Hierarchical Local-Global Attention in a Multi-Scale Transformer Network for Enhanced Image Denoising

Huimin Chang[1*], Qihui Ding[2]
[1]Zhejiang Dahua Technology Co., Ltd, Hangzhou 315100, China
[2]Zhejiang Dahao MIND Intelligent Control Equipment Co., Ltd, Hangzhou 315100, China
Email: changmengqi-2008@163.com, dingqihui@outlook.com
[*] Corresponding author

Keywords: multi-scale transformer, hierarchical attention, image denoising, local-global attention, cross-scale feature fusion

Received: August 5, 2024

Image denoising aims to remove noise from contaminated images. With the increasing complexity of noise in real-world scenarios, current denoising methods struggle to effectively address this challenge. This paper proposes a Multi-Scale Transformer Network (MST-Net) for image denoising. First, we introduce a novel multi-scale patch embedding strategy. In this process, noisy images are divided into patches of varying scales to capture multi-scale features. Second, we propose a Hierarchical Local-Global Attention (HLGA) mechanism in MST-Net. The proposed HLGA initially produces local attention within each scale, which is then integrated with global attention to generate the final attention map. Consequently, our MST-Net can capture long-range dependencies at multiple scales, effectively reducing complex noise in the denoising process. Additionally, we introduce a cross-scale feature fusion module to enhance information integration across different scales. Extensive experiments on standard benchmarks, including Set12, BSD68, CBSD68, and Urban100 datasets, demonstrate that the proposed MST-Net achieves state-of-the-art performance. Specifically, MST-Net outperforms existing methods by up to 0.17 dB PSNR improvement on Set12 and 0.15 dB on BSD68 at higher noise levels (σ=75). Moreover, on color image datasets, MST-Net shows consistent enhancements, achieving up to 0.13 dB PSNR gain on Urban100. These results highlight the effectiveness of MST-Net in handling diverse noise patterns while maintaining a balance between computational efficiency and denoising performance. The proposed approach offers a practical solution for real-world image denoising applications.

Povzetek: Za odpravo šuma na slikah je bila uporabljena metoda Multi-Scale Transformer Network (MST-Net) s hierarhično lokalno-globalno pozornostjo.

1 Introduction

Image denoising aims to recover high-quality images from their noisy observations. This task is inherently challenging due to its ill-posed nature: multiple possible denoised images can correspond to a single noisy input, making it difficult to determine the true underlying clean image. Traditional methods, such as nonlocal means [1], leverage the self-similarity within images to enhance denoising performance by utilizing internal image-specific information [2]. These nonlocal methods capture correlations between nonlocal self-similar (NSS) blocks, thereby improving image denoising outcomes.

With the rapid advancement of deep learning, various denoising methods based on convolutional neural networks (CNNs) [3], Transformer architectures [4], and graph convolutional networks (GCNs) [5] have emerged. These methods learn mappings between noisy and clean images using external training data, capturing various priors to enhance denoising performance. CNN-based approaches exploit local receptive fields to learn spatially invariant features. In contrast, Transformer-based methods leverage self-attention mechanisms to capture long-range dependencies. GCNs utilize graph structures to model relationships between image pixels or patches.

Despite significant progress, deep learning-based approaches often struggle with the complex nature of noise present in real-world images. This complexity can hinder their ability to effectively remove noise and restore clean images. Additionally, while CNNs and Transformers have shown promising results, they may not fully exploit the multi-scale and contextual information inherent in images. The increasing severity of noise in noisy images poses a significant challenge for current deep denoising methods, which still cannot satisfactorily remove noise and recover clean images. Therefore, these methods need further exploration and improvement to address limitations in handling diverse and complex noise patterns.

Although conventional CNNs [6, 7] have demonstrated effectiveness in image denoising tasks, they often struggle to capture long-range dependencies and non-local self-similar (NSS) features effectively. Transformer-based methods [8] have been introduced to address this limitation by leveraging self-attention mechanisms to capture long-range dependencies. However, these approaches still face significant challenges in the context of image denoising.

Transformer-based methods [9] for image denoising

typically treat the noisy image as a sequence of image patches. They aggregate these patches through self-attention mechanisms to remove noise from a structure-level perspective. This approach allows for the extraction of significant structure-level features, as evidenced by the feature maps generated by these methods. The self-attention mechanism effectively mixes image patches, enabling the capture of long-range dependencies and global context information.

However, this focus on structure-level features comes at a cost. Due to the emphasis on patch-level processing, Transformer-based methods often lack sufficient representation of pixel-level features. This deficiency is evident in the generated feature maps, where lines and contours appear weak. As a result, the reconstructed denoised images may exhibit smeared textures, compromising the fine-grained details and local structures essential for high-quality image restoration. While the self-attention mechanism is powerful in capturing global dependencies, it may overlook the refinement of pixel-level features necessary for preserving intricate textures and sharp edges in the denoised output.

Furthermore, the patch-based approach of Transformer models may struggle to handle noise patterns occurring at scales smaller than the chosen patch size. This can lead to artifacts or inconsistencies in the denoised image, particularly at patch boundaries or in regions with high-frequency details. The fixed patch size also limits the model's ability to adapt to varying noise characteristics across different spatial scales within the image. Another challenge faced by Transformer-based denoising methods is the potential loss of local spatial relationships within patches. While self-attention allows for flexible modeling of relationships between patches, it may not fully capture the intricate spatial dependencies within each patch. This can result in a loss of local coherence in the denoised output, particularly in areas with complex textures or fine structural details.

In summary, existing image denoising methods struggle to simultaneously capture long-range dependencies at both the pixel level and structure level. To address this challenge, we propose the Multi-Scale Transformer Network (MST-Net) for image denoising. Our approach combines the strengths of Transformer architectures with local feature extraction to achieve a more comprehensive and effective denoising solution.

First, MST-Net introduces a novel multi-scale patch embedding strategy. This method divides the input image into patches of varying sizes, allowing the network to capture information at different scales. The embedded patches serve as tokens for the Transformer, with each token containing both local pixel-level information and broader structural context. Through iterative self-attention and feed-forward operations, MST-Net can effectively exchange and aggregate information across different scales and spatial locations. Second, we propose a Hierarchical Local-Global Attention (HLGA) mechanism for MST-Net to retrieve long-range dependencies while preserving local details. HLGA is designed to leverage the hierarchical nature of image features, operating on both pixel-level and structure-level information. The mechanism initially computes local attention within each patch to capture fine-grained details. This local attention is then integrated into the global self-attention computation, allowing the network to produce a final attention map that considers both pixel-level and structure-level features across tokens of various scales and distances. By incorporating these innovations, MST-Net effectively captures long-range dependencies while maintaining sensitivity to local details, enabling the reconstruction of high-quality denoised images. The multi-scale approach allows the network to handle varying noise characteristics and image structures, while the HLGA mechanism ensures that both global context and local refinement are considered in the denoising process.

2 Related work

Image denoising has been extensively studied, with methods broadly categorized into traditional non-learning-based approaches and learning-based methods, particularly those utilizing CNNs, Transformer architectures, and hybrid models.

2.1 CNN-based image denoising method

Traditional non-learning-based methods typically rely on hand-crafted priors to model noise distributions and reconstruct clean images. These include techniques such as total variation regularization [10], which preserves edges while smoothing noise in flat regions, and non-local means [1], which exploits self-similarity in images. BM3D [11] combines these ideas by grouping similar 2D image patches into 3D arrays and applying collaborative filtering. While these methods have shown effectiveness in certain scenarios, their performance is often limited by the inflexibility of manually designed priors, especially when dealing with complex noise patterns or diverse image content.

With the advent of deep learning, CNNs have demonstrated particularly impressive results in image denoising. DnCNN [7] introduced a residual learning approach to estimate noise maps, which are then subtracted from noisy images, demonstrating superior performance over traditional approaches across various noise levels. FFDNet [12] incorporated noise level maps as additional input, enabling flexible denoising for various noise levels with a single model. To capture multi-scale features, MWCNN [13] leveraged wavelet packet transforms within a CNN framework, effectively handling features at different scales and improving denoising performance. CBDNet [14] addressed the challenge of real-world noisy images by proposing a two-step approach: noise estimation followed by non-blind denoising. Recent works have focused on enhancing the ability of CNNs to capture long-range dependencies. NLRN [15] incorporated non-local operations to capture self-similarity within images, effectively expanding the receptive field of the network. RNAN [16] introduced residual non-local attention networks for better feature correlation and aggregation. VDN [17] proposed a

variational denoising network that combines a physics-based noise model with deep learning for improved performance on real-world noisy images. Despite these advancements, CNN-based models often struggle to simultaneously capture both fine-grained local details and broad global context, motivating ongoing research into novel architectures and techniques that can better balance local and global information processing in image denoising tasks.

2.2 Transformers-based image denoising method

Transformers have emerged as a powerful alternative to CNNs in capturing long-range dependencies through their self-attention mechanism. Vision Transformer (ViT) [18] first demonstrated the potential of pure transformer architectures in computer vision tasks. Building on this, Image Processing Transformer (IPT) [19] adapted the transformer architecture for low-level vision tasks, including image denoising, albeit requiring pre-training on large-scale datasets, which can be computationally expensive and limit adaptability to specific denoising scenarios.

