Complementary Blind-Spot Network for Self-Supervised Real Image Denoising

Linwei Fan, Jin Cui, Huiyu Li, Xiaoyu Yan, Hui Liu, and Caiming Zhang

***Abstract*— Recently, self-supervised denoising methods have attracted significant attention due to the considerable challenge posed by constructing a large-scale real noise dataset for supervised training. The most representative self-supervised denoisers are based on blind-spot networks (BSNs), which exclude the central pixel of receptive field. However, excluding any input pixel potentially leads to the loss of vital information required for accurate predictions, especially when the excluded pixel corresponds to the output position. In addition, a standard BSN has struggled to effectively reduce real-world noise due to the spatial correlation of noise, though it makes the significant results with independently distributed synthetic noise. In this paper, we propose a novel self-supervised real-world image denoising framework called Complementary-BSN based on two reciprocal branches (Mask-Map branch and Enhanced-PD-BSN branch) with an efficient loss function to employ the pixels information ignored by masked convolution and provide additional optimization target for self-supervised output. Specifically, we exploit a block-wise random-placing (BRP) scheme for further weaken the noisy correlation to avoid the illusion of image structure recovery due to existing complex noise and make Complementary-BSN more suitable for real noise. Additionally, we develop an efficient strategy (multi-stride PD (MPD)) to fuse multiple PD strides for inference, narrowing the restoration gap between textural and flat regions. Extensive experiments on real-world datasets demonstrate that our method achieves superior performance to other state-of-the-art (SOTA) self-supervised denoising methods. The code is available at https://github.com/cuijin7382/Complementary-BSN.**

***Index Terms*— Self-supervised denoising, real-world noise, blind-spot network.**

I. Introduction

IMAGE denoising is a fundamental research task in low-level vision, which aims at restoring clean images from noisy observations [1]. Recently, with the advancements of convolutional neural networks (CNNs), learning-based denoising algorithms have achieved remarkable progress compared to traditional non-learning-based methods [2], [3], [4], [5], [6], [7], [8], [9]. In general, learning-based image denoising techniques can be broadly categorized into two main fields, supervised methods and self-supervised methods.

The supervised denoising methods [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28] have afforded exceptional denoising performance by using clean-noisy training image pairs. For example, IENet [23] builds a denoising network that exploits external resemblances to acquire internal and external associations. APD-Nets [26] proposals an adaptive prior denoiser that jointly employs adaptive regularisation and complementary priors. These supervised techniques typically rely on synthesized additive white Gaussian noise (AWGN) to generate massive training data. Nevertheless, acquiring authentic paired noisy-clean images poses a significant challenge. Furthermore, the notable disparities between AWGN and real-world noise undermine the performance of these models in real-world image denoising scenarios.

To circumvent the limitation of the large-scale paired datasets collection, several self-supervised image denoising (SSID) methods [10], [11], [29], [30], [31], [32], [33], [34], [35], [36], [37] are proposed that requires only noisy images without clean images. The pioneer work Noise2Noise (N2N) [32] trains the network with two fully aligned noisy images, which can effectively relieve the situation of collecting the clean-noisy pairs in the real world. However, it is still a costly work to obtain such noisy-noisy image pairs in practice. In order to further improve the practicality of self-supervised methods, more approaches intend to require a single noisy image for training, instead of pairs of observations. Among them, BSNs [35], [38], [39] show significant advancements in the recovery of clean pixels by utilizing noisy pixels in the adjacent area, with a special requirement of a blind spot receptive field (mask the central pixel of receptive field). Despite the blind-spot strategy can effectively avoid the identity mapping by learning to predict artificially the masked pixel using its surrounding pixels, the prerequisite of BSN presumes that noise is unrelated to pixel. This is definitely not accurate for real-world noise, which often exhibits a complex distribution and strong spatial correlation.

Most recently, several attempts have been launched to tackle the constraints of the aforementioned BSN-based approaches. Lee et al. [11] propose AP-BSN, a combination of asymmetric pixel-shuffle downsampling (AP) and BSN, to effectively handle real-world noise with strong spatial

Manuscript received 5 January 2024; revised 26 March 2024 and 6 May 2024; accepted 14 May 2024. Date of publication 16 May 2024; date of current version 30 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62002200 and Grant 62202268, in part by the Natural Science Foundation of Shandong Province under Grant ZR2020QF012 and Grant ZR2021QF134, in part by Shandong Provincial Science and Technology Support Program of Youth Innovation Team in Colleges under Grant 2021KJ069, in part by the Social Science Planning Project of Shandong Province under Grant 22DGLJO11, and in part by Taishan Scholars Project of Shandong Province under Grant tstp20221137. This article was recommended by Associate Editor F. Zhang. *(Corresponding author: Huiyu Li.)*

Linwei Fan, Jin Cui, Xiaoyu Yan, and Hui Liu are with the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China, and also with the Shandong Key Laboratory of Digital Media Technology, Jinan 250014, China (e-mail: lwfan129@163.com; 19863430650@163.com; yanxiaoyu12@outlook.com; liuh lh@126.com).

