

HIMOSA: Efficient Remote Sensing Image Super-Resolution with Hierarchical Mixture of Sparse Attention

Yi Liu Yi Wan[†] Xinyi Liu Qiong Wu Panwang Xia Xuejun Huang Yongjun Zhang[†]

Wuhan University

[†] Corresponding author.

{liuyiwhu28, yi.wan, zhangyj}@whu.edu.cn

Abstract

In remote sensing applications, such as disaster detection and response, real-time efficiency and model lightweighting are of critical importance. Consequently, existing remote sensing image super-resolution methods often face a trade-off between model performance and computational efficiency. In this paper, we propose a lightweight super-resolution framework for remote sensing imagery, named HIMOSA. Specifically, HIMOSA leverages the inherent redundancy in remote sensing imagery and introduces a content-aware sparse attention mechanism, enabling the model to achieve fast inference while maintaining strong reconstruction performance. Furthermore, to effectively leverage the multi-scale repetitive patterns found in remote sensing imagery, we introduce a hierarchical window expansion and reduce the computational complexity by adjusting the sparsity of the attention. Extensive experiments on multiple remote sensing datasets demonstrate that our method achieves state-of-the-art performance while maintaining computational efficiency.

1. Introduction

In recent years, with the rapid advancement of remote sensing technologies and their expanding range of applications, high-resolution remote sensing imagery (RSI) has played an increasingly critical role in fields such as urban planning, agricultural monitoring, change detection, and disaster response. In particular, during natural disasters such as earthquakes, floods, and wildfires, timely access to high-quality imagery is essential for accurate disaster assessment and rapid response, placing strict requirements on data acquisition and processing efficiency. However, due to limitations in onboard sensor capabilities, transmission bandwidth, and complex imaging conditions, directly capturing native high-resolution images often implies high costs and significant delays, which limit their effectiveness in time-sensitive scenarios.



Figure 1. Different visual patterns in remote sensing imagery. (a) multi-scale repetitive patterns; (b) repetitive patterns; (c) weak texture.

To address this challenge, single-image super-resolution (SISR) has been developed to reconstruct high-resolution images from low-resolution observations, offering a more flexible and cost-effective alternative. As a representative task in low-level vision, SISR is inherently ill-posed. Early approaches attempted to address this challenge by employing upsampling techniques or incorporating handcrafted priors to constrain the solution space. With the rise of deep learning, Convolutional Neural Networks (CNN) [8, 25] have achieved remarkable success in super-resolution by modeling local features such as edges and color patterns using small convolutional kernels (*e.g.*, 3×3). Although CNN-based methods have advantages in inference efficiency, their limited receptive fields and fixed convolutional kernels restrict global context modeling and adaptability, which are crucial for reconstructing high-resolution details.

Benefiting from powerful self-attention [40] mechanisms, Transformer-based models have demonstrated superior performance over CNN-based methods in super-resolution tasks. These models typically employ dense self-attention to aggregate global features by computing pairwise similarities among all image tokens. However, this approach suffers from quadratic computational complexity $o(N^2)$. To balance computational complexity and long-range modeling capability, some methods [24, 49] introduced window attention mechanisms, which reduce computational complexity by restricting attention to non-overlapping local windows. While this strategy significantly improves efficiency, it inevitably weakens the ability to model long-range dependencies, especially when the window size is small. Recent studies [4] further suggest that enlarging the window size can effectively expand the receptive field and enhance super-resolution performance, reinforcing the importance of capturing long-range dependencies in SR tasks. To address this issue, some methods [30, 45] use sparse attention within a large window to reduce computational complexity, while other methods [27, 51] use token clustering to model long-range dependencies through latent semantic information. Nevertheless, these methods still require all tokens in the window or cluster to participate in similarity computations, which limits their computational efficiency.

Considering the inherent large-scale Earth observation in remote sensing, remote sensing images (RSIs) exhibit several characteristics that are uncommon in natural imagery. As illustrated in Fig. 1, these include, but are not limited to, multi-scale repetitive patterns (Fig. 1(a)(b)) and large amounts of redundant information (Fig. 1(c)). The redundancy of information makes the sparse Transformer a natural choice for RSI super-resolution. However, existing sparse Transformer-based methods suffer from two major limitations: (1) they rely on manually designed token selection strategies, such as sparse intervals [50] or top- k [45]. The former lacks attention to image content, while the latter requires all tokens to participate in similarity calculations, increasing computational complexity; and (2) the adoption of fixed window sizes limits their ability to capture multi-scale repetitive patterns effectively. Therefore, resolving the above issues is essential for enhancing the performance of RSI super-resolution.

To this end, we propose a novel efficient super-resolution framework to mitigate the aforementioned issues. To reduce the computational complexity, inspired by the Mixture of Experts (MoE) [15], we propose a content-aware routing sparse attention mechanism for efficient remote sensing image super-resolution. Our approach employs Expert-Choice Routing [56] to enable dynamic, content-aware, and head-specific token selection. Specifically, we treat each expert as a head in the traditional dense multi-head self-

attention and each expert selects k specific tokens from the input. Unlike previous methods that first aggregate all tokens and subsequently enforce sparsity, our approach selects specific tokens first and then performs information aggregation, reducing the computational complexity from $o(N^2)$ to $o(k^2 + N)$. To address the prevalent multi-scale patterns in remote sensing imagery, we introduce a hierarchical window-based attention mechanism. Specifically, we replace the fixed-size windows used in previous methods with progressively enlarged windows. This design enables the network to capture informative multi-scale features and gradually expand the receptive field, thereby enhancing multi-scale and global context modeling capabilities. We highlight our main contributions as follows:

- We propose a novel super-resolution framework, named HIMOSA, boosting SR performance by exploiting multi-scale features and long-range dependencies.
- We design a content-aware routing sparse attention mechanism to selectively aggregate effective tokens within each window, enabling the efficient utilization of large window sizes.
- Extensive experiments on multiple public datasets demonstrate the superior performance of our method in achieving efficient and effective RSI super-resolution and reconstructing high-resolution details.

2. Related Work

In this section, we introduce the related work, including the single-image super-resolution and mixture of experts.

2.1. Single image super-resolution

Transformer-based methods. The introduction of Vision Transformer (ViT) [10] brought Transformer architectures into the vision domain. Owing to the powerful ability of self-attention to capture long-range dependencies, Transformer-based super-resolution methods [2, 16, 21, 27, 28, 46, 48] have achieved performance comparable to, and in many cases surpassing, CNN-based approaches [8, 9, 17–20, 25, 31, 32, 36, 38, 39, 42, 53, 54]. While the global modeling capability of self-attention is particularly well-suited for remote sensing imagery with large spatial coverage, it also incurs substantial computational complexity. Inspired by Swin-Transformer [29], SwinIR [24] restricts self-attention computation within shifted local windows, effectively reducing complexity but at the cost of limiting non-local representation due to the small window size. To alleviate this limitation, several variants have been proposed. For example, ATD [51] introduces adaptive token dictionaries with category-based attention to better capture global information; CAT [5] adopts rectangular window attention combined with axial shifting to enhance cross-window interaction for improved restoration; HAT [4] enlarges the window size and incorporates channel attention to

further strengthen inter-window communication, achieving higher-quality reconstruction. Despite these advances, the fixed-size windows employed in these methods inherently lack adaptability to the multi-scale nature of remote sensing imagery, often leading to suboptimal reconstruction quality. To address this issue, HiT-SRF [52] adopts progressively enlarged windows to capture multi-scale features. However, it does not explicitly handle redundancy within large windows, resulting in limited gains, while the increased computational complexity associated with large-scale windows remains a critical concern.

