Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gated Spatial-Frequency Fusion Enhanced Image Denoising

HENG LU
Shenyang Ligong University
WEI LIU
liuwei19781020@126.com

Shenyang Ligong University

Research Article

Keywords: Image Denoising, Dual-Domain Fusion, Attention, Wavelet Transform

Posted Date: March 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8988129/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Gated Spatial-Frequency Fusion Enhanced Image Denoising

HENG LU, WEI LIU*
School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110158, China
* Corresponding author: LiuWei19781020@126.com

Abstract: Image denoising is a fundamental challenge in low-level vision, crucial for applications such as medical imaging and remote sensing. Traditional methods often rely on single-domain modeling, limiting their ability to exploit complementary spatial and frequency characteristics. This paper introduces SFUNet, a dual-domain denoising framework that integrates spatial Transformer and wavelet convolution branches through an adaptive gated fusion mechanism. Extensive experiments on synthetic and real-world datasets demonstrate that SFUNet achieves state-of-the-art denoising quality, maintaining stable generalization under varying noise distributions. Our design provides a generic paradigm for cross-domain representation learning, shedding light on integrating spatial-frequency synergy in low-level vision tasks. The code is available at https://github.com/ywxk111/SFUNet

Key words: Image Denoising, Dual-Domain Fusion, Attention, Wavelet Transform

1. Introduction

Image denoising is a fundamental problem in low-level vision, aiming to remove noise while preserving structural fidelity and fine-grained details. It plays a crucial role in various applications, including computational photography, medical imaging, remote sensing, and video surveillance. Noise in imaging systems arises from diverse and complex sources, such as illumination conditions, sensor responses, and quantization

errors, which require denoising models to recover texture details accurately while maintaining global structural consistency. Moreover, significant statistical differences exist between synthetic and real-world noise, making cross-noise denoising particularly challenging.

Convolutional neural networks (CNNs) have long dominated deep learning-based image denoising. Representative approaches such as DnCNN[12], FFDNet[13], and DRUNet[14] exploit stacked convolutional layers and residual learning to extract local structural features, achieving strong performance on synthetic Gaussian noise removal. CNNs benefit from strong inductive biases and local connectivity, enabling efficient texture modeling and fast convergence. However, their intrinsic locality restricts their receptive field, limiting the ability to capture long-range dependencies. Additionally, fixed convolutional kernels lack adaptability to input variations, often resulting in oversmoothing and texture loss in complex or spatially variant noise scenarios.

The Transformer architecture has recently demonstrated powerful feature modeling capabilities in visual tasks. Its self-attention mechanism captures dynamic pixel-wise relationships across the entire spatial domain, effectively overcoming the locality limitations of convolution. Transformer-based models provide several advantages for low-level vision, including global context modeling for structural consistency, adaptive feature weighting based on input statistics, and multi-scale representation through windowing or hierarchical designs[15][16]. These properties enable Transformer-based denoising methods[17] to achieve state-of-the-art performance. Despite this progress, Transformers still face limitations when applied to denoising tasks. Existing architectures generally operate only in the spatial domain and lack explicit frequency modeling, making it difficult to simultaneously suppress noise

and preserve high-frequency details. Furthermore, the global self-attention operation may bias toward smooth reconstruction, weakening responses to fine textures and noise patterns. Variations between synthetic and real noise distributions also pose challenges for cross-domain generalization.

Wavelet transform offers a complementary frequency-domain perspective. Its multi-scale decomposition separates low-frequency structural components and high-frequency textures, facilitating targeted noise removal across different frequency bands. Prior studies[18][19] demonstrated that incorporating wavelet transform into CNNs enhances texture preservation and multi-scale feature representation. However, traditional wavelet-based networks rely on fixed filters and lack content adaptivity, which limits their synergy with global modeling approaches.

These observations suggest a natural complementarity between spatial-domain and frequency-domain modeling for image denoising. Spatial modeling, such as Transformers, excels in capturing long-range dependencies and global context, while frequency modeling through wavelet decomposition better distinguishes structures from fine details, enabling more precise noise suppression and texture restoration. Recent works have verified the effectiveness of spatial-frequency fusion in low-level vision tasks, such as detail-aware denoising with structure preservation[43] and sandstorm image enhancement with color-guided spatial-frequency fusion[44], which further confirms the potential of cross-domain feature synergy for image restoration. Motivated by the complementary advantages of spatial-frequency modeling, we propose a dual-domain adaptive framework termed Spatial-Frequency U-Net (SFUNet). Unlike previous hybrid architectures that merely concatenate or sum features from different domains, SFUNet establishes a learnable gated communication pathway

between the spatial and frequency branches. Through this pathway, the model dynamically regulates inter-domain information flow according to noise characteristics and content complexity. The design transforms feature fusion into a data-driven representation balancing process, allowing the network to automatically emphasize structural or textural cues as needed. Consequently, SFUNet not only enhances denoising quality but also introduces a generalized mechanism for adaptive cross-domain learning.

SFUNet consists of two parallel branches. The spatial branch employs window-based Transformer blocks (STB) to capture long-range dependencies across regions. The frequency branch leverages discrete wavelet transform (DWT) and inverse DWT (IDWT) for multi-scale frequency decomposition and reconstruction. To facilitate mutual enhancement between the two domains, a multi-scale Gated Fusion Module (GFM) is introduced, which adaptively adjusts feature interaction based on input content and noise characteristics, improving robustness under different noise distributions.

It is noteworthy that we independently train and evaluate the network on both real-world noise (SIDD) and synthetic Gaussian noise datasets, enabling a systematic assessment of its generalization capability across noise types.

The main contributions of this work are summarized as follows:

(1)A dual-domain hybrid architecture, SFUNet, is proposed to achieve collaborative spatial–frequency feature modeling within a single unified framework.

(2)A Gated Fusion Module (GFM) is developed to adaptively balance spatial and frequency-domain information, enabling dynamic cross-domain feature fusion.

(3)Extensive experiments on both real-world and synthetic noise datasets validate the stability, effectiveness, and generalization capability of the proposed method under varying noise distributions.

2. Related Works

2.1. CNN-based Image Denoising Methods

Convolutional neural networks (CNNs) have long dominated image denoising research. With strong generalization capacity on large-scale datasets, deep CNN models significantly outperform traditional handcrafted approaches such as BM3D[20]. DnCNN[12] employs residual learning to directly predict noise components, enabling efficient end-to-end denoising. FFDNet[13] incorporates noise level maps as auxiliary input, allowing the model to adapt to a wide range of noise intensities. DRUNet[14] adopts a multi-scale encoder–decoder design, improving restoration quality with a compact architecture.

To enhance network modeling capability, various structural improvements have been proposed, including non-local modules[21][22], attention mechanisms[23], and dynamic convolutions[24].Even the classic non-local mean filter has been optimized for better denoising performance in complex noise scenarios[45], further enriching the design of traditional denoising methods. Nevertheless, the locality of convolution limits receptive field expansion. These techniques enrich contextual modeling and flexibility. Nevertheless, the locality of convolution limits receptive field expansion. Even with deep stacking or dilated convolutions, CNNs struggle to capture long-range pixel dependencies. Additionally, convolution kernels remain fixed across inputs, limiting

adaptability to complex noise distributions. These issues often result in degraded performance under real-world or non-Gaussian noise conditions.