Huiyu Li is with the School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014, China (e-mail: huiyuroy@163.com).

Caiming Zhang is with the School of Software, Shandong University, Jinan 250101, China, and also with the Shandong Key Laboratory of Digital Media Technology, Jinan 250014, China (e-mail: czhang@sdu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3402095.

Digital Object Identifier 10.1109/TCSVT.2024.3402095

correlation. Although pixel-shuffle downsampling (PD) [40] can be used to meet the noise assumptions of BSN, the masking mechanism in AP-BSN only shields the central pixel within the receptive field, leading to a challenge when dealing with large areas of spatially correlated noise. After that, several BSN denoising methods [12], [13], [41] are proposed to improve upon AP-BSN through multiple techniques and obtain better results, but these methods fail to utilize the most informative central pixel, resulting in performance degradation.

In this paper, we propose a novel real-world image denoiser based on self-supervised learning, named Complementary-BSN, to address the above mentioned insufficience of BSN-based denoising methods. First, we introduce a Mask-Map branch that utilizes global-aware mask mapper [29] and non-blind training to complement information of the central pixels. Specially, the Mask-Map branch can reserve the central pixels and supplement enrich information to another branch in complementary network, in this way the two branches outputs contain more accurate image structure than single branch output. In addition, predicted result of Mask-Map branch can provide a higher quality constraint for the self-supervised strategy than single noisy image, which provides an upper limit on the optimized goal of the network. Herein, for stabilizing the training and provide new supervised target, we augment the loss function with a new revisible loss and ensure the integrity of the information. Second, in Enhanced-PD-BSN branch, we exploit a novel BRP strategy as post-processing of PD to eliminate the strongly spatial correlation between pixels, by this way, when fixing the small stride, the distance between the relevant pixels can still further enhance and without inducing more visual artifacts, which prevents the image structure from being affected by noise and artifacts. Finally, we propose an enhanced inference scheme (MPD) that processes the restored result with different PD strides and seeks for a better trade-off of noise removal between flat and textural areas. We demonstrate that Complementary-BSN outperforms several existing self-supervised blind denoisers on real-world noisy images in terms of various image quality metrics, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [42], learned perceptual image patch similarity (LPIPS) [43] and deep image structure and texture similarity (DISTS) [44]. In summary, the major contributions of our method are as follows:

- We present a novel self-supervised denoising framework, where the network can learn the ignored information of the central pixel within receptive field through controllable masked convolution and a novel revisible loss is also introduced to the network to prevent the network from over-fitting the information of central pixels.
- We provide a new BRP strategy for randomly rearranging the image in block-wise after PD to enhance independence of pixels and mitigate the aliasing artifacts.
- We design an effective inference scheme with multiple PD-strides to narrow the gap between flat and textural areas in the restored result with different strides of PD.

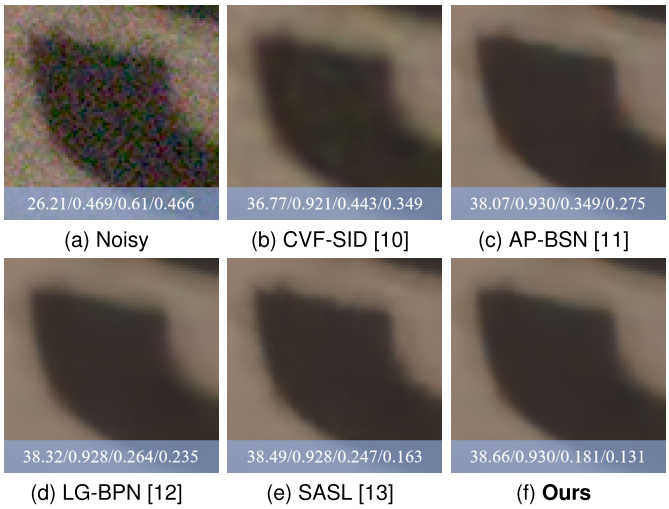


Fig. 1. Visualization of real-world image denoising on the SIDD validation dataset [14]. (a) Noisy image. (b) CVF-SID [10] still exhibits residual noise in the restoration process, indicating that it fails to completely eliminate real-world noise. (c) The main limitation of AP-BSN [11] is the loss of details in the denoised image. (d) LG-BPN [12] makes the denoised image over-smooth. (e) SASL [13] exists artifacts at the edges of denoised. (f) Our Complementary-BSN preserves better details in removing real-world noise. The PSNR↑(dB) / SSIM↑ / LPIPS↓ / DISTS↓ results in comparison to the ground-truth are marked in the figure.

