IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 February 2024, accepted 31 March 2024, date of publication 4 April 2024, date of current version 11 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3384577

RESEARCH ARTICLE

Self-Supervised Real-World Image Denoising Based on Multi-Scale Feature Enhancement and Attention Fusion

HAILIANG TANG[1], WENXIAO ZHANG[2], HAILIN ZHU[1], AND KE ZHAO[1]

[1]School of Information Science and Engineering, Qilu Normal University, Jinan 250200, China

[2]School of Finance and Economics, Shandong University of Engineering and Vocational Technology, Jinan 250200, China

Corresponding author: Hailiang Tang (20170333@qlnu.edu.cn)

ABSTRACT Deep learning denoising methods are often constrained by the high cost of acquiring real-world noisy images and the labor-intensive process of dataset construction. Our self-supervised Multi-Scale Blind-Spot Network with Adaptive Feature Fusion (MA-BSN) addresses these issues, offering an efficient solution for image denoising. MA-BSN mitigates the challenges of spatial noise correlation preservation and limited receptive fields, which are prevalent in existing self-supervised denoising approaches. The network employs a blind-spot architecture that generates sub-images at multiple scales, enhancing denoising beyond the capabilities of pixel-shuffle downsampling. A depth-wise convolutional Transformer network (DTN) extracts features across a global receptive field, addressing the convolutional neural networks' (CNNs) limitations. An adaptive feature fusion module (AFF) is introduced to refine feature learning for specific regions in the denoised images, leveraging attention mechanisms for improved performance. Our network's efficacy is validated through experiments on the SIDD and DND real-world noise benchmark datasets. Results on the DND dataset show a PSNR/SSIM of 38.41 dB/0.940, surpassing state-of-the-art self-supervised methods and underscoring our approach's superior denoising capability.

INDEX TERMS Self-supervised image denoising, multi-scale feature learning, attention mechanism.

I. INTRODUCTION

Various uncontrollable factors can degrade image quality and interfere with visual perception in image acquisition and transmission. Therefore, image denoising has become a widely adopted task in image processing. It typically serves as a preliminary operation for enhancing image clarity, accuracy, and overall success in subsequent tasks.

Early studies often assumed that noise is independent and identically distributed. Additive Gaussian White Noise (AWGN) is commonly employed to simulate noisy images. Traditional algorithms, such as filtering techniques and sparse learning, have been widely utilized to address these tasks, considering the inherent properties of natural images, such as self-similarity and low rank. NLM [1] utilizes the weighted average of all pixels within a search window in an image to achieve noise removal. BM3D [2] improves the sparse representation by iterative filtering. The learned simultaneous sparse coding (LSSC) [3] introduces the idea of nonlocal averaging and sparse coding. The weighted nuclear norm minimization (WNNM) [4] employs a low-rank approach and a priori knowledge to improve the denoising performance. MCWNNM [5] extends this idea to the field of multichannel image denoising. LRA-SVD [6] uses singular value truncation to achieve efficient noise filtering.

Although traditional denoising methods can achieve satisfactory performance, many algorithms are computationally intensive and time-consuming due to multiple iterations. Moreover, the non-convex nature of these algorithms makes it challenging to find the optimal solution. Additionally, these models often require manual adjustment of numerous parameters to achieve desirable denoising results, introducing uncertainty in their performance. Deep neural networks have been employed for image-denoising tasks to enhance the capability of learning intrinsic image features. DnCNN [8] efficiently learns noisy features by learning clean

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

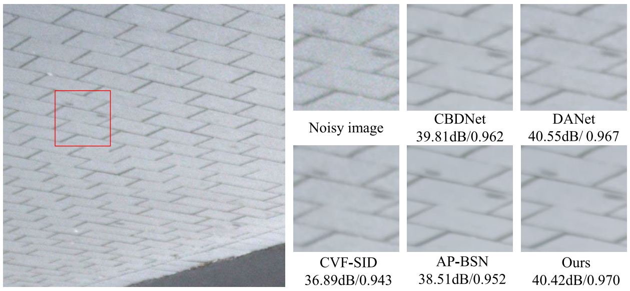


FIGURE 1. **Visual comparison of various methods on the DND benchmark dataset [50]. Compared with CBDNet [7], DANet [11], CVF-SID [60], and AP-BSN [31], Our method only requires a single noisy image as input and does not require any additional data. DANet exhibits poor noise removal performance at image edges. AP-BSN still contains noise in the output.**

image-noise image pairs in supervised batch. However, due to the fitting properties of deep neural networks, this algorithm can only solve the AWGN denoising task for a given noise level, and it is difficult to generalize to other noise levels or other types of noise. The convolutional blind denoising network (CBDNet) [7] adapts itself to images with different noise levels. VDN [9] uses a variational a posteriori optimization network with good generalization ability. DCANet [10] proposes a dual convolutional blind denoising network, which can remove a wide range of Gaussian noises. DudeNet [11] proposes a dual denoising network, which uses the image sparse properties to learn global and local features of the image.

Researchers have recently found that real-world noise is more complex than synthetic noise. Noise behaves as content-dependent rather than independent, and these forms are spatially variable. Therefore, recent work has focused on real-world noisy images. In traditional methods, NLH [12] proposes pixel-level Wiener filtering algorithms that utilize an image prior to achieving denoising of real-world noise. TWSC [13] utilizes a trilaterally weighted sparse coding scheme that describes the real-world noise and the image prior by introducing a weighting matrix. In the field of deep learning, numerous algorithms with excellent results have been proposed [14], [15], [16], [17], [19], [20]. These methods utilize the powerful inductive bias capability of deep neural networks to thoroughly learn the potential features of noisy images in the dataset and fit ground-truth images to improve the denoising performance.

Although the aforementioned deep learning algorithms have demonstrated effective denoising capabilities, they typically rely on a supervised approach for model training, which entails using clean image pairs with synthetic or real-world noise. However, image pairs synthesized using content-independent Gaussian white noise or content-dependent Poisson noise may not accurately represent the distribution of real-world noise. Consequently, denoising models that perform well on synthetic noise tasks may need to be more competent in handling real noise scenarios. The high cost of acquiring real noise image pairs limits the training and deployment of models. In this case, self-supervised denoising algorithms [21], [22], [24], [25], [26], [27], who require only noisy images to train the model, do not rely on noisy-clean image pairs. The pioneering work DIP [25] was the first attempt to leave the training set behind while not requiring the original image to be labeled. The whole process requires only a single noisy image. Although its denoising effect is worse than supervised deep learning denoising algorithms in the same period, its innovative ideas have inspired many subsequent works. Noise2Noise [21] uses pairs of noise for training, but it is difficult to obtain different pairs of noise under the same scene due to factors such as illumination. Like adding masks in MAE [28], Bernoulli sampling is introduced to Self2Self [27] to improve the denoising level. Noise2Void [24] proposes Blind-Spot Network (BSN), a practical self-supervised learning module.

The blind-spot strategy avoids constant mapping by learning to predict artificially missing pixels using neighboring pixels. Thus, denoising neural networks can only be trained with noisy images. References [29] and [30] assume that noise has content relevance and spatial independence, which verifies the feasibility of image denoising by sampling noisy images and training them jointly. Moreover, real-world noise has spatial continuity; when the noise level is high, the BSN-based denoising methods [24], [30], [31], [32] are prone to noise interference in the process of denoising with neighboring pixels. AP-BSN [31] aims to reduce the spatial correlation of noise by performing five times pixel-shuffle downsampling on the input before training, which involves using a center-masked convolutional kernel and a dilated convolutional layer to mitigate blind-spot issues during forward propagation. Due to the effect of sampling on high-frequency features, the sampling step size requirement is strict; if the step size of PD is too large, the spatial information of high frequency will be lost, and if it is too small, it is difficult to reduce the spatial correlation of noise. The above BSN denoising method adopts the CNN network for feature extraction. However, CNN's local feature extraction ability is excellent, its receptive field limits it, and its denoising ability is limited when the noise level or spatial correlation is high.

Inspired by the supervised image denoising method [17], this paper introduces the Transformer network with strong global feature extraction ability, combines it with CNN, and learns the noisy image features from both the local and global aspects so that the image feature extraction ability can be further enhanced. AP-BSN [31] adopts a sampled feature map with a step size of 5 for training, and it is easy to filter out the high-frequency image information. Therefore, this paper introduces a pyramid network for multi-scale feature extraction and an attention-based adaptive fusion network used to enhance the feature extraction capability of the network while effectively reducing the loss of image information.

Enhancing images in low-light conditions has consistently presented a challenge due to these images often being affected by additive noise and reduced contrast. In recent years, a variety of methods have been proposed to improve the quality of images captured under these conditions. Notably, Oh and Hong introduced a Retinex model based on a nonlinear mapping function in their study, which effectively enhanced the dynamic range and color restoration of images [62]. They then further explored the application of mixed deep learning techniques and a hybrid norm loss function in low-light image enhancement [63]. Moreover, the research by Duong and Hong [64] introduced EBSD-Net, a deep network specifically designed for low-light color images, aimed at enhancing brightness and suppressing image degradation. In processing images with limited dynamic range, MA-BSN can be integrated with the Retinex model to enhance the overall brightness of the image while preserving the details in shadows and highlights. Particularly in scenarios of color distortion caused by low-light conditions, the combination with the Retinex model facilitates more accurate color correction.

