Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-6, June 2025
DOI: 10.53469/jrse.2025.07(06).03

Lightweight Image Super-Resolution Via Superpixel Prior and Feature Reinforcement

Rui Xu[1], Wei Fan[2]

[1,2]Chongqing University of Technology, Chongqing 400054, China
[1]xurui@zqlgdx19.wecom.work, [2]19910039@cqut.edu.cn

Abstract: ***Single Image Super-Resolution (SISR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs, serving as a core task in computer vision. Despite recent advances, existing methods often struggle to balance structural fidelity and computational efficiency. To address this, we propose PFRNet, a lightweight superpixel-aware model integrating superpixel segmentation, local attention aggregation, and global structure modeling. The framework comprises four key modules: GASS, SPDF, SPFA, and LAP, jointly enabling multi-scale and structure-consistent feature learning. Experiments on benchmark datasets (e.g., Set5, Urban100) show that PFRNet achieves superior performance with fewer parameters. Ablation studies further verify the effectiveness of each module.***

Keywords: Image Super-Resolution, Superpixel Awareness, Lightweight, Structural Modeling.

1. Introduction

Single Image Super-Resolution (SISR) reconstructs a high-resolution (HR) image from a low-resolution (LR) input and is widely used in medical imaging, remote sensing, surveillance, and photography. Deep CNNs and Transformers have boosted SISR accuracy, yet balancing structural fidelity and efficiency remains difficult.

Studies show structural priors [1] and adaptive attention [2] improve texture and edge recovery. However, two issues persist: (1) fixed-shape attention struggles with irregular structures, reducing precision; (2) Transformer-based global attention, though effective, suffers from high quadratic complexity [3], limiting lightweight deployment.

To address these challenges, we propose PFRNet, a superpixel-driven network centered on the Superpixel Prior Aggregation Module (SPAM), which integrates four synergistic submodules for efficient, structure-aware, multi-scale image enhancement:

1) **Global-Aware Superpixel Sampling (GASS):** Uses spatial and gradient cues to generate structurally coherent superpixels.
2) **Superpixel-Guided Dual-Stream Fusion (SPDF):** Fuses pixel- and region-level features via dual attention for local-global balance.
3) **Superpixel-Focused Attention (SPFA):** Enhances regional discriminability through sparse attention on key pixels.
4) **Local Attention Partition (LAP):** Preserves textures and reduces artifacts via overlapping local attention.

By integrating these submodules, PFRNet leverages superpixel consistency to enhance detail reconstruction and model long-range dependencies. Experiments on standard benchmarks confirm its superior performance–efficiency trade-off.

2. Related Work

2.1 Super-Resolution Reconstruction Networks

CNN-based methods have achieved remarkable progress in SISR. SRCNN [4] first introduced end-to-end mapping from LR to HR. VDSR [5] leveraged deeper residual networks, while EDSR [6] simplified blocks to boost performance. RCAN [7] employed channel attention to enhance feature focus, followed by SAN [8], HAN [9], and NLSA [10], which incorporated spatial and non-local attention for structure restoration. However, these models often fail to capture fine structural details, leading to blurring or over-smoothing in textures and edges.

2.2 Lightweight Super-Resolution Methods

To meet the constraints of mobile devices, lightweight designs like FSRCNN [4] and ESPCN [11] defer upsampling to reduce computation. CARN [12] and IMDN [13] further optimize efficiency via group convolutions and feature distillation. Transformer-based models such as SwinIR [14] use sliding-window attention to balance accuracy and cost. ELAN [15] extends this with GMSA for faster global modeling. However, most rely on fixed or windowed partitions, neglecting natural boundaries and textures, which leads to fragmented features and weak contextual understanding in complex scenes.

2.3 Superpixel and Pixel Clustering Modeling

To overcome structural limitations of patch-based methods, recent work introduces superpixel segmentation for structure-aware modeling. Superpixels group pixels into perceptually coherent regions, naturally aligning with edges and semantic boundaries. SSN [16] proposed a differentiable superpixel generator via soft k-means, enabling integration with GNNs and attention modules. Other works combine pixel clustering and graph convolutions, e.g., PAN forms semantic graphs to improve texture reconstruction. However, superpixel-based priors remain underexplored in SISR due to integration complexity and sensitivity to scale, lighting, and texture. Designing lightweight, structure-aware aggregation remains a key challenge.