2.2 Transformer-based Image Denoising Methods

Originally introduced for natural language processing[25], Transformers have become a prominent architecture in computer vision[26] due to their capability of modeling global dependencies through self-attention(SA). Unlike convolutions, Transformers adaptively allocate attention weights according to image content, enabling long-range context reasoning. Significant progress has been achieved in low-level vision tasks including denoising, super-resolution, and deblurring.Efficient Transformer architectures such as attention pyramid design[50] and structure-aware attention fusion[48] have further improved the performance of Transformer in image enhancement and denoising tasks, expanding its application in low-level vision.

SwinIR[17] utilizes window-based attention for hierarchical feature extraction, achieving a balance between local and global modeling in high-resolution tasks. Restormer[27] extends attention to the channel dimension, improving global feature aggregation. Uformer[16] integrates Transformer blocks into a U-shaped architecture to support multi-scale recovery. IPT[28] adopts large-scale pretraining for enhanced generalization across multiple image restoration tasks.

Despite their advantages, challenges remain when applying Transformers to denoising. Many models perform feature interactions only in spatial or channel dimensions, lacking explicit frequency-domain representation for fine texture recovery. Self-attention tends to favor smooth reconstruction, potentially weakening high-frequency detail preservation. Furthermore, Transformer models are sensitive to noise distribution shifts, resulting in unstable generalization between real and synthetic noise.

Transformer models are also widely used in other image restoration tasks such as single image dehazing with multi-contrastive regularization[47] , but the lack of frequency-domain modeling is a common limitation in these methods. These limitations motivate hybrid approaches combining Transformer-based modeling with frequency-domain knowledge.

2.3 Frequency-domain and Wavelet-based Denoising

Frequency-domain strategies provide an alternative perspective for noise suppression by leveraging spectral differences between noise and structural components. The wavelet transform offers joint spatial–frequency analysis, facilitating multi-scale decomposition of low-frequency structures and high-frequency details. MWCNN[18] first introduced discrete wavelet transform (DWT) and inverse wavelet transform (IDWT) within deep neural networks to extract features at multiple scales without information loss. Subsequent studies[29][30] further combined convolutional representations with wavelet decomposition to enhance texture fidelity.

However, wavelet-based networks typically rely on fixed wavelet bases and lack adaptability to varying image characteristics. Moreover, pure frequency-domain methods are insufficient for capturing complex structures and long-range correlations. Frequency-domain enhancement has also been successfully applied in medical image segmentation with lightweight depthwise separable ConvNets[49], verifying the generality of frequency-domain modeling in various computer vision tasks. Recent research explores combining wavelet analysis with attention mechanisms for improved robustness, yet challenges remain regarding feature fusion strategies and training stability.

2.4 Parallel and Cross-Domain Fusion Architectures

Parallel architectures enable multiple feature streams to be learned simultaneously, improving representational diversity. DUCT[31] exemplifies this design by combining convolutional and Transformer branches, which focus respectively on local structural features and global dependencies. Similar dual-branch fusion strategies have been adopted to enhance multi-scale feature interaction and robustness[32][33].A recent dual-branch network SCNet combines Swin Transformer and ConvNeXt for strong noisy image denoising[46], which further demonstrates the effectiveness of dual-branch architecture in handling complex noise. In image restoration, parallel designs create new opportunities to integrate spatial- and frequency-domain information.

In image restoration, parallel designs create new opportunities to integrate spatial- and frequency-domain information. Spatial-domain modeling ensures global structural consistency, while frequency-domain modeling benefits fine-grained texture recovery. However, most parallel architectures target semantic-level tasks, and the exploration of spatial–frequency collaborative modeling for low-level vision remains limited.

Motivated by these developments, the proposed SFUNet combines complementary spatial and frequency representations within a unified framework. A dual-branch structure is adopted: the spatial branch applies window-based Transformer blocks for long-range dependency modeling, while the frequency branch exploits db4 wavelet-based convolutional modules for multi-scale frequency decomposition. A multi-scale Gated Fusion Module then enables adaptive cross-domain feature interaction, leading to more stable performance on both synthetic and real-world noise denoising tasks.

3. Method

This paper proposes an image denoising network named SFUNet based on dual-branch spatial-frequency domain information fusion. Through the network design that captures global structures via spatial-domain Transformer, conducts fine-grained frequency-domain modeling via Wavelet Transform, and implements adaptive gated fusion, an efficient denoising model is constructed. This section elaborates on the details of the method from three aspects: network architecture, core modules, and loss function. The overall architecture of the proposed method is shown in Fig. 1.

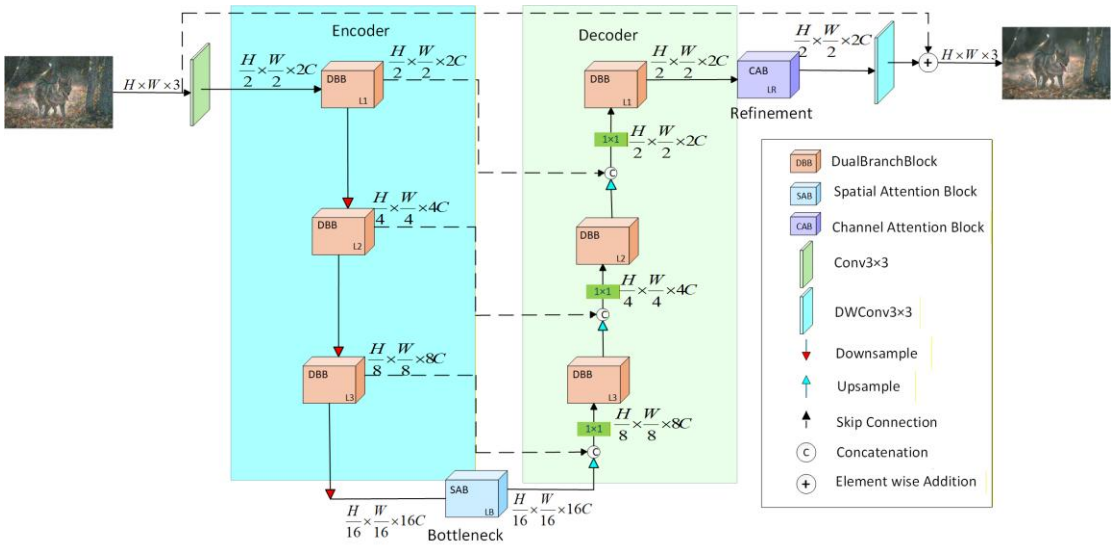


Fig. 1. Overall architecture of the proposed SFUNet. The network adopts a dual-branch encoder–decoder structure, where the Spatial Attention Branch (SAB) models long-range structural dependencies in the spatial domain, and the Wavelet Convolution Branch (WCB) performs multi-scale frequency decomposition using DWT/IDWT. Multi-level Gated Fusion Modules (GFMs) enable adaptive cross-domain feature interaction. Channel Attention Block (CAB) further refines aggregated features before residual reconstruction.

3.1. Overall Network Architecture

SFUNet adopts a hierarchical Encoder-Decoder structure, with dual-branch parallel processing and multi-scale gated fusion as its core, to achieve end-to-end

mapping from noisy images to clean images. As shown in Fig. 1, given a noisy image $x \in \mathbb{R}^{H\times W\times 3}$, a 3×3 convolution layer is used to extract shallow features $F_{shallow} \in \mathbb{R}^{H\times W\times C}$. Subsequently, the shallow features $F_{shallow}$are fed into two parallel encoder branches:

• Spatial Attention Branch: This branch employs a spatial Transformer module, focusing on learning long-range spatial dependencies and structural information.

