

Article

SymSwin: Multi-Scale-Aware Super-Resolution of Remote Sensing Images Based on Swin Transformers

Dian Jiao ¹, Nan Su ^{1,2,*} , Yiming Yan ^{1,2} , Ying Liang ³, Shou Feng ^{1,2} , Chunhui Zhao ^{1,2} and Guangjun He ³ 

¹ College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China; sameen404@tju.edu.cn (D.J.); yanyiming@hrbeu.edu.cn (Y.Y.); fengshou@hrbeu.edu.cn (S.F.); zhaochunhui@hrbeu.edu.cn (C.Z.)

² Key Laboratory of Advanced Marine Communication and Information Technology, Harbin Engineering University, Ministry of Industry and Information Technology, Harbin 150001, China

³ State Key Laboratory of Space-Earth Integrated Information Technology, Beijing Institute of Satellite, Information Engineering, Beijing 100095, China; liangying@spacestar.com.cn (Y.L.); hgjun_2006@163.com (G.H.)

* Correspondence: sunan08@hrbeu.edu.cn

Abstract: Despite the successful applications of the remote sensing image in agriculture, meteorology, and geography, its relatively low spatial resolution is hindering the further applications. Super-resolution technology is introduced to conquer such a dilemma. It is a challenging task due to the variations in object size and textures in remote sensing images. To address that problem, we present SymSwin, a super-resolution model based on the Swin transformer aimed to capture a multi-scale context. The symmetric multi-scale window (SyMW) mechanism is proposed and integrated in the backbone, which is capable of perceiving features with various sizes. First, the SyMW mechanism is proposed to capture discriminative contextual features from multi-scale presentations using corresponding attentive window size. Subsequently, a cross-receptive field-adaptive attention (CRAA) module is introduced to model the relations among multi-scale contexts and to realize adaptive fusion. Furthermore, RS data exhibit poor spatial resolution, leading to insufficient visual information when merely spatial supervision is applied. Therefore, a U-shape wavelet transform (UWT) loss is proposed to facilitate the training process from the frequency domain. Extensive experiments demonstrate that our method achieves superior performance in both quantitative metrics and visual quality compared with existing algorithms.

Keywords: super-resolution; remote sensing; multi-scale representations; Swin transformers; wavelet transform loss



Citation: Jiao, D.; Su, N.; Yan, Y.; Liang, Y.; Feng, S.; Zhao, C.; He, G. SymSwin: Multi-Scale-Aware Super-Resolution of Remote Sensing Images Based on Swin Transformers. *Remote Sens.* **2024**, *16*, 4734. <https://doi.org/10.3390/rs16244734>

Academic Editor: Salah Bourennane

Received: 11 November 2024

Revised: 9 December 2024

Accepted: 16 December 2024

Published: 18 December 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

Satellite remote sensing (RS) images are widely applied in diverse domains such as smart agriculture, environmental monitoring, and natural disaster forecasting [1]. The accurate representation of texture in remote sensing imagery is fundamental for the effective interpretation and extraction of valuable information, thus, making high-resolution remote sensing images critical for subsequent tasks such as detection [2,3], recognition [4], and segmentation [5]. However, the accessibility of high-resolution satellite imagery remains severely constrained, not only due to limitations in imaging technology and sensor equipment, but also due to commercial monopolies over such data. Consequently, super-resolution (SR) reconstruction techniques emerged as a crucial approach for enhancing the resolution of satellite imagery. This has become a prominent area of research within the field of RS image processing.

SR intends to reconstruct high-resolution (HR) images from low-resolution (LR) images, which requires full utilization of the limited information. Through the migration of SR algorithms originally designed for natural images, it is possible to achieve the reconstruction of RS images. Current SR algorithms for natural images can be broadly categorized

into two primary approaches: learning the affine relationships between LR and HR pixels, and aimed pixels, generating high-resolution pixels from a latent space. The first category encompasses both traditional methods and algorithms based on deep learning-based models, such as convolutional neural networks (CNNs) [6,7] and transformers [8,9], while the second category primarily involves generative models, including generative adversarial networks (GANs) [10–12] and diffusion models [13,14]. However, remote sensing (RS) data present distinct challenges compared to natural images, as it is captured from significant observation distances, resulting in a wide field of view, abundant content, and relatively low resolution. Consequently, generative SR methods are not ideal for RS images, as they tend to introduce artifacts that are not consistent with the ground truth [15], owing to the limited feature information present in LR RS images. CNN-based models primarily focus on local pixel neighborhoods. As a result, they struggle to capture long-range dependencies between distant pixels, which are crucial for achieving effective remote sensing image super-resolution (RSISR) [16]. In contrast, transformer-based models, due to their self-attention mechanism, are better equipped to model global dependencies and thereby enhance the performance of RSISR [17].

Although transformers demonstrated remarkable performance in super-resolution tasks, their application to remote sensing images remains limited. Transformers are capable of capturing long-range pixel dependencies. However, vision transformers [17] without optimization are not well suited for remote sensing image super-resolution (RSISR) tasks due to their high computational cost. To address this issue, SR algorithms incorporating transformers introduced the shifted window (Swin) mechanism [8], which helps alleviate the computational burden. These models generally follow a three-stage architecture: shallow feature extraction, deep feature extraction, and image reconstruction, with Swin transformers primarily responsible for implementing the function of deep feature extraction. Most models based on Swin-based models utilize a fixed window size [8,9,18–20], which results in a limited receptive field and fails to appropriately capture the multi-scale characteristics inherent in RS data. This constraint prevents the model from applying the appropriate scale of attention to various features in the image. To overcome this limitation, the hierarchical transformer [18] was proposed, offering an increasing size of neighborhood size. However, it overlooks the correlation between contextual features.

To address the unique challenges posed by remote sensing (RS) data, we propose an innovative multi-scale-aware Swin transformer-based architecture for single remote sensing image super-resolution (SRSISR), named SymSwin. We also leverage frequency representations through an auxiliary loss function.

The backbone of our network incorporates the symmetric multi-scale window (SyMW) mechanism on our backbone, which plays a crucial role in extracting accurate contextual information from multi-scale representations. This mechanism enables the network to utilize diverse receptive fields at both shallow layers and deep layers, providing sufficient context for the model to capture texture and semantics with various sizes. In addition, we introduce the cross-receptive field-adaptive attention (CRAA) module in every block of the backbone to enhance feature representation. The CRAA module consists of two components: cross-receptive field attention (CRA) and adaptive feed-forward (AFF). CRA computes the cross-covariance of hybrid-size contexts, while AFF enhances the network's ability to focus on strongly correlated features. This integration enables the model to achieve precise scale handling and detail enhancement. Furthermore, we employ a U-shape wavelet transform (UWT) loss, which supervises the training utilizing frequency domain features. While wavelet losses have been explored previously and proven to be effective for training transformer-based super-resolution models [19], our UWT loss facilitates more precise frequency comprehension, leading to improved performance.

Integrating the innovations mentioned above, we present SymSwin, a multi-scale-aware super-resolution model for a single remote sensing image based on a transformer. SymSwin demonstrates superior performance compared to existing state-of-the-art

transformer-based SR models across multiple evaluation metrics on the RESISC45 dataset and Dior dataset.

To conclude, the paper presents three key contributions that can be summarized as follows:

- We propose the symmetric multi-scale window (SyMW) mechanism to functionalize the backbone with the capability of capturing multi-scale characteristics brought by RS data and generating more precise contexts.
- We introduce the cross-receptive field-adaptive attention (CRAA) module to every block of our backbone to model the dependencies across multi-scale representations, effectively enhancing the information.
- In addition, we train SymSwin with an innovating U-shape wavelet transform (UWT) loss. The UWT aims to leverage frequency features to facilitate more effective image restoration.

The rest of this article is organized as follows: Section 2 reviews some remarkable designs that lay the foundation and inspire our work. Section 3 describes the detailed construction of the SymSwin model. Section 4 introduces the implementation of experiments and evaluates the achieved results. Section 5 provides analysis of the performance of the proposed method. Section 6 makes a conclusion of the entire paper.

2. Related Works

In this section, we first provide a general review of mainstream super-resolution (SR) algorithms based on transformers, highlighting their limitations. Consequently, we introduce some works that inspired our innovation, including the modification to backbone and the auxiliary loss.

2.1. Transformer-Based Image SR

A key advantage of transformers is their ability to adaptively capture long-range dependencies between image patches, making them highly effective for tasks that require global attention. As a result, transformers have been increasingly explored for low-level vision tasks such as super-resolution (SR). These methods often outperform traditional CNN-based approaches [20].

Current transformer-based SR methods based on transformers normally would adopt a shifted-window mechanism [21] to reduce computational complexity and introduce locality into self-attention. A notable example is SwinIR [8], which forms the foundation for image restoration using shifted-window self-attention (SW-SA). Following this, several models leveraged Swin transformers for image super-resolution, often incorporating various optimization techniques. For instance, HAT [22] and DAT [23] enhance the original model by integrating convolution operations and adding channel attention to complement the spatial attention mechanism. Meanwhile, SRformer [9] and SPIN [24] focus on improving computational efficiency, with SRformer reducing model size by permuting the window sizes into the channel dimension, and SPIN using pixel clustering to replace square windows.

However, these methods typically rely on fixed window sizes, leading to limiting the receptive field and hindering the extraction of multi-scale contextual information. RS images present unique challenges due to their complex spatial structures, dense target distributions, variable object shapes, and poor resolution. These characteristics make RS images particularly difficult to handle with conventional SR metrics [25]. As a result, the inability of these transformer-based methods to capture a multi-scale context can significantly restrict their performance when applied to RS data.

To address these challenges, we propose an innovative Swin transformer-based backbone designed specifically for SRSISR. Our model effectively exploits the multi-scale context in RS images while maintaining an efficient parameter count.

2.2. Multi-Scale Representation Mechanism in Single RS Image SR

Multi-scale processing involves sampling a signal at varying granularities, allowing the network to capture different features at each scale. Recent approaches increasingly focused on exploiting the exploration of multi-scale properties of remote sensing imagery. Enhancing the network's ability to understand features of varying sizes has been shown to significantly improve performance in SRSISR tasks [26].

TransENet [27] proposed a transformer-based multistage enhancement framework to integrate multi-scale low-dimensional and high-dimensional features. In this architecture, transformers are used to extract features at different levels. The network consists of multiple encoders, enabling the embedding of multi-level features during the feature extraction process. However, the algorithm relies on the full self-attention mechanism of vanilla transformers, which leads to quadratic increases in computational complexity with the number of patches [28]. The introduction of the Swin transformer backbone mitigates this issue to a large extent, trading more calculation redundancy for adding multi-scale frameworks.

TransENet [27] devised a transformer-based multistage enhancement structure to fuse multi-scale low-dimension and high-dimension features, where transformers are introduced to obtain features at different levels. The network is composed of multiple encoders and multiple decoders, realizing the embedding of the multilevel features in the feature extraction and fusing these encoded features with adaptive adjustment. Unfortunately, the algorithm employs the full self-attention in vanilla transformers, increasing the computational complexity quadratically with the number of patches [28]. The emergence of the Swin transformer backbone alleviates the computational burden to a great extent, trading more calculation redundancy for adding multi-scale frameworks.