- Extensive experiments demonstrate that the superiority of our Complementary-BSN over leading SSID approaches with fine texture details preservation on real-world noise removal, as visually depicted in Fig. 1.

The remainder of this paper is organized as follows. In Section II, we offer a concise overview of existing denoising approaches. Section III introduces Complementary-BSN, a novel self-supervised learning framework designed for real-world noise reduction. Section IV provides comparative experiments that demonstrate the effectiveness of our SSID approach compared to other SOTA methods, along with several ablation studies of the proposed modules. Finally, Section V summarizes our findings and outlines future directions.

II. Related Works

In this section, we review the most relevant work with this paper, including supervised, unpaired image and self-supervised denoising methods.

A. Supervised Denoising

In recent years, CNN-based supervised denoising methods outperform classical non-learning-based algorithms (e.g., BM3D [3] and WNNM [6]) both on process speed and quality. The first CNN-based method, DnCNN [15], used for supervised denoising, trains the network with noisy-clean image pairs that are synthesized by manually adding synthetic AWGN to clean images. In the following, several advance methods are proposed for AWGN removal, such as FFDNet [19] and CBDNet [20]. Although these methods bring the outstanding performance on denoising tasks with synthetic AWGN noise, their generalization ability is limited when applied to real-world scenarios, owing to the different characteristics of synthetic and real noise. To this end,

several powerful methods [18], [45], [46], [47], [48], [49], such as RIDNet [18] and VDN [45], are proposed for real image denoising, which directly use noisy-clean pairs from the real world when available. However, data collection of such real image training pairs can be extremely expensive and labor-intensive in practice.

B. Unpaired Image Denoising

To address the limitations of supervised denoising methods, unpaired methods [38], [50], [51], [52] seek to synthesize noisy samples from clean images, allowing networks to be trained with unpaired noisy-clean data. GCBD [50] is the first unpaired approach for blind denoising using a generative adversarial network (GAN) [53], but it has a misunderstanding of the properties of real noise leads to poor application performance. Recently, C2N [51] considers various noise elements, including signal-independent, dependent, and spatially-correlated noise, for simulating real-world noise more accurately. Additionally, Wu et al. [38] develop a self-supervised denoising network that learns the noise distribution using synthetic noisy-clean pairs. However, since these approaches require ground-truth to generate the corresponding targets, the mismatch in scene distribution can degrade the quality of the synthesized data.

C. Self-Supervised Denoising

To eliminate the reliance on clean images, self-supervised image denoising methods use only noisy images to train the network. For instance, N2N [32] trains a CNN with a pair of entirely aligned noisy images of the identical scene as input and target, respectively. However, these noisy-noisy pairs are still difficult to collect in practice. To utilize the assumption of N2N, Neighbor2Neighbor (NBR2NBR) [54] forms noisy-noisy pairs for self-supervision training by sampling the noisy input into two sub-noisy images. However, subsampling inevitably destroy structural continuity. Subsequently, Noise2Void (N2V) [31] and Noise2Self (N2S) [35] discover that training a denoising network with single noisy image offers advantages over paired-image denoising. Based on this insight, researchers propose integrating BSN with self-supervised learning by masking the central pixel of the receptive field. Laine19 [39] and D-BSN [38] further optimize the BSN with effective network architectures,while they suffer from information loss due to blind spots, leading to over-smoothing and damaged details in the restored image. To improve this, Blind2Unblind (B2UB) [29] utilizes a novel loss which using all the pixels of input image for self-supervised training, making blind-spot visible again. In addition, Noisy-As-Clean (NAC) [34] takes noisy images as targets, and treats as clean images for denoising model training. Nevertheless, while the above self-supervised methods are effective in pixel-wise independent noise for denoising, they fail to tackle real-world noisy images that exhibit spatially correlated noise.

