MD-BAN: Multi-Direction Mask and Detail Enhancement Blind-Area Network for Self-Supervised Real-World Denoising

Ruiying Wang and Yong Jiang (✉)

School of Computer Science and Technology, Southwest University of Science and Technology, Sichuan 621010, Mianyang, China
wangruiying@mails.swust.edu.cn, jiang_yong@swust.edu.cn

Abstract. Recently, Asymmetric PD and Blind-Spot Network (AP-BSN) has shown effectiveness for real-world image denoising. However, when the noise-related area is large, it uses the single-center pixel mask which cannot break the noise spatial correlation, therefore the blind spot recovered from the surrounding pixels still contains noise, resulting in obviously abnormal color spots in the denoised image. In addition, AP-BSN enlarges the receptive field by stacking multiple dilated convolutional layer (DCL), but these layers may lead to block artifacts and partial pixel detail information loss due to their interpolation and overlap operations. To address the above issues, we propose a multi-direction mask convolution kernel (MDMCK) to form a blind area to further destroy large-scale spatial connection noise. We also propose a detail feature enhancement (DFE) module to supplement the detail lost by MDMCK and stacking DCL. Finally, we use a robust joint loss function to train our model, generating denoised images with clean and sharp detail while alleviating the block artifacts. Extensive quantitative and qualitative evaluations of the SIDD and DND datasets show that our proposed method performs favorably.

Keywords: Self-supervised denoising, Real-world image, Multi-direction mask, Detail feature enhancement, Blind-area network.

1 Introduction

Image denoising aims to recover a clean image from its corresponding noisy observation. Learning-based methods have been widely used in image denoising in recent years due to their superior performance [1, 2]. Supervised methods usually use Additive White Gaussian Noise (AWGN) to synthesize massive noise-clean image training pairs. However, it is difficult to capture ground truth images in some cases, such as in the medical field.

To address the above issue, a series of self-supervised image denoising methods that do not require ground truth images have been proposed [4, 5, 6, 7, 8, 15]. Noise2Void [6] proposes a Blind-Spot Network (BSN) denoising method based on the assumption that pixel signals of the input image are spatially correlated, and noise signals are spatially independent with zero-mean, which can be trained with only single noisy images. Most of the existing BSN denoising methods [9, 10, 11] use a mask convolution kernel based on the single-center pixel mask, such as AP-BSN [9]. However, it cannot sufficiently destroy the spatial connection of large-scale noise, resulting in the features extracted by the single-center pixel convolution kernel still containing unconducive information for denoising. The noise that has not been removed shows obviously abnormal spots on the denoised image. Furthermore, AP-BSN has repeatedly used the dilated convolutional layer (DCL) to enlarge the receptive field, but these layers may lead to block artifacts and partial pixel detail information loss due to the interpolation and overlapping receptive fields.

In this paper, we propose a novel method, called MD-BAN, to solve these problems, including abnormal color spots, block artifacts, and the lack of detailed information. First, for breaking the spatial correlation of large-scale noise, we introduce a multi-direction mask convolution kernel (MDMCK), which contains oblique, horizontal, and vertical directions, to for blind area to mask pixels. Second, for retaining more detail information, we introduce a detail feature enhancement (DFE) module that can not only supplement the detail destroyed by the MDMCK, but also reduce the loss of pixel detail information during multiple stacking DCL. Finally, for alleviating the block artifacts and make the denoised image smoother and more realistic, we combine the more robust Charbonnier loss with the sparse L1 norm to train the model. Compared with several representative image denoising methods on the SIDD validation dataset, SIDD benchmark dataset and DND benchmark dataset, our method performs well in terms of quantitative indicators and perceptual quality. We summarize our contributions as follows:

- We propose a novel self-supervised method called MD-BAN for real-world image denoising. It retains more detail while improving denoising performance, and significantly reduces block artifacts and abnormal color spots.
- Our MDMCK comprehensively breaks the spatial correlation of large-scale noise from oblique, horizontal and vertical directions. Our DFE module focuses on supplementing the lost detail of MDMCK and stacked DCL.
- The joint loss alleviates the block artifacts caused by stacking the dilated convolution, making the denoised image smoother and more realistic. Extensive experiments show our method achieves commendable performance.

2 Related Work

Due to the advancement of deep learning technology, deep network-based methods have achieved superior denoising results and become the mainstream denoising methods. In general, deep network-based methods can be further classified according to their training manners.