The adoption of Swin transformer backbones in SRSISR methods has grown substantially. For example, TTST [29] modifies the conventional MLP layer and introduces a multi-scale feed-forward layer that incorporates convolutions with kernels of different sizes to generate a richer set of features. Similarly, MSGFormer [16] employs a group of multi-branch convolution operations to generate weights for the cascaded attention. By incorporating convolutions with varying kernel sizes, these methods can capture diverse spatial features across multiple scales and fuse the information. Notably, smaller convolution kernels excel at capturing local, detailed image information, while larger kernels, with their wider receptive fields, capture the global structure and layout of the image. Nevertheless, the strategy to parallel convolutions with large kernel size increases computational demands significantly.

To address the limitations of capturing multi-scale features through an overreliance on convolution operations, we propose a novel mechanism. Our approach extracts hierarchical information without the need for additional convolutions. Furthermore, we introduce a module to fuse hybrid-level feature maps, enhancing feature expression, which is inserted at each layer of the backbone.

2.3. SR Methods Combining with Wavelet Transform

Wavelet transform leverages its flexibility in capturing both frequency information and spatial position. It converts an image from the pixel domain to the frequency domain, resulting in four sub-bands: a low-frequency band and three high-frequency bands. The low-frequency band captures coarse-grain information, such as object contours, while the high-frequency bands preserve fine-grain information in an image as details, including edges. Due to the ability of wavelet coefficients to effectively preserve high-frequency image details, wavelet-based methods for tackling super-resolution tasks gained increasing attention [30,31].

The wavelet loss function [32] was proposed to guide autoencoders by emphasizing the space frequency characteristics of images. In this approach, the wavelet transform decomposes the image into components at different frequency levels, and the loss function constrains these components. Specifically, it uses mean squared error (MSE) loss to constrain

the low-frequency components, while also applying modulus constraints to the high-frequency components to retain fine-grain details. Wavelet-based texture adversarial loss [33] aims to reconstruct more visually plausible textures by focusing on the high-frequency sub-bands of generated images. By leveraging the multi-scale representation and invertibility of wavelet transform, this method enhances the perceptual quality of the reconstructed image. JWSGN [34] combines discrete wavelet transform with a multi-branch network to recover each frequency sub-band separately. This approach guides the super-resolution process by exploring the interactions between sub-bands. Additionally, JWSGN effectively calibrates high-frequency sub-bands, preventing the generation of incorrect textures. WFEN [35] integrates the wavelet transform mechanism into the feature extraction encoder–decoder structure, minimizing the distortion of high-frequency features during direct down-sampling and reducing fusion aliasing during up-sampling. By harnessing the characteristics of the wavelet domain, WFEN achieves robust performance without the need for excessive module stacking or redundant spatial feature collection.

Building on these advancements, we assume that incorporating frequency features could significantly benefit remote sensing (RS) image super-resolution, given the inherent spatial feature challenges in RS images. Therefore, we propose an auxiliary loss function based on wavelet transform, designed to improve the training by frequency supervision.

3. Methodology

In this section, we provide a detailed description of the proposed SymSwin network, beginning with an overview of its overall architecture. We then analyze two core components crucial to multi-scale representation capabilities: the SyMW mechanism and the CRAA module. Lastly, we present a comprehensive explanation of the UWT loss scheme.

3.1. Overview of SymSwin Architecture

Inheriting the general structure of Swin transformer based SR, our SymSwin is composed with three primary parts: shallow feature extraction G_S , deep feature extraction G_D , and image reconstruction G_R , as illustrated in Figure 1a.

The shallow feature extraction module consists of a basic convolutional layer, embedding pixels into features that transform raw pixel values into feature representations that are suitable for deep learning models. Specifically, given the low-resolution input $I_{LR} \in R^{H \times W \times 3}$, the shallow feature $F_S \in R^{H \times W \times C}$ is extracted via the convolutional layer. H and W denote the height and width for the input image, C represents the channel dimension of the embedded feature. The structure has been proven to be functional in providing stable optimization in early visual processing [36], which can be expressed as:

$$F_S = G_S(I_{LR}) = \text{Conv}_{3 \times 3}(I_{LR}). \quad (1)$$

The high-dimensional feature is subsequently passed through a deep feature extraction module comprising N cascading SymSwin groups (SymG). Each SymG group consists of a Swin transformer block, which employs symmetric multi-scale window sizes (SyMWB), along with a cross-receptive field-adaptive attention (CRAA) module. At the end of the deep feature extraction pipeline, a standard convolutional layer is applied to map the extracted features into an intermediate feature space, facilitating the reconstruction process. Therefore, the feature extraction process can be described as follows:

$$F_i = \text{SymG}_i(F_{i-1}) = \text{CRAA}_i(\text{SyMWB}_i(F_{i-1})), i = 1, 2, \dots, N, F_0 = F_S, \quad (2)$$

$$F_D = \text{Conv}_{3 \times 3}(F_N). \quad (3)$$

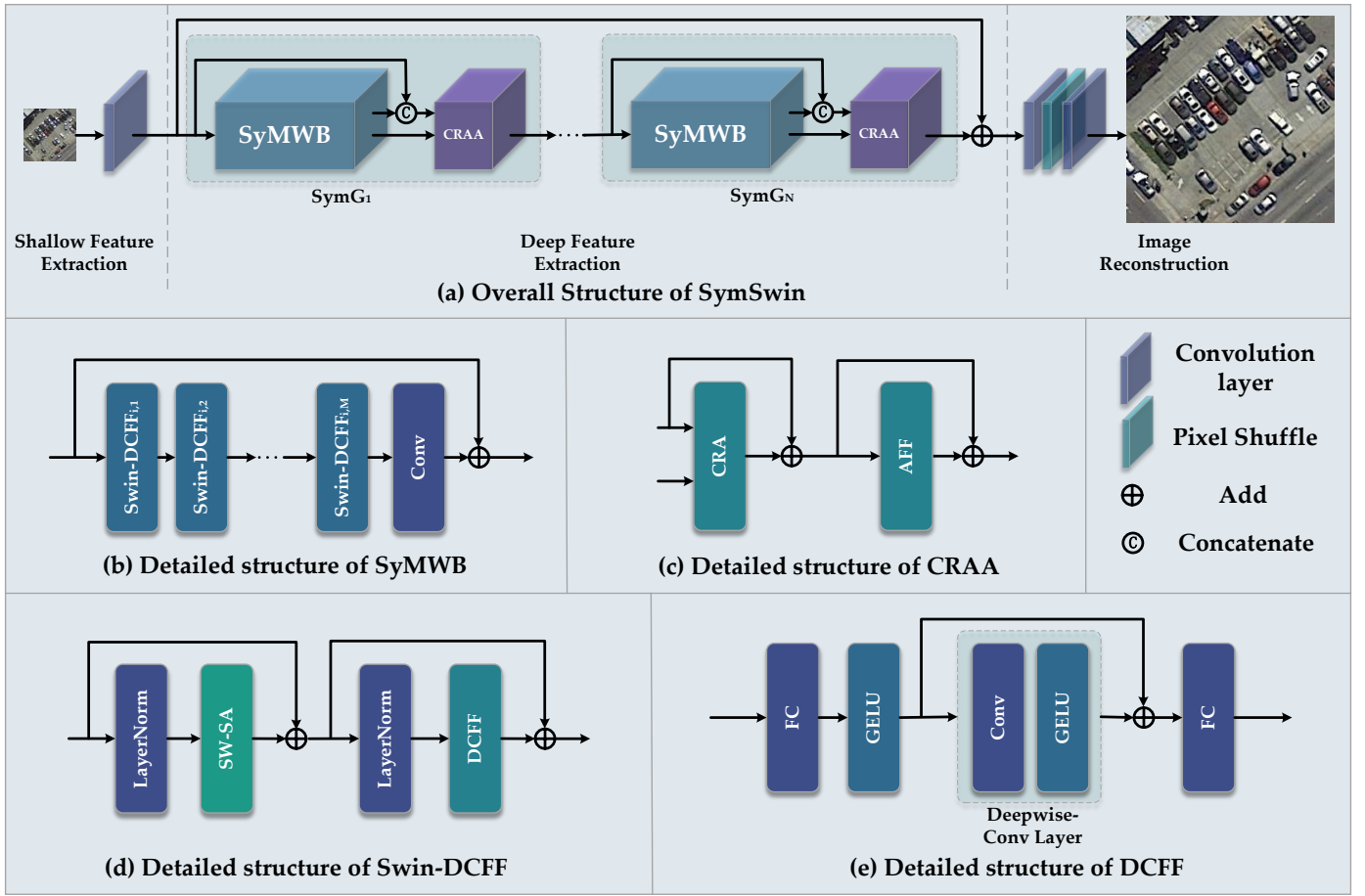


Figure 1. (a) Overall architecture of SymSwin, containing three main functional stages. The chief deep feature extraction stage involves SyMWBs and CRAAs. (b) Detailed illustration of SyMWB composition. (c) Detailed illustration of the CRAA module. (d) Detailed illustration of the Swin-DCFF layer. SW-SA denotes conventional shifted-window self-attention. (e) Detailed illustration of DCFF.

The SyMWB architecture consists of a cascade of M Swin-DCFF attention layers followed by a convolutional layer integrated with a residual connection, functioning as:

$$F_{i,j} = \text{Swin} - \text{DCFF}(F_{i,j-1}), j = 1, 2, \dots, M, F_{i,0} = F_{i-1},$$

$$F_i = \text{Conv}_{3 \times 3}(F_{i,M}) + F_{i-1},$$
(4)

$$\text{SyMWB}_i(F_{i-1}) = \text{Conv}_{3 \times 3}(F_{i,M}) + F_{i-1}.$$
(5)

Integrating a convolutional layer into the traditional feed-forward network enhances the inductive bias of the convolution operation within transformer-based architectures, thereby providing a stronger foundation for subsequent feature aggregation. Specifically, each Swin-DCFF attention layer combines vanilla shifted-window self-attention (SW-SA) with a modified multi-layer perceptron (MLP), where a depth-wise convolution is introduced between the two fully connected (FC) layers.

Eventually, the high-resolution image is generated by passing the aggregated deep feature F_D and shallow feature F_S through an image reconstruction module, which involves groups of convolutions and pixel shuffle operations [37]. The reconstruction process can be denoted as:

$$I_{SR} = G_R(F_D + F_S),$$
(6)

where $I_{SR} \in R^{HS \times WS \times 3}$ represents the recovery image, and S stands for the up-sampling scale. The structure of the image reconstruction module is determined by the up-sampling scale.

We optimize the parameters of SymSwin utilizing a hybrid weighting of L_1 loss and U-shape wavelet transform (UWT) loss as

$$L = \lambda \|I_{SR}, I_{HR}\|_1 + \mu L_{UWT}(I_{SR}, I_{HR}), \quad (7)$$

where $\|\cdot\|_1$ denotes the L_1 distance between the prediction I_{SR} and the corresponding ground truth I_{HR} , and λ, μ are hyperparameters.