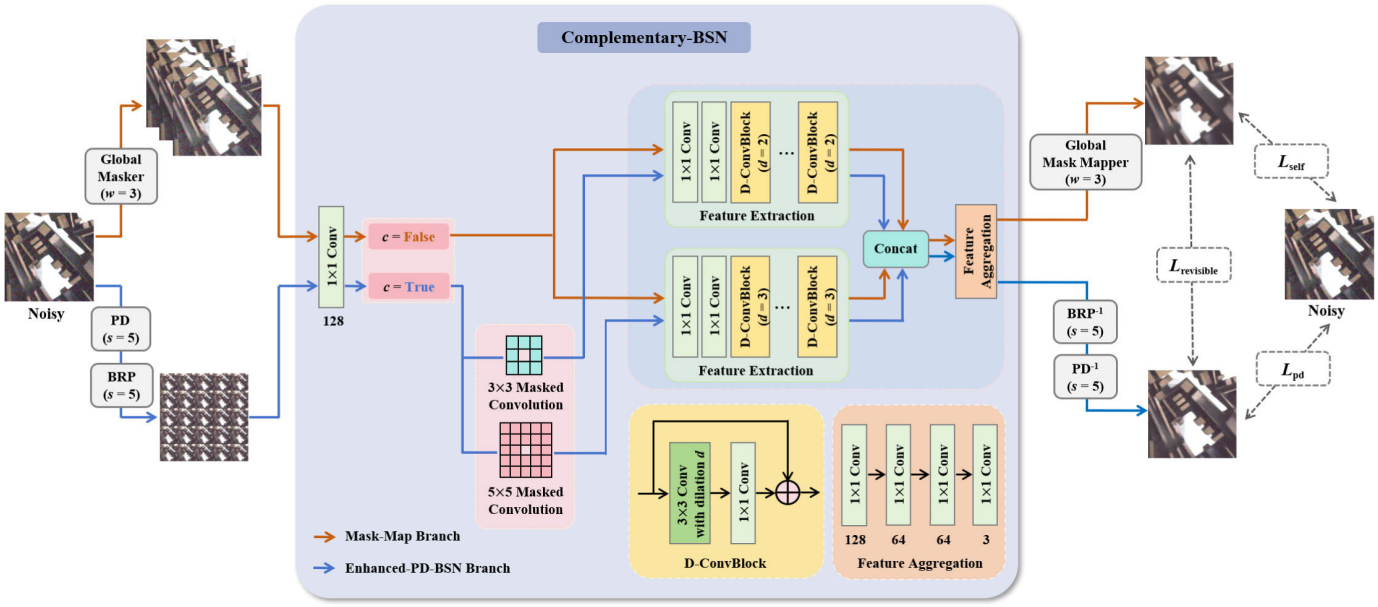
Towards real image denoising, numerous approaches have been devised to tackle spatially correlated noise through self-supervised learning. Recorrupted-to-Recorrupted (R2R) [55] uses data augmentation technique to get noisy-noisy pairs, then train a denoising model with them. However, its practicality is hampered by the requirement for prior knowledge of the noise level, which can often be difficult to obtain in the real world. Another method, CVF-SID [10], aims to separate real-world noisy images into clean images and two different noise components. Nevertheless, its reliance on the assumption of spatially independent noise limits its applicability to the real-world noise distribution. Recently, in an effort to dismantle spatially-correlated real-world noise into pixel-wise independent components, AP-BSN [11] introduces asymmetric PD [40] into the BSN framework to restore real-world noisy images. During training, using a large stride to hold the assumption of pixel-wise independence, while during inference, a small stride is used to preserve more texture information. In addition, random-replacing refinement is adopted to mitigate the negative aliasing artifacts and enhance image texture details. However, AP-BSN applies the PD with a fixed stride, which makes the denoising performance greatly reduced especially when confronted with substantial noise in the image. Unlike AP-BSN, MM-BSN [41] uses multi-masks to mask surrounding pixels at varying positions within the convolution kernel, which can effectively break the large-area spatial correlation of noise and enhance the denoising capability of the model. Li et al. [13] propose a BSN with spatially adaptive supervision, where blind-neighborhood network (BNN) deformed from BSN and locally aware network (LAN) are used to learn supervisions for flat and texture areas, respectively. Wang et al. [12] combine the strengths of BSN and Transformer applying to the recovery of real-world images. This integration not only leverages the structural details but also harnesses the global context information, thereby enhancing the overall recovery process. However, the Transformer branch increases the computational complexity of the method, limiting its practical use in certain applications. These methods do not considered the problem of causing the damage to image details by missing large important information by masking central pixels.