3. Proposed Method

To balance global semantics and local details, we propose the Superpixel Prior Aggregation Module (SPAM) as the core of PFRNet. SPAM comprises four submodules—GASS, SPDF, SPFA, and LAP—targeting global modeling, context fusion, structural refinement, and detail reconstruction. See Figure 1.

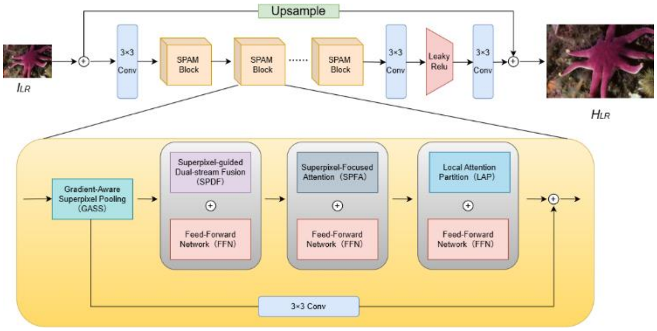
Upsample
3×3 Conv
SPAM Block
SPAM Block
SPAM Block
3×3 Conv
Leaky Relu
3×3 Conv
I_{LR}
H_{LR}
Gradient-Aware Superpixel Pooling (GASS)
Superpixel-guided Dual-stream Fusion (SPDF)
Feed-Forward Network (FFN)
Superpixel-Focused Attention (SPFA)
Feed-Forward Network (FFN)
Local Attention Partition (LAP)
Feed-Forward Network (FFN)
3×3 Conv

Figure 1: Network Architecture of the Proposed Model

The proposed Superpixel Prior Aggregation Module (SPAM) integrates four lightweight yet synergistic components to achieve structure-aware and efficient image reconstruction. GASS groups images into coherent regions to extract coarse semantic priors and facilitate cross-region flow. SPDF employs superpixels as anchors for deformable attention, aligning multi-scale context with non-rigid structures. SPFA enhances intra-region consistency through Top-K pixel attention, while LAP applies sliding-window attention to preserve textures and edges. An overlapping fusion strategy further mitigates block artifacts and promotes high-fidelity restoration.

3.1 Gradient-Aware Superpixel Pooling (GASS)

To capture structural information, we propose the Gradient-Aware Superpixel Pooling (GASS) module. It uses edge gradients to guide superpixel initialization and constructs a soft association matrix based on feature distances. An iterative optimization refines the segmentation, enhancing spatial consistency and structural perception. See Figure 2 for the workflow.

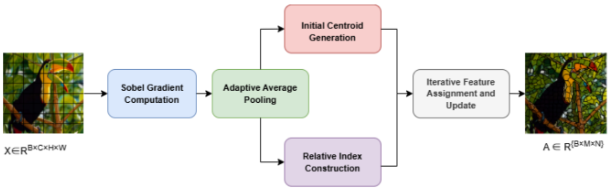
$X \in R^{B\times C\times H\times W}$
Sobel Gradient Computation
Adaptive Average Pooling
Initial Centroid Generation
Relative Index Construction
Iterative Feature Assignment and Update
$A \in R^{(B\times M\times N)}$

Figure 2: Overall Workflow of the GASS Module

The input is a batch of image features with shape $X \in R^{B\times C\times H\times W}$, where B denotes the batch size, C is the number of channels, and H and W represent the height and width of the image, respectively. The GASS module outputs two tensors: One is a soft-assignment matrix $A \in R^{B\times N\times P}$, where N is the number of superpixels and $P = H \times W$ is the total number of pixels. The other is a vector representing the number of superpixels N, which is used for subsequent region-based operations in the module. To ensure that the initial superpixel segmentation is structurally aware, GASS first applies the Sobel operator for edge detection, extracting gradient maps in the horizontal and vertical directions, denoted as G_x and $G_y \in R^{B\times C\times H\times W}$, respectively. Then, the overall gradient magnitude map is computed as:

$$G = \sqrt{G_x^2 + G_y^2} \tag{1}$$

The gradient map and the original image features are then separately subjected to adaptive average pooling, yielding the initial superpixel centroids under gradient guidance C_g and the average image feature values C_x:

$$C = \frac{1}{2}\left(AvuPool(G) + AvgPool(X)\right) \\ C \in R^{B\times C\times N} \tag{2}$$