• Wavelet Convolution Branch: This branch adopts a wavelet convolution module, which decomposes features into frequency subbands using Discrete Wavelet Transform (DWT), thereby effectively handling high-frequency details such as noise and textures.

The encoder consists of four levels, and each level halves the resolution of the feature map and doubles the number of channels through downsampling. In the middle levels of the encoder, a gated fusion module is introduced to realize feature interaction and information complementarity between the two branches. The decoder is structurally symmetric to the encoder, and gradually restores the resolution of the feature map through upsampling and skip connections. Skip connections concatenate the features of the corresponding levels in the encoder with the decoder features to retain rich low-level detail information. Finally, the outputs of the two branches are concatenated after decoding and fed into a refinement module composed of channel attention modules for deep fusion. Eventually, a 3×3 convolution layer generates the residual image I_{res}, which is added to the original input image to obtain the final restored image $I_{out} = I_{in} + I_{res}$.

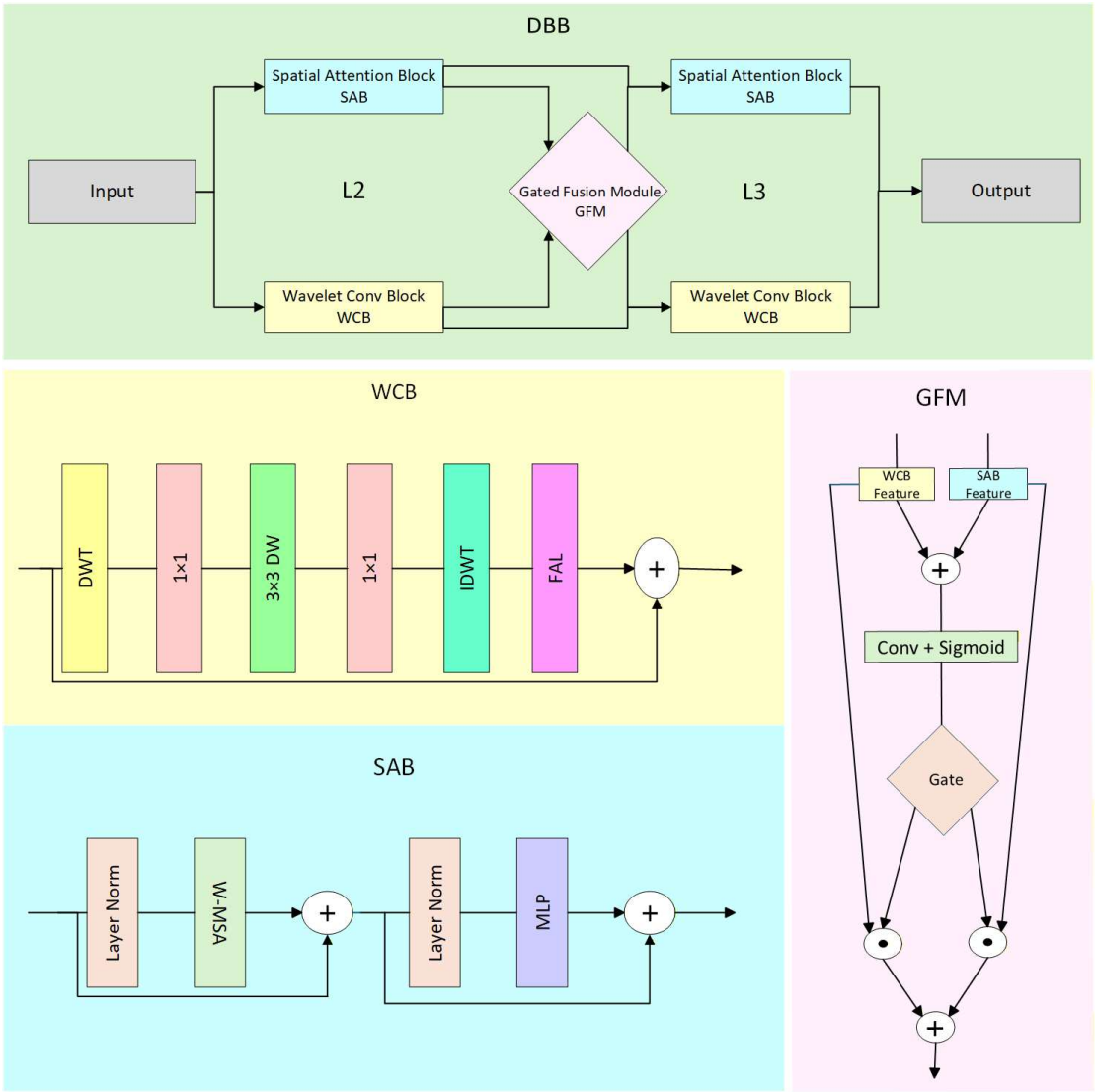


Fig. 2.Details of the dual-branch feature extraction module. The block is composed of a Spatial Attention Block (SAB), a Wavelet Convolution Block (WCB), and a Gated Fusion Module (GFM). SAB adopts window-based multi-head self-attention to capture contextual dependencies, while WCB performs frequency-domain enhancement through grouped convolution and wavelet sub-band processing. GFM adaptively regulates cross-domain information flow.

3.2. Dual-Domain Branch Module

The dual-domain branch module is mainly composed of SAB, WCB, and GFM. The structure of the dual-branch module is shown in Fig. 2.

3.2.1 Spatial Attention Block

SAB is the core of the spatial-domain branch, and its structure is based on the standard Vision Transformer Block. Each SAB includes a Window-based Multi-head Self-Attention (W-MSA) layer and a Multi-Layer Perceptron(MLP) layer, supplemented by residual connections and layer normalization. W-MSA calculates self-attention within non-overlapping local windows, which significantly reduces computational complexity. Meanwhile, it realizes cross-window information interaction through a cyclic shift mechanism, thereby effectively modeling long-range dependencies while maintaining efficient computation. The calculation process of SAB is as follows:

$$x' = W - MSA(LN(x)) + x, \tag{1}$$

$$x_{out} = MLP(LN(x')) + x'. \tag{2}$$

where W-MSA denotes Window-based Multi-head Self-Attention, MLP denotes Multi-Layer Perceptron, LN denotes layer normalization, and x denotes input feature.

•Window Attention Calculation: The feature map is divided into non-overlapping local regions using 16×16 windows, and attention is only computed within these windows. For feature reshaping, the feature $F \in \mathbb{R}^{B\times C\times H\times W}$ is transformed into a window sequence $F_{win} \in \mathbb{R}^{B\times N_{win}\times 16^2\times C}$, where $N_{win} = (H/16)(W/16)$ represents the number of windows.

•QKV Projection and Corresponding Positional Encoding: Queries (Q), Keys (K), and Values (V) are generated via linear layers, and learnable relative positional biases $B \in$

$\mathbb{R}^{(2\times16-1)^2\times N_h}$(where N_h is the number of attention heads) are introduced to capture the spatial correlation of pixels within the window.

