

# LSFFNet: Large-kernel Small-span Feature Fusion Network

**Qikai Zhou**

College of Computer Science and Technology, Qingdao University, Qingdao, 266071, China

Email: [zhouqikai@qdu.edu.cn](mailto:zhouqikai@qdu.edu.cn)

**How to cite this paper:** Zhou, Q. K. (2026). LSFFNet: Large-kernel small-span feature fusion network. *Academic Journal of Emerging Technologies*, 3(1), 48–60. ISSN Print: 3104-4417; ISSN Online: 3104-4425. <https://doi.org/10.63313/AJET.9053>

**Published: 2026-05-11**

Copyright © 2026 by author(s) and Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



## Abstract

High-resolution optical remote sensing images are crucial for improving ground object interpretation and supporting precise earth observation. However, existing mainstream super-resolution methods struggle to adapt to the inherent characteristics of optical remote sensing images, such as large scale variation, weak texture details and complex imaging degradation processes. These methods commonly suffer from blurred high-frequency details, high computational complexity and difficulty in lightweight deployment for practical applications. To address these limitations, this paper proposes a lightweight Large-kernel Small-span Feature Fusion Network (LSFFNet) for optical remote sensing image super-resolution reconstruction, targeting practical scenarios with limited computing resources. A Large-kernel Small-span Feature Extraction Block (LSBlock) is designed in the proposed model. By adopting a small number of large-sized depthwise separable convolutions, multi-scale contextual information can be captured with extremely low parameter overhead. Meanwhile, an Attention Multi-level Feature Fusion Block (AFFBlock) is constructed. Integrating channel and spatial dual attention mechanisms, it enables adaptive selective fusion of multi-layer features and mitigates feature information loss effectively. Experimental results on multiple remote sensing datasets demonstrate that compared with state-of-the-art methods, the proposed LSFFNet achieves comparable or even better quantitative performance with fewer parameters and lower computational cost, striking a favorable balance between reconstruction quality and inference efficiency. Quantitative evaluations show that our method outperforms several existing mainstream and lightweight models in terms of PSNR and SSIM. Qualitative visual comparisons further verify its superior capability in edge restoration, texture preservation and artifact suppression. The designed LSFFNet also provides a valuable reference for performance optimization of lightweight super-resolution models in remote sensing tasks.

## Keywords

Optical Remote Sensing Images; Super-Resolution; Lightweight Model; Feature Fusion

## 1. Introduction

Image Super-Resolution (SR) is a fundamental and essential research topic in the field of computer vision. Its core objective is to reconstruct high-resolution (HR) images with richer details and better visual quality from one or multiple low-resolution (LR) images. This technology not only improves the visual quality of images but also serves as a vital approach to break through the hardware limitations of physical imaging systems and fully tap the potential of image information.

With the explosive development of deep learning, especially Convolutional Neural Networks (CNNs), image super-resolution technology has achieved remarkable progress in the field of natural image processing. A large number of models with excellent performance have been proposed and widely applied in various scenarios such as medical imaging[1], remote sensing imagery[2-3], and security surveillance[4]. Current performance-oriented SR models generally suffer from high complexity and substantial computational overhead, which contradicts the practical application demands of remote sensing tasks, including massive data volume, high processing timeliness requirements, and limited resources on spaceborne and airborne platforms. For instance, rapid processing of massive disaster area images is required in emergency rescue scenarios; meanwhile, complex models are difficult to deploy on on-orbit or airborne edge computing devices due to constraints on power consumption and computing capability. Therefore, exploring lightweight and efficient remote sensing image super-resolution methods that can reduce model complexity and computational overhead while maintaining or even improving reconstruction quality possesses practical engineering application value.

Early studies such as CARN[5] reduced overall complexity by adopting cascaded lightweight residual blocks. IDN[6] and its improved version IMDN[7] systematically proposed the concept of information distillation. Through multi-stage channel splitting, feature refinement and fusion operations, these methods effectively enhance the feature utilization efficiency of lightweight models. RFDN[8] re-examined the channel splitting technique and introduced progressive optimization modules. Recently, RLFN[9] reflected on the redundancy introduced by multi-branch structures and adopted a minimalist single-path design, combined with contrastive loss to strengthen feature discriminability. As research advances, dedicated lightweight methods tailored to the characteristics of optical remote sensing images have gradually emerged. For example, CTN[10] designed context transformation layers to replace standard convolutions for parameter reduction; FeNet[11] constructed lightweight lattice blocks based on channel separation to realize lightweight nested module design. These studies explored ways to streamline model parameters and computational complexity from different perspectives.