Among them, the number of superpixels is defined as $N = H_s \times W_s$, which is determined by the user-specified token resolution (H_s, W_s). In addition, we generate the initial label mapping for each pixel using nearest-neighbor interpolation, denoted as $L_0 \in R^{B\times P}$, which is used to quickly locate the corresponding initial superpixel index for each pixel. In each iteration, GASS computes the distance between each pixel and the superpixel centroids within its 3×3 neighborhood, and obtains the similarity-normalized soft-assignment weights through the softmax function:

$$A_{ij} = \frac{exp\left(-\|x_i - c_j\|^2\right)}{\sum_{k\in N(i)} exp(-\|x_i - c_k\|^2)} \tag{3}$$

Let $x_i \in R^C$ denote the feature of the i-th pixel, and $c_j \in R^C$ denote the centroid of the j-th superpixel. $N(i)$ represents the set of neighboring superpixel centroids for pixel i, and A_{ij} is the probability that pixel i belongs to superpixel j. To improve efficiency, GASS constructs a sparse soft-assignment tensor A_{sparse}, avoiding the need to compute distances to all superpixels and significantly reducing computational and memory overhead. In each iteration, GASS updates each superpixel centroid by recalculating it based on the current soft-assignment:

$$c_j^{(k+1)} = \frac{\sum_i A_{ij}^{(k)} \cdot x_i}{\sum_i A_{ij}^{(k)} + \varepsilon} \tag{4}$$

Here, ε is a numerical stabilization term. Through iterative refinement (typically 2–3 iterations), the clustering process gradually converges, making the superpixel centroids increasingly approximate the average features of their assigned pixels, thereby improving the consistency of region aggregation. The final output soft-assignment matrix A serves as the input for region-level tasks such as region attention and region-level graph modeling.

3.2 Superpixel-guided Dual-stream Fusion (SPDF)

To exploit structural–semantic complementarity, we propose the Superpixel-guided Dual-stream Fusion (SPDF) module. It splits features into structural and semantic streams for local–global modeling. Using superpixel tokens as anchors, a bidirectional attention mechanism enables efficient pixel-level fusion. The workflow of SPDF is shown in Figure 3.

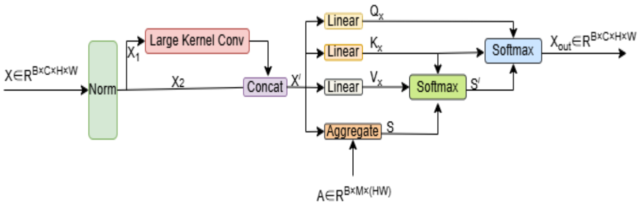
$X \in R^{B\times C\times H\times W}$
Norm
X_1
Large Kernel Conv
X_2
Concat
X'
Linear
Q_x
Linear
K_x
Linear
V_x
Aggregate
S
Softmax
S'
Softmax
$X_{out} \in R^{B\times C\times H\times W}$
$A \in R^{B\times M\times (HW)}$

Figure 3: Overall Workflow of the SPDF Module

Convolutions excel at local texture modeling but struggle with long-range dependencies. To address this, we introduce superpixels as intermediaries, enabling efficient non-local information propagation. Compared to full-image attention, superpixel tokens reduce computation while preserving spatial awareness and regional coherence. The input feature $X \in R^{B\times C\times H\times W}$ is first normalized and then split along the channel dimension into two substreams: the structural stream X_1 and the semantic stream X_2. The structural stream enhances local structural perception — such as edges and textures — through large-receptive-field convolution operations.

$$X_1' = LargeKernelConv(X_1) \tag{5}$$

The semantic stream preserves the original semantic information. The two streams are then fused to obtain the enhanced feature representation X':

$$X' = Concat(X_1', X_2) \tag{6}$$

Given the pixel-level feature X' and the superpixel-to-pixel affinity matrix $A \in R^{B\times M\times (HW)}$ (generated by the GASS module), we perform an aggregation operation to generate M superpixel-level tokens.