III. Method

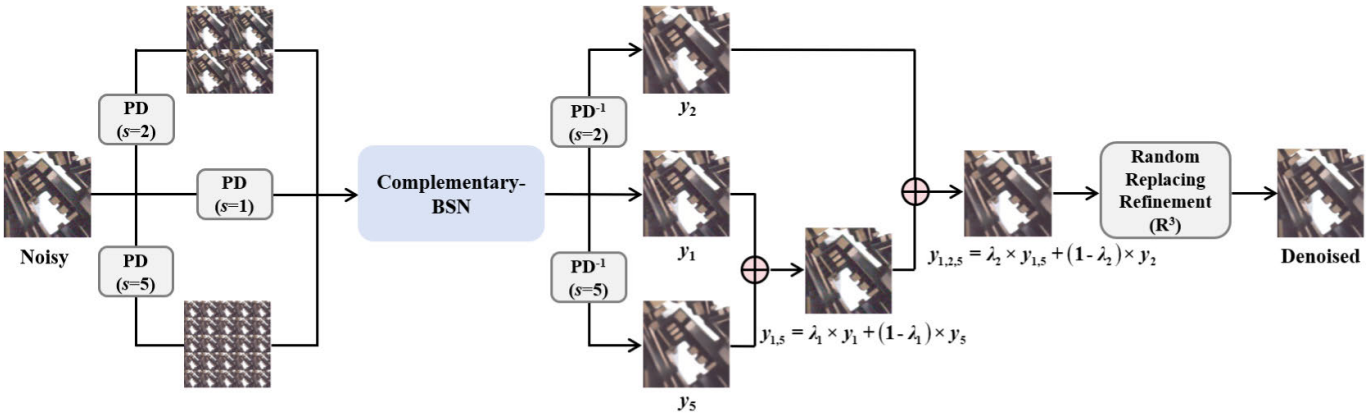
In this section, we first present an effective framework based on self-supervised learning for real-world RGB images, which is illustrated in Fig. 2. Then, we elaborate on our motivation, and introduce the details of Complementary-BSN including two branches, the proposed loss function and inference scheme (MPD).

A. Motivation and Modeling

Real-world noise is characterised by spatial correlation and pixel-wise dependence, which is affected by image signal processors (ISP). Therefore, the task of eliminating real-world noise in a self-supervised manner presents a formidable challenge. The SOTA self-supervised denoiser AP-BSN [11] deals with spatially correlated real-world noise by utilizing BSN and PD [40] asymmetrically in training and inference stages. The PD operation can make noise spatially irrelevant by separating the distance between adjacent pixels. During the



(a) Training phase



(b) Inference phase

Fig. 2. Overview of our self-supervised denoising framework. (a) In training stage, our network is composed of two branches (Mask-Map branch and Enhanced-PD-BSN branch), aiming at selectively control the network's blindness to utilize information of the central pixel. For Mask-Map branch, the global masker generates a masked volume from the noisy image as the input by introducing blind spots. Meanwhile, in Enhanced-PD-BSN branch, the input is the sub-images derived from PD_s and BRP_s operations. Then, the input goes through a 1×1 convolution and a judgement condition variable c at each branch. More specifically, the masked convolution is avoid for Mask-Map branch when $c =$ False, while the BSN adopts 3×3 and 5×5 centrally masked convolutions for the input with $c =$ True from Enhanced-PD-BSN branch. And the next step is further processed through a series of D-ConvBlocks and convolutions for the deep features. Herein, each D-ConvBlock consists of one 3×3 dilated convolution with a dilation d ($d = 2$ for the upper path and $d = 3$ for the lower path), followed by a 1×1 convolution and a residual skip connection. The number of output channels, defaulted to 128, is indicated below each convolution layer. Note that during the calculation of the total loss, the denoised volume is sampled by the global mask mapper using a width w identical to that of the global masker and the inverse operations PD_s^{-1} and BRP_s^{-1} are also performed with the same stride factor s. (b) During inference, we employ the proposed MPD technique with varying stride factors ($s = 1$, 2 and 5). This approach combines the strengths of different PD strides to optimize performance across various image regions. Specifically, a larger stride factor of 5 is employed in flat regions to achieve superior denoising results. In textured areas, a stride factor of 2 is used to effectively avoid aliasing artifacts. Additionally, a stride factor of 1 is applied to preserve texture details of the original image. Meanwhile, the same post-processing technique (R^3) as in AP-BSN [11] is incorporated to further enhances the overall performance.

training phase, this method adopts a substantially large PD stride factor ($s = 5$) to guarantee a pixel-wise independent constraint. Conversely, for inference, it utilizes a small PD stride factor ($s = 2$) to maximize reconstruction quality. Although AP-BSN employing two different stride PD was an appropriate choice for mitigating the spatial correlation of real noise and protecting texture details, its potential is still severely curtailed by inherent bottlenecks, i.e., the limitations on masked pixels resulting in abandoning the most informative central pixels and not considering the distinct attributes of PD stride on the recovery of flat and textured areas.

Specifically, as shown in Fig. 3, the traditional masked convolution principle of BSN is to exclude the pixel itself from the receptive field of each pixel by setting central pixels to zero, and then predicts the masked pixel utilizing its adjacent pixels. Although this network can avoid identity mapping, excluding any input pixels implies abandoning some information, which ultimately result in the restored image with structural discrepancy, especially when the input pixels at the

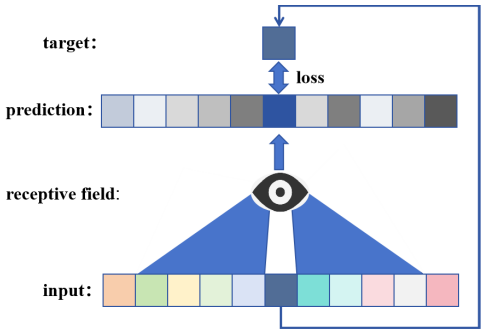


Fig. 3. Work principle of blind-spot network. A blind-spot network is trained by using the same noisy image as input and as target. Herein, each pixel will be set as a blind pixel once and predicted from other adjacent pixels. Blue patches represent receptive fields.