Meanwhile, Transformer-based architectures have attracted extensive attention due to their outstanding performance, yet their enormous computational overhead

hinders practical deployment. To address this issue, researchers have committed to improving the efficiency of Vision Transformers (ViTs), giving rise to a series of lightweight variants[12-15]. The vanilla multi-head self-attention mechanism suffers from high computational complexity, hence mechanisms such as local window attention and shifted window attention have been proposed to cut down computational costs. Convolution-attention hybrid architectures have become one of the mainstream efficient paradigms. For example, MobileViT[12] integrates MobileNet blocks and multi-head self-attention blocks, while EdgeViT[13] combines self-attention and convolution to achieve more cost-effective information interaction. In terms of efficiency optimization, EfficientFormer[14] proposed a dimension-consistent design paradigm to balance latency and model performance; FastViT[15] introduced structural reparameterization and large-kernel convolution to boost the performance of hybrid models without increasing inference latency. These in-depth efficiency optimizations lay a solid foundation for the practical deployment of Transformer architectures in resource-limited scenarios.

Nevertheless, despite the above advancements, the core feature mixing operations of existing lightweight models still mostly rely on basic modules such as self-attention and convolution[16]. Such reliance fundamentally restricts the efficiency and effectiveness of context perception and feature aggregation in lightweight networks, often making it difficult for models to balance representation capability and inference speed. Essentially, context perception and feature aggregation constitute two core processes of feature mixing, aiming to facilitate spatial information fusion[17-18]. The former is responsible for modeling contextual relationships among features, while the latter integrates features according to such relationships. Existing lightweight models adopt completely different implementation routes for self-attention and convolution: self-attention realizes perception through global feature interaction and conducts global aggregation via weighted summation of all features; convolution performs perception based on the relative positional relationship of features and implements aggregation with static convolution kernel weights. However, self-attention is plagued by high global computational complexity, and static local convolution operations struggle to model long-range dependencies. Accordingly, designing a lightweight remote sensing super-resolution network that can achieve balanced context perception and flexible feature aggregation to compensate for the deficiencies of existing self-attention and convolution mechanisms has become a key research challenge. This also provides the research direction for the subsequent exploration of lightweight models in this paper.

## **2. Related Work**

### **2.1. Lightweight optimization method**

To facilitate the deployment of deep super-resolution models in practical scenarios

with limited computational resources, a variety of lightweight optimization methods have been developed. These methods mainly follow three core directions: model compression, efficient architecture design, and remote sensing data-oriented dedicated optimization. They aim to substantially reduce model parameters, computational complexity and inference latency, while preserving reconstruction performance to the greatest extent possible.

Model compression techniques[19-22] are adopted to streamline pre-trained or newly trained networks, with mainstream approaches including network pruning and parameter quantization. Network pruning reduces parameters and computational overhead with slight performance degradation by removing redundant weight connections or entire channels within the model. Parameter quantization compresses model size and accelerates computation by lowering the numerical precision of weights and activation values. Nevertheless, such post-processing techniques may impair the representation capability of models, and are likely to cause the loss of texture details when processing remote sensing images with abundant textural features.

Designing network architectures from the model origin serves as another effective way to achieve lightweight deployment. In recent years, decomposed and efficient convolutions have been widely applied. For instance, depthwise separable convolution[23] decomposes standard convolution into depthwise convolution and pointwise convolution, which greatly reduces computational complexity and has become a fundamental component of numerous lightweight models. In addition, the lightweight reconstruction of attention mechanisms has become a popular research focus in recent years. Other approaches such as structural reparameterization technology[24-28] allow the adoption of complex multi-branch structures during training to enrich feature extraction, which can be equivalently converted into a simple single-path structure in the inference phase. This design improves model performance without increasing inference time. The introduction of dynamic feature distillation and selection mechanisms[29-31] enables networks to adaptively focus on critical features and suppress redundant information, further boosting computational efficiency. Moreover, as a model-agnostic compression paradigm, knowledge distillation enables small-scale networks to learn the behavior or feature representations of complex networks, achieving performance comparable to large models with far fewer parameters.