$$S = \frac{A\cdot X'_{flat}}{\sum A} \; where \; X'_{flat} \in R^{B\times (HW)\times C} \tag{7}$$

Each superpixel token can be regarded as a contextual representative of a local region. In SPDF, a two-stage attention pathway is introduced to enable both pixel-to-superpixel context extraction and superpixel-to-pixel context enhancement: Pixel → Superpixel: Superpixels are used as queries, while pixels serve as keys and values, enabling the transmission of pixel-level features to superpixel tokens.

$$S' = softmax\left(\frac{Q_s K_x^T}{\sqrt{D}}\right)\cdot V_x \tag{1}$$

Where: $Q_s \in R^{B\times M\times D}$ denotes the superpixel query vectors; $K_x, V_x \in R^{B\times N\times D/C}$ represent the pixel keys and values; D is the attention reduction factor; $S' \in R^{B\times M\times C}$ is the updated superpixel token representation. Superpixel → Pixel: Pixels serve as queries, while superpixels are used as keys and values, enabling the flow of contextual information back into the pixel space.

$$X_{out} = softmax\left(\frac{Q_x K_S^T}{\sqrt{D}}\right)\cdot S' \tag{2}$$

Where: $Q_x \in R^{B\times M\times D}$ denotes the pixel query vectors; $K_s \in R^{B\times M\times D}$ represents the superpixel keys; $X_{out} \in R^{B\times N\times C}$ is the final pixel output after fusing contextual information from the superpixels.

SPDF enables pixels to connect with distant regions via superpixel intermediaries, enhancing long-range dependency modeling. A lightweight Feed-Forward Network (FFN), including layer normalization, gating, and channel attention, follows the attention block to improve cross-channel interaction and local response selectivity.

3.3 Superpixel-Focused Attention (SPFA)

To capture intra-superpixel similarity and complementarity, we propose the Superpixel-Focused Attention (SPFA) module. It applies structure-aware attention on Top-K representative pixels to enable efficient, consistent feature aggregation. The SPFA workflow is shown in Figure 4.

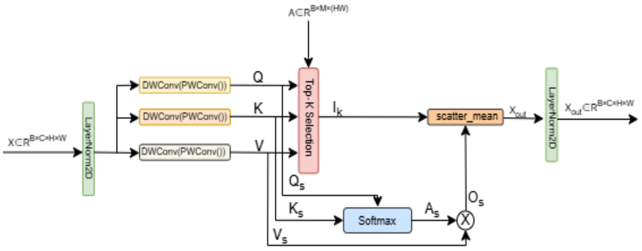

Figure 4: Overall Workflow of the SPFA Module

While global attention captures long-range dependencies, it is computationally expensive and may blur structural details. To address this, we adopt superpixel-based partitioning for fine-grained modeling on salient local regions, enhancing detail preservation and reducing redundancy. Given the input feature map $X \in R^{B\times C\times H\times W}$ and the precomputed superpixel-to-pixel affinity matrix $A \in R^{B\times M\times (HW)}$, where B denotes the batch size, C is the number of channels, M represents the number of superpixels, and HW is the total number of pixels, the processing pipeline of SPFA consists of the following four steps: (1) Query, key, and value feature extraction: The input features are first normalized, and then pointwise (1×1) convolutions combined with depthwise (3×3) convolutions are used to extract the query (Q), key (K), and value (V) features.

$$\begin{aligned} Q &= DepthwiseConv\big(Conv_{1\times 1}(X)\big) \\ K &= DepthwiseConv\big(Conv_{1\times 1}(X)\big) \\ V &= DepthwiseConv\big(Conv_{1\times 1}(X)\big) \end{aligned} \tag{3}$$

This design retains local spatial cues and rich channel features. (2) Top-K Pixel Selection: We extract the K most similar pixels per superpixel from affinity matrix A to focus attention. The indices are defined as:

$$I_k = TopK(A, k) \tag{4}$$

Where $I_k \in R^{B\times M\times k}$ denotes the indices of the Top-K most similar pixels corresponding to each superpixel. (3) Intra-superpixel Attention Computation: Based on the selected pixel indices, we extract the corresponding subsets from Q, K, V denoted as Q_s, K_s, V_s, and perform multi-head attention computation within each superpixel. The attention computation follows the standard Scaled Dot-Product Attention formulation:

$$\begin{aligned} A_s &= softmax\left(\frac{Q_s\cdot K_s^T}{\sqrt{d}}\right) \\ O_s &= A_s\cdot V_s \end{aligned} \tag{5}$$