The main contributions of our work are outlined as follows:

1) We introduce the concept of multi-scale feature extraction into the blind-spot network, which effectively mitigates the degradation of high-frequency information in images caused by pixel-shuffle downsampling. This is the first attempt to incorporate such a concept in this context.
2) We employ a depth-wise-based Transformer to jointly learn image features across channels by integrating the sampled sub-images. This approach enhances capturing and utilizing contextual information for denoising tasks.
3) We propose an adaptive feature fusion module based on attention mechanisms, which enables image stitching to focus more on the content of the images. This module improves the overall denoising performance by effectively integrating relevant information from different parts of the image.
4) The proposed MA-BSN achieves denoising tasks in a self-supervised manner, requiring only a single noisy image. It demonstrates advanced performance on real-world noise image benchmark datasets such as SIDD and DND, showcasing its effectiveness in practical scenarios.

II. RELATED WORK

A. UNSUPERVISED IMAGE DENOISING

Unsupervised image denoising is briefly introduced in the introduction to provide an overview of the methodology employed in this paper. In recent years, significant efforts have been made to address the challenges in supervised image denoising. These efforts can be broadly categorized into three general directions, as outlined below.

1) GENERATE SYNTHETIC NOISE-CLEAN IMAGE PAIRS

As external factors change in real-time, real-world noise-clean image pairs are difficult to obtain accurately. Synthetic real-world noise-clean image pairs are extended by noise domain adaptation using the generative method ADGAN [36]. UIDNet [30] utilizes a self-supervised denoising network for further feature extraction. C2N [38] considers real-world scenes with various noise generation factors to achieve more accurate noise synthesis.

2) GENERATE PSEUDO-NOISE IMAGES

Although the real noise-clean image pairs are difficult to obtain, using the zero-mean property of the noise distribution to obtain relatively independent noise images, and constructing real noise-true noise image pairs for training, can also achieve good performance. noise2noise uses multiple noise images for training without introducing the ground truth image. OCT-NGAN [37] trains the discriminator to distinguish between actual noise samples and pseudo-noise samples generated by the denoiser in the absence of clean images, and the discriminator guides the generator to denoise. Neighbor2Neighbor [26] constructs noisy image pairs by obtaining sub-images from sampling random neighborhoods. Self2self [27] utilizes Bernoulli sampling to obtain noisy image pairs and achieves good performance through iterative denoising, but a single noisy image often consumes several hours and is computationally inefficient. NAC [39] treats the noisy image used for denoising as ground truth and synthesizes it with another similar noisy image for training. A similar noisy image is synthesized with another corresponding image used for training.

3) TRAINING ON NOISY IMAGES ONLY

As in the above two points, There is also a self-supervision approach for training directly on a single noisy image without constructing image pairs for training. Noise2Void [24] proposes novel self-supervised Blind-Spot Networks (BSNs), which utilize a blind-spot strategy to avoid network identity mapping, forcing the network to learn features from pixel neighborhoods. Noise2Self [22] utilizes the independence of inter-pixel noise to recover the image. Laine19 [29] achieves the effect of a "blind spot" at the center of the receptive field by masking the receptive field in different directions. D-BSN [30] uses center blind-spot convolution and dilated convolution layer (DCL) to construct the BSN. The approach taken here involves further filtering out the noise, which is key to destroying the spatial correlation. AP-BSN [31] performs 5-pixel stepwise pixel-shuffle downsampling on the image during training and uses a blind-spot convolution kernel with a center mask to independently recover the image. Mask's blind convolutional kernel and dilated convolutional layer (DCL) to achieve feature learning. However, AP-BSN employs a large PD step size to reduce noise correlation, which results in some high-frequency information being filtered out and reduces the image recovery effect. The CNN-based network architecture also limits the learning of global image features.

B. SUPERVISED IMAGE DENOISING

The significance of self-supervised learning in image denoising lies in its ability to reduce data annotation

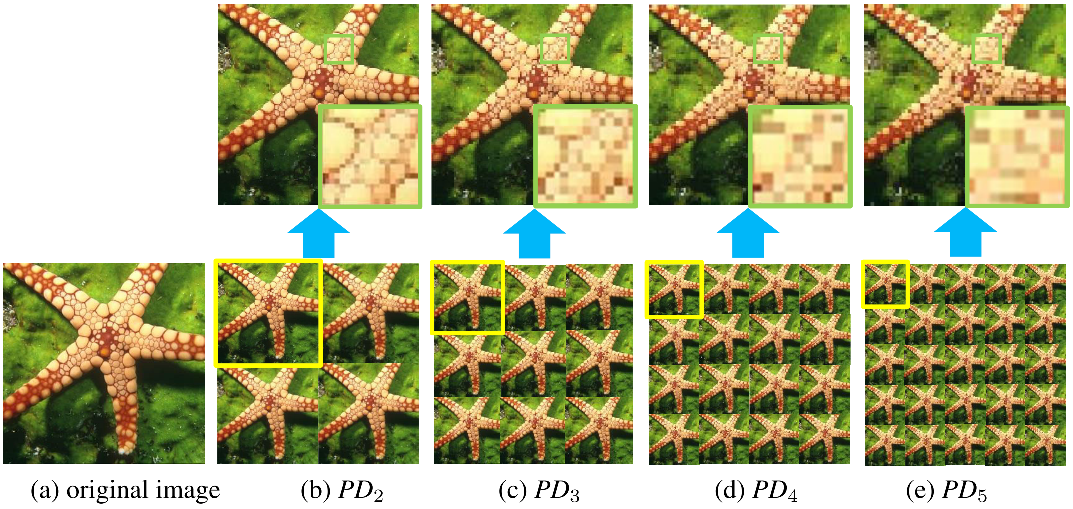


FIGURE 2. Comparison of sub-images obtained using pixel-shuffle downsampling with different strides. The subscript indicates the stride value. When the stride is 2, the image features are relatively intact. However, as the stride increases, the degree of damage to the image edges gradually increases. The figure shows that when the stride is 5, the spatial correlation of the texture on the hat is severely disrupted.

costs, enhance model generalization, adapt to various noise environments, minimize human intervention, and be adaptive to diverse tasks. This renders self-supervised methods an economically efficient, versatile, and highly adaptive solution for image denoising, with the potential to deliver outstanding performance across a range of noisy scenarios.

Like the development process of self-supervised image denoising tasks, early supervised image denoising tasks are mainly based on CNN networks [16], [33], [34], [35], which utilize the excellent feature learning ability of CNNs for local regions to improve denoising performance.

Although convolutional neural networks have achieved advanced performance, there are areas for improvement, such as the convolutional kernel being trained to form a fixed weight matrix, which cannot adapt to the input of different image contents. Its scale is often small, which makes it challenging to extract feature information in a large receptive field, and if downsampling is used to expand the receptive field, the detailed information will be lost.

The self-attention mechanism has been introduced into the field of image processing [41], [42], [43] to make up for the shortcomings of the CNN module. The self-attention mechanism uses learning to extract features from the global region of an image. It computes weights for different regions to increase the importance of the regions with larger weights so that they can play a more significant role in the learning process.

Although the receptive field of the Transformer based on attention mechanisms is more significant than that of CNN, and more effective information can be extracted, its computational cost is relatively higher, and its computational complexity grows quadratically with the increase of the receptive field, so it is more demanding on the operating equipment. The use of the scenarios is limited, and it only applies to high-computing-power equipment and low-resolution images. The image denoising task studied in this paper differs from the advanced computer vision tasks such as image classification and semantic segmentation that focus on the image as a whole, focusing more on the pixel-level scene and requiring higher image resolution. Hence, the above Transformer variant does not apply to image denoising. There are relatively few existing Transformer variants for image denoising [17], [47], and in order to reduce the computational cost, these methods decompose the whole image into non-overlapping image blocks and then perform self-attention operations on these blocks. SUNet [18] leverages the local attention mechanism of the Swin Transformer to effectively handle details and textures in images, demonstrating exceptional denoising performance under complex noise conditions. SwinIR [46], with its innovative multi-scale processing strategy and effective feature fusion technique, achieves significant results in various image restoration tasks, including denoising. Supervised deep learning denoising methods have been developed over the years and can learn features from images better than unsupervised denoising methods that started relatively late.

III. METHODS

A. MOTIVATION AND MODELING

This section begins by introducing the motivation behind our work. Subsequently, the architecture of MA-BSN is described, as shown in Figure 3.