• Attention Weighting: After normalization via Softmax, attention weights are calculated and weighted with V to obtain window attention features. The formula is:

$$Attention = Softmax(\frac{QK^T}{\sqrt{C/N_h}} + B) \cdot V \quad (3)$$

where C denotes the number of channels, N_h is the number of attention heads, and B denotes learnable relative positional biases.

3.2.2 Wavelet Convolution Branch

To efficiently process frequency information, we design a wavelet convolution module with the workflow of "channel compression – group convolution – channel restoration". First, this module uses Discrete Wavelet Transform (DWT) based on the db4 wavelet basis to decompose the input feature $x \in \mathbb{R}^{H\times W\times C}$ into four frequency subbands (LL, LH, HL, HH), resulting in $x_{wavelet} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 4C}$. To reduce computational cost without sacrificing performance, a channel compression mechanism is introduced, which reduces the number of channels from 4C to C via a 1×1 convolution. Subsequently, a 3×3 group convolution is used to process the compressed features, further reducing the number of parameters. After processing, another 1×1 convolution is used to restore the number of channels to 4C. Finally, Inverse Discrete Wavelet Transform (IDWT) is applied to reconstruct the processed frequency coefficients back to the spatial domain, and a Feature Adaptation Layer (FAL) is used to adjust the distribution of the reconstructed features to make them compatible with the original features. The entire module stabilizes the training process through residual connections, and the formula is:

$$F_{out} = IDWT(conv_{1\times1}(conv_{3\times3}(conv_{1\times1}(DWT(F_{in}))))) + F_{in} \quad (4)$$

where F_{in} is input feature, $conv_{n\times n}$ denotes convolution operation, and F_{out} is output feature.

3.2.3 Gated Fusion Module

To realize efficient interaction between spatial-domain features and frequency-domain features, we design a Gated Fusion Module (GFM) to adaptively adjust the bidirectional flow of features in the dual-branch structure. Different from traditional fusion methods based on summation or concatenation, GFM realizes adaptive adjustment of feature contributions and dynamic fusion of information through a learnable gating mechanism while ensuring numerical stability. Let the spatial-domain feature be $F_S \in \mathbb{R}^{B\times C\times H\times W}$ and the frequency-domain feature be $F_w \in \mathbb{R}^{B\times C\times H\times W}$. First, an adaptive gating coefficient $\alpha \in [0,1]$ is generated based on global feature statistics:

$$\alpha = \sigma(W_2\phi(W_1 GAP([F_s, F_w]))) \quad (5)$$

where$[F_s, F_w]$denotes feature concatenation in the channel dimension, $\mathrm{GAP}(\cdot)$ represents the Global Average Pooling operation, $\phi(\cdot)$denotes the GELU activation function, W_1 and W_2 are 1×1 convolution mappings, and $\sigma(\cdot)$ denotes the Sigmoid activation function. This gating parameter adaptively adjusts the fusion ratio of the spatial domain and frequency domain according to the input image features.

After obtaining the gating parameters, GFM performs the following bidirectional residual update:

$$F_s' = F_s + \alpha \cdot C_w(F_w), \quad (6)$$

$$F'_w = F_w + (1 - \alpha) \cdot C_s(F_s). \tag{7}$$

where $C_w(\cdot)$ and $C_s(\cdot)$represent lightweight 3×3 convolution mappings for frequency-domain and spatial-domain features, respectively. This residual fusion enables the information of the two domains to complement each other under gating adjustment while preserving the integrity of the original features.

Conceptually, the gating coefficient α can be interpreted as a domain relevance estimator that quantifies the relative importance of spatial and frequency cues. When the local region exhibits strong structural regularity, α tends to emphasize the spatial features; conversely, in areas dominated by fine textures or noise, the frequency-domain response becomes dominant. This adaptive control enables the network to perform context-aware modulation, improving robustness under spatially variant noise.

From a representation learning perspective, GFM performs a bidirectional correction process that aligns feature statistics between domains. Instead of native concatenation, it treats cross-domain fusion as a dynamic equilibrium problem: balancing global context and fine detail through learnable gating. Such formulation implicitly regularizes feature interaction, leading to more stable and interpretable denoising behavior.

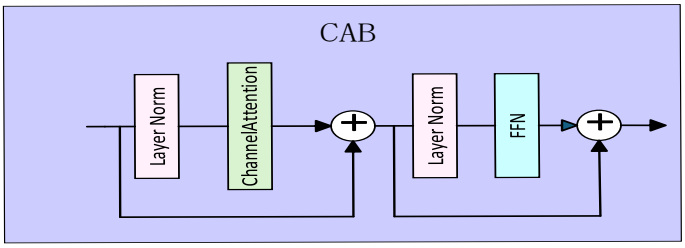


Fig. 3. Structure of the Channel Attention Block (CAB). Multi-head attention across channels captures inter-channel dependencies, followed by a feed-forward network for

feature refinement. Residual learning and layer normalization ensure stable enhancement of fused spatial–frequency information.

3.3. Channel Attention Block

At the end of the network, we use a Channel Attention Block (CAB) as the refinement module, as shown in Fig. 3. Different from SAB (which operates in the spatial dimension), CAB performs multi-head self-attention in the channel dimension. It treats each pixel position of the feature map as a sequence element and calculates attention weights between channels, thereby effectively capturing the interdependencies between different feature channels. CAB consists of layer normalization, channel attention, and a feed-forward network (FFN), which can deeply integrate and refine the concatenated dual-branch features to generate high-quality residual information, as shown in Fig. 3.

• Layer Normalization: Given the input feature $x \in \mathbb{R}^{B\times C\times H\times W}$, CAB is composed of two residual stages:

$$x' = x + Attention(LN(x)), \tag{8}$$

$$x_{out} = x' + FFN(LN(x')). \tag{9}$$

where x is input feature, FFN denotes feed-forward network.

• Channel Attention Calculation: First, queries (Q), keys (K), and values (V) are generated through 1×1 convolution and depthwise convolution (DWConv):

$$Q, K, V = Conv_{1\times 1}^{qkv}(DWConv(x_{out})). \tag{10}$$

The attention calculation process in the channel dimension is as follows:

$$Attention = Softmax(\frac{QK^T}{\sqrt{d}} \cdot \tau) \cdot V. \quad (11)$$

where τ is a learnable temperature parameter, d denotes the dimension of the Q or K vector.

3.4. Loss Function

L1 loss is adopted as the training objective, which constrains the difference between the model output and the clean image through pixel-level errors. The formula is:

$$\mathcal{L} = \left\| \hat{I}_{HQ} - I_{HQ} \right\|_1 \quad (12)$$

where $\hat{I}_{HQ}$ is the output of our model, I_{HQ} is the corresponding ground-truth image, with $\|\cdot\|_1$ denoting the L1 norm. The reasons for choosing L1 loss are as follows: it has stronger robustness to noisy samples, can effectively avoid gradient explosion caused by extreme noisy pixels, and is more consistent with the human visual perception characteristics of "pixel deviation".

4. Experiments

4.1 Experimental Setup

We evaluate the proposed SFUNet on two types of image denoising tasks: synthetic Gaussian denoising and real image denoising.

For synthetic Gaussian denoising, the training datasets includes BSD500[3], WED[4], Flickr2K[5], and DIV2K[6], while the evaluation is conducted on standard benchmarks including BSD68[7], Urban100[8], Set12[9], McMaster[10], and Kodak24[11]. All training images are randomly cropped into patches of size 128 × 128

and augmented by random horizontal flips, vertical flips, and rotations to improve generalization capability.