Sparse Transformer. Dense attention often introduces task-irrelevant features, thereby weakening the extraction of useful information. To address this issue, numerous studies have explored sparse attention mechanisms to reduce computational complexity while focusing on key features. NLSA [31] combines non-local operations with sparse representation, employing dynamic sparse selection to improve efficiency and robustness; ART [50] integrates dense and sparse attention to achieve a broader receptive field and enhanced feature representation in image restoration. Another line of work [41, 55] explicitly filters the most relevant positions in the attention map to suppress redundant information. Building on this idea, methods like TTST [45] and DRSformer [3] introduce learnable top- k operators to adaptively retain the most relevant attention values, thus improving restoration performance. PFTNet [30] leverages progressive feature aggregation (PFA) to gradually refine feature selection and reduce computational complexity; SeemoRe [47] employs an expert mining strategy with mixed low-rank experts to achieve a lightweight design.

2.2. Mixture of Experts

Mixture-of-Experts (MoE) [15] integrates multiple specialized networks (experts) with a gating mechanism that routes each input to a small subset of experts, producing outputs as a sparsely weighted combination of selected predictions. A major challenge is load balancing: without explicit constraints, routing tends to overuse a few experts while leaving others inactive, thus limiting model capacity [34]. Traditional approaches mitigate this with auxiliary balancing losses [11, 22], while expert-choice routing [56] ensures balanced utilization by allowing experts to select tokens instead. MoE has been applied to both feed-forward and attention layers, such as SwitchHead [7], which reduces the number of attention heads by replacing them with MoE, and MoA [12], which enhances multi-query attention [1]. In addition, MoSA [33] introduces sparsity into the attention mechanism through a mixture-of-experts design, thereby reducing computational complexity. Inspired by these works, we also adopt a mixture-of-experts design to introduce sparsity into the attention computation, thereby improving inference efficiency.

3. Methodology

3.1. Overall Architecture

Given an LR input image $I_{\text{lr}} \in \mathbb{R}^{h \times w \times 3}$, we aim to restore the high resolution details and reconstruct the corresponding HR results $I_{\text{sr}} \in \mathbb{R}^{H \times W \times 3}$. The overall architecture consists of three main components: shallow feature extraction, deep feature extraction, and image reconstruction, as illustrated in Fig. 2.

First, we employ a shallow feature extractor F composed of a 3×3 convolutional layer to obtain low-level features from the input image:

$$X_0 = F(I_{\text{lr}}). \quad (1)$$

Then, we cascaded N hierarchical mixture of sparse attention (HIMOSA) blocks to progressively extract deep and rich representations from the encoded features. Each HIMOSA block is composed of M hierarchical layers designed to capture both local spatial correlations and long-range contextual dependencies at multiple scales. Within each layer, the content-aware routing sparse attention (CARSA) adaptively selects informative tokens, enabling efficient but expressive token aggregation. The channel attention module (CA) further enhances the global information modeling capability, while the convolutional gated linear unit (ConvGLU) introduces nonlinearity and spatial adaptability to achieve more flexible feature modulation. Finally, the aggregated hierarchical features are fed into a lightweight reconstruction head that employs PixelShuffle [37] to upsample the features and generate the final high-resolution output with fine spatial details.

3.2. Hierarchical Mixture of Sparse Attention

In this subsection, we introduce the motivation and details of our proposed hierarchical mixture of sparse attention.

In remote sensing imagery, a common characteristic is the presence of multi-scale repetitive patterns. Unlike previous methods that use fixed-scale small windows, we propose progressively expanding the window size, allowing the network to aggregate multi-scale information while increasing its receptive field. However, as the window size increases, the computational complexity grows rapidly. How can we address this challenge? We observe that larger windows tend to contain substantial redundant information, making the attention distribution increasingly sparse. Inspired by recent works such as PFTNet [30] and MOSA [33], we note that not all tokens need to be involved in the similarity computation, especially in a large window, e.g., 64×64 . Therefore, we selectively retain the top- k tokens that contribute most to image reconstruction before performing the similarity computation with quadratic complexity, thereby significantly improving the computational efficiency of our method.

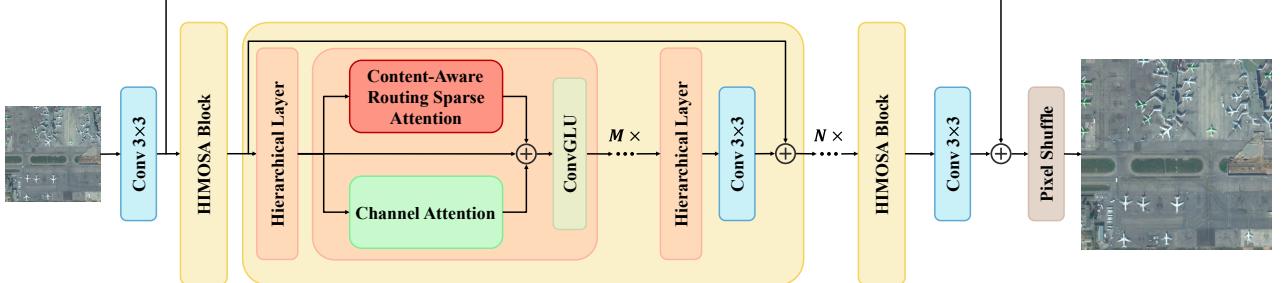


Figure 2. The overall architecture of HIMOSA. Each HIMOSA block contains M hierarchical layers, each of which includes content-aware routing sparse attention (CARSA), a channel attention module (CA) and a convolutional gated linear unit (ConvGLU).

3.2.1. Hierarchical windows

Inspired by HiTSR [52], we introduce the hierarchical windows to aggregate the multi-scale information. Specifically, given a base window size ws_B , in each HIMOSA block, we set the window size ws_i for the i -th hierarchical layer to:

$$ws_i = \alpha_i ws_B, \quad (2)$$

where $\alpha_i \in (\alpha_0, \alpha_1, \dots, \alpha_M)$ denotes the hierarchical ratio for the i -th hierarchical layer.

3.2.2. Content-aware routing sparse attention

In each hierarchical layer, we propose Content-Aware Routing Sparse Attention (CARSA) to efficiently model the relationships among tokens within the hierarchical window.

Compared with expert-level gating in MOSA [33], CARSA adopts a content-aware scoring function within each head, ranking tokens by feature similarity and keeping top- k candidates for attention computation. Then, a layer-wise sparsity ratio ρ_i controls the number of selected tokens, jointly with hierarchical windows to balance local and global dependencies. This routing scheme introduces deterministic sparsity and better spatial consistency. Specifically, for each hierarchical layer, the input features are partitioned into non-overlapping windows of size ws_i . For notational clarity, we denote the partitioned features of the i -th layer as $X_i \in \mathbb{R}^{n \times d}$. We introduce a content-aware router to help different experts select tokens. First, we calculate the selection scores $r_i \in \mathbb{R}^{n \times m}$ for each token:

$$r_i = \sigma(X_i W^r) \quad (3)$$

where $W^r \in \mathbb{R}^{d \times m}$ is the linear transform for input tokens, and σ is the sigmoid activation function. m denotes the number of experts.