Although current algorithms perform well in simple synthetic noise removal, self-supervised denoising methods have strong spatial noise correlation due to exploiting pixel spatial neighborhood features, leading to significant performance degradation when dealing with real-world noise, which

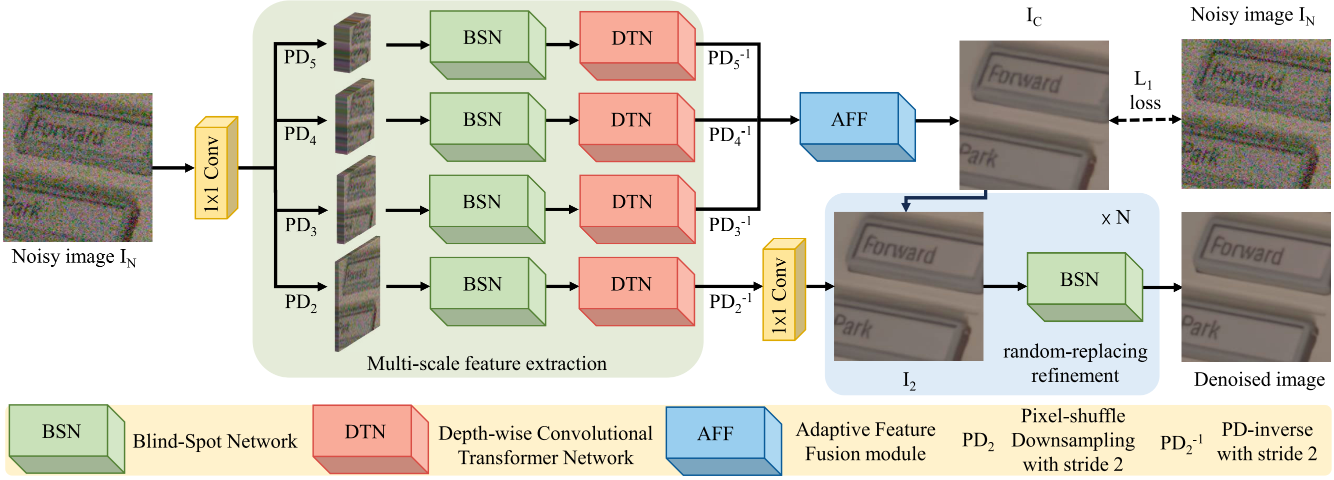


FIGURE 3. MA-BSN architecture. Downsampling the noisy image into multi-scale feature maps enhances the self-supervised learning capability. The transformer module increases the receptive field of the network, further improving the denoising performance. The Adaptive Feature Fusion (AFF) module enhances attention to the image by combining denoised images from multiple branches.

means that the assumption on which most state-of-the-art methods are based, i.e., the pixel-level spatial independence of the noise, is not justified. These methods assume that the clean signal of each pixel depends on the neighborhood, while the noise is independent. Therefore, in real scenarios, they inevitably misinterpret spatially correlated noise as a clean signal, making it difficult to recover a clean image efficiently. Therefore, considering the spatial correlation of noise is necessary for self-supervised real-world noise removal.

We employ pixel-shuffle downsampling (PD) to disrupt the correlation of authentic noise. Compared to other downsampling methods, such as bicubic interpolation and max pooling, PD maintains the integrity of image content by rearranging pixels, particularly preserving texture and edge information. Moreover, this technique is reversible, allowing the network to restore to the original resolution losslessly in later stages, which is advantageous for precise reconstruction in self-supervised learning.

While current self-supervised image denoising methods have shown advanced performance, there are still some limitations. Firstly, AP-BSN [31] utilizes only PD_2 and PD_5 (subscript represents step size) to sample sub-images, which limits the network's learning capacity when dealing with complex feature maps. Inspired by feature pyramid networks [44], we have added additional scale branches. Specifically, we attempt to fill the gap between PD_2 and PD_5 by introducing PD_3 and PD_4 for sub-image sampling, further enhancing the network's ability to learn features at different scales. Secondly, in pyramid networks, different-scale feature maps are combined through a simple concatenation, without emphasizing the importance of image features. Inspired by [40], we propose an adaptive feature fusion method. Lastly, previous self-supervised methods often rely solely on CNNs to learn image features, which can limit the network's receptive field. In recent years, Transformers have been widely adopted for their large receptive fields. Inspired by [48], we introduce Transformer suitable for image denoising [20] to enhance the network's learning capability.

Sub-images obtained through multi-scale sampling are independently processed by the network for denoising, incorporating information from various scales and angles to enhance BSN's comprehensive noise feature capture. Parallel processing of these sub-images enables BSN to learn noise distribution from multiple perspectives, enhancing denoising accuracy and robustness. The denoised sub-images are then adaptively fused to reconstruct a full-resolution image, optimizing visual quality by considering the denoising effect of each sub-image and employing a fusion strategy.

In summary, we present a self-supervised image denoising method with feature enhancement and adaptive fusion at multiple scales. Benefiting from the strong generalization ability of self-supervised learning, MA-BSN samples input noisy images into multi-scale feature sub-images and uses Transformers to self-supervisedly learn their image features in a larger receptive field. Furthermore, an attention-based feature fusion mechanism further enhances image denoising capabilities.

Based on the above explanation, we now introduce the overall architecture of our method. As shown in Figure 3. The single input noisy image is initially subjected to feature transformation using a 1 × 1 convolution. Subsequently, multiple sets of sub-images at different scales are obtained through pixel-shuffle downsampling (PD) with varying strides. In parallel, the blind self-supervised denoising network (BSN) is applied to each set of sub-images for denoising. The depth-wise transform network (DTN) is then employed to learn image features with a large receptive field. The inverse of the pixel-shuffle downsampling is used to obtain the initial denoised image. The adaptive feature fusion (AFF) module is utilized to fuse the features from the previous three branches in an adaptive manner. The final denoised image is obtained through the fusion of feature maps derived from two branches. The conventional PD-refinement [23] employs a fixed pixel replacement strategy, which leaves correlations among the replaced noise that affect the denoising performance. The Random-replacing refinement [31] addresses this issue by adopting a random pixel replacement strategy. In a manner akin to the configuration of AP-BSN, we set the quantity N for random-replacing refinement at 4. The proposed method is trained using the L_1 loss function:

$$E = \|I_N - I_C\|_1 \tag{1}$$

where I_N represents the input noise image. I_C is obtained from I_N by MA-BSN as:

$$I_C = \text{AFF}(\sum_{i=3}^{5} \text{PD}_i^{-1}(I_i^*)) \tag{2}$$

where AFF denotes the daptive feature fusion module, PD^{-1} stands for the inverse of the pixel-shuffle downsampling. The denoised images I_i^* from each branch can be combined through the Adaptive Feature Fusion (AFF) to obtain I_C.

$$I_i^* = \text{DTN}(\text{BSN}(\text{PD}_i(I_N))) \tag{3}$$

where DTN denotes the depth-wise Transformer network, BSN denotes the blind-spot network, PD indicates the pixel-shuffle downsampling. The subscript i denotes the step size of PD, where $i \in 1, 2$ and 3. Firstly, the input noise image I_N is decomposed into sub-images through PD, and then BSN is applied to the sub-images to achieve self-supervised image denoising. Finally, the PD inverse operation is performed to obtain I_C. Minimize the L_1 loss of I_N and I_C to achieve self-supervised training of the network. The denoised image I_D is obtained by:

$$I_D = \frac{1}{N} \sum_{j=1}^{N} \text{BSN}(I_{\mathcal{P}_j(x,y)}) \tag{4}$$

where N denotes the quantity of random-replacing refinement, set to 4. $I_{\mathcal{P}_j}$ represents the fused image obtained on the j-th occurrence, obtained by the following equation:

$$I_{\mathcal{P}_j(x,y)} = \mathcal{P}_j(x,y) \odot I_2 + (1 - \mathcal{P}_j(x,y)) \odot I_C \tag{5}$$

where matrix $\mathcal{P}_j(x,y)$ signifies the value of $\mathcal{P}^j \in 0, 1$ at the index (x,y) with a probability $p = 0.16$ [31]. $\odot$ denotes element-wise multiplication.

In comparison to the L_2-norm, the L_1-norm achieves superior denoising performance. This is attributed to the nature of noise present in natural images, which typically does not adhere to a normal distribution but rather approximates a Laplace distribution with a kurtosis value exceeding 3. The adoption of an L_1-norm loss function is more adept at conforming to the actual distribution of image data, given its heightened sensitivity to outliers, such as noise within images. Consequently, this facilitates the effective suppression of noise [61].

B. BLIND-SPOT NETWORK

In this task, BSN and the multi-resolution sub-images obtained through Pixel-Shuffle Downsampling (PD) together to achieve an efficient and effective denoising strategy. BSN is a powerful denoising network that does not rely on predefined noise models and does not require paired clean and noisy training samples. BSN can adaptively learn and process various noises present in input images.

The function of PD is to generate multi-resolution sub-images. Through PD, we can obtain sub-images of different resolutions from the original image, which retain important features of the original image. This method allows BSN to perform denoising on images of different resolutions, thereby better adapting to and processing various types and levels of noise.