For real-world denoising, we use the SIDD training datasets[1] for model training and evaluate performance on both the SIDD validation set and the DND dataset[2], ensuring that the robustness and generalization ability under real noise distributions are thoroughly assessed.

During training, we adopt AdamW as the optimizer with $\beta_1 = 0.9$,$\beta_2 = 0.999$, and a weight decay of 1×10^{-4}. The initial learning rate is set to 3×10^{-4} and decayed using the MultiStepRestartLR scheduler with a decay factor $\gamma = 0.7$. Gradient clipping is applied throughout the training process to stabilize optimization and avoid exploding gradients.

The model is trained for 400K iterations for Gaussian denoising and 300K iterations for real image denoising. The evaluation metrics include peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM), which assess image fidelity and perceptual consistency, respectively.

4.2 Results on Additive Gaussian Noise Removal for Color Images

We evaluate the denoising performance of the proposed SFUNet on synthetic noisy color images corrupted by additive white Gaussian noise (AWGN). Comparisons are conducted against several representative state-of-the-art methods, including BM3D[20], DnCNN[12], FFDNet[13], IRCNN[42], DRUNet[14], RNAN[34], IPT, and SwinIR. The quantitative results are summarized in the following table 1.

Dataset	CBSD68			Kodak24			McMaster			Urban100		
Method	σ = 15	σ = 25	σ = 50	σ = 15	σ = 25	σ = 50	σ = 15	σ = 25	σ = 50	σ = 15	σ = 25	σ = 50

BM3D[20]	33.52	30.17	27.38	34.28	32.15	28.46	34.06	31.66	28.51	33.93	31.36	27.93
DnCNN[12]	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
FFDNet[13]	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
IRCNN[42]	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
DRUNet[14]	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
RNAN[34]	-	-	28.27	-	-	29.58	-	-	29.72	-	-	29.08
IPT[28]	-	-	28.39	-	-	29.64	-	-	29.98	-	-	29.71
SwinIR[17]	**34.42**	**31.78**	**28.56**	35.34	32.89	29.79	35.61	33.20	30.22	35.13	32.90	29.82
SFUNet	34.40	31.77	28.55	**35.46**	**33.02**	**29.96**	**35.61**	**33.32**	**30.23**	**35.17**	**32.99**	**30.03**

Table 1:Quantitative results of the AWGN color noise removal evaluation for different noise levels σ. The underlined and ***bold*** *number indicate the second best and the best results.*

A comprehensive comparison under three noise levels (σ = 15, 25, and 50) is presented in Table 1. From the quantitative results on the Kodak24, McMaster, and Urban100 datasets, SFUNet consistently achieves the highest PSNR values across all noise levels, outperforming existing state-of-the-art methods. On the CBSD68 dataset, SFUNet also delivers highly competitive performance, with only a marginal difference of 0.01–0.02 dB compared with the best-performing approach. Notably, on Kodak24, McMaster, and Urban100, SFUNet shows clear superiority, surpassing the second-best method by more than 0.1 dB under certain noise levels. These results demonstrate that the proposed approach provides strong generalization capability and robustness under varying noise intensities and image distributions.

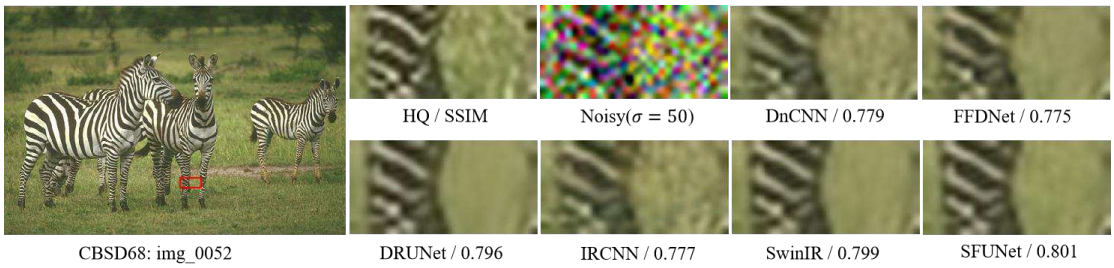


Fig. 4.Qualitative results on AWGN color images from the CBSD68 dataset.From up to down, we show the HQ image, the real noisy image,the results of DnCNN, FFDNet, DRUNet, IRCNN, SwinIR, and SFUNet.

As illustrated in Fig. 4, SFUNet achieves an SSIM of 0.801 in this example, while preserving relatively clear edge and texture structures. Compared with other competing methods, the proposed approach shows a certain advantage in noise suppression and structural fidelity, leading to a more natural visual appearance.

4.3 Results on Additive Gaussian Noise Removal for Grayscale Images

Table 2 reports the quantitative results on grayscale image denoising. Overall, SFUNet demonstrates stable performance across different datasets and noise levels. Specifically, SFUNet achieves the second-best performance on all datasets when σ = 15. When the noise level increases to σ = 25, SFUNet achieves the best results on all datasets. For σ = 50, the proposed method remains the best on Set12 and Urban100, and ranks second on BSD68. These results indicate that SFUNet maintains reliable denoising capability and adaptability under varying noise intensities.

As illustrated in Fig. 5, SFUNet preserves more detailed textures and structural edges in this example, producing a clearer reconstruction that is visually closer to the clean reference.

Dataset	Set12			BSD68			Urban100		
Method	*σ = 15*	*σ = 25*	*σ = 50*	*σ = 15*	*σ = 25*	*σ = 50*	*σ = 15*	*σ = 25*	*σ = 50*
BM3D[20]	32.37	29.97	26.72	31.08	28.57	25.60	32.35	29.70	25.95
DnCNN[12]	32.86	30.44	27.18	31.73	29.23	26.23	32.64	29.95	26.26
FFDNet[13]	32.75	30.43	27.32	31.63	29.19	26.29	32.46	29.80	26.22

IRCNN[42]	32.76	30.37	27.12	31.63	29.15	26.19	32.46	29.80	26.22
DRUNet[14]	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
NLRN[22]	33.16	30.80	27.64	31.88	29.41	26.47	33.45	30.94	27.49
RNAN[34]	-	-	27.70	-	-	26.48	-	-	27.65
DeamNet[35]	33.19	30.81	27.74	31.91	29.44	26.54	33.37	30.85	27.53
SwinIR[17]	**33.36**	31.01	27.91	**31.97**	29.50	**26.58**	**33.70**	31.30	27.98
SFUNet	33.20	**31.06**	**27.96**	31.87	**29.50**	26.57	33.35	**31.54**	**28.31**

Table 2: Quantitative results of the AWGN gray noise removal evaluation for different noise levels σ. The underlined and ***bold*** *number indicate the second best and the best results.*