Due to the varying levels of redundancy across different window sizes, processing all tokens in the initial small windows remains computationally affordable. With the expansion of window size, the level of information redundancy within each window correspondingly increases. To this end, we assign different sparsities ρ_i to different window sizes,

so that the sparsity in the attention calculation gradually increases with the window size. This design allows the model to substantially reduce computational cost while still preserving its ability to capture critical information for high-quality reconstruction. Then we select the top- k indices in the selection score r_i , which can be formally as:

$$k_i = n / \rho_i \quad (4)$$

$$r_i^{topk}, I_i = \text{TopK}(r_i, k_i) \quad (5)$$

where r_i^{topk} denotes the highest k values of r_i and $I_i \in \{0, \dots, n - 1\}^k$ is the corresponding indices. Then we use I_i to select the subset of input tokens for each head:

$$X_i^S = (X_{I_1}, X_{I_2}, \dots, X_{I_k}) \in \mathbb{R}^{k \times d} \quad (6)$$

After that, queries, keys, and values are calculated identically to the standard MHA [40]. For each expert, it has its own linear mappings $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d'}$ and the output mapping $W^O \in \mathbb{R}^{d' \times d}$, where $h \in [1, m]$:

$$Q_h^S = X_i^S W_h^Q, K_h^S = X_i^S W_h^K, V_h^S = X_i^S W_h^V, \quad (7)$$

$$e_h = \text{softmax}(Q_h^S K_h^{ST} / \sqrt{d'}) V_h^S \quad (8)$$

$$X_{\text{CARSA}} = \text{Concat}(e_1, e_2, \dots, e_h) W^O \quad (9)$$

3.2.3. Channel attention

In conventional expert-choice routing strategies, each expert selects the tokens to process based on routing scores. This “expert-selects-token” mechanism has proven effective in natural language processing (NLP) tasks [56], primarily because semantic information in language data is typically concentrated in a small number of key tokens. However, directly applying this strategy to low-level vision tasks such as image super-resolution may lead to performance bottlenecks. Due to the lack of a global coordination mechanism, multiple experts tend to select tokens from semantically salient regions of the image (such as edges or textured regions), resulting in a highly imbalanced token distribution: some tokens are repeatedly processed by multiple experts, while many other tokens (typically located

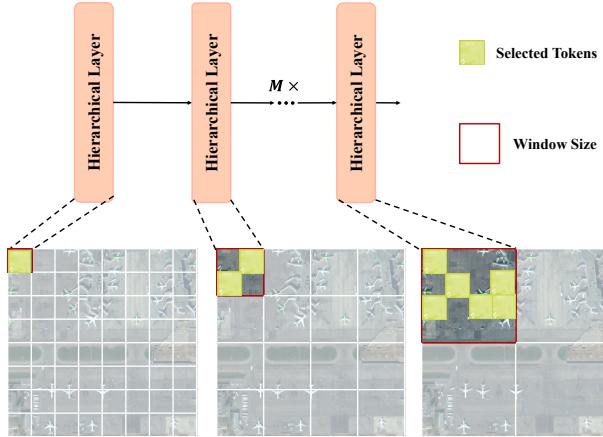


Figure 3. Hierarchical window for sparse attention. Increasing window sizes are applied to different hierarchical layers to aggregate features with expanding receptive fields.

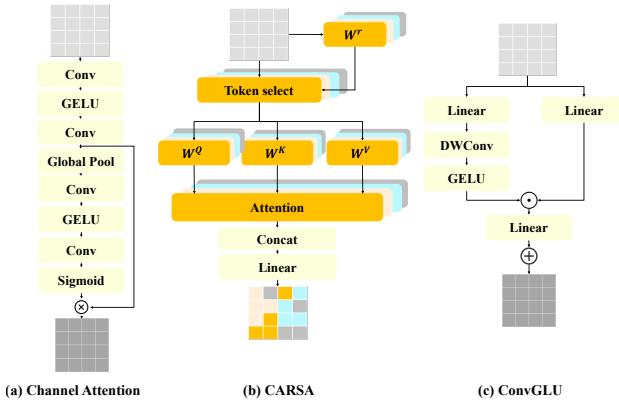


Figure 4. The structure of: (a) Channel attention; (b) CARSA; (c) ConvGLU.

in smooth or low-texture regions) are consistently ignored. For low-level tasks like super-resolution, this redundant focus on a limited subset of tokens leads to underutilization of image-wide information, thereby reducing reconstruction quality. Therefore, we further incorporate the channel attention block (CAB) [53] to enhance the network’s ability to capture global information.

3.2.4. ConvGLU

Finally, we replaced the standard convolutional feed-forward layer with a convolutional gated linear unit [35]. As shown in Fig. 4, the convolutional GLU consists of two parallel convolutional projections. One branch generates a gating signal through a nonlinear activation function, while the other performs the main feature transformation. The outputs of these two branches are then combined through element-wise multiplication. In this way, the gating branch adaptively controls the contribution of the transformed features, determining which components should be emphasized and which should be suppressed. This adaptive feature selection mechanism allows the model to highlight

informative patterns while mitigating irrelevant or redundant signals, significantly improving reconstruction performance without sacrificing computational efficiency.

4. Experiments and Analysis

4.1. Implementation Details

Datasets. In this paper, we employ four remote sensing imagery datasets, including AID [43], DOTA v2.0 [44], DIOR [23], and NWPU-RESISC45 [6]. For the AID dataset, we randomly select 100 images per scene for training and 30 images per scene for testing, resulting in a total of 3,000 training images and 900 testing images. The NWPU-RESISC45, DIOR, and DOTA v2.0 datasets are used exclusively for testing.

Training Details. There are 4 HIMOSA blocks in total, each comprising 6 hierarchical layers with a channel number of 60. In each HIMOSA block, the base window size ws_B is set to 8 and the hierarchical ratio is set to $(0.5, 1, 2, 4, 6, 8)$. For CARSA, the number of experts is set to 8, and the sparsity is set to $(1, 1, 2, 4, 8, 12)$ as the window expands. During training, we adopt the Muon [26] optimizer with an initial learning rate of 5×10^{-4} . The total number of training iterations is set to 250k, including a 10k warm-up phase to ensure training stability. In addition, we establish HIMOSA-light, which is a variant of HIMOSA that sets the number of experts to 4, significantly improving inference efficiency while ensuring model performance.

4.2. Results

Quantitative results. To validate the effectiveness of our method in remote sensing image super-resolution, we conduct comparisons with state-of-the-art methods in different remote sensing image datasets. In Tab. 1, we compare our method with other natural image super-resolution methods SwinIR-light [24], NLSA [31], ATD-light [51], HiT-SRF [52], PFT-light [30], CATANet [27], and remote sensing image super-resolution method HSENet [20], TransENet [21], ESTNet [16]. For a fair comparison, all models were retrained in the AID dataset. From Tab. 1, it can be seen that our method consistently achieves the best performance across various scene types, and outperforms the state-of-the-art methods by 0.06 dB and 0.0021 in average PSNR and SSIM, respectively. In particular, the results on scenes with highly redundant information, such as deserts and beaches, further confirm that not all tokens within a window need to be computed, verifying the effectiveness of the proposed sparse attention mechanism. In addition, the outstanding performance in dense residential areas and other scenes with evident multi-scale repetitive patterns demonstrates the capability of our method to effectively extract and leverage multi-scale information.