2.1 Supervised Image Denoising

Zhang et al. [12] first attempted to apply deep learning technology to image denoising tasks. They proposed DnCNN, which trains the model with a synthetic noise-clean image pair by manually adding AWGN to the clean image. This model not only improves the denoising performance, but also greatly reduces the amount of calculation. However, there are significant differences in the distribution between AWGN noise and real noise. To reduce the gap, Gou et al. [13] proposed a convolutional blind denoising network (CBDNet) specifically designed for real images. Under its framework, Zhao et al. [14] proposed SDNet, which achieves good results on both synthetic images and real noisy images. Unfortunately, these methods often rely on the quantity and quality of training data, and obtaining absolutely clean images is impractical in practical applications. Therefore, it makes methods that do not require clean images more valuable.

2.2 Self-supervised Image Denoising

To get rid of the dependence on clean images, many self-supervised methods [11, 18] that only use noisy images for training have been proposed. Noise2Noise [7] requires two completely aligned noise image pairs, which is difficult to obtain in practice. NAC [23] proposes to use the existing noise images and add new noise to the noisy images to form image pairs for training. Some researchers have developed denoising model that use a single noisy image for training, the most widely used is the BSN proposed in Noise2Void [6]. Subsequently, Noise2Self [4] proposes a general framework for denoising high-dimensional measurements. But a single sample training will cause the large variance. To overcome this problem, Self2Self [30] is trained with dropout on the pairs of Bernoulli-sampled instances of the input image. Blind2Unblind [17] used all the pixels for training by generating sub-masked images with pixels masked at different locations. Neighbor2Neighbor [5] synthesized two sub-noise images by randomly selecting two adjacent pixels from the neighborhood of the rawRGB image. CVF-SID [16] decomposes noisy images into clean images and noise components. Nevertheless, the denoising ability of the above method is limited by the assumption that the noise is spatially independent. Since the noise in the real world is usually spatially correlated, it does not conform to the assumption of BSN.

To break the spatial correlation of real-world noise, AP-BSN [9] asymmetrically uses pixel-shuffle downsampling (PD) and the single-center pixel mask convolution kernel. However, if large-scale noise exists in the image, blindly increasing the PD stride will cause damage to the image details. Under this restriction, continuing to use the single center pixel mask will result in poor denoising effect. Li et al. [31] proposed the blind-neighborhood network (BNN), which is deformed from BSN but have different receptive field. Luiken et al. [32] proposed a new network whose receptive field excludes an entire direction. Zhang et al. [19] combined CNN with a window-based Transformer to balance noise removal and preserve local detail. But transformer is computationally intensive. Instead, our method is more lightweight. We propose MDMCK to mask more pixels that are strongly noise correlated with the central pixel from different directions, the mask area is restored by using the pixels with weak noise correlation, resulting in the denoised image is clearer. We also propose a novel module DFE that combines the MDMCK to destroy large noise correlation while preserving texture details of the original image. Finally, the joint loss function is used to deal with the block artifacts, which makes the denoised image more real and clear.

3 Method

In this section, we describe the proposed method in detail. The overview is displayed in Fig. 1. First, the original noisy image is sampled into multiple small images after PD_5, and the sampled images are passed through a 1×1 convolution layer for linear transformation. Second, there are two branches in parallel, aiming at breaking the spatial connection and detail feature supplement, respectively. For the detail feature supplement branch, we apply 3×3 MDMCK to extract more complete information feature. For breaking the spatial connection branch, we apply 5×5 MDMCK to further break the spatial correlation. Subsequently, both branches pass through $m = 12$ stacked DFE modules to address the problem of detail loss during stacking dilated convolution layers. Finally, the model is optimized by the joint loss function.

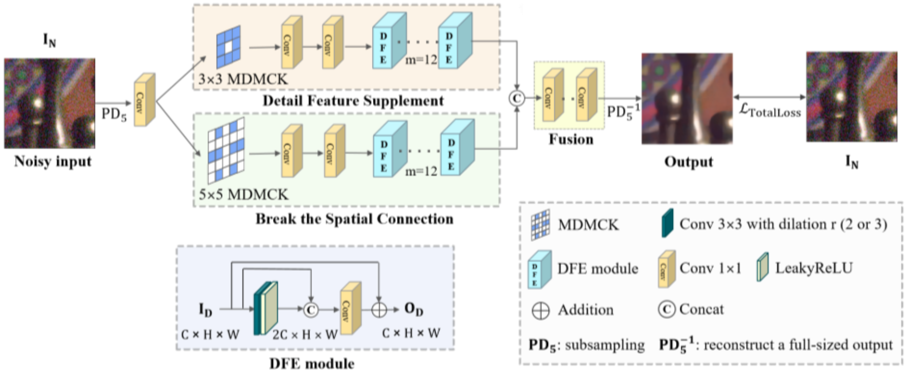

Fig. 1. The overall architecture of MD-BAN. Our method mainly contains two branches. For each branch, the input after PD_5 and linear transformation first goes through an MDMCK, then further processed by the DFE module for deep features. Finally, the output of two branches is fused. Then a full-sized output is reconstructed using PD_5^{-1} to calculate loss with noise image. Note the 3×3 branch of the MDMCK we all use a center single pixel mask convolution kernel.