Here, d is the dimensionality reduction factor for the attention heads. $A_s \in R^{B\times M\times h\times k\times k}$ represents the attention weights among pixels within each superpixel, and $O_s \in R^{B\times M\times h\times k\times d}$ is the aggregated output result. (4) Feature Backflow and Re-fusion: The aggregated superpixel features O_s are rearranged and projected back to their corresponding original pixel positions. Using the scatter_mean operation, the updated features are averaged and assigned to their corresponding Top-K pixels.

$$X_{out} = ScatterMean(O_s, I_k) \tag{6}$$

Finally, normalization yields the final output feature:

$$X_{final} = LayerNorm(X_{out}) \qquad (7)$$

SPFA enhances structural preservation and efficiency by combining superpixel priors with sparse attention over Top-K pixels, improving feature quality in fine-grained regions such as edges and textures.

4. Experimental Setup and Results Analysis

4.1 Dataset Setup

We train our model on the DIV2K dataset [17], which includes 800 training and 100 validation images. LR inputs are generated using ×2 and ×3 downsampling following RCAN [7]. For evaluation, we adopt five standard benchmarks—Set5 [18], Set14 [19], BSDS100 [20], Urban100 [21], and Manga109 [22]—covering diverse content from natural scenes to manga, ensuring comprehensive performance assessment.

4.2 Experimental Configuratio

We train the model with Adam (β_1 = 0.9, β_2 = 0.999), a learning rate of 5e-4, for 1000 epochs. Data augmentation includes 90°, 180°, 270° rotations and horizontal flips. The network comprises 8 SPAM modules, each outputting 40 channels and initialized with superpixel patches of size 12–24 for multi-scale structure extraction. Evaluation uses PSNR and SSIM on the Y channel, following RCAN [7].

4.3 Comparison with Lightweight Models

We compare PFRNet with state-of-the-art lightweight models, including CNN-based CARN [12], IMDN [13], and Transformer-based ESRT [26], SwinIR [20]. Table 1 presents PSNR/SSIM results for ×2/×3/×4 upscaling on multiple benchmarks. While Transformers capture long-range dependencies well, their fixed patching may harm structural continuity in reconstructed images.

Table 1: PSNR and SSIM of Different Methods on Multiple Benchmark Datasets under ×2, ×3, and ×4 Upscaling Factors