Dedicated lightweight optimization for remote sensing images needs to further take inherent data characteristics into account. Remote sensing images feature large size and wide coverage, and directly applying lightweight models designed for natural images may lead to insufficient modeling of long-range dependencies. Accordingly, existing studies devote efforts to designing local-global feature interaction modules, so as to enhance the modeling capability for large-scale ground object structures under lightweight constraints. Meanwhile, aiming at the common problems of weak

texture regions and mixed complex degradation in remote sensing images, lightweight design needs to avoid excessive texture smoothing or spectral distortion caused by over-compression, which puts forward higher requirements for balancing feature retention and computational simplification. At present, hybrid models combining attention mechanisms and lightweight convolutions, as well as hardware-aware neural architecture search, are emerging as cutting-edge research directions for lightweight super-resolution of remote sensing images.

## 2.2. Introduction of Lightweight Models

Lightweight model design is critical to promoting the deployment of deep neural networks in practical scenarios with constrained computational resources. This subsection reviews relevant models and their technical ideas from three aspects: general lightweight CNN models, super-resolution-oriented lightweight designs, and lightweight Vision Transformers, so as to provide theoretical and technical references for the lightweight network model design in this paper.

The exploration of CNN lightweight design has formed a mature technical system for general computer vision tasks. The introduction of depthwise separable convolution marks a landmark breakthrough. Models such as MobileNet[[22] and Xception[23] decompose standard convolution into depthwise convolution and pointwise convolution, which drastically reduces computational complexity. Subsequently, MobileNetV2[32] proposes the inverted residual structure and linear bottleneck to further optimize feature flow and information density. To facilitate information transfer between groups, the ShuffleNet series introduces channel shuffle and channel split operations. In addition, hardware-aware neural architecture search is adopted to automatically explore compact network structures with optimal latency for specific hardware platforms. Notably, to compensate for the inherent limitation of limited receptive fields in lightweight CNNs, enhancing their capability to model long-range dependencies has become a new research direction. For instance, CFSRCNN[33] acquires a more comprehensive receptive field by extending short-path features to deep long-path features; ParC-Net[34] obtains a global receptive field via position-aware circular convolution; AFFNet[35] realizes global convolution by adopting adaptive frequency filtering and circular padding; LrfSR[36] constructs an information distillation module with a large receptive field through dilated convolution. The expanded receptive field enables the network to capture more pixel correlations and fuse multi-scale information. These studies provide core components for constructing basic visual backbones with high efficiency and strong representation capability.

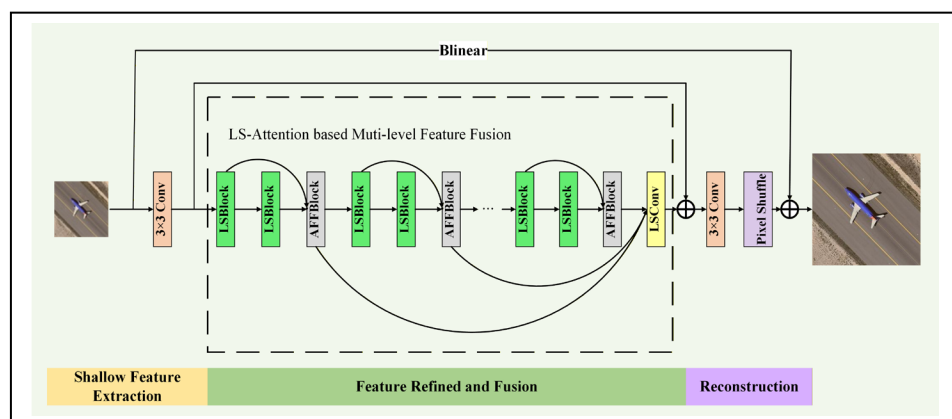
As Vision Transformers have demonstrated superior performance in visual tasks, research on their lightweight variants has flourished. RetNetSR[37] leverages Manhattan distance to obtain spatial priors for self-attention enhancement, and transfers the concept of temporal decay into the spatial domain. CATANet[38]

proposes an efficient content-aware token aggregation module to aggregate tokens with similar remote content, and further realizes long-range information interaction through intra-group self-attention. These studies lay a crucial technical foundation for the application of Transformers in computation-sensitive remote sensing super-resolution tasks.

In summary, current lightweight model design has evolved into a comprehensive system integrating multiple technical routes and collaborative multi-objective optimization. Core design concepts ranging from depthwise separable convolution and dynamic feature distillation to hybrid attention mechanisms offer a valuable technical toolkit for the lightweight network design targeted at optical remote sensing image super-resolution.