3.1 Multi-direction Mask Convolution Kernel (MDMCK)

It is mentioned above that AP-BSN is inefficient in breaking the spatial correlation of large-scale noise by only using a single-center pixel mask. The blind spot recovered from the surrounding pixels will still contain noise and show obvious abnormal color spots in the denoised image, as shown in Fig. 4.

Motivated by the visualization of spatial correlation in noise between the center pixel and other pixels in the LG-BSN [20]. By masking more parts that are strongly correlated with the central pixel, we can avoid the blind area recovered from the surrounding pixels still containing noise, and achieve the effect of breaking the spatial correlation of large-scale noise. Based on the above analysis, we propose several combinations of oblique, horizontal, and vertical MDMCK as shown in Fig. 2 (take the OHHV mask as an example) to further destroy the spatial correlation of large-scale noise. MDMCK is obtained by different multi-direction masks as shown in Fig. 3.

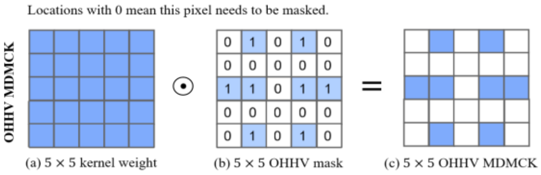

Fig. 2. The OHHV MDMCK of the 5×5 branch is shown.

Among them, the oblique refers to the noise in the diagonal direction of the image. When the noise correlation area is large, the combination of different directions can more comprehensively consider the various noise distributions in the image and break its connection, to achieve more accurate and effective large-scale noise denoising. At the same time, this MDMCK helps to improve the robustness and applicability of the image denoising model, so that it can deal with various complex noise situations.

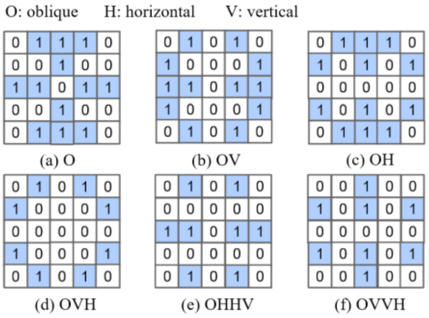

Fig. 3. The multi-direction masks are shown on 5×5 kernel. Locations with 0 mean this pixel needs to be masked. (a) represent only oblique mask. (b) represent the mask containing oblique and vertical. (c) represent the mask containing oblique and horizontal. (d) represent the mask containing oblique, vertical and horizontal. (e) represent the mask containing oblique, vertical and two horizontals. (f) represent the mask containing oblique, horizontal and two verticals.

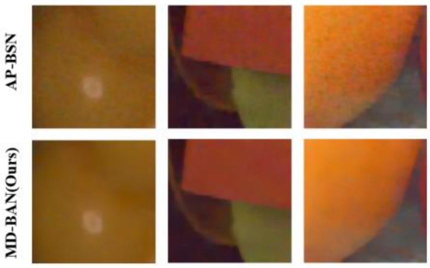

Fig. 4. The AP-BSN method fails to break the spatial correlation of large-scale noise, resulting in abnormal color spots in the denoised image. Our method can obtain cleaner denoising results through the MDMCK.

However, as the mask area increases, the pixel detail information of the image itself is more and more destroyed. How to preserve more detail information while destroying the large-scale noise spatial connection is also a challenge. Therefore, the DFE module is proposed to deal with this problem.

3.2 Detail Feature Enhancement (DFE)

The DFE module connects the features of the previous layer which contains complete detail with the more global features obtained by the dilated convolution layer, so that the model can comprehensively utilize the features of different levels. In Fig.1, because the features extracted by the 3×3 MDMCK contain complete information, then more detail can be retained by passing through several DFE modules, but this branch cannot sufficiently break the noise correlation. On the contrary, the 5×5 MDMCK masks more surrounding pixels, which can more effectively break the spatial connection of noise, but it loses more pixel detail. Therefore, we can combine the features extracted by the 3×3 branch to provide more details and the 5×5 branch to break the spatial correlation of noise and reinforce each other to obtain better denoising results.