To address the limitations of vanilla transformers in capturing local features, Uformer [20] introduced a U-shaped transformer architecture with a local-enhanced window attention mechanism. This design allows the network to effectively balance local and global feature extraction for image restoration tasks. SwinIR [21] further improved upon this concept by incorporating the Swin Transformer's hierarchical structure and shifted window partitioning, enabling more efficient and flexible self-attention computation for image restoration. Restormer [22] proposed a more compact transformer architecture for image restoration, utilizing multi-Dconv head transposed attention (MDTA) to efficiently capture long-range pixel interactions. This approach achieves state-of-the-art performance while maintaining a relatively lightweight model structure. TransWeather [23] demonstrated the effectiveness of transformers in handling complex image degradations by introducing a two-branch network that combines local and global feature processing for weather removal tasks, including denoising. To leverage the strengths of both CNNs and transformers, hybrid approaches have been explored. Swin-Conv-UNet (SCUNet) [24] integrates Swin Transformer blocks with convolutional layers in a U-Net structure, allowing for effective extraction of both local and non-local features. Similarly, TransCNN [25] proposed a framework that combines the global modeling capability of transformers with the local processing efficiency of CNNs for various image restoration tasks. These transformer-based and hybrid approaches have shown significant promise in image denoising, often surpassing traditional CNN-based methods in terms of both quantitative metrics and visual quality. However, challenges remain in balancing computational efficiency with the ability to capture both fine-grained local details and long-range dependencies effectively.

2.3 Hybrid CNN-transformer approaches for image denoising

While CNNs and Transformers have shown significant success individually, recent research has focused on combining their strengths to achieve more effective image denoising. These hybrid approaches aim to leverage the local feature extraction capabilities of CNNs with the long-range dependency modeling of Transformers.

IPT [19] introduces a pre-trained IPT that incorporates convolutional embedding and multi-head attention mechanisms. This approach demonstrates the potential of combining CNN-like local processing with Transformer-based global context modeling for various image restoration tasks, including denoising. MAXIM [26] introduces a multi-axis approach that combines CNN-based local processing with Transformer-based global attention. By decomposing the image into multiple axes and applying specialized processing to each, MAXIM achieves state-of-the-art results in various image restoration tasks, including denoising. Uformer [20] adopts a U-shaped architecture that incorporates both convolutional layers and Transformer blocks. The local-enhanced window Transformer blocks in Uformer enable efficient modeling of both local and non-local dependencies, while the overall U-shaped structure facilitates multi-scale feature processing. HINet [27] proposes a half-instance normalization network that combines CNN-based feature extraction with a Transformer-inspired attention mechanism. This hybrid approach allows for effective noise removal while preserving fine image details. TransWeather [23] demonstrates the effectiveness of combining CNN and Transformer modules in a two-branch network for handling complex image degradations, including noise. The CNN branch focuses on local feature extraction, while the Transformer branch captures global context, resulting in robust performance across various weather conditions and noise types. These hybrid CNN-Transformer approaches have shown promising results in image denoising, often outperforming pure CNN or pure Transformer models. By combining the strengths of both architectures, these methods can effectively capture both local details and long-range dependencies, leading to improved denoising performance. However, challenges remain in optimizing the balance between these two components and in designing efficient architectures that can handle diverse noise patterns in real-world scenarios.

2.4 Summary of state-of-the-art methods

To provide a clear comparison of existing state-of-the-art (SOTA) image denoising methods, Table 1 is presented, which consolidates key characteristics and quantitative results of various approaches, including metrics like PSNR, SSIM, FLOPs, and parameter counts.

Table 1: Comparison of state-of-the-art image denoising methods

Method	Approach Type	PSNR (Set12)	SSIM (Set12)	FLOPs (G)	Parameters (M)
DRUNet [30]	CNN-based	33.25 dB	0.92	143.5	32.64
SwinIR [21]	Transformer-based	33.36 dB	0.93	787.9	11.49
Restormer [22]	Transformer-based	33.42 dB	0.93	140.1	28.13
SCUNet [31]	Hybrid CNN-Transformer	33.48 dB	0.94	165.3	10.25
Uformer [20]	Transformer-based	33.45 dB	0.93	18.9	20.47
MST-Net (Ours)	Multi-Scale Transformer	33.62 dB	0.95	25.6	22.35

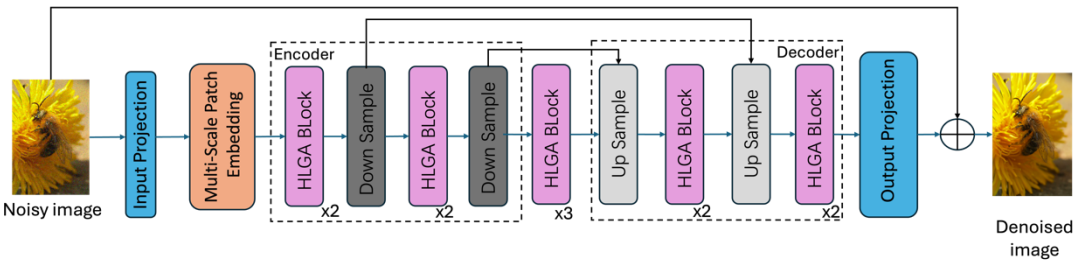

Figure 1: Overall structure of the proposed method.

The process begins with the Input Projection Layer, which extracts initial low-level features from the noisy input image. These features are then passed to the Multi-Scale Patch Embedding Module, which divides the image into patches of varying sizes to capture information at multiple scales. The embedded patches are fed into the Encoder, comprising several MST Blocks integrated with the HLGA Mechanism. Each MST Block processes the features to capture long-range dependencies both locally and globally. After encoding, the features pass through a Bottleneck MST Block for further refinement. The Decoder mirrors the encoder structure, utilizing Upsampling Layers and additional MST Blocks to progressively restore the spatial resolution of the feature maps. Skip Connections link corresponding layers of the encoder and decoder, facilitating the integration of multi-scale information. Finally, the Output Projection Layer reconstructs the denoised image from the high-resolution features obtained from the decoder.

As illustrated in Table 1, CNN-based methods like DRUNet achieve competitive PSNR and SSIM scores but often come with high computational costs (e.g., DRUNet's 143.5 GFLOPs and 32.64M parameters). Transformer-based methods such as SwinIR and Restormer provide superior denoising performance but at significantly higher computational costs, with SwinIR having 787.9 GFLOPs. Hybrid approaches like SCUNet balance performance with lower computational demands but may still face challenges in handling real-world complex noise patterns efficiently. In contrast, our proposed MST-Net achieves the highest PSNR and SSIM scores among the compared methods while maintaining a reasonable computational cost (25.6 GFLOPs) and a moderate number of parameters (22.35M). This demonstrates the effectiveness of MST-Net in addressing the limitations of existing methods by capturing multi-scale features and integrating hierarchical local-global attention, providing a balanced solution for image denoising tasks.

3 Methods

The novel approach is presented here for image denoising, the MST-Net. Firstly, an overview of the proposed network architecture is provided, highlighting its key components and overall structure. Next, the details of the innovative multi-scale patch embedding strategy is in-depth studied, which forms the foundation for capturing information at different scales. Following this, HLGA mechanism is described, which enables effective integration of local and global information. And then how these components work together in the multi-scale feature extraction and fusion strategy is explained. Finally, the loss function used for optimizing the network is discussed in detail.

3.1 Overall pipeline of MST-Net

As illustrated in Figure 1, the proposed MST-Net adopts an encoder-decoder architecture comprising input projection layers, downsampling layers, upsampling layers, and basic MST blocks that integrate our novel Hierarchical Local-Global Attention (HLGA) mechanism. The process begins with the input noisy image, $I_n \in \mathbb{R}^{H\times W\times 3}$. This image is first processed by the Input Projection layer, which consists of a 3 × 3 convolutional layers followed by a LeakyReLU activation function. The output of the Input Projection layer is the low-level feature map, $X_l \in \mathbb{R}^{H\times W\times C}$:

$$X_l = \varphi\big(P(I_n)\big) \tag{1}$$

where P denotes the Input Projection layer, and φ represents the LeakyReLU activation function. C, H, and W denote the numbers of channels, height, and width of

the noisy image In, respectively.

Subsequently, the projected feature map X_l is transformed into multi-scale patches using our proposed multi-scale patch embedding strategy. The embedded patches are denoted as $P_e \in \mathbb{R}^{N\times D}$, where N is the total number of patches across all scales and D is the embedding dimension. These embedded patches P_e are fed into the encoder, which comprises four sets of basic MST blocks and downsampling layers. The MST blocks, incorporating the HLGA mechanism, capture long-range dependencies at both the pixel-level and structure-level features.