3.2. SymSwin Backbone with Multi-Scale Representations

To enhance the network's ability to capture multi-scale features in RS images, we propose two key modifications to the original Swin transformer backbone. First, we introduce a general mechanism, SyMW, which allows the Swin transformer to capture context at multiple receptive field sizes. Second, we incorporate a CRAA module after each feature extraction block, enabling the joint processing of multi-scale representations and improving the expression of deep features.

3.2.1. Symmetric Multi-Scale Window (SyMW) Mechanism

Conventional Swin transformer-based backbones use a fixed window size across all residual Swin transformer blocks, which leads to limited feature extraction and reduced effectiveness in RS image processing. To address this issue, we propose a symmetric multi-scale window pattern that leverages informative multi-scale features and improves contextual understanding in RS images. Specifically, the distribution of window sizes in each block follows a symmetric sequence, as illustrated in Figure 2.

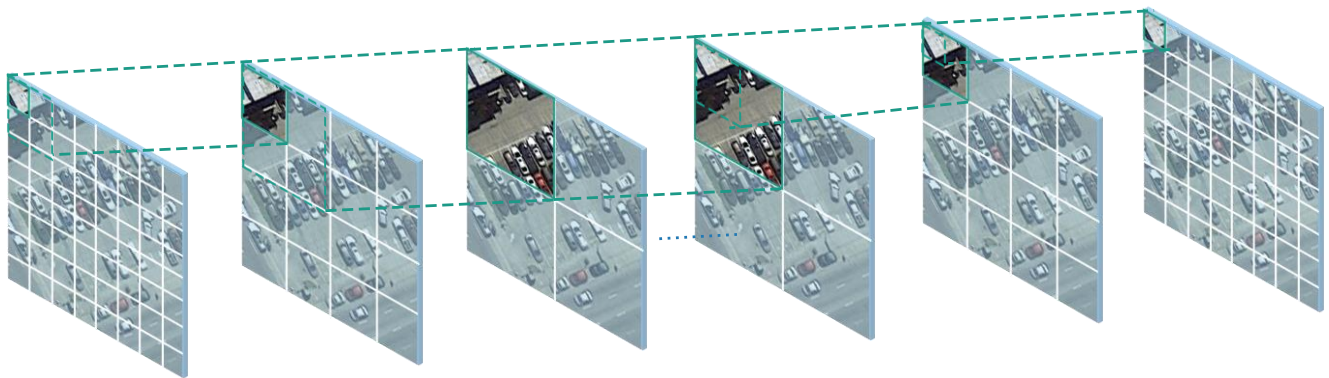


Figure 2. Indication of SyMW mechanism.  denotes window for SyMWB_{*i*},  denotes window for SyMWB_{*i*+1}, and  denotes feature map of SyMWB_{*i*}. Each feature map represents the extraction of a whole block. The grid denotes the window size used on each feature map. The illustration intuitively demonstrates the SyMW can provide multi-scale context.

The initial layers of deep learning networks typically capture shallow features, such as textures and basic shapes. Therefore, we configure smaller window sizes in these early layers to effectively capture local dependencies. Due to the ultra-long shooting distance of RS images, certain fine-grained features persist even in the later layers. To maintain the model's attention to these fine details, we use smaller window sizes in the final layers as well. Moreover, the window size should vary gradually rather than randomly, following a pattern of first increasing and then decreasing. Shallow features form the foundation, while deeper features enhance the model's discriminative ability. Since both types of features are equally important, we adopt a symmetric approach, where the window sizes expand and contract in a consistent, double-scale manner.

We adjust the window sizes on a block-by-block basis, and the window sizes in each block are complied with:

$$W_k = \begin{cases} W_s \cdot 2^{k-1}, & k \leq (N|2) \\ W_s \cdot 2^{N-k}, & k > (N|2) \end{cases}, k = 1, 2, \dots, N, \quad (8)$$

where W_s stands for the window size of the initial block, N stands for the total number of the SyMWB, and operator $|$ denotes integer division. In our case, W_s is set to 8 and N is set to 6, consistent with the configuration of mainstream SR algorithms.

Accordingly, given the feature map $F_i \in R^{H \times W \times C}$, we first split F_i into $N_{W_k} = \frac{HW}{W_k^2}$ non-overlapping square windows $X \in R^{N_{W_k} \times W_k^2 \times C}$, and the partitioned feature is then fed into the Swin-DCFF attention layer. To be more concrete, the SW-SA layer first computes the standard self-attention separately for each window. We obtain the corresponding query, key and value matrices through projection linear L_Q , L_K and L_V :

$$Q = L_Q(X), K = L_K(X), V = L_V(X). \quad (9)$$

Calculating the relevance between query and key, and multiplying with V , we obtain the attention matrix, formulated as:

$$Attn(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V, \quad (10)$$

where $\sqrt{d}$ is a defined scalar [17] and B is the learnable relative positional embedding [21]. Following the SW-SA, we cascade a MLP with an extra depth-wise convolutional layer, named DCFF, to better assist in encoding details. Since self-attention can be viewed as a low-pass filter, adding such convolution operation compensates for the loss of high-frequency information, and meanwhile, increases nearly no computations. The LayerNorm (LN) layer is inserted before both SW-SA function and DCFF function, and the residual connection is employed for both modules. The process can be expressed as follows:

$$X = SW - SA(LN(X)) + X, \quad (11)$$

$$X = DCFF(LN(X)) + X. \quad (12)$$

The non-overlapping partition strategy lacks the perception of connection across local windows, so that shifted window partitioning is introduced to enable such perception [21]. In our case, we alternatively configure the shift-window strides to window size and half of the window size at each successive layer.

3.2.2. Cross-Receptive Field-Adaptive Attention (CRAA) Module

In remote sensing imagery, large acquisition distances result in significant scale variations across the images. To effectively enhance information at different scales, inspired by SPIFFNet [38], we introduce CRAA, which is capable of modeling interdependencies across multi-scale representations. This capability makes CRAA an essential component for the remote sensing image super-resolution (RSISR) task, where precise scale handling and detail enhancement are critical for optimal performance.

While the contours and textures in RS images can vary across multiple scales, their expressions in the channel dimension remain stable. The core idea behind CRAA is to achieve multi-scale representation interaction through attention operation along channel dimension. Notably, channel-wise attention is utilized to better facilitate sub-pixel convolution [37], which is ultimately employed for high-resolution image reconstruction. The rationale behind this approach is to leverage periodic shuffling along the channel dimension to mitigate data gaps in the spatial domain.

The CRAA module consists of two components: the cross-receptive field attention (CRA) for biased fusion of multi-scale representations and the adaptive feed-forward network (AFF) to emphasize informative features. Together, the module captures the cross-covariance between multi-scale representations via channel attention, enhancing information by integrating feature maps from both current and previous groups. A detailed illustration of the CRAA architecture is provided in Figure 3.

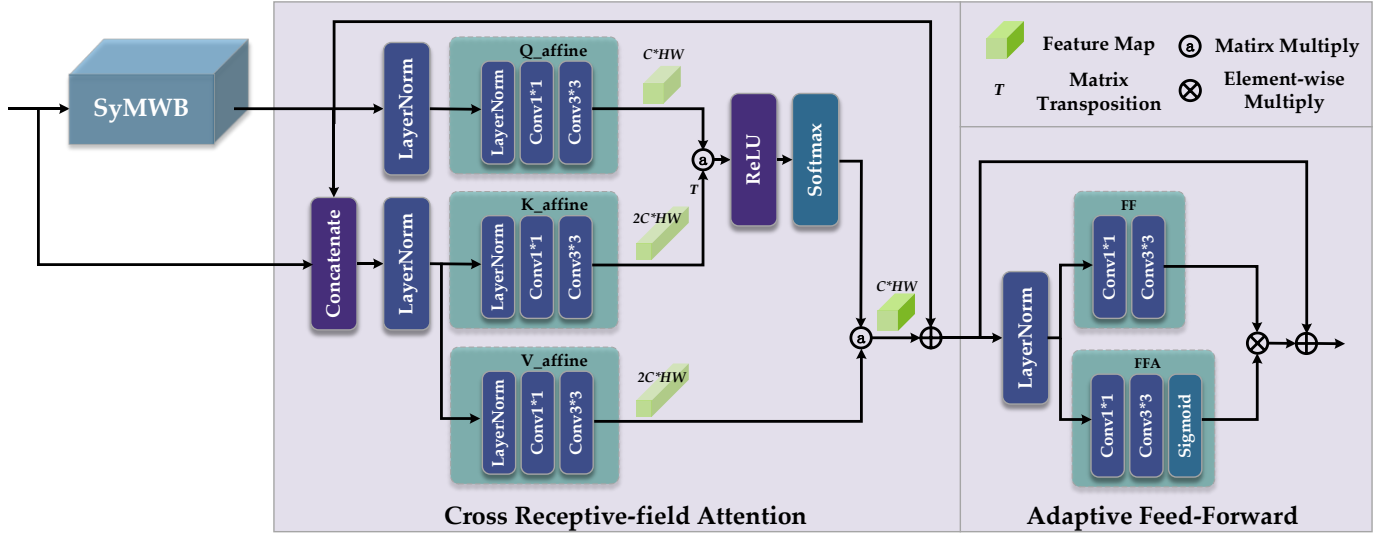


Figure 3. Illustration of the CRAA module, containing two main functional stages. During the CRA stage, we calculate the correlation between context with different receptive fields and achieve flexible fusion. During the AFF stage, we adaptively enhance the fusion feature.

By using the feature map extracted from the current SyMWB as the query, and combining it with the feature map extracted from the previous SymG as a key value pair, we achieve dynamic fusion through cross-attention. Specifically, with the support of the SyMW mechanism, these feature maps represent information at different scales, which are related to each other by a factor of two (doubling) or one-half (halving).

To elaborate, let $F_{pre} = F_{i-1} \in R^{H \times W \times C}$ and $F_{cur} = SyMWB_i(F_{i-1}) \in R^{H \times W \times C}$ represent the input feature maps. Initially, we concatenate these features along the channel dimension to obtain the fused feature map, $F_f \in R^{H \times W \times 2C}$. Both the current feature map and the fused feature map undergo a layer normalization (LN) operation before being passed through the attention embedding.