The BSN we utilized is inspired by the AP-BSN [31], Figure 4 visualizes its detailed architecture. The workflow consists of four steps. First, a linear transformation is applied to the input noisy image using a 1 × 1 convolution layer, and output features containing complete image information are obtained in parallel through several PD with different strides, resulting in sub-images of various resolutions. Second, the sub-images from each branch go through 3 × 3 and 5 × 5 centrally masked convolutions in parallel, predicting the central image from the neighborhood. Third, the predicted sub-images are processed through a Dilated Convolution (DC) module, Incorporated within each dilated convolution module is a 3 × 3 dilated convolutional layer featuring distinct stride values s. Specifically, $s = 2$ is employed for the upper path of the network, while $s = 3$ is applied to the lower path.

Different stride values to hierarchically expand the receptive field. The $s = 2$ in the upper pathway facilitates the network in capturing a broader range of contextual information, while the $s = 3$ in the lower pathway aids in acquiring more refined image details. This configuration of stride variation enables the network to effectively integrate global and local information, thereby enhancing denoising performance. Finally, the sub-images output from the two branches are merged using Adaptive Feature Fusion (AFF), yielding the output of the BSN. It is noteworthy that, we have tailored the AFF module for a scenario with only two multi-scale input feature maps, a departure from the original three. This adjustment is necessitated by the outputs from the BSN, which provides two distinct scale feature maps. The fusion operation has been streamlined to merge these two inputs, potentially reducing computational load and model complexity. The adaptive fusion mechanism employs a Softmax activation to recalibrate the feature maps, utilizing the derived attention weights to enhance the model's focus on relevant features, rather than relying on concatenation operations.

C. DEPTH-WISE CONVOLUTIONAL TRANSFORMER NETWORK

In our work, we integrate a modified Transformer structure optimized for image denoising, drawing from the

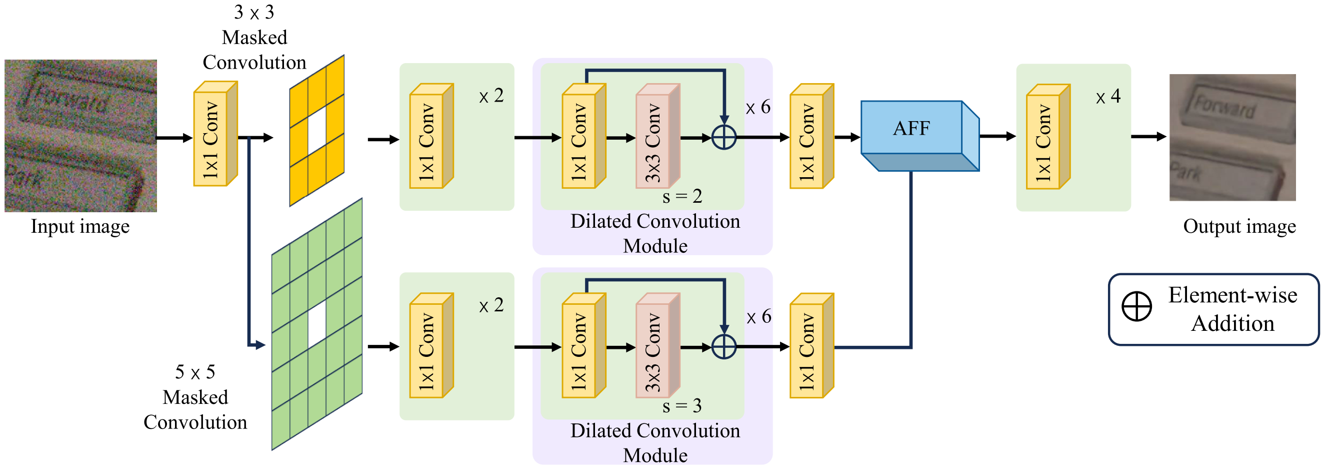


FIGURE 4. Visualization of blind-spot network architecture. Similar to our multi-scale denoising network, the input image is divided into two branches for parallel processing. Different sizes of centrally masked convolution are applied in each branch. Dilated convolution is utilized to increase the receptive field. The proposed Adaptive Feature Fusion (AFF) is employed to adaptively fuse the feature maps from both branches.

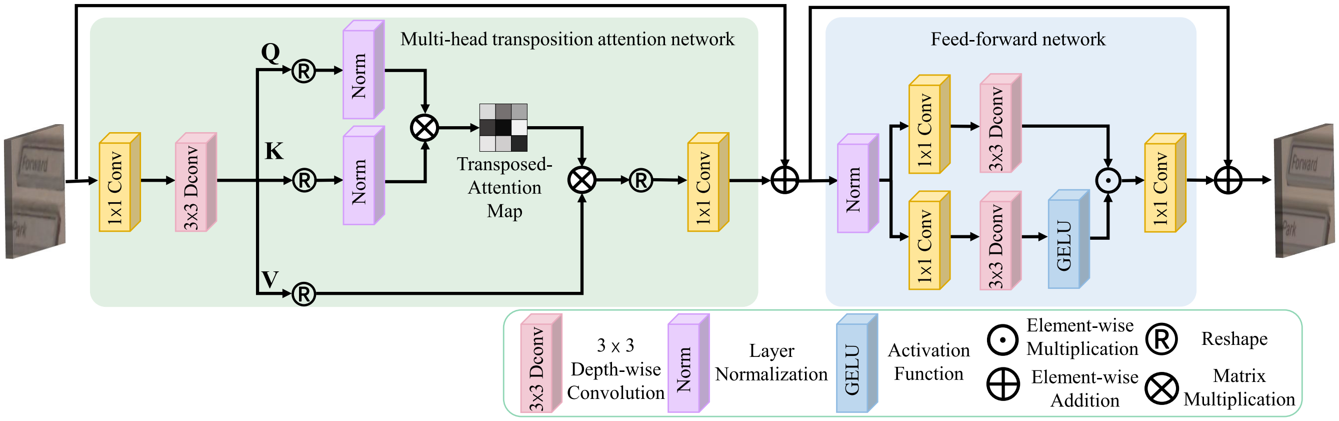


FIGURE 5. Visualization of depth-wise convolutional transformer network architecture (DTN). Using 1 × 1 convolution and 3 × 3 depth-wise convolution, image features can be learned from both spatial and channel dimensions. The multi-heads module further enhances learning ability.

advancements by Restormer [20], shown in Figure 5. This structure is designed to fully capture image features across various scales while being computationally efficient for high-resolution images. Unlike the standard Transformer, which suffers from high computational costs due to the self-attentive structure, our adapted model employs a multi-head transposition attention network that utilizes depth-wise convolution. This allows for self-attention operations to be performed in the channel dimension, thereby reducing complexity to $O(W^2H^2)$ and facilitating the extraction of global information without partitioning the image into chunks.

The self-attention mechanism is redefined to compute an implicit attention map across channels, as opposed to an explicit pixel-level interaction. This is achieved by the following operations:

$$T = \text{Reshape}(\text{Dconv}(\text{Conv1}(X))) \tag{6}$$

$$Q, K = \text{Norm}(\text{Reshape}(T)), \quad V = \text{Reshape}(T) \tag{7}$$

$$\dot{X} = \text{Conv1}(V \otimes \text{Softmax}(K \otimes Q/\alpha)) + X \tag{8}$$

where X and $\dot{X}$ denote the input and output of the attention network, respectively, while T represents the transformed feature map. The matrices Q, K, and V are derived from the input through a series of convolutions and reshape operations, with α being a learnable scaling factor that normalizes the attention weights. Norm, Conv1, Dconv, and Reshape correspond to layer normalization, 1×1 convolution, 3×3 depth-wise convolution, and feature map dimension reshaping, respectively.

Furthermore, the feed-forward network (FFN) is refined to process spatial features more efficiently:

$$Z = \text{Dconv}(\text{Conv1}(\text{Norm}(X))) \tag{9}$$

$$\widetilde{X} = \text{Conv1}(Z) \odot \text{GELU}(Z) + X \tag{10}$$

where Z represents the intermediate feature map, and $\widetilde{X}$ is the output of the FFN. The gating mechanism, indicated by element-wise multiplication with the activation function GELU, selectively enhances information flow, focusing on pertinent spatial details.

This streamlined method preserves the Transformer's capacity for global and local feature processing, optimized for high-resolution image denoising. It obviates the need for partitioning the image into blocks, utilizing depth-wise convolution for efficient feature extraction in both channel and spatial dimensions, thus reducing parameter count.

It is worth noting that layer normalization due to its independence from batch size and its focus on the normalization of individual feature channels, is preferentially employed in image denoising tasks to ensure consistent performance even in small batch processing. Compared to batch normalization, layer normalization circumvents the amplification of noise or loss of crucial details caused by intra-batch variability, which is essential for maintaining the structural integrity of denoised images. Consequently, layer normalization is more suitable for image denoising as it preserves the authenticity and consistency of image details throughout the denoising process.