Furthermore, to validate the generalizability of our

Table 1. Quantitative results ($\times 4$) achieved by different methods on the AID datasets. Here, PSNR(dB) \uparrow and SSIM \uparrow values are reported. **bold** texts indicate the best performance.

Category	SwinIR-light		ATD-light		HiT-SRF		CATANet		ESTNet		PFT-light		HIMOSA(Ours)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airport	29.77	0.8165	29.91	0.8195	30.01	0.8232	30.01	0.8226	30.02	0.8229	30.06	0.8236	30.09	0.8252
Desert	42.06	0.9517	42.10	0.9521	42.00	0.9521	42.12	0.9524	42.10	0.9521	42.13	0.9524	42.15	0.9525
Farmland	36.29	0.8989	36.46	0.9009	36.60	0.9035	36.58	0.9030	36.56	0.9026	36.57	0.9027	36.68	0.9045
Forest	29.61	0.7146	29.77	0.7211	29.80	0.7234	29.78	0.7232	29.77	0.7222	29.78	0.7224	29.81	0.7239
Industrial	28.37	0.7866	28.65	0.7921	28.74	0.7980	28.76	0.7980	28.74	0.7972	28.79	0.7989	28.86	0.8021
Meadow	36.60	0.8372	36.69	0.8381	36.72	0.8390	36.69	0.8385	36.69	0.8385	36.71	0.8386	36.74	0.8391
MediumResidential	29.51	0.7680	29.79	0.7705	29.98	0.7825	29.92	0.7805	29.92	0.7809	29.91	0.7802	30.01	0.7834
Mountain	30.69	0.7829	30.73	0.7833	30.76	0.7853	30.75	0.7849	30.75	0.7848	30.77	0.7856	30.79	0.7863
Park	28.98	0.7623	29.11	0.7665	29.18	0.7706	29.18	0.7701	29.17	0.7699	29.19	0.7705	29.25	0.7728
Parking	26.77	0.8327	27.81	0.8559	28.17	0.8634	28.01	0.8602	27.96	0.8587	28.07	0.8617	28.37	0.8679
Playground	33.89	0.8761	34.15	0.8794	34.32	0.8854	34.27	0.8832	34.24	0.8838	34.31	0.8833	34.41	0.8867
Bareland	38.48	0.8866	38.52	0.8867	38.52	0.8874	38.54	0.8876	38.56	0.8876	38.56	0.8876	38.58	0.8879
Pond	31.58	0.8357	31.70	0.8374	31.72	0.8387	31.75	0.8389	31.77	0.8389	31.76	0.8392	31.78	0.8397
Port	28.15	0.8442	28.41	0.8497	28.54	0.8543	28.51	0.8534	28.47	0.8525	28.56	0.8540	28.66	0.8569
RailwayStation	28.56	0.7601	28.90	0.7680	28.93	0.7723	28.97	0.7724	28.93	0.7717	29.04	0.7740	29.06	0.7758
Resort	28.13	0.7712	28.28	0.7751	28.42	0.7808	28.39	0.7796	28.39	0.7798	28.41	0.7804	28.47	0.7829
River	31.60	0.7891	31.68	0.7905	31.73	0.7926	31.72	0.7926	31.72	0.7923	31.73	0.7926	31.77	0.7936
School	28.74	0.7962	28.95	0.8007	29.08	0.8061	29.05	0.8045	29.04	0.8049	29.07	0.8054	29.15	0.8082
SparseResidential	27.26	0.6538	27.41	0.6592	27.49	0.6640	27.47	0.6633	27.48	0.6623	27.46	0.6626	27.51	0.6647
Square	29.05	0.7988	29.39	0.8054	29.48	0.8097	29.45	0.8086	29.48	0.8091	29.54	0.8110	29.59	0.8125
Stadium	28.44	0.8138	28.78	0.8211	28.74	0.8216	28.77	0.8222	28.79	0.8230	28.92	0.8270	28.89	0.8261
StorageTanks	27.50	0.7664	27.64	0.7697	27.76	0.7750	27.73	0.7733	27.71	0.7740	27.75	0.7741	27.79	0.7763
BaseballField	33.56	0.8903	33.78	0.8930	33.89	0.8952	33.88	0.8945	33.89	0.8945	33.91	0.8951	33.98	0.8961
Viaduct	28.94	0.7533	29.11	0.7588	29.23	0.7652	29.23	0.7646	29.23	0.7652	29.27	0.7669	29.33	0.7693
Beach	30.93	0.8063	31.04	0.8067	31.05	0.8079	31.06	0.8082	31.07	0.8079	31.07	0.8081	31.10	0.8087
Bridge	31.80	0.8089	32.03	0.8112	32.08	0.8130	32.03	0.8119	32.07	0.8131	32.14	0.8140	32.15	0.8142
Center	27.98	0.7928	28.31	0.8002	28.47	0.8064	28.42	0.8059	28.47	0.8069	28.48	0.8068	28.58	0.8101
Church	25.32	0.7065	25.62	0.7186	25.78	0.7269	25.74	0.7246	25.75	0.7252	25.77	0.7254	25.84	0.7292
Commercial	28.57	0.7944	28.78	0.7991	28.92	0.8050	28.89	0.8035	28.89	0.8039	28.90	0.8039	28.97	0.8067
DenseResidential	25.20	0.7003	25.43	0.7117	25.59	0.7220	25.55	0.7185	25.52	0.7182	25.52	0.7178	25.64	0.7248
Average	30.41	0.7999	30.63	0.8047	30.72	0.8090	30.71	0.8082	30.70	0.8082	30.74	0.8088	30.80	0.8109

method, we conducted experiments on different remote sensing image datasets, including DOTA, NWPU, and DIOR. As shown in Tab. 2, our method also achieves the best performance on these benchmarks.

Qualitative results. We present visual comparison results on the DOTA and AID datasets, as shown in Fig. 5, and Fig. 9. In the scenes with repetitive or weak-textured surfaces, our method achieves clearer and more accurate texture reconstruction compared to previous sparse Transformer-based approaches. As illustrated in Fig. 5, the playground contains pronounced repetitive texture patterns, and our method achieves a more accurate reconstruction of surface textures compared to previous approaches, primarily because it suppresses redundant information within the window during inference and guides the network’s attention toward critical boundary details. Furthermore, by leveraging multi-scale windows, the network effectively exploits similar textures from other playground regions within the scene. In addition, as shown in Fig. 9, our method also produces clearer reconstruction results in weak-texture areas. These results provide strong evidence of the effectiveness of our design in handling repetitive and weak-texture scenes.

4.3. Ablation Study

In this subsection, we perform ablation studies on HIMOSA model and train all models for 250k iterations.

Sparsity. The sparsity is an important hyperparameter to balance computational complexity and model performance. When the sparsity approaches 1, the attention becomes closer to dense attention, capturing more information but significantly increasing computational complexity. Conversely, as the sparsity increases, the model utilizes less information, thereby reducing computational cost but potentially degrading performance. Therefore, an appropriate choice of sparsity is essential to achieve an optimal trade-off between efficiency and reconstruction quality. We analyze the impact of different sparsity configurations on model performance and inference speed in Tab. 3. After a comprehensive evaluation, we select the configuration (1, 1, 2, 4, 8, 12) as the sparsity setting for our method.