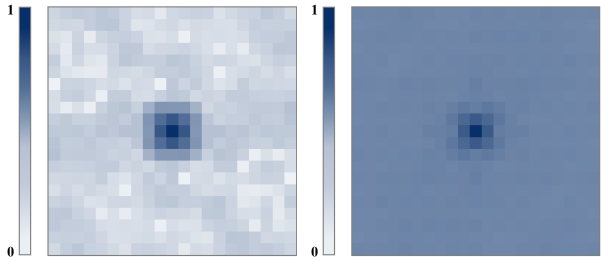


Fig. 4. Spatial correlation of real-world noise. The intensity of color represents the degree of correlation. (a) For the spatial correlation of general noisy images, the strong correlation is mainly concentrated near the central pixel. (b) For the spatial correlation of complex and densely noisy images, their long-distance pixels also exhibit strong correlation.

corresponding output positions are excluded. Thus, we aim to compensate for this deficiency by reusing these discarded central pixels and realise the maximum of data utilization. In addition, for noisy images in different real-world scenarios, the specific distribution and spatial correlation of noise are complex and diverse. From Fig. 4, we can observe that the spatial correlation of real-world noise exhibits a distribution that spreads outward from the center and decays with increasing distance. When processing the image with dense noise as shown in Fig. 4b, a small stride cannot completely break the spatial correlation of noise, while a large stride can produce more pronounced grid artifacts. This inspires us to find a flexible approach to increase the distance between dense noisy pixels without changing the PD stride, which can effectively break strong noise correlation while preventing the introduction of additional artifacts. Meanwhile, as shown in Fig. 5, we observe the characteristics of asymmetric PD strides, where different strides exhibit varying contributions to the image structure in inference. When dealing with textured areas, PD_2 acquires more prominent semantic information. However, as the stride increases, the structure becomes increasingly damaged. When restoring flat areas, PD_5 achieves the best visual and quantitative result, ensuring that local information is not separated. As a consequence, AP-BSN only uses PD_2 in the inference stage, which limits the restoration ability of the network on different image areas. In this situation, we need to balance the denoising performance of flat areas and textured areas.

Fig. 5. Visual comparison of denoising results by using different PD strides during the inference stage of AP-BSN. Quantitative results are shown with PSNR↑(dB)/SSIM↑/LPIPS↓/DISTS↓.

Hence, to tackle these challenges, we present a new self-supervised denoising framework capable of effectively removing real-world noise by utilizing disappearing pixels and multiple PD strides to compensate for the excessive blurring and loss of texture details. As shown in Fig. 2a, Complementary-BSN is mainly composed of Mask-Map branch and Enhanced-PD-BSN branch in parallel that respectively utilizes global-aware mask mapper and BRP strategy, aiming at complementing pixels information across them. After that, we utilize a novel revisible loss to minimize outputs of the above two branches in training phase. Moreover, for the inference stage, we consider fusing the PD results of different strides using weights, and thus propose an efficient inference scheme, MPD, to boost overall performance (see Fig. 2b).

B. Mask-Map Branch

In Mask-Map branch, we introduce the global-aware masker, which is shown in Fig. 6, to apply information of dropped central pixels to BSN. The global masker generates blind spots from input's pixels in a regular manner and maintains the scale of input image. Instead of leaving a zero pixel in the masked position, the global masker replaces the masked pixels with interpolation of its surrounding pixels. Then, as shown in Fig. 2a, our Complementary-BSN process the outputs of global masker without masked convolution. Thus, to selectively control the blindness of the Complementary-BSN architecture, we optimize the network by adding a judgment condition parameter c that discriminate input types. If the input of the network from the global masker (when the judgment condition parameter c is False), then it directly enters the feature extraction modules, which include dilated convolution blocks (D-ConvBlocks) under different strides with two parallel paths. Finally, the information from the two paths is fused together by feature aggregation module. Otherwise, when c = True, the image first performs center masked convolution to extract the shallow features, and then implement the same parallel feature extraction and feature aggregation operations.