D. ADAPTIVE FEATURE FUSION NETWORKS

Neurons in the human visual cortex have the feature of changing their reception area according to external stimuli [40]. Inspired by this, feature fusion and adaptive feature selection can be performed in neural networks using convolutional neural networks on multi-scale feature maps. The more common feature fusion methods are stitching or summation. However, the different scale feature maps are related to the image content, and simple stitching or summation will limit the feature representation, so neither of these approaches is the best choice. In this paper, we adopt a self-attentive mechanism for the adaptive fusion of multi-scale feature maps called Adaptive Feature Fusion Module (AFF).

The denoising method for multi-scale feature map fusion mechanism is less studied in previous studies, inspired by [40], we use a mechanism for the adaptive fusion of multi-scale feature maps for feature fusion of three parallel branches, and the network structure is shown in Figure 6. Firstly, feature maps of various scales obtained from the inverse operation of PD in Figure 3 are fused. Feature maps of varying scales carry a wealth of spatial hierarchical information, which is crucial for the understanding and interpretation of visual scenes. Feature maps at smaller scales tend to encompass high-frequency information, such as edges and textures, whereas those at larger scales provide a more expansive contextual overview, aiding in the comprehension of the overall structure of images. By combining these multi-scale feature maps through element-wise summation, we obtain:

$$m_t = (m_1 \oplus m_2 \oplus m_3) \tag{11}$$

where m_1, m_2 and m_3 are the input images of the three branches in Figure 3, $\oplus$ denotes element-wise addtion, m_t is the input feature map used to train the attention map. Subsequently, the feature map undergoes Global Average Pooling (GAP) across its spatial dimensions, a process which serves to diminish dimensionality while concurrently distilling global contextual cues:

$$v = \mathrm{GAP}(m_t) \tag{12}$$

where GAP denotes the global average pooling operation. v represents the feature vector after dimensionality reduction. The global information furnishes the network with a panoramic perspective, enabling the subsequent selection operations to adjust and enhance the feature maps with greater precision, thereby augmenting the network's sensitivity and responsiveness to pivotal information. The dimensionality-reduced feature vector $v \in \mathbb{R}^{1\times1\times C}$ is obtained, where C is the number of channels. The vector v is then futher compressed using the 1×1 convolution layer and the GELU activation function:

$$u = \mathrm{GELU}(\mathrm{Conv1}(v)) \tag{13}$$

where Conv1 denotes 1×1 convolution, GELU represents the activation function GELU, which effectively balances the propagation of salient features and the suppression of less informative ones. This selective transmission of information ensures that the subsequent channel-downscaling convolution operates within a feature space that is rich in representational quality yet compact in dimensionality. Then we can obtain the compressed feature vector $u \in \mathbb{R}^{1\times1\times n}$ from Equation (13), where $n = \frac{C}{10}$. This operation is designed to reduce the channel dimensionality of feature maps, thereby generating a more compact representation of features. Concurrently, it diminishes computational complexity and augments the network's generalization capabilities. u is made to pass through three parallel channels-upsampling convolutional layers to obtain the extended vectors g_1, g_2, and g_3:

$$g_i = \mathrm{Conv3}(u) \tag{14}$$

where Conv3 denotes 3×3 convolution, $i \in 1$, 2 and 3, denoted the three different branches. Channel-upscaling convolution effectively augments the network's capacity for feature representation by increasing the number of channels in feature maps. This operation enables the model to capture more granular information, thereby enhancing the precision of recognition for complex patterns. Concurrently, it furnishes the requisite high-dimensional feature space for multi-scale feature fusion, bolstering the network's proficiency in processing information across varying scales.

In AFF frameworks, channel reduction convolution initially decreases the feature dimensions to distill key information and reduce computational load. Subsequently, channel expansion convolution restores the feature dimensions, enhancing the expressive power of the features and providing a wealth of information for subsequent feature fusion. This process not only improves the computational efficiency of the model but also augments the model's capacity to capture and integrate essential features. Then a softmax operation is performed on each of the three extended vectors to adaptively obtain the relative weights w_1, w_2, and w_3 containing the

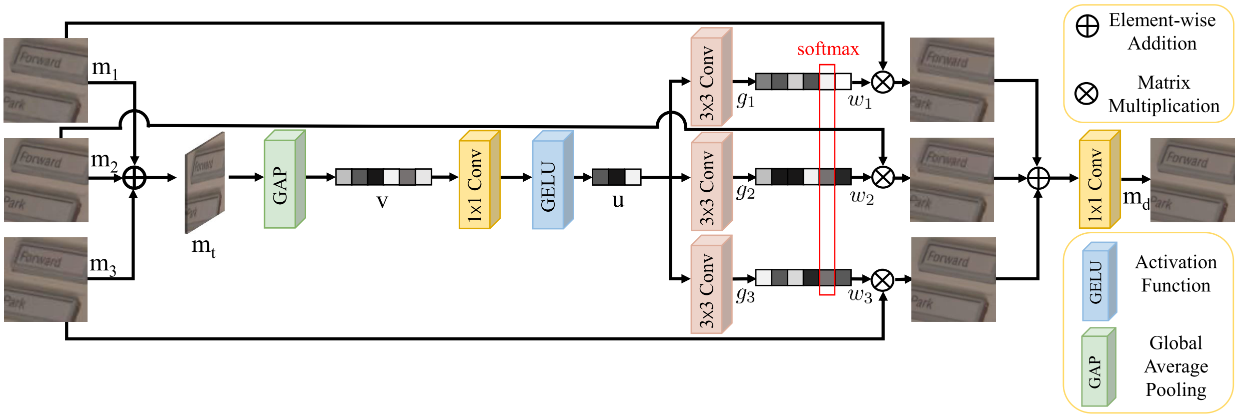


FIGURE 6. Visualization of adaptive feature fusion module architecture (AFF). We use a mechanism for the adaptive fusion of multi-scale feature maps for feature fusion of three parallel branches. Compared to concat, AFF places more emphasis on the learned image features.

attentional features:

$$w_i = \text{Softmax}(g_i) \tag{15}$$

where $i \in 1, 2$ and 3. The Softmax function facilitates adaptive recalibration of features by assigning normalized attention weights to feature maps of varying scales. Furthermore, the softmax operation's application to the feature descriptors ensures that the resulting weights are non-negative and sum to one, thus enabling a probabilistic interpretation of the importance of each feature map. This is particularly significant in the context of multi-scale feature fusion, where it is essential to balance the contribution of feature maps that encapsulate different levels of semantic information. The relative weights w_1, w_2 and w_3 effectively act as gating mechanisms that modulate the influence of each feature map in the aggregated representation. Finally, it performs elemen-twise multiplication with the corresponding input image respectively, so as to realize the adaptive feature fusion of different feature maps:

$$m_d = \text{Conv1}(\sum_{i=1}^{3} w_i \otimes m_i) \tag{16}$$

where $\sum$ represents the element-wise addition of the three branches, m_i represents the three images of the input. m_d denotes the output image of AFF.

In summary, during the fusion phase, feature maps from different branches are first merged through element-wise addition to form a comprehensive feature map. This feature map is then utilized to extract global contextual information via Global Average Pooling (GAP). Subsequently, this information is condensed into a compact feature vector through a channel-reduction convolutional layer and transformed into feature descriptors by a channel-expansion convolutional layer. In the feature adaptive selection phase, these feature descriptors are processed with a Softmax activation function to generate attention weights, which adaptively adjust the significance of the original feature maps. Finally, these recalibrated feature maps are aggregated into the final fused feature map through a weighted summation approach. Adaptive Feature Fusion (AFF) achieves an effective combination of different branch features in this manner, adaptively adjusting their contributions to generate a more favorable feature representation for downstream tasks.

IV. EXPERIMENTS

This section first describes the training method of the proposed network. Then the trained network is compared with other methods on different noisy datasets to verify the effectiveness of the network in this chapter on the image denoising task.

A. DATASET

Since the denoising performance of real-noise images determines the success or failure of the algorithm in practical applications, we select two real-world noisy image datasets that are more commonly used to verify the effectiveness of the network in this paper. The two real-noise image datasets are SIDD [49] and DND [50], and the SIDD test set is 40 high-resolution sRGB images taken in 40 scenes different from the training set. Real-world noisy sRGB image blocks and the DND dataset consist of 50 high-resolution real-noisy sRGB image blocks captured by four consumer-grade cameras and do not provide any additional noisy image-clean image pairs for training. Therefore, the network trained on the SIDD training set is directly used for denoising the DND dataset to avoid the possibility of overfitting, and the denoising performance is more convincing.

B. IMPLEMENTATION DETAILS

During training, our hyperparameters are set as follows. The batch size used in the experiments is 8. Patches measuring 64 × 64 are stochastically sampled from images afflicted with noise, and each image used in training is subjected to augmentation through random flips and 90-degree rotations. We use the L1 loss between the noisy image and the output for training. The learning rate starts from 0.0001 with the

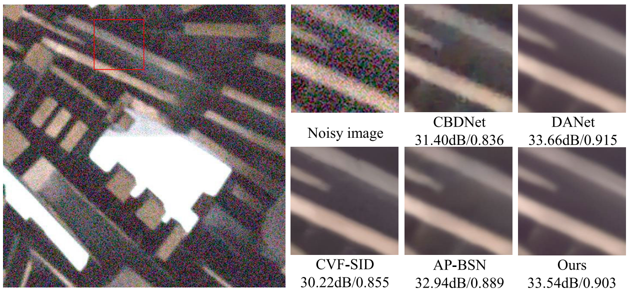


FIGURE 7. Visual comparison of various methods on the DND benchmark dataset.