Based on the above analysis. The input feature $I_D \in R^{C\times H\times W}$ is sequentially passed through one dilated convolution layer and LeakyReLU, $Conv_{r=2}(\cdot)$ or $Conv_{r=3}(\cdot)$ with a convolution kernel size of 3×3, where r denotes the dilated rate, and the feature information $O_{cat} \in R^{2C\times H\times W}$ is obtained by using $Concat(\cdot)$ to channel-summing from the dilated convolution layer and the features of the previous layer. The complete calculation process is shown in Eq. (1):

$$O_{cat} = Concat(I_D, leakyReLU(Conv_{r=stride}(I_D))). \tag{1}$$

Then, feature fusion and channel processing are performed using 1×1 convolution layer, $Conv(\cdot)$ on $O_{cat} \in R^{2C\times H\times W}$. Finally, the skip connection is used to transfer the shallow feature information to the deeper convolution layer to output the feature $O_D \in R^{C\times H\times W}$. The complete calculation process is shown in Eq. (2):

$$O_D = I_D + (Conv(O_{cat})) \tag{2}$$

In addition, partial pixel detail information loss due to the interpolation and overlapping receptive fields can also be alleviated by introducing the DFE module. The ablation experiments of the DFE module are detailed in Section 4.5 Table 3.

3.3 Loss function

Since the L1 norm is linear, it is relatively insensitive to outliers. Self-supervised with the Charbonnier loss function in Eq. (3) can better deal with outliers and improve performance, introducing a constant ϵ to better deal with outliers. When the $diff$ is small, the value of the loss function is mainly determined by the constant ϵ, that is, the loss remains smooth near ϵ. On the contrary, When the $diff$ is large, the loss is mainly determined by the square root of the $diff$, which can slow down the impact of outliers on the loss.

$$\mathcal{L}_{Charbonnier\ Loss} = \sqrt{diff^2 + \epsilon^2} \tag{3}$$

Where $\mathcal{L}_{Charbonnier\ Loss}$ is the Charbonnier penalty function, $diff$ is the difference between the denoising output and the noisy image of the model; We empirically set ϵ to $1\mathrm{e}-3$.

Combined with the sparsity of L1 norm, it can help to extract important features. At the same time, the smoothness of Charbonnier Loss and the robustness of outliers are used. By combining these two loss functions, the denoising performance and generalization of the model can be improved. Furthermore, the block artifacts caused by overlapping dilated convolution layers are weakened, which has been verified in ablation in the fourth and fifth rows of Fig. 5. The formula is as follows in Eq. (4):

$$\mathcal{L}_{TotalLoss} = w\mathcal{L}_{L1\ norm} + (1-w)\mathcal{L}_{Charbonnier\ Loss} \tag{4}$$

Where the hyperparameter w imply the contribution weight of each loss function.

4 Experiments

4.1 Datasets and Evaluation Metrics

Smartphone Image Denoising Dataset (SIDD) [3]. Contains five smartphone cameras that captured 10 different scenes under 4 specific settings and conditions, each of which captured 150 consecutive image sequences. For training, we used 320 noisy sRGB image pairs from the SIDD Medium dataset. For validation and evaluation, we used sRGB images from the SIDD validation set and SIDD benchmark set, respectively. Both contain 1280 patches with a size of 256×256, which also provide ground truth images for the validation set.

Darmstadt Noise Dataset (DND) [21]. Contains 50 noisy images for benchmarking, including indoor and outdoor scenes without the ground truth provided. The denoising results can only be obtained through the online system. Since our method does not need to consider ground truth images. Therefore, we directly use DND as the training and test set.

Metrics. Two metrics are used to evaluate the performance of the method, including peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [22]. The larger value of PSNR and SSIM implies better fidelity. LPIPS [33] results on SIDD validation dataset to measure the perceptual quality, which the lower the better and shows that denoised images are most perceptually similar to the clean images.

4.2 Training Detail

The computer hardware environment used in the whole experiment is NVIDIA Titan V GPU; the software environment is Windows 10 operating system; the running environment is Python3.8, PyTorch1.8.2, and Pycharm2022.3.3. The learning rate is initially set to $1e$-4, where the Adam optimizer is adopted. The number of epochs is set to 30 and the batch size is set to 8. We adopt a joint loss function between noisy image and denoising output for training. We set the PD stride as 5 for training and 2 for testing, and the same post-processing as AP-BSN [9].

4.3 Quantitative Results

We validate the effectiveness of our method for real-world image denoising on the commonly used SIDD validation, SIDD benchmark, and DND benchmark datasets. In Table 1, we compare our method with some recent works, our method delivers competitive results. Compared with unsupervised methods trained on unpaired clean-noisy data, our method does not rely on additional data for synthesizing training pairs. For self-supervised methods, Noise2Void [6] and Noise2Self [4] cannot handle the noise in sRGB images due to their spatially independent noise assumptions. CVF-SID does not consider the strong spatial correlation in real noise; AP-BSN cannot break the spatial connection of large-scale noise because the PD stride factor cannot be too large; I2V requires more model parameters; Moreover, AP-BSN and CA-BSN still use a single-center pixel mask convolution kernel to generate blind-spot.