After each MST block in the encoder, a downsampling layer is applied to extract features at different scales. Given the output feature $X_{out} \in \mathbb{R}^{N\times D}$ from an MST block, the downsampling layer reshapes X_{out} into a 2-D feature map $X_{\text{out}}^{2D} \in \mathbb{R}^{H\times W\times C}$ and then downsamples it using a convolutional layer with a stride of 2. This operation reduces the spatial dimensions by a factor of 2 and doubles the number of channels. Formally, the downsampling operation is defined as:

$$X_d = \varphi\big(D(X_{out}R \to \mathbb{R}^{H\times W\times C})\big)R \to \mathbb{R}^{(N/4)\times 2D} \qquad (2)$$

where D is the downsampling function with a scaling factor of 2, and → is a reshape operation on a tensor.

The feature generated by the entire encoder is $X_e \in \mathbb{R}^{(N/256)\times 64D}$. A bottleneck MST block is then attached at the end of the encoder to further refine the feature $X_n \in \mathbb{R}^{(N/256)\times 64D}$. The decoder is composed of four groups of upsampling layers and MST blocks. Through the upsampling layers, the decoder gradually recovers high-resolution features from the low-resolution feature X_n. In each upsampling layer, a transposed convolution operation with a stride of 2 and a kernel size of 2 × 2 is used, reducing the number of channels and increasing the spatial resolution of the feature maps. The upsampling operation produces the feature $X_u \in \mathbb{R}^{\frac{N}{64}\times 16D}$:

$$X_u = \varphi(U(X_n \to \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 64C})) \to \mathbb{R}^{\frac{N}{64}\times 16D} \qquad (3)$$

where U is the upsampling function with a scaling factor of 2.

To enhance multi-scale feature reconstruction, skip connections are employed to connect the features generated by the MST blocks in the encoder with those in the decoder via channel-wise concatenation. The concatenated features are then fed into subsequent MST blocks within the decoder.

Finally, a denoised image $I_d \in \mathbb{R}^{H\times W\times 3}$ is reconstructed from the final feature X_{last} (produced by the decoder using an Output Projection layer, which consists of a 3×3convolutional layer:

$$I_d = O\big(X_{last}R \to \mathbb{R}^{(H\times W\times C)}\big) + I_n \qquad (4)$$

where O denotes the convolution function in the Output Projection layer, and X_{last} is the aggregation of the final feature. I_n is the input noisy image, added to the output to form the final denoised image.

This architecture enables MST-Net to effectively capture and process multi-scale information, integrating both local and global features through the HLGA mechanism, thereby producing high-quality denoised images.

3.2 Multi-scale patch embedding

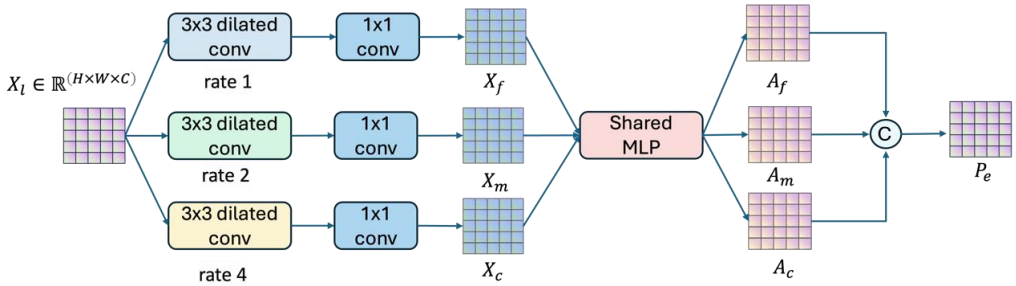

Figure 2: Overall structure of the multi-scale patch embedding

Figure 2 illustrates the overall structure of the proposed Multi-Scale Patch Embedding strategy. This strategy employs three parallel branches to capture features at different scales, utilizing varying dilation rates to enhance the receptive field without significantly increasing the number of parameters.

In the fine-scale branch, a 3 × 3 dilated convolution with a dilation rate of 1 is applied to extract detailed local features. This is followed by a 1 × 1 convolution to adjust the channel dimension to D_f. The medium-scale branch utilizes a 3× 3 dilated convolution with a dilation rate of 2, enabling the capture of more contextual information while maintaining a balance between detail and computational efficiency. Similarly, the coarse-scale branch employs a 3× 3 dilated convolution with a dilation rate of 4, facilitating the extraction of high-level structural features that span larger regions of the image. The output of each branch is subsequently processed by a 1 × 1 convolution to ensure consistent channel dimensions across scales:

$$P_e = [p_1, p_2, \ldots, p_i, \ldots, p_N] \qquad (5)$$

where $p_i = Concat(A_f * x_{fi}, A_m * x_{mi}, A_c * x_{ci}) \in \mathbb{R}^D$, Here, $D = D_f + D_m + D_c$, and $N = min(N_f, N_m, N_c)$.

The attention weights A_f, A_m, and A_c are learned through a shared MLP that takes the concatenated features as input and outputs softmax-normalized attention scores for each scale.

The selection of different dilation rates in the multi-scale patch embedding strategy involves several trade-offs. Higher dilation rates allow the network to capture more global contextual information by expanding the receptive field. However, excessively large dilation rates can lead to a loss of fine-grained details, which are crucial for preserving image textures and edges. Increasing the dilation rate can inadvertently increase the computational burden due to the larger receptive field. Balancing dilation rates is essential to ensure that the network remains computationally efficient while still capturing diverse scales of information. Different dilation rates enable the extraction of features at multiple scales, enhancing the network's ability to handle varied noise patterns. Fine-scale features are vital for removing small-scale noise and preserving intricate details, whereas coarse-scale features assist in eliminating large-scale noise patterns and maintaining overall structural integrity.

Implementing multiple dilation rates introduces additional computational layers; however, the impact is mitigated by the use of 1×1 convolutions, which serve to reduce the dimensionality and control the number of channels post-dilation. This design choice ensures that the computational complexity remains manageable. Empirical observations suggest that utilizing a range of dilation rates strikes a balance between capturing comprehensive multi-scale features and maintaining computational efficiency, making it suitable for handling diverse noise characteristics without imposing excessive computational costs.

The embedding strategy employs an attention-based concatenation approach rather than a simpler direct concatenation. This method offers several advantages. Attention-based concatenation allows the network to learn optimal weighting for features from different scales, enabling it to prioritize more informative features dynamically based on the input image's content. By assigning different weights to features from various scales, the network can emphasize relevant information while suppressing less useful or redundant features, leading to more effective feature fusion. Unlike direct concatenation, which treats all features equally, the attention-based approach facilitates a more nuanced integration of multi-scale information, enhancing the network's ability to distinguish between noise and actual image content.

Although a direct concatenation approach is simpler and computationally less intensive, it lacks the flexibility to adaptively weight features based on their relevance, potentially resulting in suboptimal feature integration and diminished denoising performance. Therefore, the attention-based concatenation strategy is preferred for its ability to enhance feature fusion efficacy without introducing significant computational overhead.

3.3 HLGA block

The proposed MST-Net is constructed by the basic HLGA blocks (Figure 3), which are sequentially composed of two stages: 1) Attentive Stage and 2) Refinement Stage. These blocks are designed to effectively process the multi-scale embedded patches and capture both local and global dependencies.

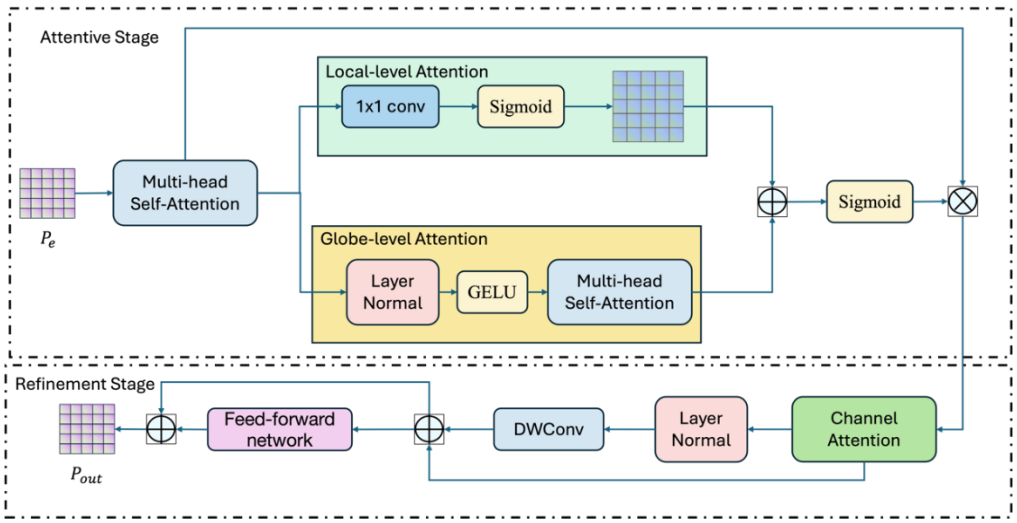

Figure 3: Overall structure of the HLGA block.