Methods	*Scale*	*Params*	*Set5*	*Set14*	*BSDS100*	*Urban100*	*Manga109*
			PSNR/SSIM	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*
CARN[12]	X2	1592K	37.76/0.9590	33.52/0.9166	32.09/0.8978	31.92/0.9256	38.36/0.9765
IMDN[13]		694K	38.00/0.9605	33.63/0.9177	32.19/0.8996	32.17/0.9283	38.88/0.9774
ESRT[24]		677K	38.03/0.9600	33.75/0.9184	32.25/0.9001	32.58/0.9318	39.12/0.9774
RFDN-L[25]		626K	38.08/0.9606	33.67/0.9190	32.18/0.8996	32.24/0.9290	38.95/0.9773
FMEN[26]		748K	38.10/0.9609	33.75/0.9192	32.26/0.9007	32.41/0.9311	38.95/0.9778
DRSAN[23]		690K	38.11/0.9609	33.64/0.9185	32.21/0.9005	32.35/0.9304	-
SwinIR[14]		878K	38.14/0.9611	33.86/0.9206	32.31/0.9012	32.76/**0.9340**	39.12/0.9783
PFRNet(ours)		445K	**38.19/0.9614**	**33.88/0.9213**	**32.32/0.9015**	**32.78/0.9340**	**39.19/0.9785**
CARN[12]	X3	1592K	34.29/0.9255	30.29/0.8407	29.06/0.8034	28.06/0.8493	33.43/0.9427
IMDN[13]		703K	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
ESRT[24]		770K	34.42/0.9268	30.43/0.8433	29.15/0.8063	28.46/0.8574	33.95/0.9455
RFDN-L[25]		633K	34.47/0.9280	30.35/0.8421	29.11/0.8053	28.32/0.8547	33.78/0.9458
FMEN[26]		757K	34.45/0.9275	30.40/0.8435	29.17/0.8063	28.33/0.8562	33.86/0.9462
DRSAN[23]		740K	34.50/0.9278	30.39/0.8437	29.13/0.8065	28.35/0.8566	-
SwinIR[14]		886K	34.62/0.9289	30.54/0.8463	29.20/0.8082	28.66/0.8624	33.98/0.9478
PFRNet(ours)		517K	**34.66/0.9292**	**30.56/0.8464**	**29.24/0.8087**	**28.72/0.8628**	**34.21/0.9487**
CARN[12]	X4	1592K	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837	30.42/0.9070
IMDN[13]		715K	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
ESRT[24]		751K	32.19/0.8947	28.69/0.7833	27.69/0.7379	26.39/0.7962	30.75/0.9100
RFDN-L[25]		643K	32.28/0.8957	28.61/0.7818	27.58/0.7363	26.20/0.7883	30.61/0.9096
FMEN[26]		769K	32.24/0.8952	28.70/0.7839	27.63/0.7379	26.28/0.7908	30.70/0.9107
DRSAN[23]		730K	32.30/0.8954	28.66/0.7838	27.61/0.7381	26.26/0.7920	-
SwinIR[14]		897K	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980	30.92/0.9151
PFRNet(ours)		503K	**32.46/0.8981**	**28.81/0.7861**	**27.72/0.7413**	**26.53/0.7996**	**30.97/0.9153**

Unlike fixed-grid methods, PFRNet leverages superpixel-guided attention for region-consistent modeling, preserving structural boundaries. On complex datasets like Set14, Urban100, and Manga109, it achieves high ×4 upscaling scores: 28.81/0.7861, 26.53/0.7996, and 30.97/0.9153 (PSNR/SSIM), outperforming SwinIR and ESRT with only 503K parameters. Figure 5 shows visual results on Urban100. For example, PFRNet better reconstructs textures in "X4_Urban100_img_011," while DRSAN [23] and IMDN [13] fail in corrupted areas. Compared to attention-based models like ESRT [24] and SwinIR [14], PFRNet retains finer details and sharper edges, validating its structural priors and adaptive regional modeling.

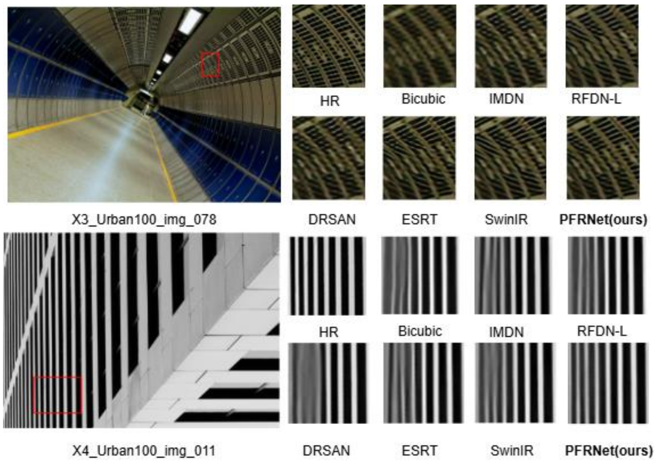

Figure 5: Comparison of Reconstruction Results Across Different Models

4.4 Ablation Stud

To assess the contribution of each submodule, we conduct ablation by progressively removing GASS, SPDF, SPFA, and LAP from the full PFRNet. The resulting variants are evaluated on Set5 and Urban100 under ×2 upscaling. Table 2 reports the corresponding PSNR and SSIM scores.