Specifically, our proposed method improves the PSNR by 1.18, 1.24 and 0.3 dB, and the SSIM also grows by 2 %, 1 % and 0.2 %, compared with the AP-BSN.

Table 1. Quantitative results on the SIDD validation, SIDD benchmark, and DND benchmark datasets. Supervised denoising and unpaired image denoising approaches leverage paired clean-noisy images while self-supervised learning methods rely on only noisy images in SIDD Medium dataset. The "AP-BSN" here is consistent with value of "AP-BSN+R^3" in the paper [9], and "-" indicates that the result was not reported in the related paper. The best and the second-best results among self-supervised methods are pointed out in bold and underlined (tilted), respectively.

	Method	SIDD validation PSNR/SSIM/LPIPS$_{\downarrow}$	SIDD benchmark PSNR$_{\uparrow}$/SSIM$_{\uparrow}$	DND benchmark PSNR/SSIM
Non-learning based	BM3D [24]	31.75/0.706/0.657	25.65/0.685	34.51/0.851
	WNNM [25]	26.31/0.524/0.635	25.78/0.809	34.67/0.865
Supervised Synthetic pairs	DnCNN [12]	26.20/0.441/0.712	23.66/0.583	32.43/0.790
	CBDNet [13]	30.83/0.754/0.288	33.28/0.868	38.05/0.942
Supervised Real pairs	DnCNN [12]	35.34/0.885/0.245	35.34/0.885	37.83/0.929
	RIDNet [26]	38.76/0.913/-	37.87/0.943	39.25/0.952
	N2C [27]	38.98/0.954/0.201	38.92/0.953	39.37/0.954
Unpaired	C2N [28] + DIDN [29]	35.39/0.891/0.192	35.35/0.930	38.14/0.941
Self-supervised	Noise2Void [6]	27.48/0.664/0.592	27.68/0.668	-
	Noise2Self [4]	29.94/0.782/0.556	29.59/0.808	-
	NAC [23]	-	-	36.20/0.925
	CVF-SID [16]	34.17/0.872/0.423	34.71/0.917	36.50/0.924
	AP-BSN [9]	36.02/0.872/0.281	35.97/0.925	38.09/0.937
	CA-BSN [11]	-	36.92/0.932	*38.24*/**0.939**
	I2V [15]	36.63/0.888/-	36.52/0.931	38.08/*0.938*
	Li *et al.* [31]	**37.39**/**0.934**/*0.176*	**37.41**/*0.934*	38.18/*0.938*
	MD-BAN(Ours)	*37.20*/*0.892*/**0.140**	*37.21*/**0.935**	**38.39/0.939**

4.4 Quantitative Results

In Fig. 5, we show five large-scale noise images from the SIDD benchmark processed by different denoising models in Table 1. C2N+DIDN denoising image still contains obvious noise in Fig. 5 (b). CVF-SID does not consider the spatial correlation of real-world noise, and the edge of the denoised image is very blurred in Fig. 5 (c). AP-BSN generates denoising results with more color spots and block artifacts in Fig. 5 (d). In contrast, the denoised image by our method contains relatively clean and sharp detail in Fig. 5 (e). Visualization results of AP-BSN and MD-BAN on the SIDD validation dataset and DND benchmark dataset can be found in Fig. 6.

In general, our proposed method can break the spatial connection of noise from multi-direction and preserve more detail. The color spots in the image are removed more thoroughly in the first and second rows of Fig. 5, the detail is displayed more clearly in the third row of Fig. 5, and the block artifacts are also weakened in the fourth and fifth rows of Fig. 5.