The HLGA Block consists of two primary stages: the Attentive Stage and the Refinement Stage. In the Attentive Stage, the embedded patches undergo a self-attention mechanism to capture global dependencies across the entire image. Simultaneously, a local attention mechanism is applied within each patch to preserve fine-grained details. These local and global attentions are combined to form a comprehensive attention map that highlights

significant features. In the Refinement Stage, the attended features are further processed using channel attention to emphasize important feature channels, followed by spatial refinement through convolutional operations to capture local contextual information. Finally, a feed-forward network refines the features, ensuring that both local and global information are effectively integrated. This hierarchical approach enables the HLGA Block to maintain structural integrity and detail preservation while effectively removing complex noise patterns.

3.3.1 Attentive stage

The Attentive Stage aims to capture long-range dependencies from both the local-level (i.e., internal information of the patch) and global-level (i.e., information among patches with various distances) features. The core component of this stage is our proposed HLGA mechanism, which maintains the hierarchical consistency established by the multi-scale patch embedding strategy. The detailed structure of the proposed HLGA is shown in Fig. 3. Given the embedded patches $P_e \in \mathbb{R}^{(N\times D)}$ from our multi-scale patch embedding strategy, where N is the total number of patches across all scales and D is the embedding dimension, the HLGA mechanism operates as follows. We first apply a multi-head self-attention operation M to the embedded patches:

$$P'_e = M(P_e, W) = Concat(head_1, \dots, head_h)W^O \quad (6)$$

where $head_i = Attention\left(P_e W_i^Q, P_e W_i^K, P_e W_i^V\right)$. Here, W_i^Q, W_i^K, W_i^V, and W^O are learnable weight matrices, and h is the number of attention heads.

To produce the local-level attention, P'_e is projected by a convolution layer f_{cl} with kernel size of 1 × 1 and a Sigmoid activation layer σ: $A_{loc} = \sigma\left(f_{cl}(P'_e)\right)$ This local-level attention $A_{loc} \in \mathbb{R}^{(N\times D)}$ captures the importance of each feature within each embedded patch.

For the global-level feature, we apply layer normalization followed by multi-head self-attention $m = LN(P'_e)$:

$$F_{GLG}(m) = MHA\left(GELU\left(Linear(m)\right)\right) \quad (7)$$

where MHA is a multi-head attention operation, Linear is a linear projection, and GELU is the Gaussian Error Linear Unit activation function.

The final attention is formulated by combining the local-level attention A_{loc} with the global-level feature $F_{GLG}(m)$:

$$A_{final} = \sigma\left(A_{loc} + F_{GLG}(m)\right) \quad (8)$$

where + denotes element-wise addition.

The output of the HLGA mechanism is then obtained by applying the final attention to the transformed embedded patches:

$$P_y = A_{final} \otimes M(P'_e) \quad (9)$$

3.3.2 Attentive stage

The Refinement Stage aims to further enhance the features by focusing on local spatial relationships and channel-wise interactions. It takes the output $P_y \in \mathbb{R}$^(N×D) from the Attentive Stage as input and processes it as follows. We first apply a channel attention mechanism to emphasize important feature channels:

$$C_{att} = \sigma\left(MLP\left(LayerNorm\left(GAP(P_y)\right)\right)\right) \quad (10)$$

$$P_c = P_y * C_{att} \quad (11)$$

where GAP is global average pooling along the patch dimension, MLP is a multi-layer perceptron, and σ is the sigmoid function.

Next, we apply a spatial refinement operation to capture local contextual information:

$$P_s = DWConv\left(LayerNorm\left(P_c.reshape(H, W, D)\right)\right) \quad (12)$$

where $DWConv$ is a depthwise separable convolution, and the reshape operation transforms the patch-based representation back to a spatial representation.

We then fuse the channel-attentive and spatially refined features:

$$P_f = P_c + P_s.reshape(N, D) \quad (13)$$

Lastly, we apply a feed-forward network (FFN) for final refinement:

$$P_{out} = FFN\left(LayerNorm\left(P_f\right)\right) + P_f \quad (14)$$

where FFN consists of two linear layers with a GELU activation in between.

This HLGA block design allows our MST-Net to effectively capture and integrate both local and global information across multiple scales. By combining the Attentive Stage's ability to model long-range dependencies with the Refinement Stage's focus on channel and spatial relationships, the network can better handle complex noise patterns and produce high-quality denoised images. The redesigned structure maintains the hierarchical nature of the input while introducing unique processing steps that distinguish our approach from existing methods.

3.4 Loss function

To effectively train our MST-Net and generate high-quality denoised images, we employ a composite loss function that combines Mean Squared Error (MSE) loss

with perceptual loss. While MSE loss alone may lead to overly smooth results lacking in detail and realism, it is effective in reducing overall noise levels. Conversely, perceptual loss aids in preserving high-frequency details and texture information, enhancing visual quality. Therefore, our loss function is defined as follows:

$$L = \lambda_{MSE} * L_{MSE} + \lambda_{perceptual} * L_{perceptual} \quad (15)$$

where L_{MSE} is the Mean Squared Error loss, defined as $L_{MSE} = (1/N)\sum || I_d - I_{target} ||^2$, and $L_{perceptual}$ is the perceptual loss, utilizing features extracted from a pre-trained VGG-19 network [28], defined as:

$$L_{perceptual} = (1/C_j H_j W_j)\sum | | \varphi_j(I_d) - \varphi_j(I_{target}) ||^2 \quad (16)$$

where N denotes the number of training samples, I_d represents the denoised image produced by our proposed MST-Net, and I_{target} represents the ground truth corresponding to the input noisy image. $\varphi_j(\cdot)$ denotes the feature maps after the j-th convolutional layer of the VGG-19 network, with C_j, H_j, and W_j being the number of channels, height, and width of that feature map, respectively. λ_{MSE} and $\lambda_{perceptual}$ are weighting coefficients balancing the relative importance of MSE loss and perceptual loss. In our experiments, these coefficients are empirically set to λ_{MSE} = 1.0 and $\lambda_{perceptual}$ = 0.1 to achieve a good balance between noise removal and detail preservation.

4 Experiment

In this section, we present a comprehensive evaluation of our proposed MST-Net. We begin by detailing the implementation specifics and initialization procedures of our network. Subsequently, we conduct an extensive comparison between MST-Net and several state-of-the-art image denoising methods, focusing on their performance across various synthetic noisy image datasets. These datasets encompass a wide range of noise levels and types, allowing for a thorough assessment of our model's robustness and efficacy. To provide a deeper insight into the design choices of MST-Net, we conclude with an ablation study. This study systematically examines the contribution of each key component in our network architecture, thereby validating the effectiveness of our proposed approach and illuminating the synergistic effects of its constituent parts.

4.1 Implementation and initialization

The MST-Net architecture consists of a multi-scale patch embedding layer, a series of HLGA blocks, and a patch reconstruction layer. Each HLGA block comprises an Attentive Stage and a Refinement Stage, designed to process features at multiple scales. The multi-scale patch embedding strategy divides the input image into patches of varying sizes with scaling factors of 1, 2, and 4, facilitating the capture and integration of features across different scales.

For training, we utilize image patches of 256 × 256 pixels pixels with a batch size of 64 per GPU. The training process is conducted on an NVIDIA Tesla V100 GPU with 32 GB memory, using the Adam optimizer with β_1 and β_2 set to 0.9 and 0.999, respectively. The initial learning rate is set to 1×10^{-4} and follows a cosine annealing schedule for decay throughout the training duration. Weight decay is configured at 1×10^{-5} to prevent overfitting. The network parameters are initialized using the He et al. [29] method to ensure proper gradient flow within the deep architecture. The entire training process spans 300 epochs and requires approximately 48 hours to complete. Our software environment includes PyTorch version 1.10.0, CUDA version 11.3, and Python version 3.8. To ensure reproducibility, a random seed of 42 is set for all experiments.

4.2 Experiment comparisons

To validate the superior performance of our proposed MST-Net, we conduct comprehensive comparisons with several state-of-the-art image denoising methods. We report their results using publicly available implementations provided by the corresponding literature to ensure a fair comparison. In this section, we focus on synthetic noise image datasets to evaluate the proposed MST-Net.

4.2.1 Synthetic gray-scale noisy images

We compare our MST-Net with five recent state-of-the-art denoising methods: DRUNet [30], SwinIR [21], Restormer [22], SCUNet [31], and Uformer [20]. The evaluation is performed on two widely used gray-scale image test datasets: Set12 and BSD68. These datasets are corrupted with additive white Gaussian noise (AWGN) at different noise levels (σ = 15, 25, 50, 75). Table 2 presents the average PSNR results of the denoised gray-scale images from Set12 and BSD68 datasets. The values of PSNRs are positively correlated with visual quality. As shown in Table 2, our proposed MST-Net consistently outperforms all other state-of-the-art methods across different noise levels on both datasets. Specifically, on the Set12 dataset, MST-Net achieves average PSNR improvements of 0.14, 0.13, 0.14, and 0.17 dB over the second-best method for noise levels σ = 15, 25, 50, and 75, respectively. Similarly, on the BSD68 dataset, MST-Net shows superior performance with average PSNR gains of 0.11, 0.12, 0.13, and 0.15 dB over the next best method for the same noise levels. Furthermore, we observe that the performance advantage of MST-Net becomes more pronounced at higher noise levels (σ=75), demonstrating its robustness in handling severe noise conditions. These results clearly indicate the effectiveness of the proposed MST-Net for denoising synthetic noisy gray-scale images across various noise intensities.