### 3. Methodology

To address the insufficient modeling of long-range dependencies and the problems of excessive texture smoothing or distortion caused by over-compression existing in dedicated lightweight models for remote sensing images, this section elaborates on the overall framework and workflow of the proposed Large-kernel Small-span Feature Fusion Network (LSFFNet). Tailored specifically for lightweight remote sensing image super-resolution tasks, the complete processing pipeline of the model is illustrated in **Figure 1**. The model architecture is sequentially composed of three core components: the Large-kernel Small-span Feature Extraction Block (LSBlock), the Attention Multi-level Feature Fusion Block (AFFBlock), and the image reconstruction module.



**Figure 1.** The network architecture of LSFFNet consists of shallow feature extraction, feature reconstruction and fusion, as well as image reconstruction modules.

#### 3.1. Overall framework of the model

Given an LR remote sensing image  $I_{lr} \in \mathbb{R}^{H \times W \times C_{in}}$ , the model first employs a  $3 \times 3$  convolutional layer  $E_{FE}(\cdot)$  to extract shallow features  $F_0' \in \mathbb{R}^{H \times W \times C}$ . These features carry basic image information and serve as the key input for subsequent

deep processing, which is formulated as Eq.(1):

$$\mathbf{F}_0' = E_{FE}(\mathbf{I}_{lr}'). \quad (1)$$

Subsequently, the shallow feature  $\mathbf{F}_0'$  is fed into the deep feature extraction backbone network  $E_{DFBN}(\cdot)$ , as shown in Eq.(2):

$$\mathbf{F}_D = E_{DFBN}(\mathbf{F}_0'), \quad (2)$$

The network  $E_{DFBN}(\cdot)$  consists of two core modules, LSBlock and AFFBlock. By simulating the dynamic scale visual perception capability of the human visual system, LSBlock is able to capture local features and contextual information. After obtaining essential local features and global contextual information, the proposed AFFBlock fuses multi-level local features into global contextual information, which improves the representation capability of deep features to a favorable extent.

Finally, the model generates the ultimate SR image  $\mathbf{I}_{sr}'$  through the reconstruction module combined with linear interpolation and residual connection. Specifically, residual feature fusion is implemented by summing the original shallow feature  $\mathbf{F}_0$  and deep feature  $\mathbf{F}_{DF}$ . The fused features are then upsampled and reconstructed via a high-quality image reconstruction module  $E_{HQR}(\cdot)$  containing a sub-pixel convolutional layer. Ultimately, the reconstruction result is aggregated with the bilinear interpolation output  $E_{Bilinear}(\mathbf{I}_{lr}')$ . The overall process is expressed in Eq. (3):

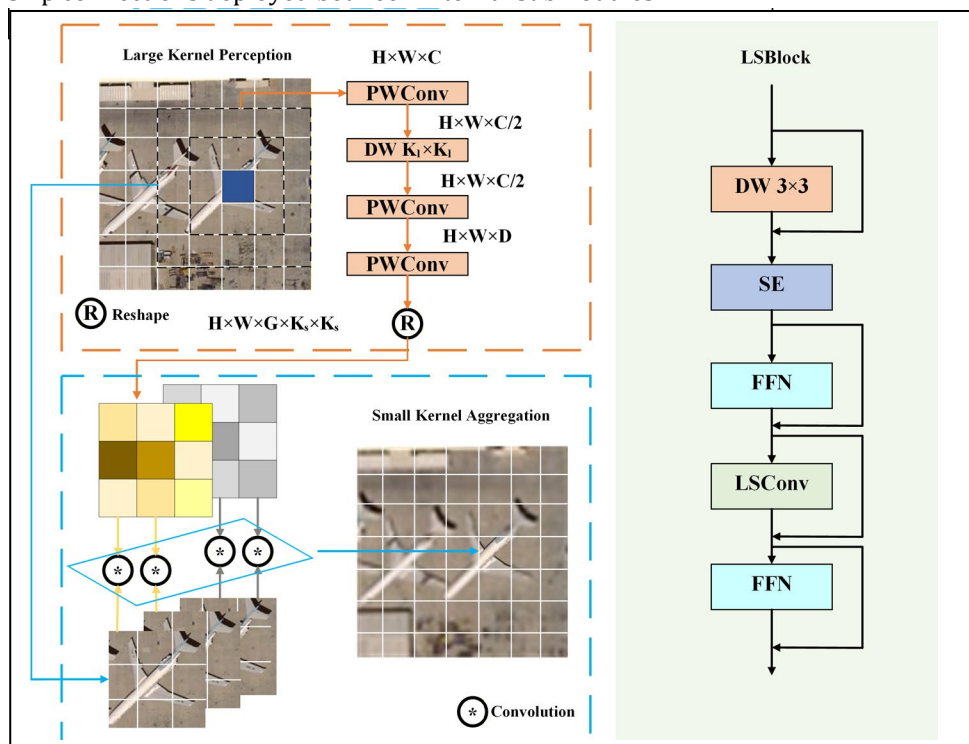
$$\mathbf{I}_{sr} = E_{HQR}(\mathbf{F}_0 + \mathbf{F}_{DF}) + E_{Bilinear}(\mathbf{I}_{lr}'). \quad (3)$$