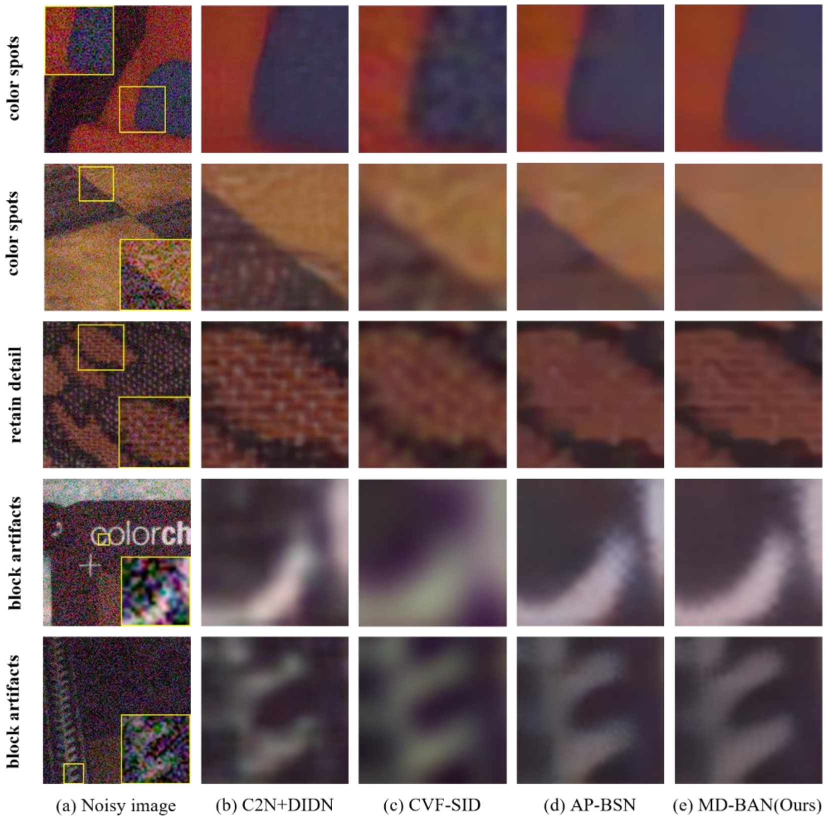

Fig. 5. Visual comparison of SIDD benchmark. In the SIDD benchmark, PSNR and SSIM of the image is not available. We zoomed in locally for a more visual comparison with the original noisy image, C2N+DIDN, CVF-SID, AP-BSN and MD-BAN(Ours).

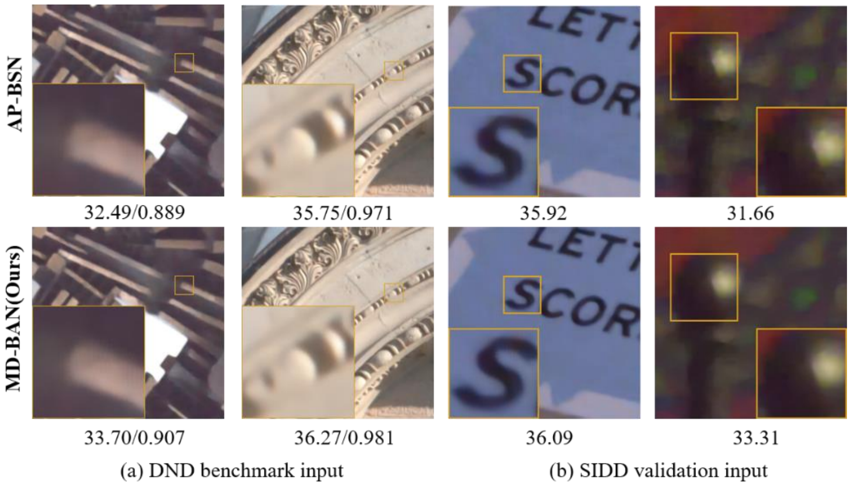

Fig. 6. Visual comparison of SIDD validation dataset and DND benchmark dataset. In SIDD benchmark dataset, the PSNR and SSIM of each image are reported below.

4.5 Ablation Studies

Effect of MDMCK. To demonstrate the effect of MDMCK, we remove the joint loss function. Since the 5×5 MDMCK will cause the loss of some pixel information, and the DFE module can complement this loss, the combination of MDMCK and DFE for ablation study can more intuitively and accurately find the most suitable MDMCK method. As shown in Table 2, it was found that OHHV had the best effect, we speculate that this is because it combines the three directions of O, V and H, and the dataset contains more horizontal noise distribution. In addition, we also performed ablation of OV and OHHV separately, and found that OHHV had a better effect, which confirmed our thought.

Table 2. Ablation study of DFE module with different MDMCK methods and separate experiment of the OV and OHHV, on SIDD validation dataset and SIDD benchmark dataset with PSNR (dB)/SSIM. The best result is marked in bold.

3×3 MDMCK	5×5 MDMCK	DFE	SIDD validation	SIDD benchmark
			PSNR (dB)/SSIM	PSNR (dB)/SSIM
center	OV	×	31.31/0.778	31.20/0.837
center	OHHV	×	34.23/0.832	34.15/0.892
center	O	√	37.15/**0.880**	37.13/0.934
center	OV	√	37.17/0.878	37.15/0.934
center	OH	√	37.04/**0.880**	37.01/0.933
center	OVH	√	37.09/0.877	37.09/0.933
center	OHHV	√	**37.20/0.880**	**37.18/0.935**
center	OVVH	√	37.15/**0.880**	37.14/0.934

Effect of DFE. To validate the effect of the DFE module. We performed ablation study on the two branches respectively in Table 3. The results of case 1 and case 2 show that the supplement of 5×5 branch detail mainly comes from the branch of 3×3. The result of case 3 is the best, since the DFE module not only complements the detail destroyed by the OHHV mask, but also alleviates the pixel information lost in the process of using dilated convolution. Note that DCL is the module used by AP-BSN.