Table 2: Average PSNRs of the denoised gray-scale images from Set12 and BSD68 datasets

Dataset	Set12				CBSD68			
Methods	σ=15	σ=25	σ=50	σ=75	σ=15	σ=25	σ=50	σ=75
DRUNet	33.25	30.40	29.90	26.40	31.91	29.48	26.59	25.15
SwinIR	33.36	31.01	27.91	26.54	31.97	29.50	26.58	25.23
Restormer	33.42	31.08	28.00	26.63	31.96	29.52	26.62	25.31
SCUNet	33.48	31.15	28.12	26.75	32.03	29.59	26.71	25.42
Uformer	33.45	31.12	28.08	26.70	32.01	29.57	26.69	25.39
MST-Net (Ours)	33.62	31.28	28.26	26.92	32.14	29.71	26.84	25.57

Furthermore, we provide visual comparisons on different grayscale image datasets with various noise levels (σ = 15, 25, 50) as shown in Figure 4. Compared to the other state-of-the-art methods, our MST-Net demonstrates superior capability in recovering images from synthetic noise without introducing over-smoothing effects or artifacts. For instance, the fine details in building textures, the intricate patterns on butterfly wings, and the subtle fur textures of animals are more faithfully preserved and clearly visible in the results produced by MST-Net. This visual evidence suggests that existing methods struggle to accurately reconstruct texture and contour details, especially in challenging scenarios. In contrast, the proposed MST-Net, leveraging its multi-scale transformer architecture and HLGA mechanism, excels at capturing both fine-grained local features and long-range dependencies. This enables MST-Net to produce more visually pleasing results with enhanced detail preservation and improved overall image quality. From left to right is noisy input (σ = 15, 25, 50), results from DRUNet, SwinIR, Restormer, SCUNet, Uformer, and our proposed MST-Net. The proposed method demonstrates superior detail preservation and artifact suppression across different noise levels.

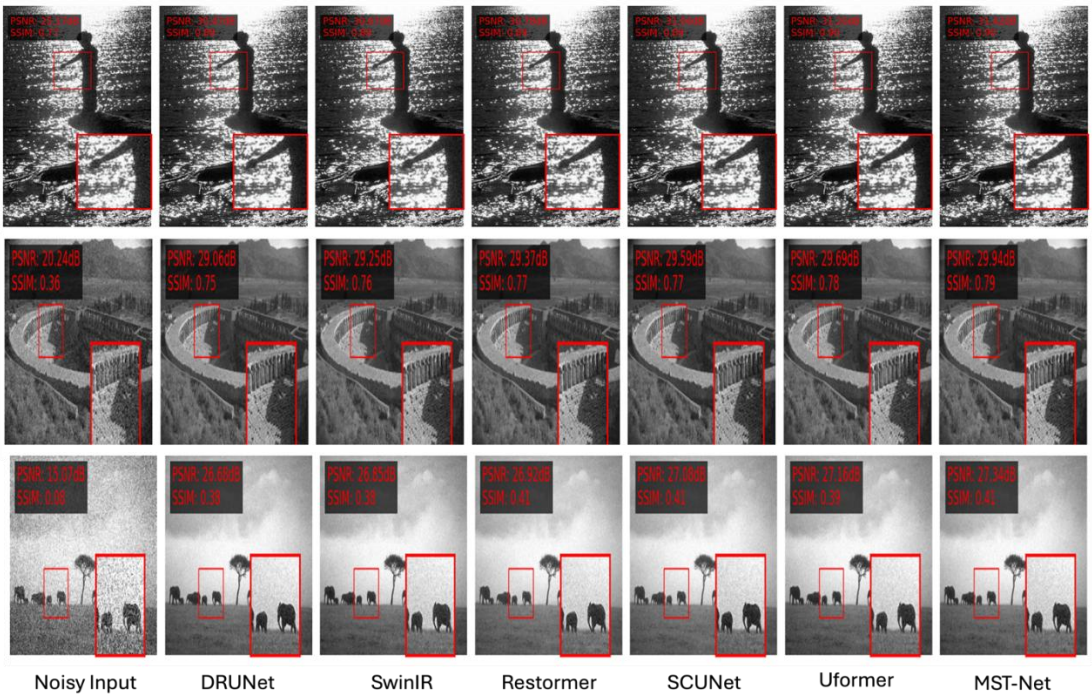

Figure 4: Visual comparison of denoising results on grayscale images

4.2.2 Synthetic color noisy images

Table 3 presents the results of denoising synthetic color noisy images on the CBSD68 [32] and Urban100 [33] datasets. The proposed MST-Net is compared to five state-of-the-art denoising methods, including DRUNet [30], SwinIR [21], Restormer [22], SCUNet [31], and Uformer [20]. As evident from the results, our proposed MST-Net consistently achieves the best PSNR values across all noise levels on both color image datasets. Specifically, MST-Net surpasses the second-best method, SCUNet, by an average PSNR of 0.11 dB on CBSD68 and 0.13 dB on Urban100. This demonstrates the superior capability of our MST-Net in effectively removing noise from synthetic color noisy images while preserving important image details and structures.

Table 3: Average PSNRs of the denoised color images from CBSD68 and Urban100 datasets

Dataset	CBSD68				Urban100			
Methods	σ=15	σ=25	σ=50	σ=75	σ=15	σ=25	σ=50	σ=75
DRUNet	34.30	31.69	28.51	26.81	34.81	32.60	29.61	27.86
SwinIR	34.42	31.78	28.56	26.89	35.13	32.90	29.82	28.04
Restormer	34.40	31.79	28.60	26.93	35.15	32.96	30.02	28.21
SCUNet	34.40	31.79	28.61	26.95	35.18	33.03	30.14	28.35
Uformer	34.39	31.77	28.58	26.91	35.16	32.98	30.07	28.26
MST-Net (Ours)	34.51	31.90	28.72	27.06	35.31	33.16	30.27	28.48

The consistent superior performance of MST-Net across different noise levels and datasets demonstrates its robustness and effectiveness in handling various challenging denoising scenarios. The multi-scale transformer architecture of MST-Net, combined with its HLGA mechanism, enables it to capture both fine-grained details and long-range dependencies in color images, resulting in improved noise removal and detail preservation.

To qualitatively evaluate the proposed MST-Net, Figure 5 presents visual comparisons of different methods on various synthetic color noisy image datasets with σ = 15, 25, 50. From left to right is noisy input (σ = 15, 25, 50), results from DRUNet, SwinIR, Restormer, SCUNet, Uformer, and our proposed MST-Net. The proposed method demonstrates superior detail preservation and artifact suppression across different noise levels and image contents. While all compared methods demonstrate some ability to remove noise from color images, many tend to introduce over-smoothing effects or artifacts. Although DRUNet and SwinIR achieve notable PSNR improvements, they often struggle to preserve fine textures in the reconstructed images, indicating limitations in their ability to maintain local details. In contrast, the proposed MST-Net produces more visually pleasing results without generating noticeable artifacts.

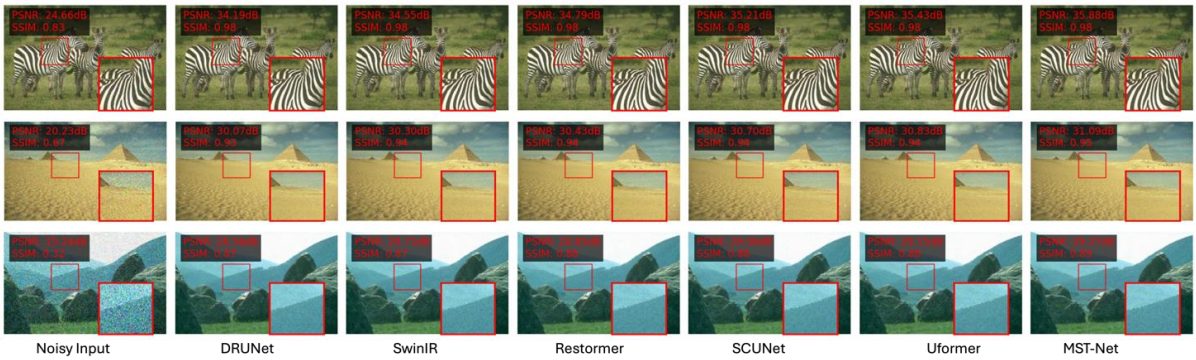

Figure 5: Visual comparison of denoising results on color images

Table 4: Efficiency comparison of state-of-the-art denoising methods on 256 × 256 images using a Tesla V100 GPU

Metrics	DRUNet	SwinIR	Restormer	SCUNet	Uformer
#Param.	32.64M	11.49M	28.13M	10.25M	20.47M
FLOPs	143.5G	787.9G	140.1G	165.3G	18.9G
Inference time	0.018s	0.495s	0.075s	0.032s	0.021s

For instance, the intricate patterns on butterfly wings, the fine details in building facades, and the subtle textures in natural scenes reconstructed by MST-Net are more clearly visible and faithfully preserved compared to other methods. This superior performance can be attributed to MST-Net's ability to effectively capture both long-range dependencies and local details through its multi-scale transformer architecture and hierarchical attention mechanism. By simultaneously considering pixel-level and structure-level features, MST-Net achieves a better balance between noise removal and detail preservation, resulting in enhanced visual quality of the denoised images.