Remote sensing images differ greatly from natural images in scene composition, where each scene generally contains complex target distributions. Moreover, targets in remote sensing images usually occupy smaller scales compared with those in natural images, posing great challenges for remote sensing image super-resolution. The proposed LSFFNet addresses these difficulties via joint local and global processing. At the local level, LSBlock is utilized for feature extraction and enhancement to preserve the detail fidelity of all ground objects. At the global level, the AFFBlock architecture integrates multi-scale information and facilitates the fusion of diverse features. With the increase of network depth, LSBlock refines feature representations and recovers detailed contours. Meanwhile, AFFBlock acts as a perceptive feature filter, guiding the model to focus on the most relevant feature information. This ensures that the generated super-resolved images maintain global coherence and local precision. In addition, AFFBlock requires fewer parameters and multiplication operations, providing a feasible strategy to boost spatial resolution without sacrificing computational efficiency.

### 3.2. Large-kernel Small-span Feature Extraction Block

To achieve efficient feature extraction and fusion within lightweight models, this section presents the LSBlock. As illustrated in **Figure 2**, the LSBlock is mainly composed of depthwise convolution, Squeeze-and-Excitation (SE) block, Feed Forward Network (FFN), and Large-kernel Small-span Convolution (LSConv), with

skip connections deployed between internal submodules.



**Figure 1.** Implementation Framework and Pipeline of Large-kernel Small-span Feature Extraction Block and Large-kernel Small-span Convolution

This module is capable of modeling relational dependencies through large-range perception, followed by extracting and fusing key features within small-range regions with high feature coherence, which further enhances the model's ability to perceive and capture image features. Specifically, for the extracted token  $x_i$ , the contextual regions for perception and fusion are denoted as  $\Gamma_P(x_i)$  and  $\Gamma_A(x_i)$ , respectively, where the spatial coverage of  $\Gamma_P(x_i)$  is larger than that of  $\Gamma_A(x_i)$ . This process can be formulated as Eq.(4):

$$y_i = A\left(P(x_i, \Gamma_P(x_i)), \Gamma_A(x_i)\right), \quad (4)$$

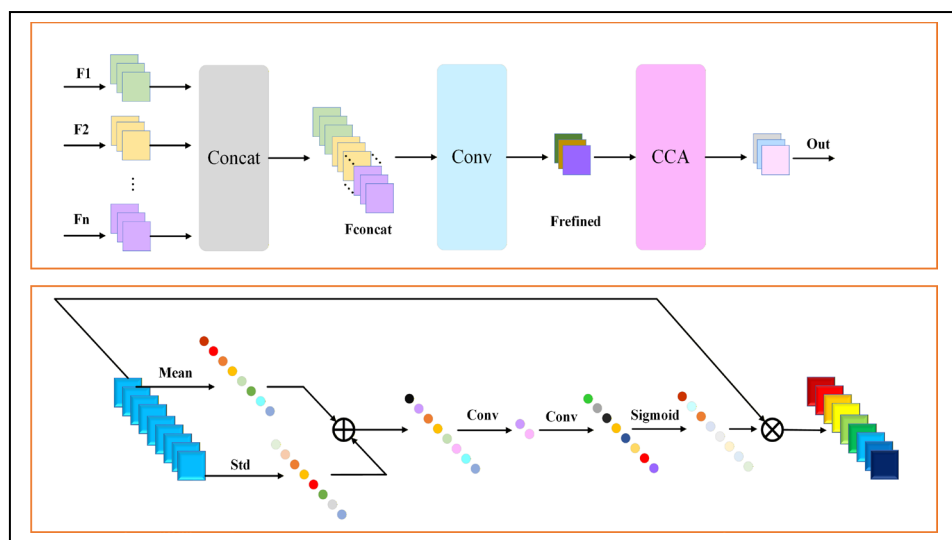
Here,  $P$  denotes the perception operation, which extracts contextual information and captures dependencies among tokens;  $A$  represents the fusion operation, which integrates features based on perceptual results and enables information aggregation from other tokens. There exists a discrepancy in the contextual scope involved in perception and fusion procedures. Such discrepancy facilitates the utilization of multi-scale contextual information, allowing the simultaneous capture of global contextual semantics and local texture details. For perception tasks requiring a large spatial scope, large-kernel depthwise convolution is adopted to reduce computational overhead. By contrast, adaptive weighted feature summation is more suitable for fusion tasks confined to small surrounding regions. A limited fusion range not only guarantees computational efficiency at a low computational cost but