In contrast to standard linear projection techniques, the CRA mechanism utilizes a sequence of LN and convolutional layers to perform the mapping function, as described below:

$$Q = Conv_{3 \times 3}(Conv_{1 \times 1}(LN(LN(F_{cur})))), \quad (13)$$

$$K = Conv_{3 \times 3}(Conv_{1 \times 1}(LN(LN(F_f)))), \quad (14)$$

$$V = Conv_{3 \times 3}(Conv_{1 \times 1}(LN(LN(F_f)))). \quad (15)$$

Then the matrices are reshaped to $Q \in R^{C \times HW}$ and $K, V \in R^{2C \times HW}$ to facilitate their production, resulting in a cross-context enhanced attention map $F_A \in R^{C \times HW}$, which can be expressed as:

$$F_A = CAttn(Q, K, V) = Softmax(ReLU(QK^T))V, \quad (16)$$

where $CAttn(\cdot)$ denotes attention operating channel wise, weighting the importance of image features through channel information. This attention mechanism enables the computation of cross-covariance across channels, producing an attention map that implicitly

encodes the global context [39]. A *ReLU* activation function is introduced before the *softmax* normalization to enhance feature control and promote the development of sophisticated image attributes. Its non-linear nature provides a sparse constraint, focusing on more informative regions.

Subsequently, the attention map is normalized and then passed into the AFF module to adaptively enhance the relevant features while suppressing irrelevant ones. The AFF consists of two parallel convolutional branches: the first performs a standard feed-forward operation, while the second computes a dynamic weight map. Element-wise multiplication is used to apply these weights, a simple yet effective operation that regulates the flow of complementary features and facilitates feature transformation [40]. Additionally, a residual connection is incorporated within the AFF, as is common in feed-forward networks integrated with attention mechanisms. Thus, given the input attention feature F_A , the manipulation performed within the AFF is defined as:

$$F_i = \text{Conv}_{1 \times 1}(\text{Conv}_{3 \times 3}(\text{Conv}_{1 \times 1}(\text{LN}(F_A)))) \cdot \sigma(\text{Conv}_{3 \times 3}(\text{Conv}_{1 \times 1}(\text{LN}(F_A)))) + F_A, \quad (17)$$

where $\sigma(\cdot)$ means sigmoid activation function.

3.3. U-Shape Wavelet Transform (UWT) Loss

Factors such as ultra-long shooting distances, adverse weather conditions, severe noise interference, and relative motion during image acquisition contribute significantly to the degradation of RS images. In light of these challenges, leveraging frequency domain representations has been shown to facilitate more effective image restoration. We integrate wavelet transform into the training process to better support the regression and improve image restoration performance. The wavelet transform, in particular, demonstrated its efficacy in enhancing transformer-based SR models, owing to its distinctive capability to capture frequency features while preserving spatial positional information. We employ the classic stationary wavelet transform (SWT) [41] to realize the transformation from the pixel domain to the spectral domain. Additionally, the Symlet wavelet is selected for the SWT decomposition to ensure the preservation of image structures and smoothness.

Explicitly, given an RGB domain image $I \in R^{H \times W \times 3}$, we first convert it into the YCrCb domain to extract luminance information $Y \in R^{H \times W \times 1}$ before frequency transformation. What is worth mentioning is that only the luminance feature is required for wavelet transform since color information is less practical in the frequency domain. The process of SWT is illustrated in Figure 4.

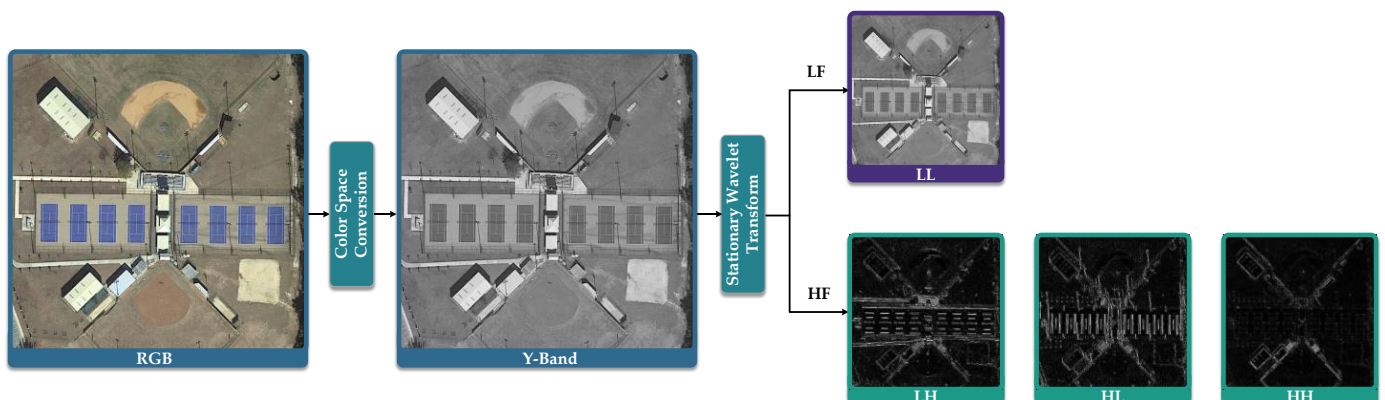


Figure 4. Illustration of the SWT process. The color space conversion converts an image from RGB space to YCrCb space, and we select the Y-band value, representing the luminance information. LF denotes the low-frequency sub-band, and HF denotes high-frequency sub-bands. The sketches of HF directly depict the horizontal, vertical, and diagonal direction edges.

SWT first employs two filters, decomposition filter and reconstruction filter, to divide the spectral composition of the input into high-frequency sub-bands and a low-frequency (LL) sub-band. LL represents general content in the image with coarse grain. High-frequency components include LH sub-band, HL sub-band, and HH sub-band, among which the LH sub-band depicts horizontal-direction detailed information, HL sub-band depicts vertical-direction detailed information, and HH sub-band depicts diagonal-direction detailed information.

When applying SWT in UWT loss function, we calculate the distance between the reconstruction image and the high-resolution image in each sub-band and sum the distances by weights, which can be formulated as:

$$L_{UWT} = \mu_t D_t, t = LL, LH, HL, HH, \quad (18)$$

where μ_t is a hyperparameter set empirically, for instance, we set $\mu_t = [0.05, 0.025, 0.025, 0.02]$ following SWT loss [19].

Notably, we define the distance here as:

$$D_t = \begin{cases} 0.5(\|sub_t(Y_{SR}), sub_t(Y_{HR})\|_1)^2, & \|sub_t(Y_{SR}), sub_t(Y_{HR})\|_1 < 1 \\ \|sub_t(Y_{SR}), sub_t(Y_{HR})\|_1 - 0.5, & \|sub_t(Y_{SR}), sub_t(Y_{HR})\|_1 \geq 1 \end{cases}, \quad (19)$$

where $sub_t(\cdot)$ stands for the function of obtaining the t sub-band in SWT operation, and $\|\cdot\|_1$ stands for L_1 distance between the prediction and the ground truth in each sub-band. As is well known, the L_1 loss function exhibits superior convergence properties compared to the quadratic function in cases where the differences are large [39]. Hence, we reserve the proportional character of L_1 loss when the difference is larger than 1, alleviating the problem of unstable training at the early stage when the reconstruction image is far from a high-resolution image. However, when the difference is between 0 and 1, the gradient of L_1 loss becomes too fast for the late training period; moreover, it is not differentiable at the 0 point, which may lead to gradient explosion. Therefore, we modified the linear character into exponential characteristics within the interval from 0 to 1, namely L_2 loss, offering a progressively smaller derivative and smooth curve at the 0 point. Additionally, to ensure that the two segments intersect and remain differentiable at the point 1, we applied an offset to the linear function and performed a scaling operation on the exponential function, which are solved through the following equations:

$$\begin{cases} y_1 = \alpha x^2 \\ y_2 = x + \beta \\ y_1(1) = y_2(1) \\ \frac{dy_1}{dx}|_{x=1} = \frac{dy_2}{dx}|_{x=1} \end{cases}. \quad (20)$$

The overall shape of the entire function resembles the letter U; hence, we designate it as UWT. With the advantage mentioned above, the proposed UWT can better facilitate the regression process through a more stable loss convergence and force the model to find a more precise global optimum.

We combine the proposed UWT loss with L_1 loss to compose the whole loss for training, which can be formulated as:

$$L = \lambda_1 L_1 + \lambda_U L_{UWT}, \quad (21)$$

where λ_1 and λ_U are both hyperparameters.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

In this paper, we conduct experiments on two public satellite remote sensing datasets: the NWPU-RESISC45 dataset [42] and the DIOR dataset [43]. The aforementioned datasets are commonly used in SR tasks, thereby making the experimental results more compelling.

The NWPU-RESISC45 dataset was published in 2017 by Northwestern Polytechnical University, containing 45 categories of varied scenes in a baseball diamond, circular farmland, dense residential, mobile home park, railway station, thermal power, etc. Each category involves 700 samples with a size of 256×256 pixels. The spatial resolution of the dataset ranges from 0.2 m to 30 m each pixel. We randomly choose nine images in each category to compose the test set, leaving the rest images serving as the train set.

The DIOR dataset was published in 2019, also by Northwestern Polytechnical University, containing 20 categories of objects in a basketball court, expressway service area, ground track field, storage tank, windmill, harbor, etc. All categories collectively include 23,463 samples with a size of 800×800 pixels. The spatial resolution of the dataset ranges from 0.5 m to 30 m each pixel. According to the official configuration, we directly utilize the pre-defined train set containing 11,725 images and the pre-defined test set containing 11,738 images.

We primarily focus on super-resolution reconstruction with the specific upscaling factors $\times 2$, $\times 3$, and $\times 4$. Particularly, the low-resolution images with corresponding sizes are down-sampled through bilinear interpolation.

4.1.2. Evaluation Metrics

In this paper, we employ four metrics to realize quantitative evaluation and comparison, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [44], learned perceptual image patch similarity (LPIPS) [45], and CLIPscore [46]. Among which, PSNR and SSIM are mathematical standards calculated with pixel values, and LPIPS and CLIPscore are perceptual standards inferred by developed pretrained deep learning models. The metrics mentioned above are performed on RGB channels as:

$$SSIM(x, y) = [l(x, y)]^{\gamma_l} \cdot [c(x, y)]^{\gamma_c} \cdot [s(x, y)]^{\gamma_s},$$

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \gamma_l = \gamma_c = \gamma_s = 1, \quad (22)$$

$$LPIPS(I_{SR}, I_{HR}) = (VGG(I_{SR}) - VGG(I_{HR}))^2, \quad (23)$$

$$CLIPscore(I_{SR}, I_{HR}) = Cossim(CLIP(I_{SR}), CLIP(I_{HR})), \quad (24)$$

where x represents the pixel value in the reconstructed image, while y denotes the corresponding pixel value in the high-resolution image. L is the maximum possible pixel value, which is 255 for images in the uint8 format. μ_x and μ_y refer to the mean values of the predicted and ground truth images, respectively, while σ_x and σ_y denote their variances. σ_{xy} represents the covariance between the two images. The constants C_1 , C_2 , and C_3 are small values introduced to avoid division by zero in the denominator. The notation $MES(\cdot)$ represents the mean squared error between two images, $Cossim(\cdot)$ indicates the cosine similarity function, and $VGG(\cdot)$ refers to the feature extraction function of the VGG16 model [47]. Similarly, $CLIP(\cdot)$ denotes the image encoder function from the CLIP model [48].