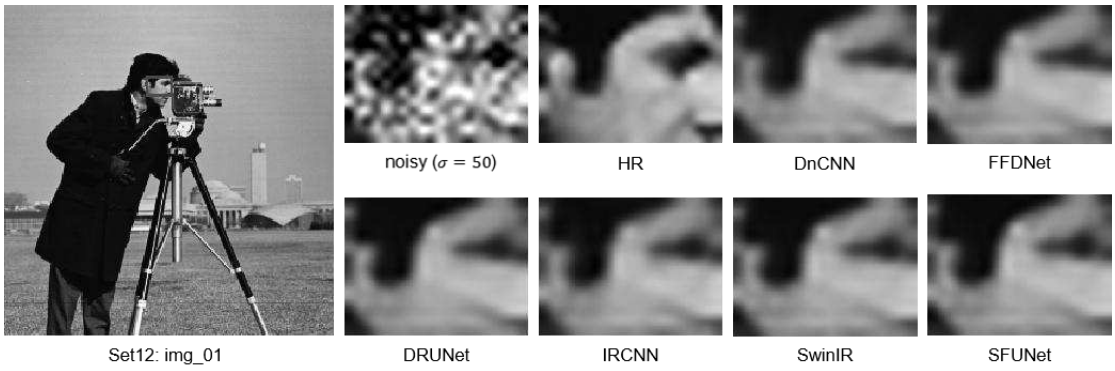


Fig. 5.Qualitative results on AWGN gray images from the Set12 dataset. From up to down,we show the HQ image, the real noisy image, the results of DnCNN, FFDNet ,DRUNet, IRCNN, SwinIR, and SFUNet.

Overall, the experimental results demonstrate the effectiveness and generalization capability of SFUNet across different datasets and noise levels. The proposed method consistently achieves competitive performance, often ranking as the best or second-best among all compared approaches, while maintaining structural integrity and perceptual quality even under high noise conditions.

4.4 Results on Real-World Image Denoising

We further evaluate the performance of SFUNet on two real-world noise benchmarks, SIDD and DND. SFUNet is compared with several representative state-

of-the-art denoising methods, including BM3D[20], DnCNN[12], CBDNet[38], RIDNet[39], VDN[40], CycleISP[41], MPRNet[37], MIRNet[36], and Uformer[16]. Peak signal-to-noise ratio (PSNR, in dB) and structural similarity index (SSIM) are adopted as evaluation metrics to quantitatively assess both fidelity and perceptual quality of the reconstructed images.

Table 3 summarizes the quantitative results on real noisy datasets. On the SIDD dataset, SFUNet achieves the best performance with a PSNR of 39.98 dB and an SSIM of 0.960. On the DND dataset, our method also attains the top performance, reaching 40.01 dB in PSNR and 0.955 in SSIM. These results indicate that SFUNet demonstrates strong robustness and generalization ability in both synthetic and real-world noise scenarios.

Dataset	Method	BM3D [20[20]	DnCNN [12]	CBDNet [38]	RIDNet [39]	VDN [40]	CycleISP [41]	MIRNet [36]	Uformer [16]	SFUNet
SIDD	PSNR	25.65	23.66	30.78	38.71	39.28	39.52	39.72	<u>39.89</u>	**39.98**
	SSIM	0.685	0.583	0.801	0.951	0.956	0.957	0.959	<u>0.960</u>	**0.960**
DND	PSNR	34.51	32.43	38.06	39.26	39.38	39.56	39.88	<u>39.98</u>	**40.01**
	SSIM	0.851	0.790	0.942	0.953	0.952	0.956	**0.956**	0.955	<u>0.955</u>

*Table 3: Quantitative results (PSNR and SSIM) of the real noise removal evaluation.The <u>underlined</u> and **bold** number indicate the second best and the best results.*

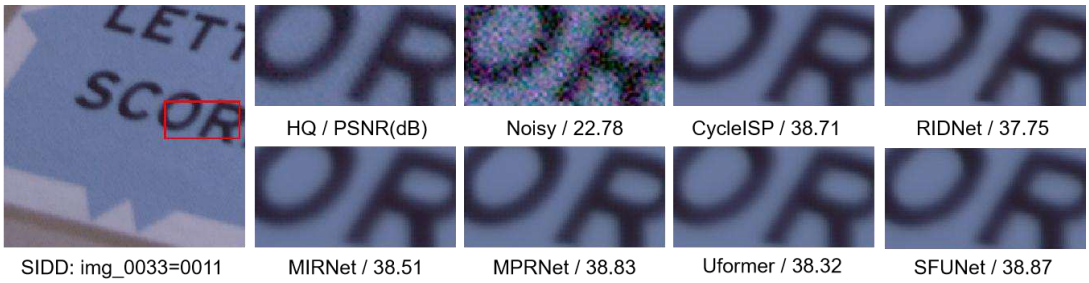


Fig. 6. Qualitative results on real noisy images from the SIDD validation dataset. From up to down, we show the HQ image, the real noisy image, the results of CycleISP, RIDNet, MIRNet, MPRNet, Uformer, and SFUNet.

As shown in the visual comparisons in Fig. 6, SFUNet achieves a more favorable balance between noise removal and structural detail preservation. Compared with existing approaches, our method suppresses noise more effectively while avoiding excessive smoothing, resulting in reconstructed images with more natural textures and clearer edges.

4.5 Ablation Study

To thoroughly investigate the contribution of each key component in SFUNet, we conduct a two-part ablation study on four benchmark datasets: CBSD68, Kodak24, McMaster, and Urban100, under AWGN with $\sigma = 15$.

The first part examines the impact of the spatial branch, frequency branch, and the Gated Fusion Module (GFM). The second part evaluates the effectiveness of deploying GFM at different hierarchical stages within the network.

As shown in Table 4, removing either branch results in a noticeable performance drop, confirming the complementary nature of spatial- and frequency-domain representations.

Furthermore, eliminating GFM leads to additional degradation, indicating that the

dynamic fusion mechanism is essential for balancing detail preservation and noise removal.

Table 5 further reveals that:

• GFM provides more significant benefits when applied in the encoder, while its impact in the decoder is relatively limited;

• Deploying GFM only in the intermediate layers (e.g., Stage 2 or Stage 3) yields moderate improvement, yet still falls short of the full configuration with GFM applied throughout the entire network.

Overall, these results validate the necessity and effectiveness of the proposed architectural components from multiple perspectives.

Method	CBSD68	Kodak24	McMaster	Urban100
	PSNR\SSIM	PSNR\SSIM	PSNR\SSIM	PSNR\SSIM
w/o Wavelet	34.35\0.935	35.38\0.930	35.49\0.933	35.00\0.951
w/o SAB	34.17\0.932	35.06\0.926	35.12\0.928	34.45\0.947
w/o GFM	34.35\0.934	35.38\0.929	35.50\0.933	35.00\0.951
all	**34.36\0.935**	**35.40\0.930**	**35.52\0.933**	**35.04\0.952**

*Table 4. Component ablation study evaluating the contributions of each core branch and the GFM (σ = 15). The **bold** numbers indicates the best results.*

Method	CBSD68	Kodak24	McMaster	Urban100
	PSNR\SSIM	PSNR\SSIM	PSNR\SSIM	PSNR\SSIM
level2	34.35\0.935	35.39\0.930	35.51\0.933	35.02\0.951
level3	34.35\0.935	35.39\0.930	35.51\0.933	35.02\0.951
encoder	34.35\0.935	35.39\0.930	35.51\0.933	35.02\0.951
decoder	34.21\0.933	35.15\0.9269	35.26\0.930	34.62\0.948

Table 5. Ablation results on the placement of the GFM at different network stages under AWGN with σ = 15.