Table 3. Ablation study of DFE module on SIDD benchmark dataset, where 3×3 represents the branch using a 3×3 central single pixel mask convolution kernel, 5×5 indicates that the branch uses the OHHV mask convolution kernel. The best result is marked in bold.

Case	Method	PSNR (dB)	SSIM	Params (M)
1	Replacing 3×3 DCL with DFE	37.16	**0.935**	4.8
2	Replacing 5×5 DCL with DFE	37.10	0.934	4.8
3	3×3 and 5×5 both DFE	**37.18**	**0.935**	5.0
4	3×3 and 5×5 both DCL	34.15	0.892	3.7

Effect of Loss Function. To validate the effect of joint loss function, we performed ablation study of hyperparameter w in Table 4, which can observe that the best result is achieved when $w = 0.9$. As illustrated in Fig. 7, the network optimized with L1 norm (green point curve) requires more iterations to achieve comparable performance with our model (red solid curve). In the fourth and fifth rows of Fig. 5, we show that the network trained with only L1 norm generates denoised results with more block artifacts. In contrast, using joint loss function contains relatively clean and sharp detail.

Table 4. Ablation study of hyperparameter w in SIDD benchmark dataset.

w	0.9	0.8	0.7	0.6	0.5
PSNR (dB)/SSIM	**37.21/0.935**	37.15/0.934	37.10/0.934	37.15/0.934	37.10/0.934

Effect of Full Model. To validate the effect of the MDMCK, DFE module and joint loss function. As illustrated in Table 5 and Fig. 7, we replace each component with the existing method. From the results, we can observe that incorporating each component has a clear contribution. Moreover, in Fig. 8 we vividly illustrate our method can effectively remove large-scale noise.

Table 5. Ablation study of MDMCK, DFE module and joint loss function in SIDD validation dataset and SIDD benchmark dataset.

Contribution			SIDD validation	SIDD benchmark
MDMCK	DFE	Joint loss function	PSNR (dB)/SSIM	PSNR (dB)/SSIM
center			36.02/0.872	35.97/0.925
√			34.23/0.832	34.15/0.892
	√		36.98/0.879	36.96/0.932
		√	36.86/0.878	36.83/0.932
√	√		**37.20**/0.880	37.18/**0.935**
	√	√	37.02/0.881	37.00/0.933
√	√	√	**37.20/0.892**	**37.21/0.935**

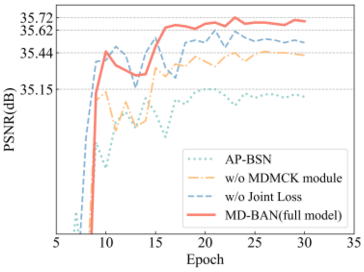

Fig. 7. Convergence analysis on the MDMCK, DFE module, and joint loss function. Our model (red solid curve) converges faster and achieves improved performance.

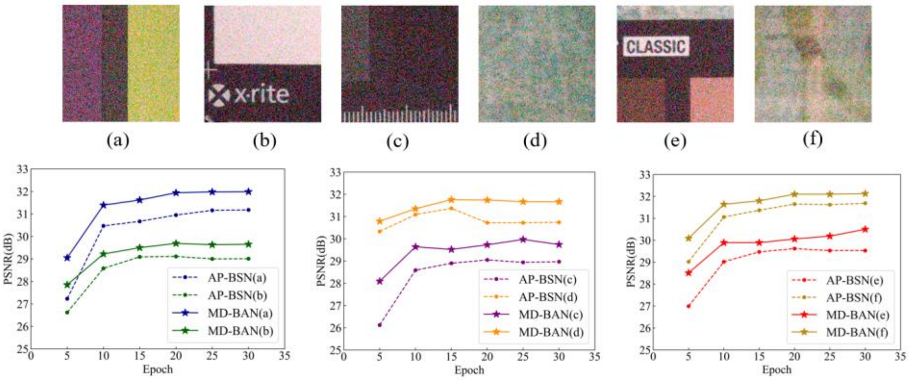

Fig. 8. Quantitative comparison of large-scale noise removal. Among them, (a)-(f) are images with large-scale noise, by comparing PSNR values of epoch from 5 to 30 of AP-BSN and MD-BAN (full model) in the above images. MD-BAN has higher PSNR than AP-BSN at each epoch, this phenomenon proves the effectiveness of MD-BAN which can better remove large-scale noise.