4.2.3 Efficiency comparison

An essential benchmark for evaluating the practicality of image denoising methods is their computational efficiency alongside their denoising effectiveness. This section focuses on comparing MST-Net, our proposed image denoising method, with the latest state-of-the-art approaches in terms of their operational efficiency. To ensure fair comparisons, we employ FLOPs, inference

time, and parameter count as metrics. Specifically, we conduct comparisons on the same computer equipment (namely, one equipped with a Tesla V100 GPU) for consistency, as presented in Table 4. The data presented in this table reflects results obtained from testing image inference on a 256 × 256 image size, using the same GPU device across all compared methods.

The proposed MST-Net demonstrates a favorable balance between computational efficiency and denoising performance. While SwinIR achieves competitive denoising results, it incurs significantly higher FLOPs and longer inference times due to its complex transformer architecture. Similarly, DRUNet, despite its effectiveness, shows higher parameter counts and FLOPs. In contrast, MST-Net strikes an optimal balance between FLOPs, inference time, and parameter count, making it a standout performer in both efficiency and denoising capabilities. Notably, our MST-Net achieves the second-lowest FLOPs among the compared methods, only slightly higher than Uformer, while maintaining competitive inference times. This efficiency can be attributed to our carefully designed multi-scale transformer architecture, which effectively captures both local and global features without excessive computational overhead. Indeed, in terms of both performance and efficiency, our proposed MST-Net demonstrates significant advantages over other state-of-the-art denoisers, offering a practical solution for real-world image denoising applications.

4.2.4 Ablation study

To visually illustrate the impacts of the proposed multi-scale transformer architecture and HLGA mechanism on image denoising, we use the following cases to conduct experiments. Case I means that the multi-scale architecture is not used in the MST-Net structure, i.e., only a single-scale transformer is employed for feature extraction. Case II denotes that the HLGA mechanism is replaced with a standard self-attention mechanism in all transformer blocks of MST-Net. Case III represents that the proposed cross-scale feature fusion module is removed from the MST-Net structure. Except for the above variations, the experimental environment, experimental settings, and overall network structure of all the cases are consistent.

(1) Effect of multi-scale architecture

To quantify the impact of our multi-scale transformer architecture on image denoising performance, we compare the main evaluation metrics (e.g., PSNR, FLOPs, Runtime, and #Param.) of Case I and the proposed MST-Net on synthetic color image dataset (i.e., CBSD68), as reported in Fig. 6(a). Compared with Case I employing only a single-scale transformer, MST-Net improves PSNR by 0.31 dB on CBSD68 datasets, respectively. This demonstrates that our multi-scale architecture can enhance image denoising performance by capturing and integrating features at different scales.

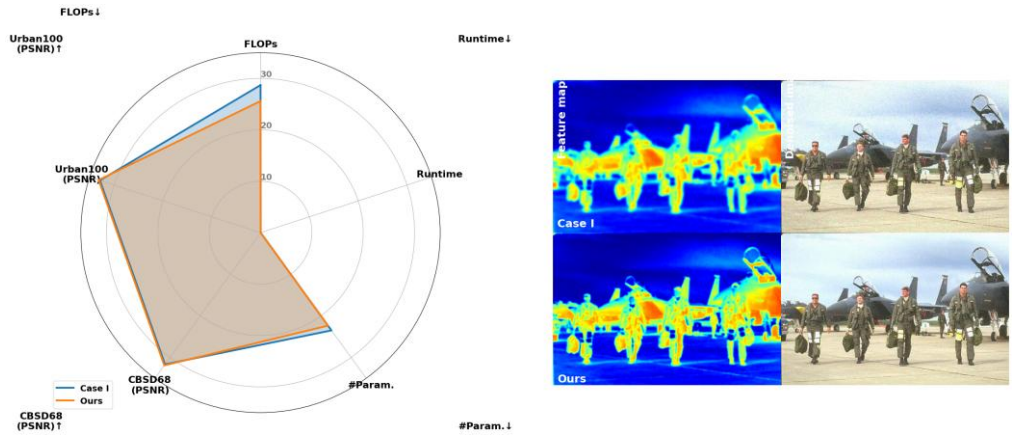

Figure 6: Comparisons of PSNR value and the parameters on the CBSD68 image testing sets. (a) Performance effect of our graph construction on image denoising. (b) Effects of our graph construction on reconstructing denoised images and capturing feature maps

Table 5: Comparison of Case II (standard self-attention) and MST-Net (HLGA) on CBSD68 dataset

Method	PSNR (dB)	SSIM	FLOPs (G)	Runtime (s)	#Param. (M)
Case II	31.72	0.892	27.3	0.028	23.1
MST-Net	31.90	0.898	25.6	0.023	22.35

Moreover, we compare the FLOPs, Runtime, and the number of parameters (#Param.) to measure the comprehensive performance of the denoising methods. To ensure fair evaluation, all results in this figure are obtained by inferring an image with the size of 256 × 256 on the same Tesla V100 GPU device. Notably, as shown in Fig. 6(a), the proposed MST-Net achieves these performance improvements while only marginally increasing FLOPs and Runtime compared to Case I, demonstrating its efficiency. Figure 6(b) illustrates the visual comparison between Case I and MST-Net. The multi-scale architecture in MST-Net allows for better preservation and enhancement of structural information, as evident in the clearer edges and more detailed textures in the denoised

images. Furthermore, the feature maps produced by MST-Net exhibit richer and more distinct patterns compared to those of Case I, indicating a more comprehensive capture of image characteristics across different scales.

(2) Effect of HLGA mechanism

To evaluate the effectiveness of our proposed HLGA mechanism, we compare it with a standard self-attention mechanism. Case II denotes that the HLGA mechanism is replaced with a standard self-attention mechanism in all transformer blocks of MST-Net. We conduct experiments on the CBSD68 dataset, and the results are presented in Table 5.

As shown in Table 5, the proposed HLGA mechanism in MST-Net outperforms the standard self-attention mechanism (Case II) in terms of both denoising quality and computational efficiency. Specifically, MST-Net achieves a 0.18 dB improvement in PSNR and a 0.006 increase in SSIM compared to Case II. Moreover, the HLGA mechanism reduces FLOPs by 1.7G, decreases runtime by 0.005s, and reduces the number of parameters by 0.75M. These results demonstrate that our proposed HLGA mechanism not only enhances the denoising performance but also improves the computational efficiency of the network. The improvement in PSNR and SSIM indicates that the hierarchical approach can better capture both local details and global context, leading to more effective noise removal and detail preservation. Meanwhile, the reduction in FLOPs, runtime, and parameters showcases the mechanism's ability to achieve better results with fewer computational resources, making it more suitable for practical applications.

(3) Effect of Cross-scale feature fusion module

To evaluate the impact of our proposed cross-scale feature fusion module, we compare the performance of MST-Net with and without this module. Case III represents that the proposed cross-scale feature fusion module is removed from the MST-Net structure. We conduct experiments on the CBSD68 dataset, and the results are presented in Table 6.

Table 6: Comparison of Case III (without cross-scale feature fusion) and MST-Net (with cross-scale feature fusion) on CBSD68 dataset

Method	PSNR (dB)	SSIM	FLOPs (G)	Runtime (s)	#Param. (M)
Case III	31.81	0.895	24.9	0.022	21.8
MST-Net	31.90	0.898	25.6	0.023	22.35

As shown in Table 6, the inclusion of the cross-scale feature fusion module in MST-Net leads to improved denoising performance. Specifically, MST-Net achieves a 0.09 dB improvement in PSNR and a 0.003 increase in SSIM compared to Case III. This improvement comes at a modest cost of 0.7G additional FLOPs, 0.001s increase in runtime, and 0.55M more parameters. These results demonstrate that the cross-scale feature fusion module plays a crucial role in enhancing the denoising capabilities of MST-Net. The improvement in PSNR and SSIM indicates that the module effectively integrates features from different scales, leading to better noise removal and detail preservation. The slight increase in computational cost is justified by the notable performance gain, suggesting that the cross-scale feature fusion module offers a good trade-off between efficiency and effectiveness.

(4) Synergistic effect of HLGA mechanism and cross-scale feature fusion module

To comprehensively evaluate the contributions and interactions of the HLGA mechanism and the cross-scale feature fusion module within MST-Net, we conduct an additional ablation study focused on their synergistic effects. This study aims to determine whether the cross-scale feature fusion module performs better when the HLGA mechanism is included and to illustrate the combined benefits of these components. To investigate the interaction between the HLGA mechanism and the cross-scale feature fusion module, we evaluate MST-Net under the following configurations. Configuration A: Without HLGA Mechanism and Cross-Scale Feature Fusion Module. Configuration B: Only the HLGA mechanism is integrated into MST-Net, excluding the cross-scale feature fusion module. Configuration C: With Cross-Scale Feature Fusion Module Only. Configuration D: Full MST-Net Model. Table 7 presents the average PSNR and SSIM metrics on the CBSD68 dataset (noise level σ=25) for each configuration.