also alleviates the redundant fusion operations introduced by self-attention mechanisms. On this basis, the LSConv is proposed, as shown in **Figure 2**. It consists of two sequential steps: large-kernel perception  $P_{LS}$  and small-kernel aggregation  $A_{LS}$ . The large-kernel static convolution simulates and expands the receptive field to model neighborhood relationships, while the small-kernel dynamic convolution adaptively fuses surrounding contextual features.

### 3.3. Attention Multi-level Feature Fusion Block

Hierarchical information fusion is an essential step in the field of optical remote sensing image super-resolution. Accordingly, an Attention Multi-level Feature Fusion Block (AFFBlock) is designed to further enhance multi-layer feature information obtained from LSBlock at different hierarchical levels. The core of AFFBlock is an attention-based fusion unit. As shown in **Figure 3**, it starts with the concatenation of  $n$  input feature maps, each containing  $C$  channels. This process is expressed in Eq.(5):

$$F_{concat} = \text{Concat}(F_1, F_{i+1}, \dots, F_{i+n}), \quad (5)$$



**Figure 1.** Implementation Framework and Pipeline of Attention Multi-level Feature Fusion Block and Contrast-Aware Channel Attention

where  $F_{concat}$  represents the concatenated features, and  $F_{i+n}$  denotes the  $n$ -th input feature. Subsequently, pointwise convolution (PW) is adopted for dimensionality reduction, and the corresponding process is defined in Eq. (6)

$$F_{refined} = PW(F_{concat}), \quad (6)$$

where  $F_{refined}$  denotes the feature after dimensionality reduction. Subsequently, the Contrast Perception Channel Attention (CCA) module is adopted to enhance  $F_{refined}$  and obtain the final output  $Out_i$ , with the process expressed in Eq.(7).

$$Out_i = E_{CCA}(F_{refined}). \quad (7)$$

As shown in **Figure 3**, given a set of feature maps, the sum of their standard deviation and mean value is calculated. The summed result is then fed into a nonlinear function sequence consisting of  $PW \rightarrow \text{ReLU} \rightarrow PW \rightarrow \text{Sigmoid}$ . Finally, the input feature maps are multiplied by the generated combination coefficients to obtain the output feature maps. This structural design enables AFFBlock to adapt to different channel dimensions and enhance feature representation capability. Conventional channel concatenation tends to cause excessive accumulation of redundant features, while standard  $1 \times 1$  convolutional layers often fail to sufficiently highlight critical features. To address these limitations, the grouping strategy is first applied to block-wise feature fusion, followed by the introduction of attention-based fusion units. This configuration is designed to achieve superior multi-level feature fusion performance.

#### 4. Experiment

To verify the effectiveness of the proposed LSFFNet model, comprehensive quantitative comparisons with a variety of state-of-the-art super-resolution methods are conducted on the UCMerced and RSSCN7 datasets in this subsection. The comparison methods include CFSRCNN[29], CTN[10], FENet[11], RLFN[9], RetNetSR[33], CATANet[32] and LrfSR[34]. The evaluation is implemented from two dimensions of reconstruction accuracy and model complexity.

**Table 1.** Results on the UCMerced test dataset. The best and second-best quantitative results are marked in bold and underlined, respectively.