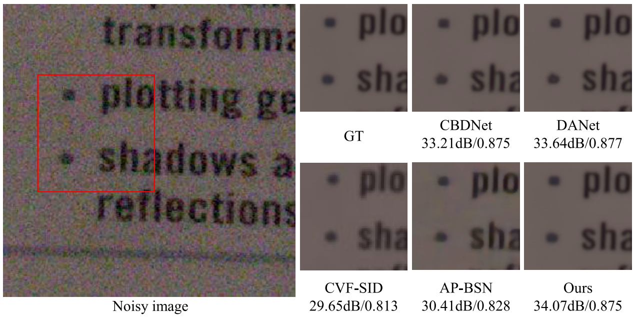


FIGURE 8. Visual comparison of various methods on the SIDD dataset.

Adam optimizer. The network was trained with 20 epochs until complete convergence. We implemented the method in PyTorch 1.9.0 and trained the model on an Nvidia RTX 3090. We use the widely used metrics of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate the method's performance.

C. RESULTS ON REAL-WORLD DENOISING

Validating the effectiveness of our proposed MA-BSN method for real-world image denoising is one approach to conducting tests on the widely used and highly persuasive SIDD [49] and DND [50] benchmark datasets. Table 1 presents the quantitative results of recent supervised and unsupervised image denoising methods on SIDD and DND, using the commonly used PSNR and SSIM evaluation metrics. Our method, MA-BSN, is shown in the last column, and it achieves the best performance among various denoising methods, as observed through comparisons.

Through comparisons, it is evident that traditional methods such as BMD and WNNM, which have not been trained on the dataset, perform poorly in handling real-world noise, highlighting the necessity of training. For supervised methods, using authentic noisy images for training outperforms using synthetic noise. For instance, the PSNR of DnCNN is approximately 15 dB lower than that of AINDNet, indicating that although the acquisition cost of synthetic noisy images is lower, their distribution does not match the real-world noise. These findings further emphasize that real-world noise is not spatially independent but exhibits specific correlations with image distribution.

In unsupervised methods, overall performance is superior to traditional methods without training and deep learning methods trained on synthetic noise images but slightly inferior to supervised deep learning methods trained on authentic noisy images. Specifically, methods that learn the image noise distribution and construct noisy-clean pairs achieve similar denoising performance as methods that construct noisy-noisy pairs. Overall, these methods outperform those that do not utilize real noise for training. There is a clear trend of rapid development in methods that train the network using only simple noisy images. Pioneering work such as Noise2Void achieved a PSNR of only 27.68 dB and SSIM of 0.668 on the SIDD dataset. However, there has been significant improvement in recent years, with approximately a 10 dB increase in PSNR, representing a one-third improvement, and a 0.3 increase in SSIM, representing a 40% improvement. These advancements indicate that the field of self-supervised image denoising, although relatively late to start, is progressing rapidly.

Supervised deep learning methods trained on real-world noisy images, often utilizing the training set provided by the SIDD benchmark dataset, typically exhibit superior denoising performance, which can be attributed to the similarity between the image distribution in the training set and the test set and the abundance of images that provide rich prior information. Conversely, self-supervised methods demonstrate competitive denoising performance by leveraging a single noisy image and hold tremendous potential for generalization tasks. Furthermore, their characteristic of not relying on a training set aligns better with the requirements of denoising tasks in practical applications. Our proposed method, which leverages advanced feature learning structures, achieves the highest performance among self-supervised methods.

We also compared the denoising results of different methods from a visual perspective. We conducted a thorough evaluation of MA-BSN's effectiveness by comparing it against a range of supervised and unsupervised methods. Specifically, the compared methods include CBDNet [7], DANet [11], CVF-SID [60], and AP-BSN [31]. Among them, CBDNet and DANet are supervised methods, while CVF-SID and AP-BSN are unsupervised.

A comparison of denoised images from different methods reveals noticeable improvements in the low-frequency region for our proposed method, as shown in Figure 1. In contrast, AP-BSN still exhibits visible noise in the low-frequency region. Additionally, the high-frequency textures appear more straightforward, indicating that the extension of pixel-shuffle downsampling with different scales yields better preservation of high-frequency features than the PD method with a stride of 2 alone. In Figure 7, CBDNet performs poorly in handling high-frequency regions, resulting in distorted edges and noise in the low-frequency region. CVF-SID exhibits artifacts in the edge regions. AP-BSN

TABLE 1. Quantitative comparison of denoising performance on sRGB images of real-world noise benchmark datasets SIDD and DND. Results with * mean these are reported from R2R [56].We obtained the results of other methods in benchmark from the official websites of SIDD and DND.

Type of supervision	Training data	Method	SIDD		DND	
			PSNR(dB)	SSIM	PSNR(dB)	SSIM
Non-learning based	None	BM3D [2] WNNM [4]	25.65 25.78	0.685 0.809	34.51 34.67	0.851 0.865
Supervised	Synthesized Pairs	DnCNN [8] CBDNet [7]	23.66 33.28	0.583 0.868	32.43 38.05	0.790 0.942
	Real pairs	AINDNet [51] SwinIR [46] SUNet [18]	38.84 39.43 39.68	0.951 0.948 0.958	39.34 39.68 39.87	0.952 0.954 0.955
Self-supervised	Noisy-clean pairs	UIDNet [53] C2N [38]+DIDN [52]	32.48 35.35	0.897 0.937	- 36.38	- 0.887
	Noisy-noisy pairs	Noise2Self [22] R2R [56]	29.56* 34.78	0.808* 0.898	- -	- -
	Single noisy observation	CVF-SID [60] AP-BSN [31] SS-BSN [58] MA-BSN(Ours)	34.71 35.97 36.73 **37.33**	0.917 0.925 0.923 **0.929**	36.50 38.09 37.72 **38.41**	0.924 0.937 0.928 **0.940**

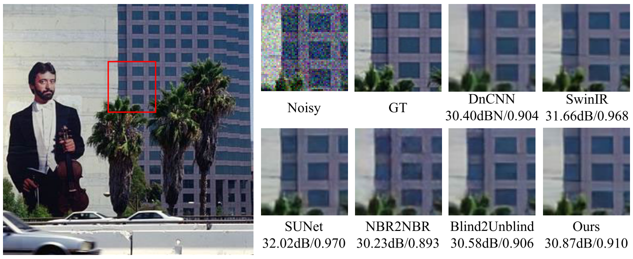


FIGURE 9. Visual comparison on synthetic noisy image tasks with noise level $\sigma = 30$.

shows curved edges in the upper-left bright area, suggesting limitations in the receptive field of AP-BSN. The denoising performance is significantly compromised when there is an insufficient correlation among neighboring pixels. Our method incorporates Transformers, enabling the learning of image features from a larger context, effectively addressing this challenge. Figure 8 demonstrates the excellent denoising performance of our method across different datasets, which shows the strong generalization ability of MA-BSN.

D. RESULTS FOR SYNTHETIC DENOISING

The generalization level of the proposed method was validated through further evaluation on the synthetic noise image datasets CBSD68 [58] and Urban100 [59] at different noise levels. Commonly utilized noise levels of 10, 30, and 50 were selected for this analysis. A comparison was conducted among image denoising methods based on image priors, supervised learning, and self-supervision, with the quantitative performances delineated in Table 2. In the realm of self-supervised methods, the proposed approach demonstrated competitive efficacy. Nonetheless, there remains room for enhancement in self-supervised methods when juxtaposed with supervised denoising methods like SUNet.

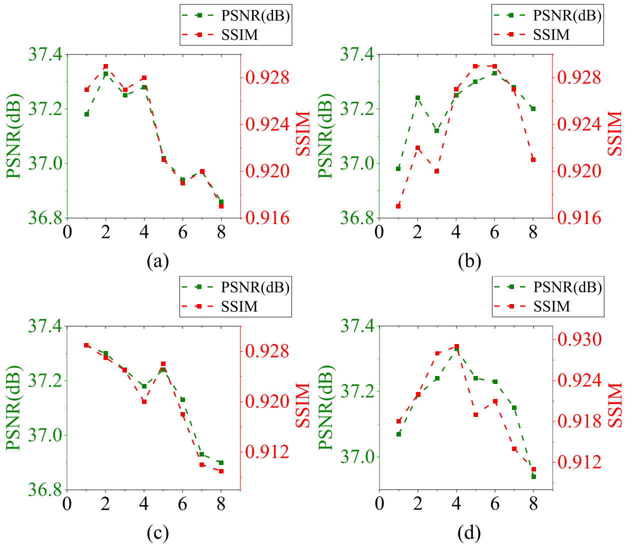


FIGURE 10. The analysis of convolutional blocks on BSN. The abscissa of the four line charts represents the number of convolutional modules. (a) The first 1 × 1 convolutional module after entering the branch. (b) The dilated convolution in each branch. (c) The second 1 × 1 convolutional module after entering the branch. (d) The 1 × 1 convolutional module after merging the branches.