Method	Params(M)	MACs(G)	FLOPs(G)
w/o SAB	16.91	17.39	34.79
w/o Wavelet	203.57	60.35	120.69
w/o GFM	213.20	68.73	137.45
level2	214.26	71.26	142.52
level3	214.26	71.26	142.52
encoder	214.26	71.26	142.56
decoder	214.26	71.26	142.56

Table 6: Ablation study on model components and hierarchical levels of SFUNet. The table reports the number of parameters (Params), multiply–accumulate operations (MACs), and floating-point operations (FLOPs). The removal or modification of key modules (SAB, Wavelet, GFM) and structural levels (level2, level3, encoder, decoder) demonstrates their impact on model complexity. Results are measured on input images of 128×128 resolution.

Table 6 summarizes the computational characteristics of SFUNet under different ablation configurations. Removing the Spatial Attention Block (w/o SAB) leads to the lowest parameter count (16.91M) and FLOPs (34.79G), indicating a substantial reduction in model complexity due to the elimination of Transformer-based computations. In contrast, removing the Wavelet branch (w/o Wavelet) or the Gated Fusion Module (w/o GFM) increases the overall complexity (120.69G and 137.45G FLOPs, respectively), as these components contribute to efficient multi-branch feature interaction rather than direct parameter reduction. The configurations "level2," "level3," "encoder," and "decoder" maintain similar computational costs (approximately 71.3G MACs and 142.5G FLOPs), reflecting a balanced design in terms of scale distribution. Overall, the results suggest that SFUNet's modular design achieves stable computational behavior across levels, and that each component plays a distinct role in balancing efficiency and representational power rather than merely reducing computational load.

5. Conclusions

The proposed gated spatial-frequency interaction represents a general mechanism for cross-domain information modulation. Rather than serving merely as an architectural variant, SFUNet embodies a dynamic learning paradigm in which spatial and frequency representations collaboratively evolve during optimization. This paradigm could be extended to other image restoration tasks such as deblurring, super-resolution, and low-light enhancement, demonstrating its potential as a universal cross-domain learning strategy.

In summer, SFUNet effectively integrates spatial Transformers and wavelet-based frequency modeling under a unified gated framework. The adaptive fusion mechanism enables robust noise removal with strong generalization to real-word scenarios, as verified by extensive benchmarks. Future work will explore lightweight and sparsity-aware extensions of the proposed paradigm to further improve computational efficiency and deployment flexibility.

Extensive experiments on both synthetic Gaussian noise (AWGN) and real-world noisy images (SIDD and DND) demonstrate the effectiveness of the proposed method. SFUNet achieves superior quantitative results and visual restoration quality, and exhibits strong robustness and generalization capability in challenging real-noise scenarios.

Nevertheless, the incorporation of multi-head attention and wavelet convolution increases the overall model complexity, which may limit efficiency on resource-constrained devices such as mobile platforms. Future work will focus on architectural lightweighting and dynamic sparse attention mechanisms to improve computational

efficiency while maintaining performance. Additionally, we plan to further explore the dual-domain fusion paradigm in broader image restoration tasks, including deblurring, super-resolution, and compressed sensing reconstruction.

Supporting Data Availability Statement

The DIV2K, Flickr2K, WED, BSD, Set12, BSD68, CBSD68, Kodak, McMaster, Urban100, SIDD, and DND datasets used in this study are all open-source benchmark datasets in the field of image denoising, which can be directly accessed through the following channels: DIV2K, a core benchmark dataset for image restoration, contains 1,000 high-resolution color images (800 training images, 100 validation images, 100 test images) supporting various denoising task scenarios. Its open-source resources include raw images, noise annotations, and preprocessing tools (https://github.com/JingyunLiang/SwinIR); Flickr2K is a high-quality image dataset selected from the Flickr platform, containing 2,650 high-resolution color images widely used in joint image denoising and super-resolution tasks. Its open-source resources provide uncompressed raw images and format conversion scripts (https://github.com/JingyunLiang/SwinIR); WED (Waterloo Exploration Dataset) focuses on image quality assessment and denoising research, containing 4,744 color images covering natural, texture, and artificial scenes with complete noise type annotations (https://github.com/JingyunLiang/SwinIR); BSD (Berkeley Segmentation Dataset) is a denoising subset of the classic image analysis benchmark dataset, containing 500 color/grayscale images of diverse scenes to support the generalization test of denoising algorithms (https://github.com/JingyunLiang/SwinIR); Set12 is a classic small-scale test set in the field of image denoising, containing 12 representative grayscale images (covering natural scenes, textures, figures, etc.) commonly used for rapid algorithm verification (https://github.com/JingyunLiang/SwinIR); BSD68 is a

dedicated test set for grayscale image denoising, containing 68 standardized grayscale images, which is a common benchmark for peak signal-to-noise ratio (PSNR) evaluation of denoising algorithms (https://github.com/JingyunLiang/SwinIR); CBSD68 (Color BSD68) is a color extended version of BSD68, containing 68 high-fidelity color images focusing on the performance evaluation of color image denoising algorithms (https://github.com/JingyunLiang/SwinIR); Kodak is a classic color image quality assessment dataset, containing 24 high-resolution standard color images known for rich color levels and details, commonly used for subjective visual quality assessment after denoising (https://github.com/JingyunLiang/SwinIR); McMaster (McMaster University Texture Dataset) focuses on texture image denoising research, containing 117 high-resolution texture images (covering natural textures and artificial textures) with annotated samples of various noise intensities (https://github.com/JingyunLiang/SwinIR); Urban100 is a dedicated image denoising dataset for urban scenes, containing 100 high-resolution color images of urban buildings and street scenes characterized by complex edges and detailed features (https://github.com/JingyunLiang/SwinIR); SIDD (Smartphone Image Denoising Dataset) is a benchmark dataset for real-scene noise denoising, containing more than 1,000 pairs of "noisy image-clean image" samples (from 10 smartphone models) with real noise model annotations (https://www.eecs.yorku.ca/~kamel/sidd/); DND (Denoising in the Dark Challenge Dataset) is a benchmark dataset for low-light environment denoising competitions, containing 50 low-light noisy images and corresponding ground-truth clean images, focusing on the evaluation of blind denoising algorithms (https://noise.visinf.tu-darmstadt.de/).

Acknowledgements

Heng Lu: Primarily responsible for writing the manuscript, methodology, formal analysis, data curation, and conceptualization.

Wei Liu: Primarily responsible for manuscript review and editing, supervision, software development, project administration, methodology, formal analysis, resources, and validation.

Meiting Song: Primarily responsible for manuscript revision and software development.

Zongzhuo Jin: Primarily responsible for manuscript revision and software development.

Wang Yi: Primarily responsible for formal analysis and validation.

Jiatong Shang: Primarily responsible for formal analysis and validation.

All authors have read and approved the final version of the manuscript.

Funding

No funds were received to suppport the preparation of this manuscript.

Declarations

Competing Interests

The authors declare that they have no competing interests.

Reference

[1]Abdelhamed, S. Lin, M. S. Brown, A High-Quality Denoising Dataset for Smartphone Cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1692-1700.

[2]T.Plötz, S. Roth, Benchmarking Denoising Algorithms with Real Photographs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2750-2759.

[3]P. Arbeláez, M. Maire, C. Fowlkes, J. Malik, Contour Detection and Hierarchical Image Segmentation, IEEE transactions on Pattern Analysis and Machine Intelligence 33 (5) (2011) 898-916.