Number of experts. We compared the impact of different numbers of experts. The results are shown in Tab. 4. Because more experts can provide a more refined distribution of image content, better performance is achieved. However, as the number of experts increases, the amount of token

Table 2. Quantitative results achieved by different methods on the AID, DOTA, NWPU and DIOR datasets. Here, PSNR(dB)↑, SSIM↑ and LPIPS↓ values are reported. **bold** and underline texts indicate the best and the second-best performance, respectively.

Method	Scale	AID			DOTA			NWPU			DIOR		
		PSNR	SSIM	LPIPS									
NLSA (2021 CVPR)	×2	36.74	0.9439	0.1087	39.75	0.9603	0.0881	34.71	0.9311	0.1189	37.41	0.9500	0.1083
SwinIR-light (2021 ICCVW)	×2	36.83	0.9428	0.1077	40.03	0.9598	0.0855	34.66	0.9229	0.1440	37.38	0.9490	0.1229
ATD-light (2024 CVPR)	×2	36.97	0.9449	0.1101	40.10	0.9605	0.0879	34.83	0.9318	0.1129	37.59	0.9504	0.1099
ESTNet (2024 TIP)	×2	37.00	0.9449	0.1071	40.14	0.9607	0.0854	34.82	0.9318	0.1110	37.60	0.9505	0.1075
HiT-SRF (2024 ECCV)	×2	36.96	0.9456	0.1017	40.10	0.9605	0.0845	34.83	0.9318	0.1090	37.59	0.9504	0.1069
PFT-light (2025 CVPR)	×2	36.90	0.9441	0.1075	40.19	0.9610	0.0855	34.84	0.9321	0.1427	37.63	0.9508	0.1211
CATANet (2025 CVPR)	×2	36.97	0.9445	0.1093	40.07	0.9607	0.0878	34.83	0.9318	0.1126	37.60	0.9505	0.1104
HIMOSA-light (Ours)	×2	37.19	0.9466	0.1010	<u>40.22</u>	0.9610	0.0808	<u>34.87</u>	0.9325	0.1058	37.69	0.9512	0.1057
HIMOSA (Ours)	×2	37.20	0.9467	0.1005	40.26	0.9613	0.0788	34.90	0.9328	0.1043	37.70	0.9513	0.1044
HSENet (2021 TGRS)	×4	29.51	0.7692	0.4024	31.31	0.8151	0.3482	28.31	0.7411	0.4265	29.46	0.7771	0.3977
TransENet (2021 TGRS)	×4	29.53	0.7718	0.3528	31.56	0.8228	0.3036	28.29	0.7417	0.3767	29.37	0.7786	0.3509
NLSA (2021 CVPR)	×4	30.55	0.8053	0.3100	32.92	0.8495	0.2681	29.23	0.7774	0.3254	30.54	0.8085	0.3223
SwinIR-light (2021 ICCVW)	×4	30.41	0.7999	0.3227	32.85	0.8465	0.2790	29.09	0.7716	0.3413	30.37	0.8041	0.3312
ATD-light (2024 CVPR)	×4	30.63	0.8047	0.3231	33.17	0.8522	0.2755	29.28	0.7785	0.3279	30.67	0.8088	0.3191
ESTNet (2024 TIP)	×4	30.70	0.8082	0.3153	33.28	0.8548	0.2694	29.33	0.7807	0.3265	30.76	0.8119	0.3183
HiT-SRF (2024 ECCV)	×4	30.72	0.8090	0.3129	33.29	0.8551	0.2689	29.36	0.7820	<u>0.3227</u>	30.74	0.8122	<u>0.3154</u>
PFT-light (2025 CVPR)	×4	30.74	0.8088	0.3165	33.34	0.8555	0.2703	29.35	0.7815	0.3364	30.77	0.8125	0.3252
CATANet (2025 CVPR)	×4	30.70	0.8082	0.3151	33.30	0.8549	0.2704	29.33	0.7811	0.3252	30.75	0.8119	0.3164
HIMOSA-light (Ours)	×4	<u>30.78</u>	<u>0.8103</u>	0.3118	<u>33.37</u>	<u>0.8563</u>	<u>0.2673</u>	<u>29.40</u>	<u>0.7835</u>	0.3269	<u>30.80</u>	<u>0.8136</u>	0.3214
HIMOSA (Ours)	×4	30.80	0.8109	<u>0.3108</u>	33.38	0.8567	0.2670	29.42	0.7840	0.3217	30.82	0.8141	0.3149

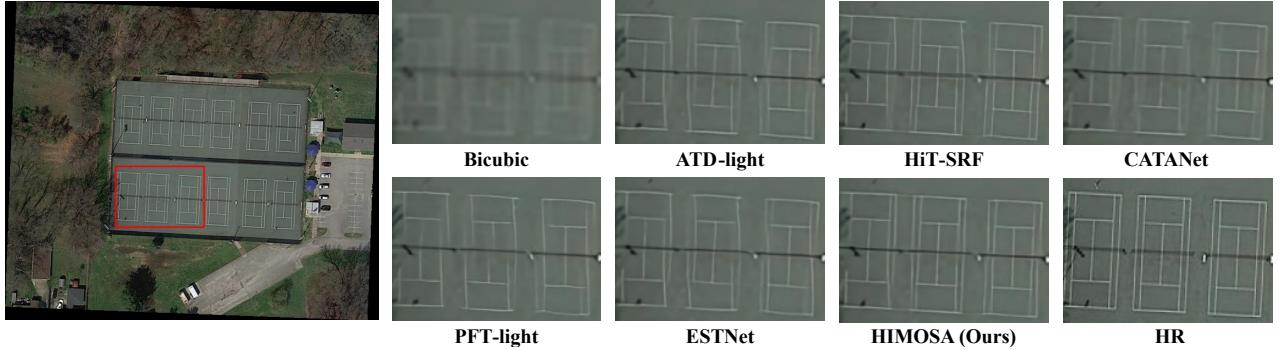


Figure 5. Visualization results (×4) achieved by different methods in DOTA datasets.

Table 3. Quantitative results (×4) achieved by methods with different sparsity.

Sparsity	AID		DIOR		Inference time (ms)
	PSNR	SSIM	PSNR	SSIM	
$\rho_{(1,1,2,4,8,12)}$	30.80	0.8109	30.82	0.8141	93.30
$\rho_{(1,1,2,4,8,8)}$	30.80	0.8110	30.83	0.8143	96.79
$\rho_{(1,1,2,4,8,16)}$	30.80	0.8109	30.80	0.8136	92.58
$\rho_{(1,2,4,8,12,16)}$	30.75	0.8094	30.78	0.8129	70.56

computation increases, resulting in a decrease in inference efficiency. Therefore, in order to balance computational efficiency and model performance, we set the number of experts to 8. In addition, we extend HIMOSA to HIMOSA-light, adjusting the number of experts to 4.

Content-aware routing. We analyze the content-aware routing mechanism. Specifically, we set different token selection strategies and analyze their impact on the performance of our method. We compare content-aware routing

Table 4. Quantitative results (×4) achieved by methods with different numbers of experts.