Table 2: Impact of Sequential Module Removal on Model Performance

Model Variant	*Set5*	*Urban100*
	PSNR/SSIM	*PSNR/SSIM*
Full model (all modules)	**38.19/0.9614**	**32.78/0.9340**
w/o GASS	37.92/0.9596	32.42/0.9308
w/o SPDF	37.87/0.9592	32.34/0.9297
w/o SPFA	37.90/0.9595	32.40/0.9301
w/o LAP	37.88/0.9593	32.36/0.9299

Ablation results show that GASS has the greatest impact, with 0.27dB and 0.36dB PSNR drops on Set5 and Urban100, respectively, confirming its importance for edge preservation via structure-aligned sampling. SPDF significantly affects detail fusion, and its removal degrades both PSNR and SSIM, indicating the value of superpixel-guided multi-scale interaction. SPFA, though slightly less impactful, improves coherence by integrating regional and global features. LAP mainly enhances local reconstruction; while its effect is modest, it helps preserve textures and reduce noise. Overall, all four modules contribute to PFRNet's performance, each addressing different aspects of structure-aware and detail-preserving reconstruction.

5. Conclusion

This paper presents PFRNet, a lightweight super-resolution network that enhances structural restoration via structure-aware, multi-scale fusion. It integrates four modules—GASS, SPDF, SPFA, and LAP—for texture modeling and feature enhancement. Experiments show superior performance across benchmarks, especially on complex datasets like Urban100, with under 500K parameters. Ablation confirms 8 SPAM modules as optimal. Future work will refine superpixel–attention synergy to boost adaptability across diverse image types.

References

[1] Ma C, Rao Y, Cheng Y, Chen C, Lu J, Zhou J. Structure-Preserving Super Resolution with Gradient Guidance. arXiv:2003.13081, 2020.

[2] Tang M, Gan Y, Zhang Y, Gan X. A Lightweight Super-Resolution Network for Infrared Images Based on an Adaptive Attention Mechanism. Computers, Materials & Continua, 2025.

[3] Choromanski K, Likhosherstov V, Dohan D, et al. Rethinking attention with performers[J]. arXiv preprint arXiv:2009.14794, 2020.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In European Conference on Computer Vi sion, pages 184–199. Springer, 2014.

[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional net works. In CVPR, pages 1646–1654, 2016. 1, 4, 6, 7

[6] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, andKyoungMuLee.Enhanceddeepresidualnetworksforsingle image super-resolution. In CVPRW, pages 136–144, 2017. 1, 2, 6

[7] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In European Confer ence on Computer Vision, pages 286–301, 2018.

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In CVPR, 2019. 2, 5, 6

[9] 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 ECCV, 2020. 1, 2, 5, 6

[10] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super resolution with non-local sparse attention. In CVPR, 2021. 1, 2, 5, 6, 9

[11] Wenzhe Shi, Jose Caballero, Ferenc Husz´ar, 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 IEEE Conference on Computer Vision and Pattern Recogni tion, pages 1874–1883, 2016.

[12] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In European Conference on Computer Vi sion, pages 252–268, 2018.

[13] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi distillation network. In Proceedings of the ACM Interna tional Conference on Multimedia, pages 2024–2032, 2019.

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration us ing swin transformer. In ICCV, 2021. 1, 2, 4, 5, 6, 7

[15] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super resolution. In European Conference on Computer Vision, pages 649–667. Springer, 2022.

[16] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In European Conference on Computer Vision, pages 352–368, 2018.

[17] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 114–125, 2017.

[18] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the British Machine Vision Conference, pages 135.1–135.10, 2012.

[19] Roman Zeyde, Michael Elad, and Matan Protter. On sin gle image scale-up using sparse-representations. In Proceed ings of the International Conference on Curves and Surfaces (ICCS), pages 711–730, 2010.

[20] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In IEEE International Con ference on Computer Vision, pages 416–423, 2001.

[21] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Sin gle image super-resolution from transformed self-exemplars. In IEEE Conference on Computer Vision and Pattern Recog nition, pages 5197–5206, 2015.

[22] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. Mul timedia Tools and Applications, 76(20):21811–21838, 2017.

[23] Karam Park, Jae Woong Soh, and Nam Ik Cho. Dynamic residual self-attention network for lightweight single image super-resolution. IEEE Transactions on Multimedia, 2021.

[24] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Lin lin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition, pages 457–466, 2022.

[25] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distil lation network for lightweight image super-resolution. In Eu ropean Conference on Computer Vision, pages 41–55, 2020.

[26] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In CVPR, pages 853–862, 2022. 8