Figure 9 exhibits a visual comparison of different methods at a noise level of 30, highlighting the enhanced texture preservation capability attributed to the integration of the Transformer. The rapid advancement of self-supervised methods is narrowing the gap with supervised approaches. Considering the flexibility and superior generalization ability of self-supervised methods, they hold significant research value.

E. THE IMPACT OF CONVOLUTIONAL BLOCKS ON BSN

The BSN shown in Figure 4 is the core of the self-supervised method. The approach taken here involves

TABLE 2. Comparative quantitative analysis of various denoising methods on synthetic noise image datasets CBSD68 and Urban100 at selected noise levels of 10, 30, and 50.

Type of supervision	Method	CBSD68 [59]			Urban100 [60]		
		10	30	50	10	30	50
		PSNR(dB)/SSIM	PSNR(dB)/SSIM	PSNR(dB)/SSIM	PSNR(dB)/SSIM	PSNR(dB)/SSIM	PSNR(dB)/SSIM
Non-learning based	BM3D [2]	35.66/0.951	29.07/0.835	25.87/0.736	35.65/0.958	29.00/0.881	25.41/0.803
	WNNM [4]	35.75/0.952	29.11/0.833	25.96/0.730	35.69/0.958	28.93/0.880	25.73/0.811
Supervised	DnCNN [8]	35.47/0.949	29.19/0.832	26.45/0.729	35.07/0.961	28.60/0.876	25.22/0.768
	ADNet [54]	34.59/0.892	29.07/0.820	26.33/0.756	34.76/0.921	28.85/0.809	25.69/0.808
	AINDNet [51]	35.43/0.951	29.76/0.866	27.02/0.778	35.86/0.969	29.74/0.894	26.95/0.837
	SwinIR [46]	35.99/0.954	30.18/0.862	27.73/0.789	35.82/0.967	29.87/0.904	27.05/0.843
	SUNet [18]	35.94/0.958	30.28/0.870	27.85/0.799	35.79/0.970	29.98/0.909	27.11/0.851
Self-supervised	NBR2NBR [26]	34.50/0.889	29.33/0.819	26.51/0.752	34.68/0.914	28.62/0.788	25.33/0.792
	Blind2Unblind [55]	34.57/0.899	29.65/0.837	26.86/0.771	34.72/0.924	28.69/0.797	25.54/0.801
	CVF-SID [60]	34.55/0.899	29.68/0.840	26.52/0.767	34.83/0.925	28.73/0.799	25.62/0.804
	AP-BSN [31]	34.61/0.903	**30.23/0.858**	26.46/0.763	34.57/0.921	28.69/0.798	25.65/0.801
	MA-BSN(ours)	**34.88/0.913**	30.09/0.854	**27.11/0.781**	**34.91/0.933**	**28.98/0.811**	**25.88/0.824**

improving denoising performance. We explored the impact of different numbers of convolutional modules on the results. Specifically, we controlled variables using the number of BSN modules used in AP-BSN as a baseline. We investigated the effects of two 1 × 1 convolutional modules after entering the branch, one dilated convolution, and the 1 × 1 convolutional module after merging the branches. The 1 × 1 convolutional module before the branch primarily adjusts the channels and thus was not studied. Figure 10 shows that the optimal number is required to achieve the best denoising performance. It is worth noting that all 1 × 1 convolutional layers have 128 channels except the last convolutional layer, which is used to generate denoised images with three channels.

TABLE 3. Ablation experiments on different modules in the SIDD validation set.

	Multi-scale	AFF	DTN	BSN w/o AFF	PSNR(dB)	SSIM
(a)	✗	✓	✓	✓	36.72	0.923
(b)	✓	✗	✓	✓	36.79	0.925
(c)	✓	✓	✗	✓	36.58	0.924
(d)	✓	✓	✓	✗	37.12	0.928
(e)	✓	✓	✓	✓	**37.33**	**0.929**

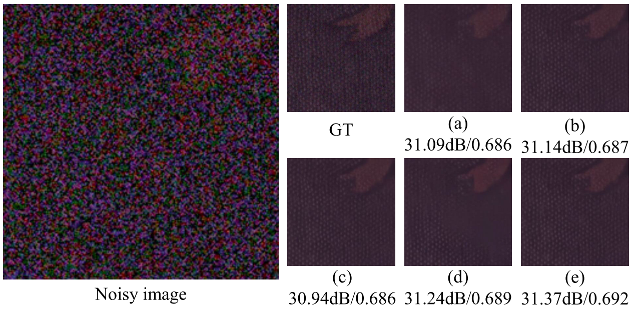
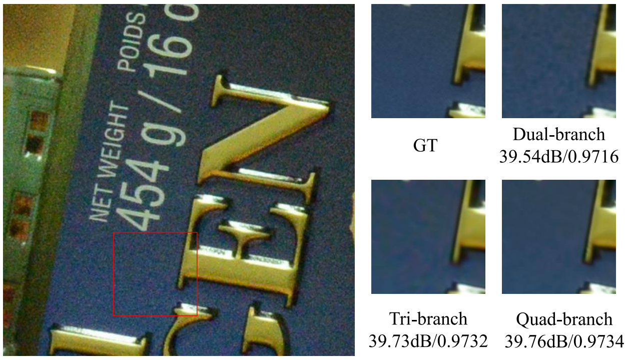

Noisy image
GT
(a) 31.09dB/0.686
(b) 31.14dB/0.687
(c) 30.94dB/0.686
(d) 31.24dB/0.689
(e) 31.37dB/0.692

FIGURE 11. Visual comparison of ablation experiments. The serial numbers under the figure correspond to the serial numbers in the Table 3.



GT
Dual-branch 39.54dB/0.9716
Tri-branch 39.73dB/0.9732
Quad-branch 39.76dB/0.9734

FIGURE 12. Visual comparison of different branch fusion strategies.

F. ABLATION STUDY

Furthermore, we conducted ablation experiments on the proposed method, focusing on the Multi-scale architecture, AFF, DTN, and BSN modules. The ablation experiment on BSN primarily aimed to test the impact of replacing the commonly used "concat" operation with AFF. As shown in Table 3, when multiple-scale branches are not constructed, the ability to extract different types of features decreases significantly. While the AFF is removed from the backbone and replaced with the concat operation, there is an inevitable decrease in denoising performance. Furthermore, removing the Transformer module leads to a significant drop in denoising effectiveness, confirming the necessity of the Transformer for the denoising network. When using BSN without the inclusion of AFF, there is a slight decrease in denoising performance, indicating that AFF plays a role in fusing different feature maps. Figure 11 is a visual comparison of ablation experiments on the SIDD validation dataset.

The impact of different branch fusion strategies on image denoising efficacy was similarly validated. A dual-branch strategy employed stride of 3 and 5. Conversely, the quad-branch strategy incorporated stride of 2, 3, 4, and 5 for branch fusion. Figure 12 illustrates that the adopted tri-branch strategy significantly outperformed the dual-branch approach. However, the enhancement from the

TABLE 4. Comparison of computational complexities.

Methods	Params(M)	FLOPs(G)
DnCNN [8]	0.6	36
AP-BSN [31]	3.7	29
SwinIR [46]	12	44
SUNet [18]	99	30
Ours	10.3	37

tri-branch to the quad-branch strategy was limited, yet accompanied by a non-negligible increase in computational cost. Considering these factors, the tri-branch strategy was ultimately selected.

G. COMPUTATIONAL COMPLEXITY

We calculated the number of parameters and computational cost of the proposed method, as shown in the Table 4. The computational efficiency demonstrates a significant advantage over supervised Transformer methods. This enhances the performance of self-supervised image denoising methods at a lower computational expense.

V. DISCUSSION

In this study, we introduce the Multi-scale Blind-spot Network (MA-BSN), employing adaptive feature fusion techniques to tackle the challenges inherent in self-supervised image denoising. Despite MA-BSN's commendable performance in self-supervised denoising tasks, we acknowledge certain limitations that are crucial for a comprehensive understanding of our approach and its applicability.

Firstly, the incorporation of the Transformer module in MA-BSN introduces computational complexity, potentially rendering it unsuitable for real-time application scenarios. While the Transformer module enhances denoising capabilities, its substantial computational demands may limit the model's deployment in environments requiring rapid processing.

Secondly, although our method is capable of addressing a broad spectrum of noise levels, its performance may deteriorate under conditions of extreme noise, particularly where the signal-to-noise ratio is exceptionally low. Potential failure scenarios include images with highly structured noise patterns, which may surpass our self-supervised method's capture range, and images containing complex textures akin to noise patterns, potentially resulting in excessive smoothing.