Dataset	Scale	Metric	CFSRCNN	CTN	FENet	RLFN	RetNetSR	LrfSR	CATANet	LSFFNet
UCMerced	x2	PSNR	32.09	32.12	31.76	32.20	32.23	32.21	32.24	32.32
		SSIM	0.9017	0.9024	0.8947	0.9026	0.9029	0.9031	0.9030	0.9043
		Params	1310.24K	349.16K	351.33K	455.18K	481.91K	389.39K	477.13K	298.43K
		FLOPS	232.20G	29.45G	38.60G	49.40G	42.73G	28.31G	23.40G	30.10G
	x3	PSNR	27.13	27.30	27.24	27.25	27.28	27.33	27.29	27.41
		SSIM	0.8055	0.8094	0.8083	0.8112	0.8120	0.8128	0.8125	0.8136
		Params	1495.01K	349.16K	357.95K	461.00K	488.40K	389.39K	550.75K	305.25K
		FLOPS	169.95G	18.33G	17.52G	22.30G	31.68G	19.77G	16.28G	18.74G
	x4	PSNR	27.08	26.88	26.99	27.02	27.06	27.09	27.11	27.19
		SSIM	0.7456	0.7198	0.7271	0.7432	0.7463	0.7475	0.7487	0.7502
		Params	1458.23K	360.00K	366.12K	470.13K	497.57K	389.39K	535.43K	328.39K
		FLOPS	156.32G	12.66G	10.16G	12.81G	14.98G	13.58G	11.47G	7.95G

**Table 1** lists the experimental results on the UCMerced dataset. In terms of reconstruction accuracy, LSFFNet achieves the optimal PSNR and SSIM values among all comparative methods under the  $\times 2$ ,  $\times 3$  and  $\times 4$  super-resolution scaling factors, delivering stable overall performance. Compared with the second-best approach, both metrics attain steady marginal improvements. The

reconstruction performance shows minor discrepancies across different scaling factors, reflecting well-balanced adaptability to various super-resolution application scenarios.

From the perspective of model complexity, LSFFNet obtains the smallest parameter volume among all contrastive models, with its parameter quantity kept below 330K at all scaling scales. In terms of computational overhead, the FLOPs of LSFFNet are only 7.95G at the  $\times 4$  scaling factor, which is the lowest among all competing methods. Meanwhile, its FLOPs under  $\times 2$  and  $\times 3$  scaling factors stay at a moderately low level, proving that the adopted lightweight design effectively controls the overall model scale.

## 5. Conclusion

This paper elaborates on LSFFNet, a lightweight optical remote sensing image super-resolution method designed for resource-constrained scenarios. Firstly, it analyzes the prevalent dilemma faced by existing lightweight models, which struggle to balance reconstruction quality and lightweight deployment performance. Secondly, the overall architecture of the proposed LSFFNet is introduced, with its core composed of two key modules: LSBlock and AFFBlock. LSBlock combines large-receptive-field depthwise separable convolution and dynamic convolution. It extracts local high-frequency features and global contextual information with relatively low parameters, and further optimizes local features through small-scale dynamic convolution. AFFBlock leverages contrast perception channel attention to adaptively fuse multi-level features, thereby enhancing the structural consistency and detail restoration capability of reconstructed images.

Finally, comparative experiments with existing mainstream lightweight super-resolution models verify that LSFFNet delivers competitive performance in quantitative indicators including PSNR and SSIM. In summary, built upon elaborate lightweight design, the LSFFNet proposed in this chapter achieves optical remote sensing image super-resolution reconstruction that balances reconstruction performance and computational efficiency. It provides a viable lightweight solution for remote sensing image super-resolution tasks in resource-limited application scenarios.

## References

- [1] Qin J, Xiong J, Liang Z. CNN - Transformer gated fusion network for medical image super-resolution[J]. Scientific Reports, 2025, 15(1): 15338.
- [2] Kang X, Duan P, Li J, et al. Efficient swin transformer for remote sensing image super-resolution[J]. IEEE Transactions on Image Processing, 2024, 33: 6367-6379.
- [3] Zhu C, Liu Y, Huang S, et al. Taming a diffusion model to revitalize remote sensing image super-resolution[J]. Remote Sensing, 2025, 17(8): 1348.
- [4] Li W, Guo H, Liu X, et al. Efficient face super-resolution via wavelet-based feature enhancement network[C]//Proceedings of the 32nd ACM international conference on multimedia. 2024: 4515-4523.