Notably, the CLIPscore metric is a perceptual quality evaluation specifically designed for RS image applications. Among the metrics, PSNR, SSIM, and CLIPscore have positive correlation with the image reconstruction quality. On the contrary, LPIPS has negative correlation with the image reconstruction quality.

4.1.3. Implementation Details

Conclusively, we set our SymSwin model integrated with six SymSwin groups, each SymSWB contains six Swin-DCFF layers. The embedding dimension is set to 180 for SW-SA operation, the window sizes for every block are 8, 16, 32, 32, 16, and 8, respectively, and the number of attention heads in every block is designated as 6 equally. Regarding the embedding dimension for the convolutional layers in the feed-forward network, we set it to twice as much as is used in the attention mechanism. To ensure that the image size can be divided by different window sizes at various scales, we configure the low-resolution patch size to 64×64 pixels.

As for the training process, we utilize an Adam optimizer [49] with the initial learning rate set to 10^{-4} to facilitate regression. A multi-step learning rate scheduler, reducing the learning rate by half at milestones, is also employed for better assistance, where the milestones are set to 2.5×10^5 , 4×10^5 , 4.5×10^5 , and 4.75×10^5 . The total number of training iterations is 5×10^5 . In terms of the training loss, we set both λ_1 and λ_U to 1.0.

The proposed network and training strategy are implemented using PyTorch. All experiments are conducted on one NVIDIA GeForce RTX 3090 graphics card.

4.2. Comparative Experiments

To better demonstrate the superior performance of the present model, we conducted comparative experiments on the NWPU-RESISC45 dataset and DIOR dataset, comparing our approach with other mainstream SR algorithms based on transformers, namely, SwinIR [8], DAT [23], HAT [22], NGswin [50], SRformer [9], HiT-SR [18], TransENet [27], and TTST [29], in terms of quantitative metrics and visual outcomes. Notably, TransENet and TTST are methods focusing on RS data, others are originally designed for natural images. We retrained all the models applying exactly the same training configuration for a fair evaluation.

4.2.1. Quantitative Results

The quantitative results encompassing various metrics and zooming scales on both datasets are illustrated in Table 1. From the experimental data presented in the table, we observe that our SymSwin model achieves the highest interpretation accuracy on both the NWPU-RESISC45 dataset and the DIOR dataset for SR tasks at $\times 2$, $\times 3$, and $\times 4$ magnifications, as measured by the PSNR and SSIM numerical metrics.

To be specific, on the NWPU-RESISC45 dataset, SymSwin reaches 28.044 dB/0.747 in the $\times 4$ SR task, exceeding the second-best with an increment of 0.285 dB/0.001. Regarding the $\times 2$ SR task and $\times 3$ SR task, our model demonstrates an advantage of 0.007 dB/0.002 and 0.012 dB/0.003 over the second-ranked TransENet. On the DIOR dataset, SymSwin achieves 32.810 dB/0.893 in the $\times 2$ task, surpassing the second-ranked TransENet with a growth of 0.384 dB/0.001. Additionally, in $\times 4$ reconstruction, the proposed model outperforms the second-best method by 0.033 dB/0.001 in PSNR/SSIM. In terms of $\times 3$ magnification, our model achieves a peak PSNR value of 24.994 dB, matching that of TransENet, while also outperforming it by 0.006 in the SSIM metric.

For the $\times 2$ and $\times 4$ tasks, sub-pixel convolutions with the factor of 2 are appended in the model to achieve the desired magnification. In the $\times 3$ task, however, sub-pixel convolution with a factor of 3 is used, which requires more parameters and is more complex than cascading two sub-pixel convolutions with a factor of 2. Consequently, as the reconstruction scale increases from $\times 2$ to $\times 4$, the difficulty of the task follows a non-linear pattern, with the $\times 3$ task being the most challenging, followed by $\times 4$ and then $\times 2$. These trends are reflected in the quantitative metrics shown in the table, where the performance of all algorithms deteriorates accordingly. However, the performance degradation of our algorithm remains minimal. This suggests that encouraging the model to focus on multi-scale contextual information enhances its robustness among multi-magnification tasks. Notably, SymSwin achieves the best performance in both PSNR and SSIM simultaneously, a distinction that is rarely matched by other algorithms.

Table 1. The comparison of PSNR, SSIM, LPIPS, and CLIPscore metrics on the NWPU-RESISC45 dataset and DIOR dataset. All methods are estimated with $\times 2$, $\times 3$, and $\times 4$ reconstruction ratios. The best is highlighted in **bold red** font.

Algorithms	Scales	NWPU-RESISC45				DIOR			
		PSNR	SSIM	LPIPS	CLIPscore	PSNR	SSIM	LPIPS	CLIPscore
SwinIR	2	31.562	0.906	0.117	0.974	32.354	0.892	0.121	0.971
DAT	2	31.210	0.899	0.123	0.973	32.337	0.892	0.121	0.971
HAT	2	31.502	0.905	0.118	0.974	32.362	0.892	0.121	0.971
NGswin	2	31.520	0.906	0.117	0.974	32.325	0.892	0.121	0.971
SRFormer	2	31.531	0.906	0.118	0.975	32.334	0.891	0.121	0.970
HiT-SR	2	31.498	0.905	0.117	0.975	32.351	0.892	0.120	0.971
TransENet	2	31.605	0.904	0.120	0.973	32.426	0.892	0.121	0.969
TTST	2	31.571	0.906	0.117	0.974	32.335	0.892	0.123	0.971
SymSwin(Ours)	2	31.612	0.906	0.118	0.973	32.810	0.893	0.120	0.971
SwinIR	3	23.269	0.679	0.304	0.911	24.766	0.709	0.287	0.934
DAT	3	23.233	0.678	0.301	0.915	24.707	0.709	0.284	0.936
HAT	3	23.169	0.675	0.305	0.913	24.656	0.708	0.287	0.935
NGswin	3	23.099	0.680	0.312	0.888	24.611	0.706	0.289	0.934
SRFormer	3	23.326	0.679	0.305	0.911	24.766	0.709	0.288	0.934
HiT-SR	3	23.194	0.680	0.300	0.914	24.595	0.708	0.284	0.936
TransENet	3	23.591	0.685	0.311	0.902	24.994	0.712	0.293	0.932
TTST	3	23.205	0.678	0.302	0.913	24.713	0.709	0.285	0.935
SymSwin(Ours)	3	23.603	0.688	0.296	0.913	24.994	0.718	0.283	0.936
SwinIR	4	27.694	0.744	0.303	0.876	27.784	0.769	0.267	0.914
DAT	4	27.715	0.745	0.303	0.875	27.988	0.773	0.263	0.916
HAT	4	27.708	0.744	0.303	0.876	27.910	0.771	0.266	0.915
NGswin	4	27.684	0.743	0.303	0.876	27.913	0.770	0.266	0.915
SRFormer	4	27.656	0.743	0.304	0.875	27.868	0.769	0.268	0.917
HiT-SR	4	27.759	0.746	0.299	0.875	24.043	0.674	0.331	0.894
TransENet	4	27.531	0.736	0.308	0.874	27.709	0.764	0.274	0.907
TTST	4	27.716	0.745	0.303	0.875	27.964	0.772	0.264	0.914
SymSwin(Ours)	4	28.044	0.747	0.283	0.891	28.021	0.774	0.264	0.915

Moreover, we performed per-category averaging of the PSNR metric in the $\times 4$ reconstruction task for several algorithms that incorporate typical innovative technologies, according to the official classification of scenes in the NWPU-RESISC45 dataset. The experimental results are presented in Table 2.

It is easily implied that the presented SymSwin have obvious advantages in PSNR over the other five algorithms among all classes in the $\times 4$ SR task. In scenes containing more complex multi-scale textures and objects, our algorithm achieved more significant improvements compared to those in other scenes. For example, SymSwin outperforms the second-best approach by 0.750, 0.730, 0.610, 0.480, and 0.433 in the “Intersection”, “Parking_lot”, “Airplane”, “Harbor”, and “Freeway” scenes, respectively. However, as for the scenes with flat structure information, such as “Mountain”, “Snowberg”, “Sea_ice”, “Lake”, and “Wetland”, our method only achieved enhancements of 0.100, 0.104, 0.107, 0.113, and 0.150 separately.

Table 2. Taking the $\times 4$ SR task as an example, the mean value of PSNR is calculated for each category of scene in the NWPU-RESISC45 dataset for some typical algorithms. The best performance is highlighted in **bold** font.

Class	SwinIR	DAT	HAT	TransENet	TTST	SymSwin
Airplane	27.701	27.714	27.725	27.410	27.725	28.335
Airport	27.951	27.968	27.946	27.822	27.943	28.235
Baseball_diamond	28.809	28.814	28.802	28.555	28.815	29.248
Basketball_court	26.158	26.189	26.169	25.923	26.181	26.660
Beach	29.688	29.708	29.727	29.632	29.722	29.874
Bridge	29.654	29.670	29.666	29.519	29.679	29.861
Chaparral	25.420	25.402	25.450	25.172	25.447	25.753
Church	26.467	26.502	26.514	26.317	26.508	26.859
Circular_farmland	33.179	33.208	33.180	32.913	33.187	33.527
Cloud	33.956	33.969	33.967	33.841	33.978	34.100
Commercial_area	25.045	25.093	25.075	24.927	25.083	25.347
Dense_residential	22.883	22.959	22.895	22.732	22.906	23.410
Desert	30.996	30.998	31.019	30.951	30.997	31.133
Forest	26.180	26.184	26.185	26.119	26.183	26.265
Freeway	27.169	27.163	27.189	26.982	27.191	27.622
Golf_course	30.964	30.978	30.977	30.797	30.967	31.235
Ground_track_field	25.814	25.835	25.828	25.690	25.828	26.068
Harbor	21.758	21.801	21.781	21.585	21.771	22.281
Industrial_area	26.182	26.219	26.187	25.970	26.197	26.643
Intersection	23.077	23.114	23.093	22.926	23.111	23.864
Island	33.053	33.046	33.047	32.906	33.059	33.238
Lake	28.689	28.688	28.695	28.628	28.687	28.808
Meadow	31.773	31.773	31.760	31.714	31.776	31.840
Medium_residential	26.050	26.072	26.079	25.916	26.087	26.459
Mobile_home_park	23.501	23.542	23.517	23.316	23.541	24.024
Mountain	28.925	28.936	28.930	28.854	28.915	29.036
Overpass	26.435	26.520	26.478	26.140	26.525	27.073
Palace	24.440	24.458	24.456	24.262	24.474	24.768
Parking_lot	23.254	23.307	23.268	23.026	23.314	23.998
Railway	25.783	25.805	25.811	25.605	25.810	26.081
Railway_station	26.600	26.621	26.613	26.419	26.616	26.974
Rectangular_farmland	30.696	30.731	30.703	30.474	30.706	31.000
River	29.515	29.525	29.517	29.407	29.518	29.675
Roundabout	25.179	25.214	25.191	25.026	25.207	25.461
Runway	33.034	33.049	33.025	32.733	33.087	33.959
Sea_ice	31.522	31.517	31.529	31.388	31.528	31.736
Ship	29.676	29.677	29.667	29.478	29.694	29.987
Snowberg	23.856	23.862	23.893	23.771	23.875	23.997
Sparse_residential	27.241	27.248	27.261	27.141	27.262	27.398
Stadium	27.405	27.422	27.442	27.184	27.439	27.819
Storage_tank	27.846	27.893	27.867	27.624	27.858	28.363
Tennis_court	26.143	26.170	26.152	26.010	26.176	26.487
Terrace	28.195	28.196	28.219	28.016	28.219	28.508
Thermal_power	27.534	27.615	27.565	27.342	27.578	27.997
Wetland	30.822	30.821	30.821	30.709	30.829	30.979
Average	27.694	27.715	27.708	27.531	27.716	28.044

4.2.2. Qualitative Results

To intuitively evaluate the performance of our proposed model, the visual results of different models with multiple reconstruction scales on both datasets are presented in this section. We select some typical examples to explain the problem more clearly. Due to the insufficient visual distinction in $\times 2$ reconstruction, we opt to present the super-resolution results at $\times 3$ and $\times 4$ magnification levels, respectively, in Figures 5 and 6.