Num of experts	AID		DIOR		Inference time (ms)
	PSNR	SSIM	PSNR	SSIM	
12	30.81	0.8113	30.84	0.8145	133.14
8	30.80	0.8109	30.82	0.8141	93.30
6	30.79	0.8107	30.81	0.8139	77.97
4	30.78	0.8103	30.80	0.8136	63.82

with two alternative selection strategies: random k indices and the first k indices in the token sequence. As shown in Tab. 5, the proposed content-aware routing achieves significant performance improvements.

In addition, we conducted a visualization analysis of the proposed content-aware routing mechanism. As shown in Fig. 6, we visualized the token selection of the first HIMOSA block at the 4th, 5th, and 6th layers for the first four experts, where the red regions indicate the selected tokens. It can be observed that each expert adaptively selects

Table 5. Quantitative results ($\times 4$) achieved by methods with different token selection strategies.

Token selection strategy	AID		DIOR	
	PSNR	SSIM	PSNR	SSIM
content-aware routing selection	30.80	0.8109	30.82	0.8141
random indices selection	30.74	0.8091	30.78	0.8129
sequential indices selection	30.73	0.8087	30.76	0.8121

Table 6. Quantitative ablation study of different modules (PSNR \uparrow /SSIM \uparrow).

Method	CA	ConvGLU	AID		DIOR	
			PSNR	SSIM	PSNR	SSIM
HIMOSA	✓	✓	30.80/0.8109	30.82/0.8141		
	✗	✓	30.75/0.8094	30.80/0.8131		
	✓	✗	30.78/0.8104	30.81/0.8137		

	Expert 1	Expert 2	Expert 3	Expert 4
Layer 4 $\rho = 4$ $ws = 32$				
Layer 5 $\rho = 8$ $ws = 48$				
Layer 6 $\rho = 12$ $ws = 64$				

Figure 6. Visualization of sparse attention. The red region indicates the selected token.

tokens according to different image content. Furthermore, with the increase in sparsity, the method effectively avoids interference from redundant information within large windows while alleviating unnecessary computational cost.

CA and ConvGLU. We perform ablation studies on the CA and ConvGLU modules. Specifically, we remove CA or replace ConvGLU with a simple feed-forward layer. As shown in Tab. 6, both modules played a significant role in improving our model performance.

4.4. Other results

Efficiency. To demonstrate the efficiency of our method, we compare our method with other lightweight methods in terms of parameters, FLOPs, and inference time. The evaluation was conducted using 256×256 image patches under the same experimental settings, and the results are presented in Tab. 7. HIMOSA achieves comparable inference speed to other lightweight methods while significantly outperforming them in reconstruction performance. Moreover, although HIMOSA-light sacrifices a certain degree of performance, it achieves the fastest inference speed, exceeding

Table 7. Model parameters, FLOPs, and inference time of different methods ($\times 4$).

Methods	Params (M)	FLOPs (G)	Inference time (ms)
SwinIR-light	0.93	60.31	96.23
ATD-light	1.26	109.09	164.49
HiT-SRF	0.87	56.53	97.90
CATANet	0.53	46.77	87.84
PFT-light	1.10	98.32	371.90
ESTNet	3.57	230.71	96.12
HIMOSA	3.26	139.58	93.30
HIMOSA-light	2.14	66.71	63.82

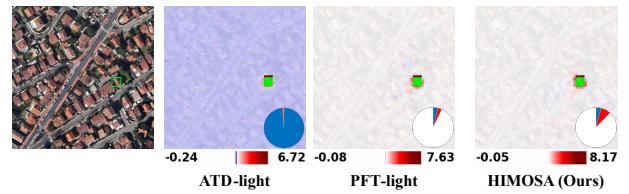


Figure 7. Causal effect maps (CEMs) of different methods on the AID [43] dataset. The patches with positive or negative causal effects and the ROI are indicated in red, blue, and green, respectively. The pie chart records the percentage of patches with different causal effects.

the current best-performing method by 27.34%.

Causal inference. In order to analyze the impact of redundant information on different networks, we used Causal Effect Map (CEM) [14] to conduct attribution analysis on the networks. As illustrated in Fig. 7, while ATD-light can utilize a wider range of relevant information through its token dictionary, most of the information has been shown to have a negative impact. In contrast, our method effectively suppresses the participation of irrelevant or negative information and utilizes more beneficial information with hierarchical window design, resulting in superior performance.

5. Conclusion

In this work, we propose a lightweight framework for remote sensing image super-resolution. Specifically, HIMOSA exploits the inherent redundancy of remote sensing imagery by introducing a content-aware sparse attention mechanism, achieving a lightweight design while preserving strong reconstruction capability. To further tackle the challenge of multi-scale repetitive patterns, HIMOSA incorporates a progressive window expansion strategy, where the sparsity of the attention mechanism is adaptively adjusted to reduce computational cost. Moreover, we extend our framework to HIMOSA-light, which attains significant improvements in inference efficiency with minimal performance degradation. Extensive experiments conducted on multiple remote sensing datasets demonstrate that our method achieves state-of-the-art performance while maintaining superior computational efficiency.

References

- [1] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghi. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. 3
- [2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2
- [3] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image de-raining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5896–5905, 2023. 3
- [4] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, 2023. 2
- [5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Cross aggregation transformer for image restoration. *Advances in Neural Information Processing Systems*, 35:25478–25490, 2022. 2
- [6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 5
- [7] Róbert Csordás, Piotr Piękos, Kazuki Irie, and Jürgen Schmidhuber. Switchhead: Accelerating transformers with mixture-of-experts attention. *Advances in Neural Information Processing Systems*, 37:74411–74438, 2024. 3
- [8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 2
- [9] Chao Dong, Chen Change Loy, and Xiaou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 2
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [11] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 3
- [12] Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, et al. Moa: Mixture of sparse attention for automatic large language model compression. *arXiv preprint arXiv:2406.14909*, 2024. 3
- [13] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 1
- [14] Jinfan Hu, Jinjin Gu, Shiyao Yu, Fanghua Yu, Zheyuan Li, Zhiyuan You, Chaochao Lu, and Chao Dong. Interpreting low-level vision models with causal effect maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 8
- [15] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2, 3
- [16] Xudong Kang, Puhong Duan, Jier Li, and Shutao Li. Efficient swin transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 33:6367–6379, 2024. 2, 5
- [17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2
- [18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.
- [19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.
- [20] Sen Lei and Zhenwei Shi. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, 2021. 2, 5
- [21] Sen Lei, Zhenwei Shi, and Wenjing Mo. Transformer-based multistage enhancement for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 2, 5
- [22] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 3
- [23] Ke Li, Gang Wan, Gong Cheng, Lijiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 5
- [24] Jingyun Liang, Jie Zhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2, 5
- [25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2
- [26] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu,

- Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025. 5
- [27] Xin Liu, Jie Liu, Jie Tang, and Gangshan Wu. Catanet: Efficient content-aware token aggregation for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17902–17912, 2025. 2, 5
- [28] Yi Liu, Xinyi Liu, Yi Wan, Panwang Xia, Qiong Wu, and Yongjun Zhang. Stereoinr: Cross-view geometry consistent stereo super resolution with implicit neural representation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1003–1012, 2025. 2
- [29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2
- [30] Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2279–2288, 2025. 2, 3, 5
- [31] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3517–3526, 2021. 2, 3, 5
- [32] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 2
- [33] Piotr Piękos, Róbert Csordás, and Jürgen Schmidhuber. Mixture of sparse attention: Content-based learnable sparse attention via expert-choice routing. *arXiv preprint arXiv:2505.00315*, 2025. 3, 4
- [34] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 3
- [35] Dai Shi. Transnext: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17773–17783, 2024. 5
- [36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2
- [37] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3
- [38] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 2
- [39] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, pages 4799–4807, 2017. 2
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4
- [41] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvvt: k-nn attention for boosting vision transformers. In *European conference on computer vision*, pages 285–302. Springer, 2022. 3
- [42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2
- [43] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 5, 8, 1
- [44] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beßongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 5
- [45] Yi Xiao, Qiangqiang Yuan, Kui Jiang, Jiang He, Chia-Wen Lin, and Liangpei Zhang. Ttst: A top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 33:738–752, 2024. 2, 3
- [46] Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee, and Tae Hyun Kim. Enriched cnn-transformer feature aggregation networks for super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4956–4965, 2023. 2
- [47] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yulun Zhang, and Radu Timofte. See more details: Efficient image super-resolution by experts mining. In *International Conference on Machine Learning*. PMLR, 2024. 3
- [48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 2
- [49] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022. 2
- [50] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 2, 3

- [51] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2856–2865, 2024. [2](#), [5](#)
- [52] Xiang Zhang, Yulun Zhang, and Fisher Yu. Hit-sr: Hierarchical transformer for efficient image super-resolution. In *European Conference on Computer Vision*, pages 483–500. Springer, 2024. [3](#), [4](#), [5](#)
- [53] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. [2](#), [5](#)
- [54] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. [2](#)
- [55] Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. Explicit sparse transformer: Concentrated attention through explicit selection. *arXiv preprint arXiv:1912.11637*, 2019. [3](#)
- [56] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022. [2](#), [3](#), [4](#)

HIMOSA: Efficient Remote Sensing Image Super-Resolution with Hierarchical Mixture of Sparse Attention

Supplementary Material

A. Training details

The details of the datasets used in our experiments are summarized in Tab. 8. Specifically, we evaluate the proposed HIMOSA framework on several representative remote sensing image super-resolution benchmarks to ensure the generalization and robustness of our method. These datasets cover diverse scenes, including urban areas, agricultural regions, coastal zones, and mountainous terrains, thereby providing a comprehensive assessment under various spatial structures and texture complexities. Each dataset contains high-resolution (HR) images and their corresponding low-resolution (LR) counterparts generated through bicubic downsampling with different scale factors. During training, we adopt random cropping, rotation, and horizontal flipping for data augmentation to improve the model’s robustness and generalization ability.

We adopt a multi-step learning rate decay strategy, initializing the learning rate at 5×10^{-4} and reducing it by a factor of 0.5 at iterations 150K, 200K, 225K, and 240K. We set the input patch size to 64×64 and use random rotation and horizontally flipping for data augmentation. The mini-batch size is set to 16. During training, we employ the L1 loss function to optimize the network. All experiments are conducted on four NVIDIA RTX 4090 GPUs.

B. Local attribute analysis

To further interpret the internal behavior of our model, we employ the Local Attribution Map (LAM) [13], an explainability tool designed for super-resolution tasks. LAM identifies which pixels in the low-resolution (LR) input contribute most to the generation of the high-resolution (HR) output, thereby revealing how effectively a model utilizes spatial information. As shown in Fig. 8, our HIMOSA exhibits a broader and denser distribution of informative regions, indicating that it can capture a larger set of relevant pixels while suppressing noise and redundant information. Compared to previous methods, HIMOSA demonstrates a higher Diffusion Index (DI), reflecting its superior ability to focus on meaningful contextual cues. These observations confirm that our content-aware sparse attention not only improves quantitative metrics but also enhances the interpretability and reliability of the reconstruction process.

C. Other visualization results

To further demonstrate the qualitative advantages of our proposed HIMOSA framework, this section presents addi-

Algorithm 1 HIMOSA Block

Input: X input feature; ws_B base window size; $(\alpha_0, \alpha_1, \dots, \alpha_M)$ hierarchical ratio; M num of hierarchical layers.
Output: X_M output feature;
1: **for** $i = 0$ to M **do**
2: Channel Attention: $X_{CAB} = CAB(X_i)$;
3: Partition windows with $ws_i = \alpha_i ws_B$;
4: Content-aware routing sparse attention in window
 ws_i : $X_{CARSA} = CARSA(X_i, A_i)$;
5: $X_i = X_i + X_{CARSA} + X_{CAB}$;
6: $X_{i+1} = \text{ConvGLU}(X_i) + X_i$;
7: **end for**



Figure 8. Local attribution maps (LAMs) of different methods on the AID [43] dataset. DI (Diffusion Index) indicates the range of involved pixels, with a higher DI indicating a wider receptive field.

tional visual comparison results across various datasets and scenes. These visualizations showcase the ability of our method to reconstruct fine structural details, preserve edge sharpness, and restore realistic textures in complex remote sensing environments. Compared with existing approaches, HIMOSA generates visually clearer and more natural results, particularly in regions with dense textures or repetitive patterns. These results further confirm the strong reconstruction capability and generalization performance of our model in diverse real-world scenarios. The visual results are shown in Fig. 10, and Fig. 11

Table 8. Description of Different Datasets Composition.

Phase	Name	Used (Total) Samples	Scene Classes	Spatial resolution(m)	Image sizes	Usage
Train	AID	3000 (10,000)	30	0.5 ~ 8	600 × 600	Scene classification
Test	AID	900 (10,000)	30	0.5 ~ 8	600 × 600	Scene classification
	NWPU-RESISC45	315 (31,500)	45	0.2 ~ 30	256 × 256	Scene classification
	DIOR	1000 (23,463)	20	0.5 ~ 30	800 × 800	Object Detection
	DOTA	900 (2,806)	14	0.3 ~ 10	800 ~ 4000	Object Detection