- [5] Ahn N, Kang B, Sohn K A. Fast, accurate, and lightweight super-resolution with cascading residual network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 252-268.
- [6] Hui Z, Wang X, Gao X. Fast and accurate single image super-resolution via information distillation network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 723-731.
- [7] Hui Z, Gao X, Yang Y, et al. Lightweight image super-resolution with information multi-distillation network[C]//Proceedings of the 27th acm international conference on multimedia. 2019: 2024-2032.
- [8] Liu J, Tang J, Wu G. Residual feature distillation network for lightweight image super-resolution[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 41-55.
- [9] Kong F, Li M, Liu S, et al. Residual local feature network for efficient super-resolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 766-776.
- [10] Wang S, Zhou T, Lu Y, et al. Contextual transformation network for lightweight remote-sensing image super-resolution[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-13.
- [11] Wang Z, Li L, Xue Y, et al. Feature enhancement network for lightweight remote-sensing image super-resolution[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-12.
- [12] Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv preprint arXiv:2110.02178, 2021.
- [13] Pan J, Bulat A, Tan F, et al. Edgevits: Competing light-weight cnns on mobile devices with vision transformers[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 294-311.
- [14] Li Y, Yuan G, Wen Y, et al. Efficientformer: Vision transformers at mobilenet speed[J]. Advances in neural information processing systems, 2022, 35: 12934-12949.
- [15] Vasu P K A, Gabriel J, Zhu J, et al. Fastvit: A fast hybrid vision transformer using structural reparameterization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 5785-5795.
- [16] Tolstikhin I O, Houshy N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in neural information processing systems, 2021, 34: 24261-24272.
- [17] Fan Q, Huang H, Zhou X, et al. Lightweight vision transformer with bidirectional interaction[J]. Advances in Neural Information Processing Systems, 2023, 36: 15234-15251.
- [18] Yang J, Li C, Dai X, et al. Focal modulation networks[J]. Advances in Neural Information Processing Systems, 2022, 35: 4203-4217.
- [19] Yang D, Solihin M I, Zhao Y, et al. Model compression for real-time object detection using rigorous gradation pruning[J]. Iscience, 2025, 28(1).
- [20] Zawish M, Davy S, Abraham L. Complexity-driven model compression for resource-constrained deep learning on edge[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(8): 3886-3901.
- [21] Liu C Y, Kuo E J, Abraham Lin C H, et al. Quantum-train: Rethinking hybrid quantum-classical machine learning in the model compression perspective[J]. Quantum Machine Intelligence, 2025, 7(2): 80.
- [22] Tian J, Solgi R, Lu J, et al. Flat-llm: Fine-grained low-rank activation space transformation for large language model compression[C]//Findings of the Association for Computational Linguistics: EACL 2026. 2026: 2988-3002.
- [23] Ma X, Zhai K, Luo N, et al. Gearbox fault diagnosis under noise and variable operating conditions using multiscale depthwise separable convolution and bidirectional gated recurrent unit with a squeeze-and-excitation attention mechanism[J]. Sensors, 2025, 25(10): 2978.

- [24] Ding X, Zhang X, Ma N, et al. Repvgg: Making vgg-style convnets great again[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13733-13742.
- [25] Ding X, Guo Y, Ding G, et al. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1911-1920.
- [26] Setyawan N, Sun C C, Hsu M H, et al. FaceLiVT: Face Recognition using Linear Vision Transformer with Structural Reparameterization For Mobile Device[C]//2025 IEEE International Conference on Image Processing (ICIP). IEEE, 2025: 1720-1725.
- [27] Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11963-11975.
- [28] Ding X, Chen H, Zhang X, et al. Re-parameterizing your optimizers rather than architectures. arXiv 2022[J]. arXiv preprint arXiv:2205.15242, 4.
- [29] Tian H, Xu B, Li S. Distillation dynamics: Towards understanding feature-based distillation in vision transformers[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2026, 40(11): 9520-9528.
- [30] Zuo R, Li Y, Wei S, et al. Calibration-augmented and mechanism-driven deep learning hybrid framework for modeling actual distillation processes[J]. Industrial & Engineering Chemistry Research, 2025, 64(7): 3856-3870.
- [31] Liu Y, Feng W, Liu Z, et al. Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 21679-21688.
- [32] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.
- [33] Tian C, Xu Y, Zuo W, et al. Coarse-to-fine CNN for image super-resolution[J]. IEEE Transactions on Multimedia, 2020, 23: 1489-1502.
- [34] Zhang H, Hu W, Wang X. Parc-net: Position aware circular convolution with merits from convnets and transformer[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 613-630.
- [35] Huang Z, Zhang Z, Lan C, et al. Adaptive frequency filters as efficient global token mixers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 6049-6059.
- [36] Wang W, Che S, Liu W, et al. A lightweight large receptive field network LrfSR for image super-resolution[J]. Scientific Reports, 2025, 15(1): 12535.
- [37] Gendy G, Sabor N, Al Marzouqi H. Lightweight image super-resolution based on retentive network[J]. Neural Computing and Applications, 2026, 38(5): 130.
- [38] Liu X, Liu J, Tang J, et al. Catanet: Efficient content-aware token aggregation for lightweight image super-resolution[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 17902-17912.