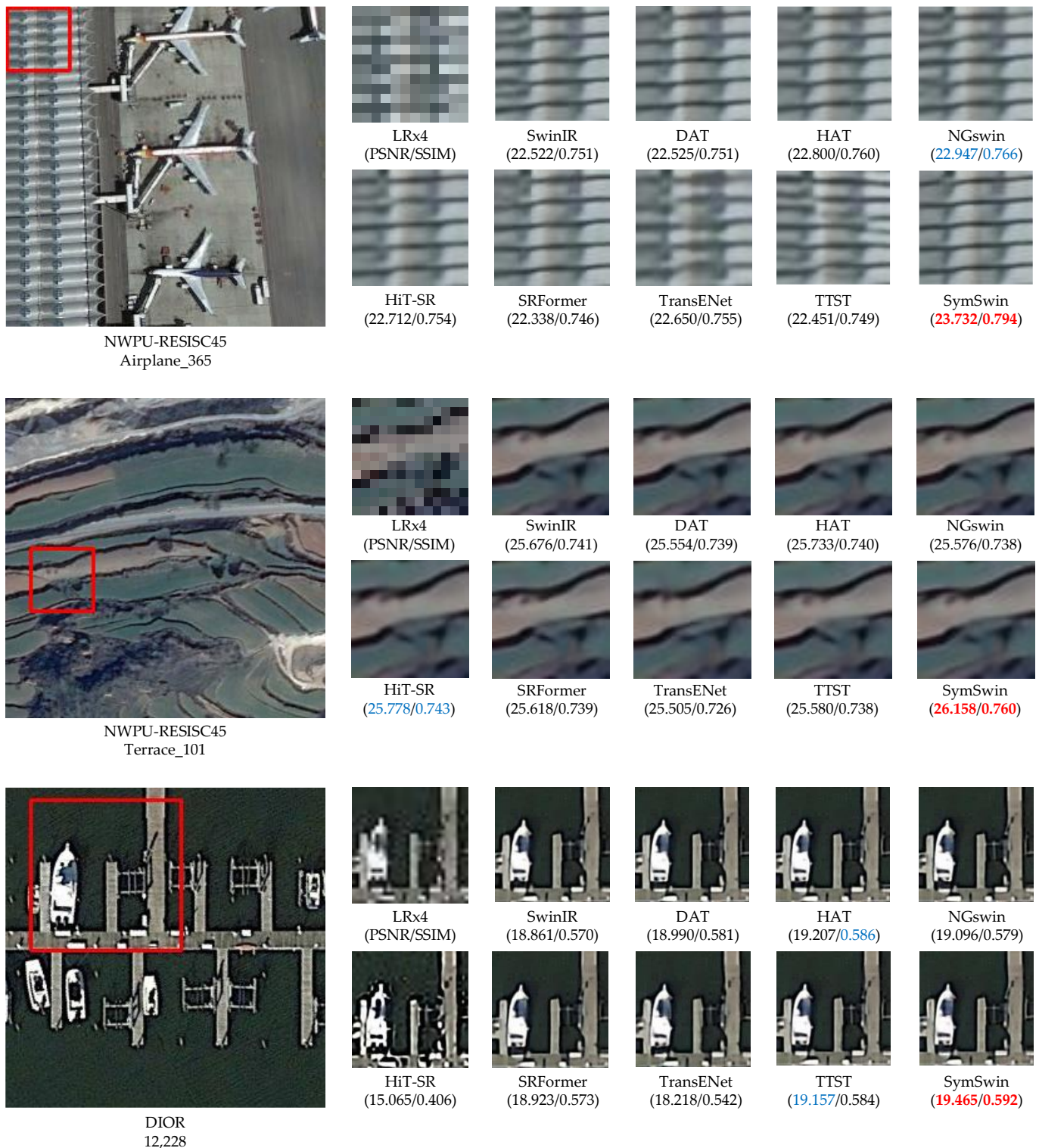


Figure 5. Cont.

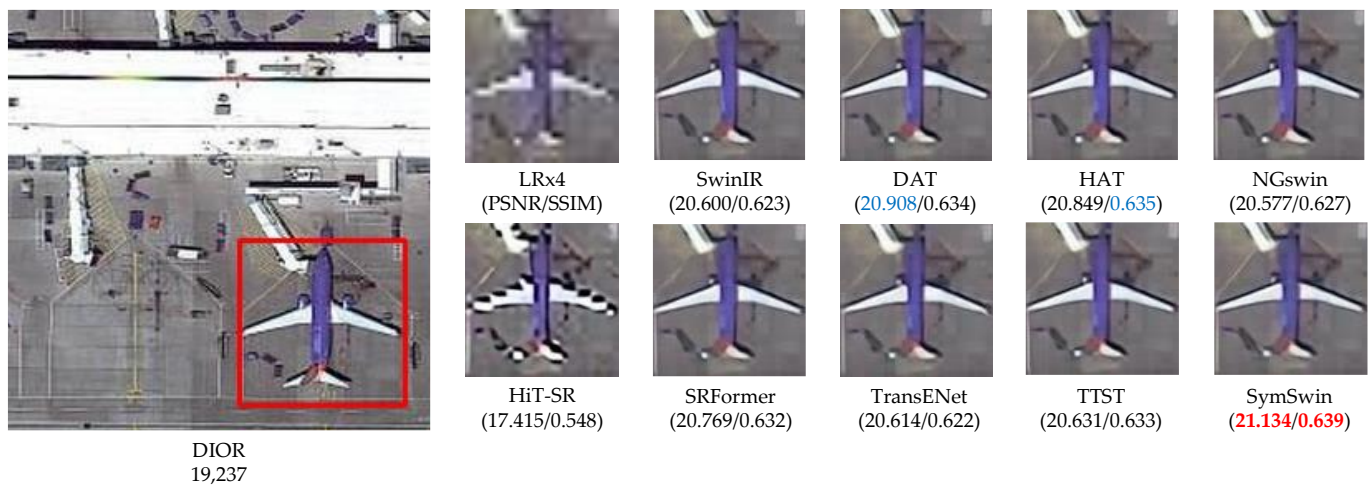


Figure 5. The visualization examples of the $\times 4$ super-resolution reconstruction inference results for the algorithms mentioned in the quantitative experiments on datasets NWPU-RESISC45 and DIOR. The values PSNR and SSIM are listed below each patch, the best performance is highlighted in **bold red** font and the second-ranked is highlighted in **blue** font. The inset on the right is a magnified view of the region enclosed by the red bounding box in the main image. Zoom in for better observation.

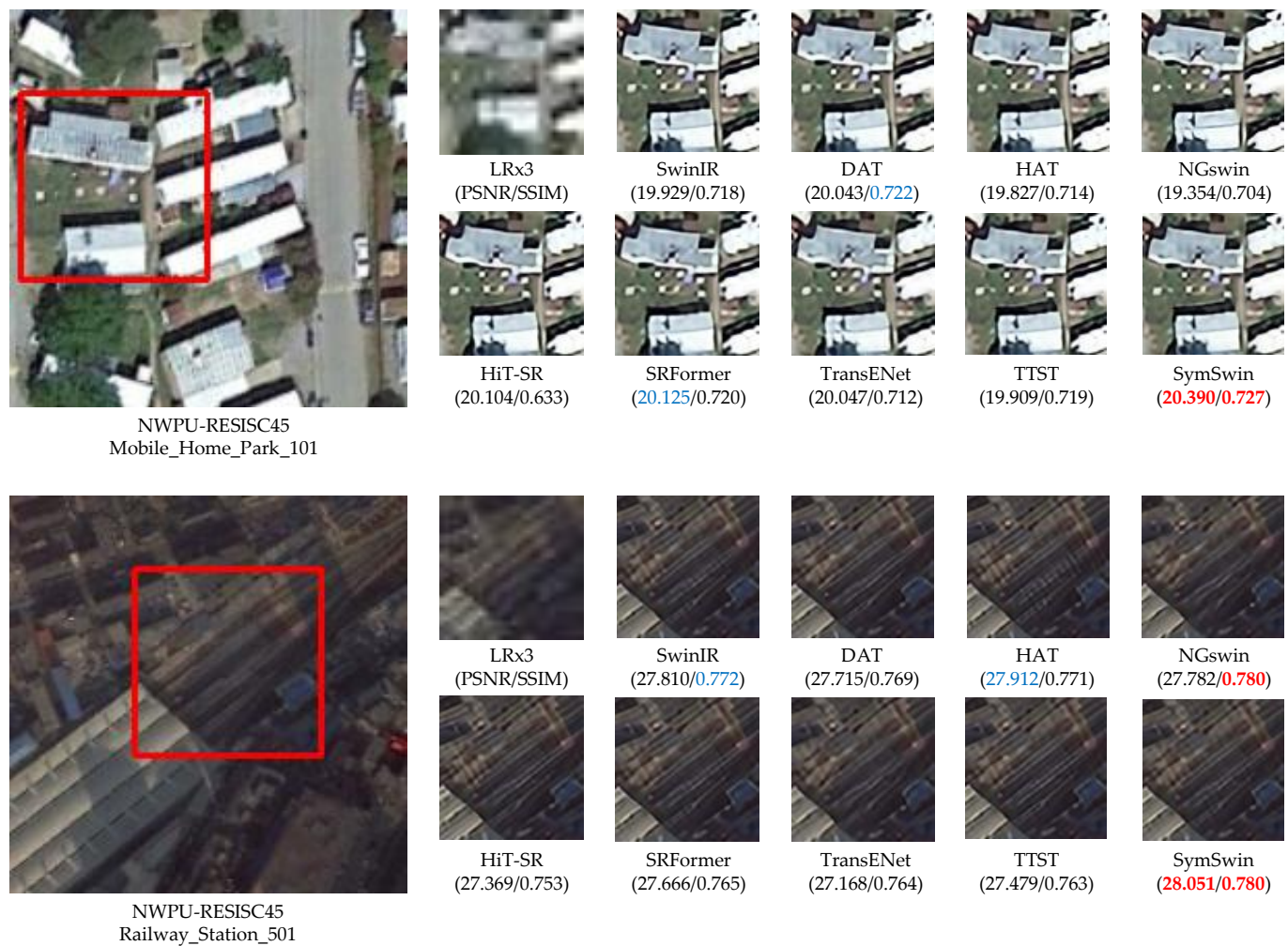


Figure 6. Cont.