Due to Mask-Map branch set masked spots to global positions by the global masker instead of central masked convolution, thus our Complementary-BSN can accept global

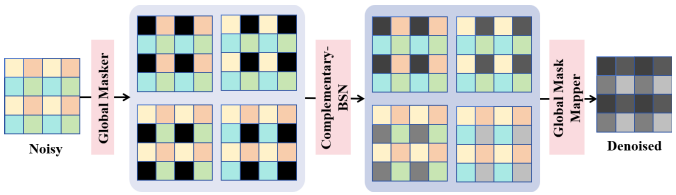


Fig. 6. Illustration of the global masker and global mask mapper. Taking a width w with 2 as an example, the size of cell is 2×2, the global masker divide noisy into blocks with fixed width and the specially appointed position of each cell are masked, the black block represent masked pixels. Then these masked images fed into Complementary-BSN in the same batch, and the gray pixels of outputs represent the network predicted pixels. Finally, the global mask mapper projects the denoised pixels onto the same plane to obtain the globally denoised output.

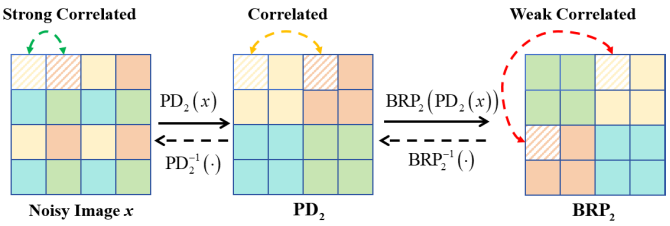


Fig. 7. Examples of PD [40] and BRP for stride factor $s = 2$. PD decomposes the image into several sub-images and BRP randomly rearranges these sub-images with the same stride after PD. BRP^{-1} and PD^{-1} restore the sub-images to their previous locations.

pixels to complement ignored central pixels in BSN and enrich the learning pixel information.

C. Enhanced-PD-BSN Branch

Enhanced-PD-BSN branch is initially motivated by AP-BSN [11]. Herein, we also utilize the original PD technique [40] to preliminarily disrupt the spatial connection between the noises of neighboring pixels in the input image. However, since PD sequentially arranges the sub-images, the pixels in the rearranged images are not fully spatially independent, as shown in Fig. 7. Instead, they still exhibit correlation within a specific distance, which can potentially impact the denoising performance.

Our solution to this difficulty is a novel BRP strategy $\mathrm{BRP}_s(\cdot)$ in Enhanced-PD-BSN branch, further suppressing the spatial correlation inherent in real-world noise. Fig. 7 shows the details of our BRP, using a stride factor of $s = 2$ as a straightforward example for clarity. First, a 4×4 input image is divided into four 2×2 sub-images by PD. Then, BRP rearranges these sub-images in a random sequence. It can be observed that BRP effectively reduces the correlation between pixels after PD by randomly shuffling the positions of sub-images. Therefore, as a result of random rearrangement, the distance of correlated pixels is not a constant, which helps to further eliminate the correlation of real noise. After PD and BRP with the stride factor s, we employ these pre-processed images as inputs to Enhanced-PD-BSN branch.

D. Total Loss Function

In this subsection, we present the comprehensive loss function that underlies our approach. For Enhanced-PD-BSN branch, we employ a self-supervised training manner by minimizing the loss function L_{pd} defined as follows:

$$L_{\mathrm{pd}} = \|I_{\mathrm{EPB}} - I_N\|_1 \tag{1}$$

$$I_{\mathrm{EPB}} = \mathrm{PD}_s^{-1}(\mathrm{BRP}_s^{-1}(f(\mathrm{BRP}_s(\mathrm{PD}_s(I_{\mathrm{N}})), c{=}\ \mathrm{True}))) \tag{2}$$

where I_{EPB} is an output from Enhanced-PD-BSN branch, I_N is the noisy input, $f(\cdot, c = \mathrm{True})$ is the BSN with masked convolutions under the condition variable c is True, PD_s is the PD operation with a stride factor s, and PD_s^{-1} is the corresponding inverse operation. Additionally, BRP_s and BRP_s^{-1} are the BRP operation with stride of s and the inverse operation of BRP, respectively. Following the previous work [11], we use L_1 norm for better generalization and set the stride factor ($s = 5$) for training to ensure pixel-wise independent constraint.

Also, for Mask-Map branch, we provide a self-supervised loss function to enhance training stability and guarantee integrity of the restoration information. The self-supervised loss L_{self} can be described as follows:

$$L_{\mathrm{self}} = \|h(f(\Omega(I_N), c = \mathrm{False})) - I_N\|_1 \tag{3}$$

where $\Omega(\cdot)$ represents a function that generates a masked volume by introducing blind spots into an image, $h(\cdot)$ is the global mask mapper responsible for selectively sampling the denoised pixels at the positions where these blind spots are present and $f(\cdot, c = \mathrm{False})$ is the denoiser without masked convolutions under c is False.