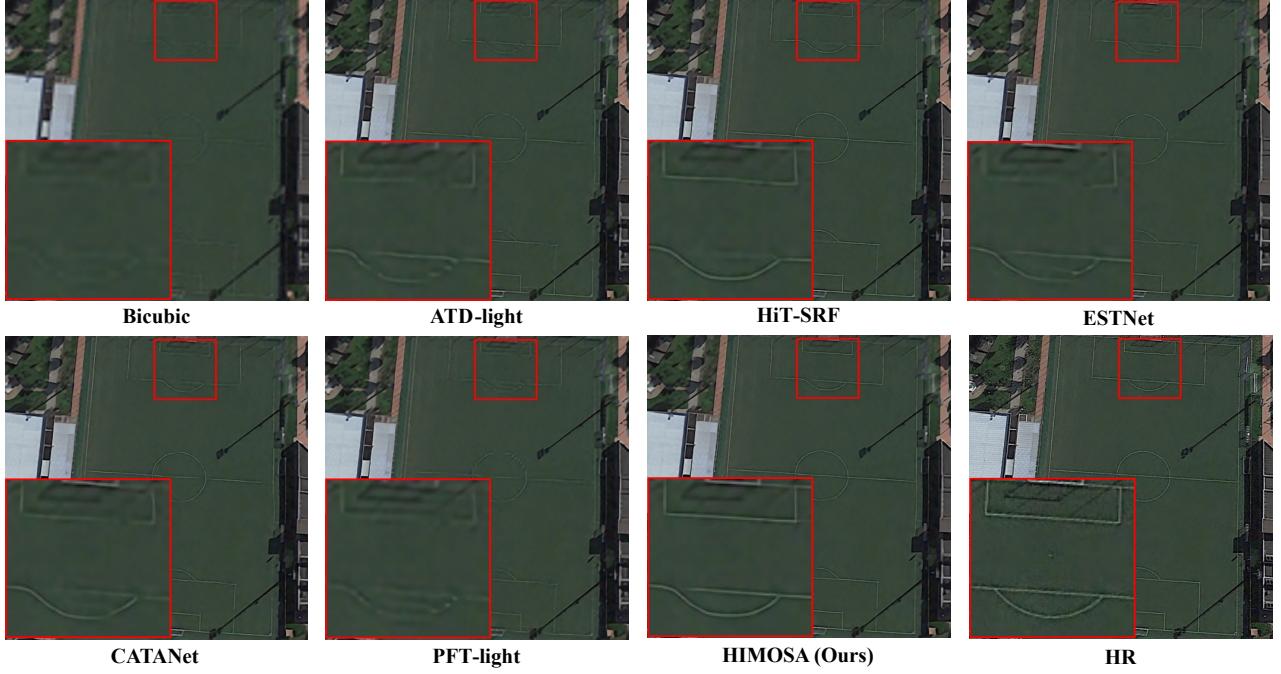


Figure 9. Visualization results ($\times 4$) achieved by different methods in AID datasets (zoom in for details).

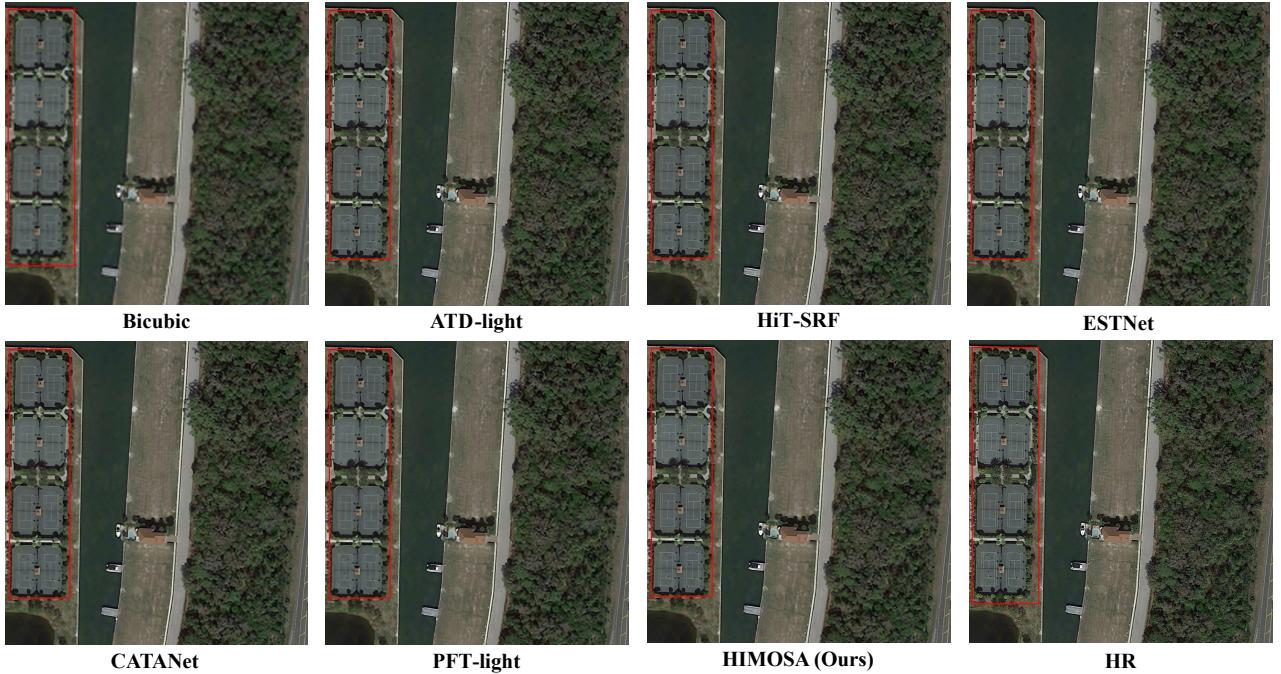


Figure 10. Visualization results ($\times 4$) of different methods in DIOR (zoom in for details).

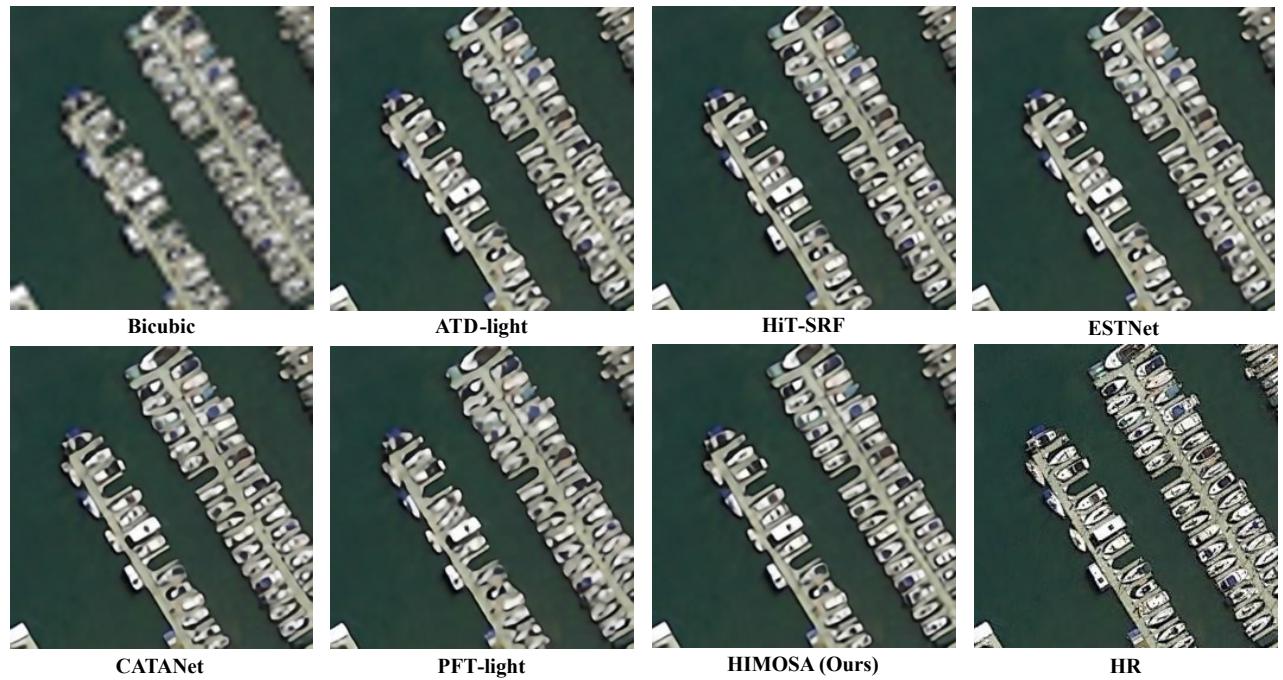


Figure 11. Visualization results ($\times 4$) of different methods in NWPU (zoom in for details).