Figure 6. The visualization examples of the $\times 3$ super-resolution reconstruction inference results for the algorithms mentioned in the quantitative experiments on datasets NWPU-RESISC45 and DIOR. The values PSNR and SSIM are listed below each patch, the best performance is highlighted in **bold red** font and the second-ranked is highlighted in **blue** font. The inset on the right is a magnified view of the region enclosed by the red bounding box in the main image. Zoom in for better observation.

We found our model demonstrating an advantage in fine-grained texture, making less mistakes in reconstructing slender and dense shapes. The present algorithm is also capable of offering more smooth edges. In example NWPU-RESISC45 Airplane 365, SymSwin managed to maintain the structure of the roof, while other algorithms result in blurry edges or wrong constructions. Example DIOR 18,337 can also prove the efficiency of SymSwin, where other models all generate an extra line in the middle of the jetty, which is not in accordance with the truth. Furthermore, our network indicates a certain denoising capability, as illustrated in example DIOR 19,840, where the inference results of our algorithm exhibit fewer ringing artifacts in the circular region of the image compared to other algorithms.

4.3. Ablation Studies

To further illustrate the effectiveness of the three innovations proposed in our method in improving the network's interpretation accuracy, we conducted ablation studies on the $\times 4$ SR task, using the NWPU-RESISC45 dataset. We evaluate the contribution of components SyMW, CRAA, and UWT loss, utilizing the same four evaluation metrics (PSNR, SSIM, LPIPS, and CLIPscore). The experimental results are presented in Table 3.

Table 3. Ablation experiments on NWPU-RESISC45 of the two multi-scale context focusing strategies and a wavelet transform-based loss function. Parameters of networks with different compositions are counted and listed below. Among the table, *w/o* denotes without, and *w* denotes with.

Model Params	Model0 (Base) 11.900 M	Model1 12.560 M	Model2 14.713 M	Model3 15.035 M	Model4 (SymSwin) 15.035 M
SyMW	<i>w/o</i>	<i>w</i>	<i>w/o</i>	<i>w</i>	<i>w</i>
CRAA	<i>w/o</i>	<i>w/o</i>	<i>w</i>	<i>w</i>	<i>w</i>
UWT	<i>w/o</i>	<i>w/o</i>	<i>w/o</i>	<i>w/o</i>	<i>w</i>
PSNR	27.694	27.749	27.707	27.776	28.044
SSIM	0.744	0.747	0.744	0.747	0.747
LPIPS	0.303	0.301	0.303	0.300	0.283
CLIPscore	0.876	0.874	0.876	0.875	0.891

According to the statistics, all three innovations are proved to be beneficial for elevating the reconstruction ability of the model, with each innovation contributing to varying degrees across different metrics. Taking PSNR as an example, the SyMW mechanism can improve the performance with a 0.055 dB increment, the CRAA module can increase 0.013 dB, and the UWT loss can boost an additional 0.244 dB based on the backbone with multi-scale representations, while the SSIM metric is only raised by the SyMW mechanism with a 0.003 increment, since SyMW provides extra structure context for the network. On the other hand, the UWT loss shows its significant utility in perceptual indicators, as CLIPscore grows 0.016 and LPIPS lowers 0.017.

4.4. Visual Demonstration of Multi-Scale Representation

To more intuitively demonstrate the role that integrating multi-scale representations to the feature extraction network plays in the understanding of context, we extracted the outputs of each RSTB [8] and SymG and visualized the feature maps. For the $\times 4$ SR task, considering the reconstruction mechanism of pixel shuffle, we uniformly sampled the feature maps along the channel dimension to produce five attention heatmaps for each, thereby representing the network's interpretation methodology. The visualization result is depicted as Figure 7.

Among the six feature extraction layers, our SymSwin focused on both background and objects since the network distributes more attention on background in layer0, layer1, and layer4, while distributing more attention on objects in other layers. On the contrary, the base network seems to stick to the objects with similar shape and feature during the whole inference process. This unique character of spreading attention on various scales of features is due to the supplementary information captured by multi-scale representations and cross-context perceiving strategies. Therefore, the proposed model denotes the capability to better understand the information of the image, leading to stronger reconstruction ability.

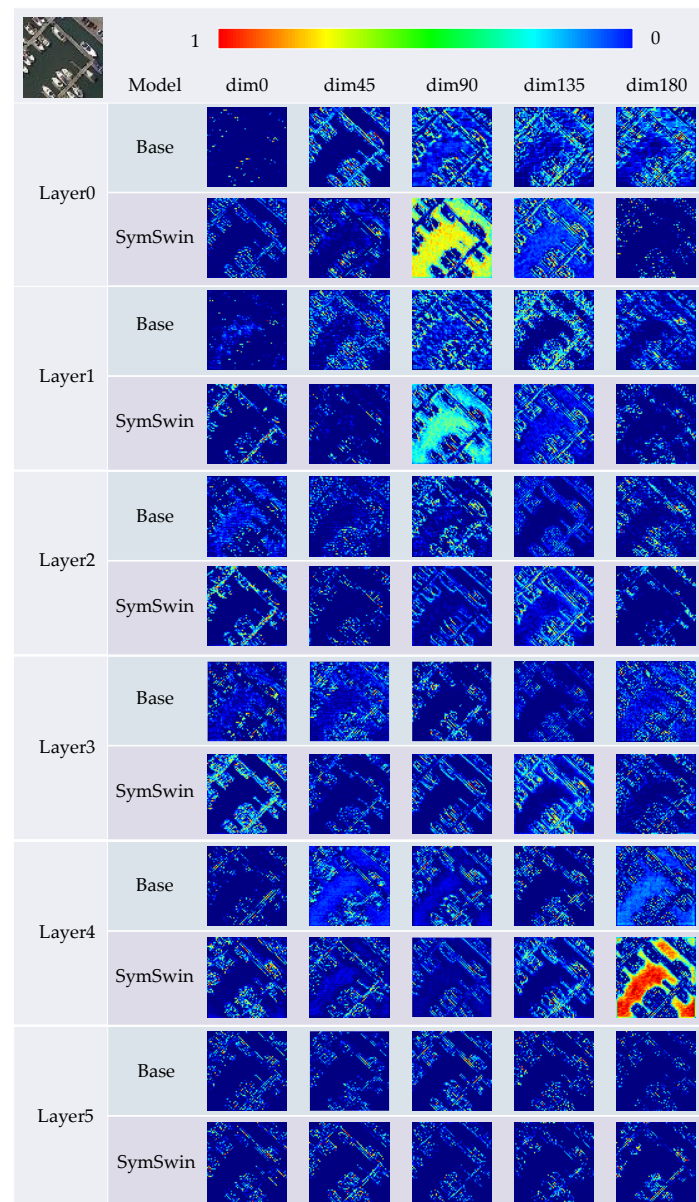


Figure 7. A comparison of the visualized feature maps extracted by each layer of the backbone with and without multi-scale representations, illustrating the different regions of interest the nets tend to focus on. The color closer to red denotes the stronger attention.

5. Discussion

Despite the promising performance SymSwin achieves, there exhibits limitations in terms of calculation efficiency, which is listed in Table 4. Thus, to further optimize the functionality of SymSwin, based on the structural characteristics of the Swin transformer backbone, we introduced the lightweight permute window strategy [9] to reduce the number of network parameters and computational complexity. We conduct a comparative experiment on the NWPU-RESISC45 dataset to observe the efficiency of the lightweight version of the proposed method, as demonstrated in Table 5. By integrating the permute-window mechanism, we managed to reduce the parameters of the network by about 14% for each scale while remaining the PSNR fluctuate within 0.8%. As for the $\times 4$ and $\times 3$ situation, the light version of SymSwin manages to keep superiority to other comparison methods, which even elevates PSNR/SSIM for 0.190 dB/0.004 under the $\times 3$ circumstance.

Table 4. The parameters and the FLOPs of all comparison algorithms with all scales. The differences in model sizes and calculation costs among different methods mainly lie in the pixel shuffle operation.

Model	Parameters ($\times 4/\times 3/\times 2$)	FLOPs ($\times 4/\times 3/\times 2$)
SwinIR	11.900 M/11.937 M/11.752 M	50.546 G/48.836 G/48.045 G
DAT	11.212 M/11.249 M/11.064 M	46.618 G/44.907 G/44.117 G
HAT	20.572 M/20.609 M/20.424 M	85.707 G/83.997 G/83.207 G
NGswin	14.672 M/14.709 M/14.524 M	51.246 G/49.536 G/48.745 G
SRFormer	10.440 M/10.477 M/10.292 M	44.580 G/42.869 G/42.079 G
HiT-SR	10.418 M/10.455 M/10.270 M	47.300 G/45.590 G/44.800 G
TransENet	9.404 M/9.441 M/9.256 M	12.536 G/8.804 G/6.569 G
TTST	18.367 M/18.403 M/18.219 M	76.842 G/75.132 G/74.341 G
SymSwin	15.035 M/15.072 M/14.887 M	63.046 G/61.336 G/60.545 G

Table 5. The performances of both SymSwin and SymSwin-Light on the NWPU-RESISC45 dataset among different SR scales, evaluated with PSNR, SSIM, LPIPS, and CLIPscore metrics. The parameters and FLOPs of different versions of SymSwin are also listed for a more intuitive comparison.

Model	Scale	Parameters	FLOPs	PSNR	SSIM	LPIPS	CLIPscore
SymSwin	4	15.035 M	63.046 G	28.044	0.747	0.283	0.891
	3	15.072 M	61.336 G	23.603	0.688	0.296	0.913
	2	14.887 M	60.545 G	31.612	0.906	0.118	0.973
SymSwin-Light	4	12.905 M	54.988 G	28.019	0.758	0.284	0.887
	3	12.942 M	53.277 G	23.793	0.692	0.301	0.906
	2	12.757 M	52.487 G	31.415	0.905	0.118	0.973

Though the light version of SymSwin made progress in computational expense, the parameters of the model and the FLOPs of the inference are still larger than some transformer-based SR algorithms. Additionally, the weight reduction strategy leads to losing the optimality in $\times 2$ SR task. Future efforts will continue to focus on the lightweight optimization of the proposed algorithm to enhance the inference efficiency and adaptability to a broader range of application scenarios.

On the other hand, SymSwin failed to achieve the satisfying performance in perceptual metrics sometimes while reaching the best on mathematical standards, indicating that the generation of the model may be overly smooth in the low-frequency regions and overly sharpened in the high-frequency regions. There are developed solutions to this issue in other tasks. Generative tasks based on the generative adversarial network (GAN) [51] tend to add random noise to introduce rich details [52], inspiring us to explore how to integrate this approach into reconstruction models based on transformers.

