146–162.
42. Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; Van Gool, L. Efficient and Explicit Modelling of Image Hierarchies for Image Restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 18278–18289.
43. Shi, W.; Caballero, J.; Huszar, 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.
44. Kong, X.; Zhao, H.; Qiao, Y.; Dong, C. ClassSR: A General Framework to Accelerate Super-Resolution Networks by Data Characteristic. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12011–12020.
45. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
46. 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. 9195–9204.