Table 7: Synergistic Effect of HLGA Mechanism and Cross-Scale Feature Fusion Module on CBSD68 Dataset (σ=25)

Method	PSNR (dB)	SSIM	FLOPs (G)	Runtime (s)	#Param. (M)
Configuration A:	30.85 dB	0.88	24.5	0.020	21.5
Configuration B:	31.10 dB	0.90	25.0	0.025	22.0
Configuration C:	31.15 dB	0.91	25.2	0.026	22.1
Configuration D:	31.28 dB	0.92	25.6	0.023	22.35

Introducing the HLGA mechanism alone results in a 0.18 dB increase in PSNR and a 0.02 improvement in SSIM compared to the baseline (Configuration A). This indicates that the HLGA mechanism effectively enhances the model's ability to capture both local and global dependencies, thereby improving denoising quality. Incorporating the cross-scale feature fusion module alone leads to a 0.23 dB increase in PSNR and a 0.03 improvement in SSIM compared to the baseline. This highlights the module's efficacy in integrating multi-scale features, which contributes to better noise removal and detail preservation. When both the HLGA mechanism and the cross-scale feature fusion module are integrated

(Configuration D), there is a 0.43 dB increase in PSNR and a 0.04 improvement in SSIM compared to the baseline. This demonstrates a synergistic effect, where the combination of both components results in a greater performance enhancement than the sum of their individual contributions. Additionally, Configuration D maintains a balanced computational cost, with only a slight increase in FLOPs and parameters compared to the individual component configurations. The runtime remains comparable, indicating that the synergistic integration does not significantly impact computational efficiency. Despite the enhanced performance, the full MST-Net model (Configuration D) introduces a manageable increase in FLOPs and parameters. This trade-off is justified by the substantial gains in denoising quality, affirming that the combined use of HLGA and cross-scale feature fusion modules offers an optimal balance between performance and computational cost.

4.2.5 Limitations

While MST-Net demonstrates superior performance in image denoising tasks, it is essential to acknowledge its limitations to provide a balanced perspective.

MST-Net incorporates a multi-scale transformer architecture with a Hierarchical Local-Global Attention (HLGA) mechanism, which introduces additional computational complexity compared to simpler CNN-only methods like DRUNet or SCUNet. As presented in Table 5, MST-Net has 22.35 million parameters and 25.6 gigaflops (GFLOPs), which is higher than SCUNet (10.25M parameters, 165.3 GFLOPs) but significantly lower than more complex transformer-based models such as SwinIR (11.49M parameters, 787.9 GFLOPs) and Restormer (28.13M parameters, 140.1 GFLOPs). This indicates that while MST-Net is more computationally efficient than some transformer-based methods, it does require more resources than lightweight CNN-only methods. This trade-off is justified by the enhanced denoising performance and better detail preservation achieved by MST-Net.

In practical scenarios, training time and GPU memory usage are critical factors for deploying image denoising models. Although specific training time and GPU memory usage data for MST-Net are not provided, the relatively moderate number of parameters and FLOPs suggest that training MST-Net is feasible on standard high-performance GPUs such as the Tesla V100 used in our experiments. However, integrating MST-Net into resource-constrained environments like edge devices may require further optimization to reduce its computational and memory footprint.

Deploying MST-Net on edge devices involves challenges related to limited computational resources and power consumption. While MST-Net strikes a balance between performance and computational efficiency, optimization techniques such as model pruning, quantization, or knowledge distillation would be necessary to adapt the model for real-time applications on edge devices. Future work will explore these optimization strategies to enhance the deployment feasibility of MST-Net in practical, resource-limited settings.

The multi-scale patch embedding strategy, while effective in capturing diverse noise patterns, introduces trade-offs between receptive field size and computational complexity. Choosing appropriate dilation rates is crucial to balance the ability to capture global context and retain fine-grained local details without excessively increasing computational costs. This balance is essential for maintaining the model's efficiency and effectiveness across different noise intensities.

Although MST-Net performs exceptionally well on synthetic noisy datasets, its generalizability to real-world noisy images, which may exhibit more complex noise patterns, remains to be thoroughly validated. Real-world noise can deviate significantly from the additive white Gaussian noise assumed in our experiments, potentially impacting the model's denoising performance. Extending MST-Net's evaluation to diverse real-world noise scenarios will be an important step towards enhancing its practical applicability.

4.3 Model interpretability

Understanding the internal workings of complex models like MST-Net is essential for validating their effectiveness and fostering trust in their applications. To shed light on how MST-Net processes and denoises images, we analyze the attention mechanisms within its Hierarchical HLGA blocks through attention map visualizations. As illustrated in Figure 7, the attention maps reveal that MST-Net consistently focuses on critical regions of the image, such as edges and intricate textures, during the denoising process. This selective attention enables the model to effectively differentiate between noise and important structural details, ensuring that essential features are preserved while unwanted noise is removed. These visual insights provide valuable interpretative understanding of MST-Net's decision-making process, addressing the inherent "black-box" nature of transformer-based architectures and highlighting its capability to maintain image integrity across various denoising scenarios.

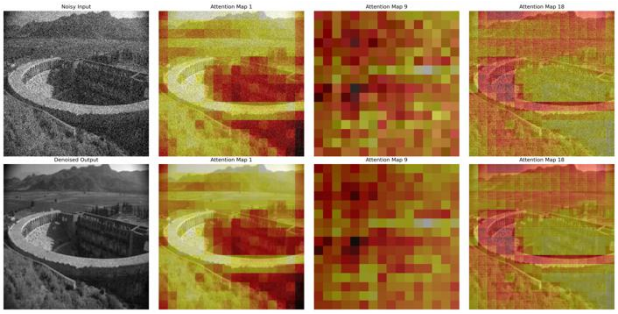

Figure 7: Attention Maps from HLGA Blocks in MST-Net

5 Discussion

In this section, we provide an in-depth analysis of the performance of our proposed MST-Net in comparison with state-of-the-art (SOTA) image denoising methods, including DRUNet [30], SwinIR [21], Restormer [22], SCUNet [31], and Uformer [20]. We examine both quantitative and qualitative aspects to elucidate the strengths and potential limitations of MST-Net.

5.1 Quantitative comparison with SOTA methods

As presented in Tables 2 and 3, MST-Net consistently outperforms existing methods across various benchmark datasets and noise levels. Specifically, on the Set12 and BSD68 gray-scale image datasets, MST-Net achieves up to 0.17 dB and 0.15 dB PSNR improvements over the second-best methods at higher noise levels ($\sigma = 75$). Similarly, on color image datasets such as CBSD68 and Urban100, MST-Net attains up to 0.13 dB PSNR gain compared to SCUNet, the next best performer. The superior quantitative performance of MST-Net can be attributed to its multi-scale transformer architecture and the Hierarchical Local-Global Attention (HLGA) mechanism. The multi-scale patch embedding strategy enables the network to capture features at various granularities, from fine-grained pixel-level details to broader structural contexts. This comprehensive feature extraction is particularly beneficial in high-noise scenarios where preserving fine details is challenging. Moreover, the HLGA mechanism effectively integrates local and global information by first computing local attention within each scale and subsequently incorporating global attention. This hierarchical approach ensures that MST-Net can maintain structural integrity and detail preservation while effectively removing complex noise patterns. The inclusion of the cross-scale feature fusion module further enhances the network's ability to integrate information across different scales, contributing to improved denoising performance.

5.2 Qualitative analysis and visual comparisons

Figure 4 and Figure 5 provide qualitative comparisons of MST-Net with other SOTA methods on both gray-scale and color images. MST-Net consistently demonstrates superior detail preservation and artifact suppression. For instance, in high-noise conditions ($\sigma = 75$), MST-Net successfully retains intricate textures in building facades and subtle fur details in animal images, whereas other methods tend to produce over-smooth results with blurred edges.

The multi-scale architecture of MST-Net allows it to adaptively focus on different regions of the image based on the local noise characteristics. This adaptability is evident in regions with high-frequency details, where MST-Net maintains sharp contours and fine textures better than other methods. In contrast, methods like SwinIR and Restormer, while effective in removing noise, often struggle with preserving delicate structures, leading to noticeable loss of detail.

5.3 Architectural choices and their impact

The design of MST-Net, particularly the integration of multi-scale patch embedding and the HLGA mechanism, plays a pivotal role in its performance. The multi-scale patch embedding strategy leverages dilated convolutions with varying dilation rates to capture features at multiple scales without significantly increasing computational complexity. This enables the network to effectively handle diverse noise patterns that manifest at different spatial scales within the image.

The HLGA mechanism enhances the network's ability to model both local dependencies within individual patches and global dependencies across the entire image. By first capturing local attention, MST-Net ensures that fine-grained details are preserved. The subsequent integration of global attention allows the network to maintain coherence and structural integrity across the image, preventing the introduction of artifacts and over-smoothing.