6. Conclusions

In this paper, we propose SymSwin, a novel transformer-based SR network designed to enhance multi-scale representations and cross-context perception in RS data. SymSwin demonstrates superior performance in both visual effects and interpretation accuracy. By integrating the SyMW mechanism into a Swin transformer structure, the network is capable of leveraging informative multi-scale features, improving contextual understanding in RS images. Subsequently, CRAA is incorporated to achieve multi-scale representation interaction through attention operation along channel dimension, facilitating precise scale handling and detail enhancement. In the domain of remote sensing, where images often contain large-scale variations and diverse content, the ability to interpret multi-scale contextual information leads to significantly improved reconstruction quality. To further

improve training, we introduce a U-shaped wavelet transform loss function that operates supervision in the frequency domain, addressing the challenge of limited high-definition supervision in RS imagery. Experiments on two publicly available satellite RS datasets validate the effectiveness of the proposed algorithm, demonstrating its ability to outperform existing methods both qualitatively and quantitatively.

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

Funding: This work is supported by the National Natural Science Foundation of China: No. 62271159, No. 62071136, No. 62002083, No. 61971153, No. 62371153; Excellent Youth Foundation of Heilongjiang Province of China: YQ2022F002; Fundamental Research Funds for the Central Universities: 3072024XX0805; Heilongjiang Province key research and development project: GA23B003; Key Laboratory of Target Cognition and Application Technology: 2023-CXPT-LC-005.

Data Availability Statement: The training codes and the structure codes of the algorithm will be available at https://github.com/SamJ404/SymSwin_master, accessed on 10 October 2024.

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

References

- Wang, X.; Yi, J.; Guo, J.; Song, Y.; Lyu, J.; Xu, J.; Yan, W.; Zhao, J.; Cai, Q.; Min, H. A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sens.* **2022**, *14*, 5423. [\[CrossRef\]](#)
- Tang, X.; Zhang, H.; Mou, L.; Liu, F.; Zhang, X.; Xiang, X.; Zhu, X.; Jiao, L. An Unsupervised Remote Sensing Change Detection Method Based on Multiscale Graph Convolutional Network and Metric Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5609715. [\[CrossRef\]](#)
- Liu, C.; Zhang, S.; Hu, M.; Song, Q. Object Detection in Remote Sensing Images Based on Adaptive Multi-Scale Feature Fusion Method. *Remote Sens.* **2024**, *16*, 907. [\[CrossRef\]](#)
- Wu, F.; Duan, J.; Chen, S.; Ye, Y.; Ai, P.; Yang, Z. Multi-target recognition of bananas and automatic positioning for the inflorescence axis cutting point. *Front. Plant Sci.* **2021**, *12*, 705021. [\[CrossRef\]](#)
- Li, X.; Yong, X.; Li, T.; Tong, Y.; Gao, H.; Wang, X.; Xu, Z.; Fang, Y.; You, Q.; Lyu, X. A Spectral-Spatial Context-Boosted Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2024**, *16*, 1214. [\[CrossRef\]](#)
- 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.
- 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 Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
- Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L.; Timofte, R. SwinIR: Image Restoration Using Swin Transformer. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1833–1844.
- Zhou, Y.; Li, Z.; Guo, C.-L.; Bai, S.; Cheng, M.-M.; Hou, Q. SRFormer: Permuted Self-Attention for Single Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 12734–12745.
- Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Real-worldistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
- Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Chen, C.-L.; Qiao, Y.; Tang, X. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW), Munich, Germany, 8–14 September 2018; Volume 11133, pp. 63–79.
- 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. [\[CrossRef\]](#)
- Chitwan, S.; Jonathan, H.; William, C.; Tim, S.; David, J.-F.; Mohammad, N. Image Super-Resolution via Iterative Refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 4710.
- Hshmat, S.; Daniel, W.; Chitwan, S.; David, F. Denoising Diffusion Probabilistic Models for Robust Image Super-Resolution in the Wild. *arXiv* **2023**, arXiv:2302.07864.

15. Xie, L.-B.; Wang, X.-T.; Chen, X.-Y.; Li, G.; Shan, Y.; Zhou, J.-T.; Dong, C. DeSRA: Detect and Delete the Artifacts of GAN-based Real-World Super-Resolution Models. In Proceedings of the Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023; pp. 8561–8572.
16. Lu, Y.-T.; Wang, S.-Z.; Wang, B.-L.; Zhang, X.; Wang, X.-X.; Zhao, Y.-Q. Enhanced Window-Based Self-Attention with Global and Multi-Scale Representations for Remote Sensing Image Super-Resolution. *Remote Sens.* **2024**, *16*, 2837. [[CrossRef](#)]
17. Alexey, D.; Lucas, B.; Alexander, K.; Dirk, W.; Zhai, X.-H.; Thomas, U.; Mostafa, D.; Matthias, M.; Georg, H.; Sylvain, G.; et al. An Image is Worth 16x16 Words Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
18. Zhang, X.; Zhang, Y.-L.; Yu, F. HiT-SR: Hierarchical Transformer for Efficient Image Super-Resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024.
19. Cansu, K.; Ahmet Murat, T. Training Transformer Models by Wavelet Losses Improves Quantitative and Visual Performance in Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), Seattle, WA, USA, 17–21 June 2024; pp. 6661–6670.
20. Boah, K.; Jeongsol, K.; Jong, C.-Y. Task-Agnostic Vision Transformer for Distributed Learning of Image Processing. *IEEE Trans. Image Process.* **2023**, *32*, 203.
21. Liu, Z.; Lin, Y.-T.; Cao, Y.; Hu, H.; Wei, Y.-X.; Zhang, Z.; Lin, S.; Guo, B.-N. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021.
22. Chen, X.-Y.; Wang, X.-T.; Zhou, J.-T.; Qian, 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, 18–22 June 2023; pp. 22367–22377.
23. Chen, Z.; Zhang, Y.-L.; Gu, J.-J.; Kong, L.-H.; Yang, X.-K.; Yu, F. Dual Aggregation Transformer for Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 12312–12321.
24. Zhang, A.-P.; Ren, W.-Q.; Liu, Y.; Cao, X.-C. Lightweight Image Super-Resolution with Superpixel Token Interaction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 12728–12737.
25. Qin, Y.; Wang, J.-R.; Cao, S.-Y.; Zhu, M.; Sun, J.-Q.; Hao, Z.-C.; Jiang, X. SRBPSwin: Single-Image Super-Resolution for Remote Sensing Images Using a Global Residual Multi-Attention Hybrid Back-Projection Network Based on the Swin Transformer. *Remote Sens.* **2024**, *16*, 2252. [[CrossRef](#)]
26. Xiao, Y.; Su, X.; Yuan, Q.-Q.; Liu, D.-H.; Shen, H.-F.; Zhang, L.-P. Satellite Video Super-Resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5610819. [[CrossRef](#)]
27. Lei, S.; Shi, Z.-W.; Mo, W.-J. Transformer-Based Multiscale Enhancement for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5615611.
28. 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](#)] [[PubMed](#)]
29. Xiao, Y.; Yuan, Q.-Q.; Jiang, K.; He, J.; Lin, C.W.; Liang, P.-Z. TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-Resolution. *IEEE Trans. Image Process.* **2024**, *33*, 738. [[CrossRef](#)] [[PubMed](#)]
30. Zhang, D.-F.; Huang, F.-Y.; Liu, S.-Z.; Wang, X.-B.; Jin, Z.-Z. Swinir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv* **2023**, arXiv:2208.11247v3.
31. Cansu, K.; Ahmet Murat, T.; Zafer, D. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 5926–5936.
32. Zhu, Q.-Y.; Wang, H.; Zhang, R.-X. Wavelet Loss Function for Auto-Encoder. *IEEE Access* **2021**, *9*, 27101. [[CrossRef](#)]
33. Li, Z.; Kuang, Z.-S.; Zhu, Z.-L.; Wang, H.-P.; Shao, X.-L. Wavelet-based Texture Reformation Network for Image Super-Resolution. *IEEE Trans. Image Process.* **2022**, *31*, 2647. [[CrossRef](#)]
34. Zou, W.-B.; Chen, L.; Wu, Y.; Zhang, Y.-C.; Xu, Y.-X.; Shao, J. Joint Wavelet Sub-Bands Guided Network for Single Image Super-Resolution. *IEEE Trans. Multimedia* **2023**, *25*, 4623. [[CrossRef](#)]
35. Li, W.-J.; Guo, H.; Liu, X.-N.; Liang, K.-M.; Hu, J.-N.; Ma, Z.-Y.; Guo, J. Efficient Face Super-Resolution via Wavelet-based Feature Enhancement Network. *arXiv* **2024**, arXiv:2407.19768.
36. Xiao, T.-T.; Mannat, S.; Eric, M.; Trevor, D.; Piotr, D.; Ross, G. Early Convolutions Help Transformers See Better. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QU, Canada, 6–14 December 2021; pp. 30392–30400.
37. Shi, W.-Z.; 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.
38. Lu, Y.-T.; Min, L.-T.; Wang, B.-L.; Zheng, L.; Wang, X.-X.; Zhao, Y.-Q.; Long, T. Cross-Spatial Pixel Integration and Cross-Stage Feature Fusion Based Transformer Network for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625616. [[CrossRef](#)]
39. 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, 19–24 June 2022; pp. 5728–5739.

40. Chen, L.-Y.; Chun, X.-J.; Zhang, X.-Y.; Sun, J. Simple Baselines for Image Restoration. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
41. Jawerth, B.-D.; Wim, S. An Overview of Wavelet Based Multiresolution Analyses. *SIAM Rev.* **1994**, *36*, 377. [[CrossRef](#)]
42. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [[CrossRef](#)]
43. Li, K.; Wan, G.; Cheng, G.; Meng, L.-Q.; Han, J.-W. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ISPRS J. Photogram. Remote Sens.* **2020**, *159*, 296. [[CrossRef](#)]
44. 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](#)]
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, 18–22 June 2018; pp. 586–595.
46. Wolters, P.; Bastani, F.; Kembhavi, A. Zooming Out on Zooming In: Advancing Super-Resolution for Remote Sensing. *arXiv* **2023**, arXiv:2311.18082v1.
47. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Reconstruction. *arXiv* **2015**, arXiv:1409.1556v5.
48. Radford, A.; Kim, J.-W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the Conference on Machine Learning (ICML), Vienna, Austria, 18–24 July 2021; pp. 8748–8763.
49. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
50. Haram, C.; Jeongmin, L.; Jihoon, Y. N-Gram in Swin Transformer for Efficient Lightweight Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 2071–2081.
51. Ian, J.G.; Jean, P.-A.; Mehdi, M.; Xu, B.; David, W.-F.; Sherjil, O.; Aaron, C.; Yoshua, B. Generative Adversarial Nets. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QU, Canada, 8–13 December 2014; pp. 2672–2680.
52. Tero, K.; Samuli, L.; Timo, A. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Angeles, CA, USA, 16–19 June 2019.

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.