[4]K. Ma et al., Waterloo Exploration Database: New Challenges for Image Quality Assessment Models, IEEE transactions on Image Processing 26 (2) (2017) 1004-1016.

[5]Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced Deep Residual Networks for Single Image Super-Resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1132-1140.

[6]E. Agustsson, R. Timofte, NTIRE 2017 Challenge on Single Image Super-Resolution:

Methods and Results, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1122–1131.

[7]S. Roth, M. J. Black, Fields of Experts: a framework for learning image priors, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 860–867.

[8]J. B. Huang, A. Singh, N. Ahuja, Single Image Super-resolution from Transformed Self-Exemplars, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5197-5206.

[9]S. Roth, M. J. Black, Fields of experts: A framework for learning image priors, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, IEEE, 2005, pp. 860–867.

[10]Zhang, L., Wu, X., Buades, A., Li, X., 2011. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. Journal of Electronic Imaging 20(2), 023016.

[11]Franzen, R., 1999. Kodak lossless true color image suite. Available online: http://r0k.us/graphics/kodak.

[12]K.Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a Gaussian Denoiser:Residual Learning of Deep CNN for Image Denoising, IEEE transactions on Image Processing 26 (7) (2017) 3142-3155.

[13]K. Zhang, W. Zuo, L. Zhang, FFDNet: Toward a Fast and Flexible Solution for CNN Based Image Denoising, IEEE transactions on Image Processing 27 (9) (2018) 4608-4622.

[14]K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, R. Timofte, Plug-and-Play Image Restoration with Deep Denoiser Prior, IEEE transactions on Pattern Analysis and Machine Intelligence, 44 (10) (2022) 6360-6376.

[15]Z. Liu, Y. Lin, Y. Cao, H. Han, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012-10022.

[16]Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A General U-Shaped Transformer for Image Restoration, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17662-17672.

[17]J. Liang, J. Cao, G. Sun, K. Zhang, L. Gool, R. Timofte, SwinIR: Image Restoration Using Swin Transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1833-1844.

[18]P. Liu, H. Zhang, K. Zhang, L. Lin, W. Zuo, Multi-level Wavelet-CNN for Image Restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 773-782.

[19]W. Bae, J. Yoo, J. Chul Ye, Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1141-1149.

[20]K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, 2007. Image denoising with block-matching and 3D filtering, in: Image Processing: Algorithms and Systems, Vol. 6064, SPIE, pp. 354-365.

[21]X. Wang, R. Girshick, A. Gupta, K. He, Non-local Neural Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794-7803.

[22]Liu, B. Wen, et al.,Non-local Recurrent Network for Image Restoration, in: Proceedings of

the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 1680–1689.

[23]J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141.

[24]Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic Convolution: Attention over Convolution Kernels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11030-11039.

[25]A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[26]Dosovitskiy, A., et al., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations.

[27]S.Waqas Zamir, A. Arora, S. Khan, M. Hayat, F. Shahbaz Khan, M. Hsuan Yang, Restormer: Efficient Transformer for High-Resolution Image Restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728-5739.

[28]H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, W. Gao, Pre-Trained Image Processing Transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12299-12310.

[29]Z. Sheng, X. Liu, S. -Y. Cao, H. -L. Shen and H. Zhang, Frequency-Domain Deep Guided Image Denoising, in IEEE Transactions on Multimedia, vol. 25, 2023，pp. 6767-6781.

[30]L. Chen, C. Wang, X. Zhou and Z. Qin, Robust and Compatible Video Watermarking via Spatio-Temporal Enhancement and Multiscale Pyramid Attention, in IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, 2025, pp. 1548-1561.

[31]H. Duan, Y. Long, S. Wang, H. Zhang, C. G. Willcocks and L. Shao, Dynamic Unary Convolution in Transformers, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, 2022, pp. 12747-12759.

[32]X. Li, W. Wang, X. Hu, J. Yang, Selective Kernel Networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510-519.

[33]Hu, Y., Ge, Y., Qi, M. et al. Dual branch attention network for image super-resolution. Sci Rep 15, 29019 (2025). https://doi.org/10.1038/s41598-025-97190-1.

[34]Zhang, Y., Li, K., Li, K., Zhong, B., Fu, Y., 2019. Residual non-local attention networks for image restoration. In: International Conference on Learning Representations.

[35]Ren, X. He, C. Wang, Z. Zhao, Adaptive Consistency Prior based Deep Network for Image Denoising, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8596-8606.

[36]Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L., 2020. Learning enriched features for real image restoration and enhancement. In: European Conference on Computer Vision, pp. 492–511.

[37]S.W. Zamir et al., Multi-Stage Progressive Image Restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14816-14826.

[38]S. Guo, Z. Yan, K. Zhang, W. Zuo, L. Zhang, Toward Convolutional Blind Denoising of Real Photographs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1712-1722.

[39]S. Anwar, N. Barnes, Real Image Denoising With Feature Attention, in: Proceedings of the

IEEE/CVF International Conference on Computer Vision, 2019, pp. 3155-3164.

[40]Z. Yue, H. Yong, Q. Zhao, L. Zhang, D. Meng,Variational Denoising Network: Toward Blind Noise Modeling and Removal, in: Advances in Neural Information Processing Systems, 2019, pp. 1694-1701.

[41]S. W. Zamir et al., CycleISP: Real Image Restoration via Improved Data Synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2693-2702.

[42]K. Zhang, W. Zuo, S. Gu, L. Zhang, Learning Deep CNN Denoiser Prior for Image Restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2808-2817.

[43]Wu, J., Wu, H. & Yuan, G, Detail-aware image denoising via structure preserved network and residual diffusion model, The Visual Computer, 2025, 41: 639-658.

[44]Guo, Z., Wang, B, Enhancing sandstorm images via color-guided spatial-frequency fusion network, The Visual Computer, 2025, 41: 8041-8053.

[45]X. Y. Liu, Z. K. Wu, X. C. Wang, Validity of non-local mean filter and novel denoising method, Virtual Reality & Intelligent Hardware, 2023, 5(4): 338-350.

[46]Lin C , Zou C , Xu H, SCNet: A Dual-Branch Network for Strong Noisy Image Denoising Based on Swin Transformer and ConvNeXt, Computer Animation and Virtual Worlds, 2025, 36(3): e70030.

[47]Xu, H., Zou, C., & Lin, C, UTMCR: 3U-Net Transformer With Multi-Contrastive Regularization for Single Image Dehazing, Computer Animation and Virtual Worlds, 2025, 36(3): e70029.

[48]Y. Wen, B. Luo, W. Shi, J. Ji, W. Cao, X. Yang, and B. Sheng, SAT-Net: Structure-Aware Transformer-Based Attention Fusion Network for Low-Quality Retinal Fundus Images Enhancement, IEEE Transactions on Multimedia, 2025, 27: 6198-6210.

[49]Y. Wen, S. Shen, W. Shi, W. Cao, L. Bi, X. Yang, and B. Sheng, A Lightweight Depthwise Separable ConvNet with Frequency-domain Enhancement for Retinal Vessel Segmentation, ACM Transactions on Multimedia , 2025, 21(336):1-23.

[50]X. Lin, S. Sun, W. Huang, B. Sheng, P. Li and D. D. Feng, EAPT: efficient attention pyramid transformer for image processing, IEEE Transactions on Multimedia, 2023, 25: 50-61.