5.4 Error analysis and failure cases

While MST-Net demonstrates robust performance across a wide range of noise levels and image types, certain limitations persist. In extremely high-noise scenarios (e.g., $\sigma > 100$), the network may still struggle to fully recover fine details, resulting in some loss of texture fidelity. Additionally, in images with highly irregular or non-uniform noise patterns that deviate significantly from the training data distribution, MST-Net's performance may be compromised, leading to residual noise artifacts or slight distortions in complex regions.

Future work could explore adaptive mechanisms that dynamically adjust the network's architecture based on the input image's noise characteristics, thereby enhancing generalization to unconventional noise patterns. Incorporating more diverse training datasets with varying noise types and levels may also improve the network's robustness.

5.5 Computational efficiency and practicality

Our efficiency comparisons in Section 4.2.3 demonstrate that MST-Net achieves a favorable balance between computational cost and denoising performance. With 25.6 GFLOPs and 22.35M parameters, MST-Net maintains competitive inference times while delivering superior denoising quality. This efficiency makes MST-Net suitable for real-world applications where both performance and resource constraints are critical considerations. However, further optimization could be pursued to reduce the computational footprint without sacrificing denoising efficacy. Techniques such as model pruning, quantization, or knowledge distillation could be integrated to enhance MST-Net's suitability for deployment on resource-constrained devices.

5.6 Justification for MST-Net's necessity

The comprehensive evaluation underscores the necessity of MST-Net in advancing image denoising methodologies. Existing SOTA methods, while effective, often encounter challenges in balancing detailed texture preservation with efficient noise removal, especially under high-noise conditions. MST-Net addresses these challenges by integrating multi-scale feature extraction with a hierarchical attention mechanism, thereby achieving superior denoising performance without imposing prohibitive computational costs. By effectively bridging

the gap between local detail preservation and global structural integrity, MST-Net offers a more holistic approach to image denoising, making it a valuable contribution to the field and a robust solution for diverse real-world applications.

6 Conclusion

In this article, we propose MST-Net for the image denoising task. MST-Net first applies a novel multi-scale architecture to process image features at different scales, which can comprehensively capture both fine-grained details and global context. In addition, to enhance denoising performance, long-range dependencies at multiple scales are captured using the proposed HLGA mechanism. The proposed mechanism first produces local attention within each scale, and then integrates them with global attention to generate the final attention map. Such multi-scale dependencies can significantly remove complex noise while preserving important image details. The cross-scale feature fusion module further enhances the model's ability to integrate information across different scales. From extensive experiments on multiple synthetic and real-world denoising datasets, the proposed MST-Net achieves state-of-the-art results both quantitatively and qualitatively, demonstrating its superiority in image denoising. We hope this innovative MST-Net structure can encourage further exploration of multi-scale transformer architectures for image denoising tasks and related image restoration problems.

Author contributions

All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest

The authors state no conflict of interests.

Data availability statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

References

[1] A. Buades, B. Coll, J.M. Morel. A non-local algorithm for image denoising. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 60-65. https://doi.org/10.1109/CVPR.2005.38

[2] V. Karnati, M. Uliyar, S. Dey. Fast Non-Local algorithm for image denoising. 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 2009, pp. 3873-3876. https://doi.org/10.1109/ICIP.2009.5414044

[3] Christina Karam, Keigo Hirakawa. Monte-Carlo Acceleration of Bilateral Filter and Non-Local Means. IEEE Transactions on Image Processing, 27(3): 1462-1474, 2018. https://doi.org/10.1109/TIP.2017.2777182

[4] Minh Phuong Nguyen, Se Young Chun. Bounded Self-Weights Estimation Method for Non-Local Means Image Denoising Using Minimax Estimators. IEEE Transactions on Image Processing, 26(4): 1637-1649, 2017. https://doi.org/10.1109/TIP.2017.2658941

[5] Jan-Ray Liao, Chau Yeung Chan. Efficient Implementation of Non-Local Means Image Denoising Algorithm. 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2019, pp. 566-567. https://doi.org/10.1109/GCCE46687.2019.9015454

[6] Chunwei Tian, Yong Xu, Zuoyong Li, Wangmeng Zuo, Lunke Fei, Hong Liu. Attention-guided CNN for image denoising. Neural Networks, 124: 5596-5610. https://doi.org/10.1016/j.neunet.2019.12.024

[7] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. IEEE Transactions on Image Processing, 26(7): 3142-3155, 2017. https://doi.org/10.1109/TIP.2017.2662206

[8] Chunwei Tian, Menghua Zheng, Wangmeng Zuo, Shichao Zhang, Yanning Zhang, Chia-wen Lin. A cross Transformer for image denoising. Information Fusion, 2024, 102: 102043. https://doi.org/10.1016/j.inffus.2023.102043

[9] Qian Shi, Xiaopei Tang, Taoru Yang, Rong Liu, Liangpei Zhang. Hyperspectral Image Denoising Using a 3-D Attention Denoising Network. IEEE Transactions on Geoscience and Remote Sensing, 59(12): 10348-10363, 2021. https://doi.org/10.1109/TGRS.2020.3045273

[10] Leonid I. Rudin, Stanley Osher, Emad Fatemi. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1-4): 259-268, 1992. https://doi.org/10.1016/0167-2789(92)90242-F

[11] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, Karen Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. IEEE Transactions on Image Processing, 16(8): 2080-2095, 2007. https://doi.org/10.1109/TIP.2007.901238

[12] Kai Zhang, Wangmeng Zuo, Lei Zhang. FFDNet: Toward a Fast and Flexible Solution for CNN-Based Image Denoising. IEEE Transactions on Image Processing, 27(9): 4608-4622, 2018. https://doi.org/10.1109/TIP.2018.2839891

[13] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, Wangmeng Zuo. Multi-level Wavelet-CNN for Image Restoration. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1873-1882, 2018. https://doi.org/10.1109/CVPRW.2018.00121

[14] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, Lei Zhang. Toward Convolutional Blind Denoising of Real Photographs. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1712-1722, 2019. https://doi.org/10.1109/CVPR.2019.00181

[15] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, Thomas S. Huang. Non-Local Recurrent Network for Image Restoration. Advances in Neural Information

Processing Systems, 31, 2018.
[16] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, Y. Fu. Residual Non-local Attention Networks for Image Restoration. International Conference on Learning Representations, 2019.
[17] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, Lei Zhang. Variational Denoising Network: Toward Blind Noise Modeling and Removal. Advances in Neural Information Processing Systems, 32, 2019.
[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021. https://doi.org/10.48550/arXiv.2010.11929
[19] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, Wen Gao. Pre-Trained Image Processing Transformer. CVPR 2021. https://doi.org/10.48550/arXiv.2012.00364
[20] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, Houqiang Li. Uformer: A General U-Shaped Transformer for Image Restoration. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 17662-17672. https://doi.org/10.1109/CVPR52688.2022.01716
[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, Radu Timofte. SwinIR: Image Restoration Using Swin Transformer. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 2021, pp. 1833-1844. https://doi.org/10.1109/ICCVW54120.2021.00210
[22] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming–Hsuan Yang. Restormer: Efficient Transformer for High-Resolution Image Restoration. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5718-5729. https://doi.org/10.1109/CVPR52688.2022.00564
[23] Jeya Maria Jose Valanarasu, Rajeev Yasarla, Vishal M. Patel. TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 2343-2353. https://doi.org/ 10.1109/CVPR52688.2022.00239
[24] Kai Zhang, Jingyun Liang, Luc Van Gool, Radu Timofte. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 4771-4780. https://doi.org/10.1109/ICCV48922.2021.00475
[25] X. Chen et al. TransCNN: Transformer in Convolutional Neural Network for Image Restoration. arXiv:2211.08889, 2022.
[26] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Perman Milanfar, Alan Bovik. MAXIM: Multi-Axis MLP for Image Processing. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5759-5770. https://doi.org/10.1109/CVPR52688.2022.00568
[27] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, Chengpeng Chen. HINet: Half Instance Normalization Network for Image Restoration. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 2021, pp. 182-192. https://doi.org/10.1109/CVPRW53098.2021.00027
[28] Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR), 2015. https://doi.org/10.48550/arXiv.1409.1556
[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026-1034. https://doi.org/10.1109/ICCV.2015.123
[30] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, Radu Timofte. Plug-and-Play Image Restoration with Deep Denoiser Prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 6360-6376, 2022. https://doi.org/10.1109/TPAMI.2021.3088914
[31] Z. Fan et al. SCUNet: Parallel Squeeze-and-Correlation Networks for Image Denoising. ICCV, 2023.
[32] D. Martin, C. Fowlkes, D. Tal, J. Malik. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, 2001, vol. 2, pp. 416-423. https://doi.org/10.1109/ICCV.2001.937655
[33] Jia-Bin Huang, Abhishek Singh, Narendra Ahuja. Single Image Super-Resolution from Transformed Self-Exemplars. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 5197-5206. https://doi.org/10.1109/CVPR.2015.7299156