Future research will focus on exploring lightweight Transformer architectures to mitigate computational demands while preserving denoising efficacy. Efforts will also be directed towards enhancing MA-BSN's robustness against highly structured noise and extreme noise levels through the introduction of more sophisticated noise modeling techniques. Additionally, we plan to refine the self-attention mechanisms within the adaptive feature fusion module to more effectively discriminate between noise and complex image textures. These research directions aim not only to advance the development of self-supervised image denoising methods but also to overcome the current limitations of MA-BSN, paving new pathways for future image processing technologies.

VI. CONCLUSION

In this paper, we propose a novel multi-scale blind-spot network with adaptive feature fusion (MA-BSN) to address the real-world image denoising problem self-supervised. MA-BSN is designed with four distinct branches that sample images at various scales, enabling the BSN to effectively extract image features, with each branch capturing unique image characteristics. Self-supervised learning of neighborhood pixel features is performed on each branch, striking a balance between disrupting noise spatial correlation and preserving image high-frequency features. By employing a Transformer module, our method enhances the CNNs' limited receptive field. This module learns image features across channels, which facilitates global receptive field feature extraction and minimizes the loss of high-frequency content that typically results from pixel-shuffle downsampling. Additionally, an adaptive feature fusion module is proposed to merge denoised images obtained at different scales using self-attention, further enhancing the denoising performance of the network. Compared to other state-of-the-art image denoising methods on various benchmark datasets, MA-BSN achieves superior performance among self-supervised denoising methods, validating the advancement of the MA-BSN approach. In future work, we will explore meta-learning to improve hyperparameters and extend MA-BSN to other domains for self-supervised solutions to various image-processing tasks.

REFERENCES

[1] A. Buades, B. Coll, and J.-M. Morel, "Non-local means denoising," *Image Process. On Line*, vol. 1, pp. 208–212, Sep. 2011.

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[3] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2272–2279.

[4] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2862–2869.

[5] J. Xu, L. Zhang, D. Zhang, and X. Feng, "Multi-channel weighted nuclear norm minimization for real color image denoising," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1096–1104.

[6] Q. Guo, C. Zhang, Y. Zhang, and H. Liu, "An efficient SVD-based method for image denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 868–880, May 2016.

[7] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1712–1722.

[8] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[9] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, "Variational denoising network: Toward blind noise modeling and removal," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[10] C. Tian, Y. Xu, W. Zuo, B. Du, C.-W. Lin, and D. Zhang, "Designing and training of a dual CNN for image denoising," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 106949.

[11] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 41–58.

[12] Y. Hou, J. Xu, M. Liu, G. Liu, L. Liu, F. Zhu, and L. Shao, "NLH: A blind pixel-level non-local method for real-world image denoising," *IEEE Trans. Image Process.*, vol. 29, pp. 5121–5135, 2020.

[13] J. Xu, L. Zhang, and D. Zhang, "A trilateral weighted sparse coding scheme for real-world image denoising," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 20–36.

[14] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2022, pp. 17–33.

[15] J. Prost, A. Houdard, A. Almansa, and N. Papadakis, "Learning local regularization for variational image restoration," in *Proc. Int. Conf. Scale Space Variational Methods Comput. Vis.* Springer, 2021, pp. 358–370.

[16] S. Cheng, Y. Wang, H. Huang, D. Liu, H. Fan, and S. Liu, "NBNet: Noise basis learning for image denoising with subspace projection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 4896–4906.

[17] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17683–17693.

[18] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer UNet for image denoising," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 2333–2337.

[19] S. Anwar and N. Barnes, "Real image denoising with feature attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3155–3164.

[20] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[21] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning image restoration without clean data," 2018, *arXiv:1803.04189*.

[22] J. Batson and L. Royer, "Noise2Self: Blind denoising by self-supervision," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 524–533.

[23] Y. Zhou J. Jiao and H. Huang, "When AWGN-based denoiser meets real noises," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 13074–13081.

[24] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2Void—Learning denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2129–2137.

[25] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[26] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2neighbor: Self-supervised denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14781–14790.

[27] Y. Quan, M. Chen, T. Pang, and H. Ji, "Self2self with dropout: Learning self-supervised denoising from single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1890–1898.

[28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 16000–16009.

[29] S. Laine, T. Karras, J. Lehtinen, and T. Aila, "High-quality self-supervised deep image denoising," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[30] X. Wu, M. Liu, Y. Cao, D. Ren, and W. Zuo, "Unpaired learning of deep image denoising," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 352–368.

[31] W. Lee, S. Son, and K. M. Lee, "AP-BSN: Self-supervised denoising for real-world images via asymmetric PD and blind-spot network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17704–17713.

[32] D. Honzátko, S. A. Bigdeli, E. Türetken, and L. A. Dunbar, "Efficient blind-spot neural network architecture for image denoising," in *Proc. 7th Swiss Conf. Data Sci. (SDS)*, Jun. 2020, pp. 59–60.

[33] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14821–14831.

[34] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 492–511.

[35] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3587–3596.

[36] K. Lin, T. H. Li, S. Liu, and G. Li, "Real photographs denoising with noise domain adaptation and attentive generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1717–1721.

[37] A. Guo, L. Fang, M. Qi, and S. Li, "Unsupervised denoising of optical coherence tomography images with nonlocal-generative adversarial network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[38] G. Jang, W. Lee, S. Son, and K. M. Lee, "C2N: Practical generative noise modeling for real-world denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 2350–2359.

[39] J. Xu, Y. Huang, M.-M. Cheng, L. Liu, F. Zhu, Z. Xu, and L. Shao, "Noisy-as-clean: Learning self-supervised denoising from corrupted image," *IEEE Trans. Image Process.*, vol. 29, pp. 9316–9329, 2020.

[40] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[43] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 1–10.

[44] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 1–11.

[46] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[47] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 12299–12310.

[48] D. Zhang and F. Zhou, "Self-supervised image denoising for real-world images with context-aware transformer," *IEEE Access*, vol. 11, pp. 14340–14349, 2023.

[49] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1692–1700.

[50] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1586–1595.

[51] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3482–3492.

[52] S. Yu, B. Park, and J. Jeong, "Deep iterative down-up CNN for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2095–2103.

[53] Z. Hong, X. Fan, T. Jiang, and J. Feng, "End-to-end unpaired image denoising with conditional adversarial networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 4140–4149.

[54] C. Tian, Y. Xu, Z. Li, W. Zuo, L. Fei, and H. Liu, "Attention-guided CNN for image denoising," *Neural Netw.*, vol. 124, pp. 117–129, Apr. 2020.

[55] Z. Wang, J. Liu, G. Li, and H. Han, "Blind2unblind: Self-supervised image denoising with visible blind spots," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2027–2036.

[56] T. Pang, H. Zheng, Y. Quan, and H. Ji, "Recorrupted-to-recorrupted: Unsupervised deep learning for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2043–2052.

[57] Y.-J. Han and H.-J. Yu, "SS-BSN: Attentive blind-spot network for self-supervised denoising with nonlocal self-similarity," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, May 2023, pp. 1–10.

[58] D. Martin, C Fowlkes, D Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, vol. 2, Jul. 2001, pp. 416–423.

[59] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[60] R. Neshatavar, M. Yavartanoo, S. Son, and K. M. Lee, "CVF-SID: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17583–17591.

[61] MT. Duong, S. Lee, and MC. Hong, "DMT-Net: Deep multiple networks for low-light image enhancement based on retinex model," *IEEE Access*, vol. 11, pp. 132147–132161, 2023.

[62] J. Oh and M.-C. Hong, "Adaptive image rendering using a nonlinear mapping-function-based retinex model," *Sensors*, vol. 19, no. 4, p. 969, Feb. 2019.

[63] J. Oh and M.-C. Hong, "Low-light image enhancement using hybrid deep-learning and mixed-norm loss functions," *Sensors*, vol. 22, no. 18, p. 6904, Sep. 2022.

[64] M.-T. Duong and M.-C. Hong, "EBSD-Net: Enhancing brightness and suppressing degradation for low-light color image using deep networks," in *Proc. IEEE Int. Conf. Consum. Electronics-Asia (ICCE-Asia)*, Oct. 2022, pp. 1–4.

HAILIANG TANG received the M.A. degree in computer application from Shandong Normal University, China, in 2017.

He is currently a Lecturer with the College of Information Science and Engineering, Qilu Normal University, China. His current research interests include machine learning, neural networks, and data mining.

WENXIAO ZHANG received the M.S. degree in business administration from Shandong Normal University, China, in 2020.

She is currently a Lecturer with the College of Digital Finance, Shandong University of Engineering and Vocational Technology, China. Her current research interests include machine learning, deep learning, pattern recognition, and data mining.

HAILIN ZHU received the M.A. degree in computer software and theory from Shandong Normal University, China, in 2009.

He is currently a Lecturer with the College of Information Science and Engineering, Qilu Normal University, China. His current research interests include machine learning, pattern recognition, and the Internet of Things.

KE ZHAO received the M.S. degree in software engineering from Shandong University, China, in 2006.

He is currently a Lecturer with the College of Information Science and Engineering, Qilu Normal University, China. His main research interests include machine learning, pattern recognition, and intelligent systems.