5 Conclusion

In this paper, we propose MD-BAN for self-supervised real-world image denoising, aiming to remove more large-scale noise image while retaining the detail information of the image. First, we propose MDMCK to comprehensively break the spatial connection of large-scale noise from multi-direction, which cannot be commendably removed by PD with a stride factor of 5 and the single-center pixel mask. Second, we propose the DFE module, which effectively combines MDMCK and supplements detail lost by MDMCK and stacking dilated convolution layers. Finally, we train our model with a robust joint loss function and generate denoised images with cleaner and sharper detail. Extensive results on real-world sRGB benchmark datasets reveal the superior denoising performance of MD-BAN.

References

1. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 60–65 (2005)
2. Dabov, K., Foi, A., Katkovnik, V., et al.: Image denoising by sparse 3-D transform-domain collaborative filtering. In: IEEE Transactions on Image Processing, pp. 2080–2095 (2007)
3. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1692–1700 (2018)
4. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. In International Conference on Machine Learning, pp. 524–533 (2019)
5. Huang, T., Li, S., Jia, X., et al.: Neighbor2neighbor: Self-supervised denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14781–14790 (2021)
6. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2129–2137 (2019)
7. Lehtinen, J., Munkberg, J., Hasselgren, J., et al.: Noise2noise: Learning image restoration without clean data. arXiv preprint arXiv, 1803, 04189 (2018).
8. Zhang, Y., Li, D., Law, K.L., et al.: Idr: Self-supervised image denoising via iterative data refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2098–2107 (2022)
9. Lee, W., Son, W., Lee, K.M.: Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 17725-17734 (2022)
10. Jang, Y.I., Lee, K., Park, G.Y., et al.: Self-supervised image denoising with downsampled invariance loss and conditional blind-spot network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17725-17734 (2023)
11. Cai, X., Liu, Y., Liu, S., et al.: CA-BSN: Mural Image Denoising Based on Cross–Attention Blind Spot Network. Applied Sciences, 14(2), 741 (2024).
12. Zhang, K., Zuo, W., Chen, Y., et al.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing, 26(7), 3142–3155 (2017).
13. Guo, S., Yan, Z., Zhang, K., et al.: Toward convolutional blind de-noising of real photographs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1712–1722 (2019)
14. Zhao, H., Shao, W., Bao, B., et al.: A simple and robust deep convolutional approach to blind image denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)
15. Lee, K., Lee, K., Jeong, W.K.: I2V: Towards Texture-Aware Self-Supervised Blind Denoising using Self-Residual Learning for Real-World Images. arXiv preprint arXiv, 2302, 10523 (2023).
16. Neshatavar, R., Yavartanoo, M., Son, S., et al.: Cvf-sid: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17583–17591 (2022)
17. Wang, Z., Liu, J., Li, G., et al.: Blind2unblind: Self-supervised image denoising with visible blind spots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2027–2036 (2022)

18. Zhang, D., Zhou, F., Jiang, Y., et al.: Mmbsn: Self-supervised image denoising for real-world with multi-mask based on blind-spot network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4188–4197 (2023)
19. Zhang, D., Zhou, F.: Self-supervised image denoising for real-world images with context-aware transformer. IEEE Access, 11, 14340–14349 (2023).
20. Wang, Z., Fu, Y., Liu, J., et al.: Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18156–18165 (2023)
21. Plotz, T., Roth, S.: Benchmarking denoising algorithms with real photographs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1586–1595 (2017)
22. Wang, Z., Bovik, A.Z., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. IEEE TIP, 13(4), 600–612 (2004).
23. Xu, J., Huang, Y., Cheng, M.M., et al.: Noisy-as-clean: Learning self-supervised denoising from corrupted image. IEEE TIP, 29, 9316–9329 (2020).
24. Dabov, K., Foi, A., Katkovnik, V., et al.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE TIP, 16(8), 2080–2095 (2007).
25. Gu, S., Zhang, L., Zuo, W., et al.: Weighted nuclear norm minimization with application to image denoising. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2862–2869 (2014)
26. Anwar, S., Barnes, N.: Real image denoising with feature attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3155–3164 (2019)
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, pp. 234–241 (2015)
28. Jang, G., Lee, W., Son, S., et al.: C2N: Practical generative noise modeling for real-world denoising. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2350–2359 (2021)
29. Yu, S., Park, B., Jeong, J.: Deep iterative down-up CNN for image denoising. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp: 0–0 (2019)
30. Quan, Y., Chen, M., Pang, T., et al.: Self2self with dropout: Learning self-supervised denoising from single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1890-1898 (2020).
31. Li, J., Zhang, Z., Liu, X., et al.: Spatially adaptive self-supervised learning for real-world image denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9914-9924 (2023)
32. Luiken, N., Ravasi, M., Birnie, C.: Integrating self-supervised denoising in inversion-based seismic deblending. Geophysics, 89(1), WA39-WA51 (2024).
33. Zhang, R., Isola, P., Efros, A. A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586-595 (2018).