In addition, since PD and masked convolution could bring unexpected artifacts and over-smoothing to the denoised image, we consider to utilize the restored image from Mask-Map branch, which preserves the central pixel information, as a complement. Based on this observation, to minimize the differences between the outputs of Mask-Map and Enhanced-PD-BSN branches, we propose a novel revisible loss is defined as follows:

$$L_{\mathrm{revisible}} = \|h(f(\Omega(I_N), c = \mathrm{False})) \text{ - } I_{\mathrm{EPB}}\|_1 \tag{4}$$

Without the self-supervised loss, the function $f(\cdot, c = \mathrm{True})$ exhibits randomness during the initial stages of training, potentially leading to incorrect guidance for the denoised image. In general, to further optimize the primary denoised image and facilitate the transition of BSN from blind to non-blind, we define the comprehensive loss function as follows:

$$L_{\mathrm{total}} = L_{\mathrm{pd}} + L_{\mathrm{revisible}} + L_{\mathrm{self}} \tag{5}$$

E. Multi-Stride PD

During the inference of AP-BSN [11], although adopting the PD stride factor with $s = 2$ represents a delicate balance between removing the spatial correlation of noise and preserving structural information, the non-stationarity of natural image signals may alter the optimal stride factor pixel-wise. Similar to [11], we also employ stride factors during both the training and inference stages of Complementary-BSN, denoted as $\mathrm{PD}_{a/b}$. Here, a and b represent the stride factors used

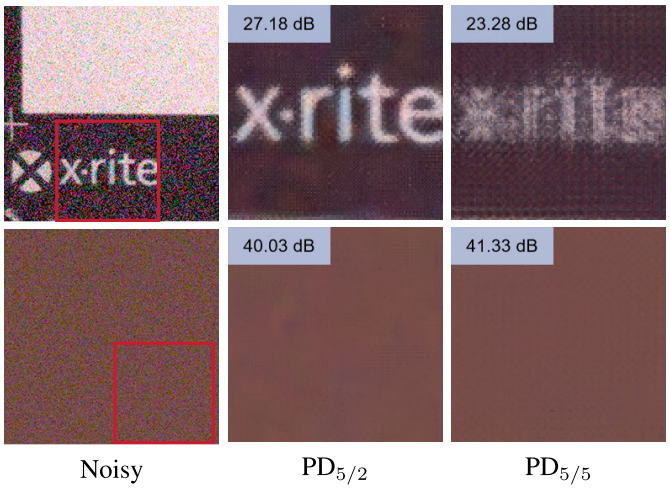


Fig. 8. Visual comparison of $PD_{5/2}$ and $PD_{5/5}$ on the textured and flat regions. The performance of inference stride factors differs with the type of region.

for training and inference, respectively. When applying $PD_{5/b}$ in Complementary-BSN, the visual comparison of $PD_{5/2}$ and $PD_{5/5}$ for two stride factors ($b = 2, 5$) in inference phase illustrated in Fig. 8. We note that as the inference stride factor b increases, the model performs more favourably in flat regions. In contrast, a larger b leads to a more pronounced performance decline in regions with complex textures. This is because, when utilizing large-stride PD, high-frequency details in the downsampled images are erroneously interpreted as noise, resulting in strong aliasing artifacts, as shown in Fig. 9. Thus, a smaller inference stride factor can retain more detailed texture and a larger inference stride factor is more suitable for the flat area.

Inspired by the above observation, we propose a novel inference scheme, MPD, to take advantages of multiple PD strides. As shown in Fig. 2b, during the inference process, we first utilize three PD stride factors ($s = 1, 2, 5$) to downsample and shuffle the noisy image, respectively. Then, we reconstruct each corresponding output of our Complementary-BSN utilizing the PD-inverse operator PD^{-1} with the same stride of its downsampling phase. Finally, for mitigating visual artifacts, we fuse these outputs and utilize the post-refinement processing to obtain the final prediction output of the noisy image. Herein, since a stride of 1 means the absence of pixel-shuffle downsampling to image, this branch would not produce any aliasing artifacts. Additionally, an inference stride of 2 helps to achieve the optimal average performance in overall denoising, as reported in [11], especially for the image with textured regions. On the other hand, an inference stride of 5 provides better performance for the image with flat regions. Therefore, given a set of the denoised images $y' = \{y_1, y_2, y_5\}$, we define the denoised predictions of MPD as follows:

$$\mathrm{MPD}(y', \lambda_1, \lambda_2) = \lambda_2 \cdot [\lambda_1 \cdot y_1 + (1 - \lambda_1) \cdot y_5] + (1 - \lambda_2) \cdot y_2 \tag{6}$$

$$y_s = \mathrm{PD}_s^{-1}(f(\mathrm{PD}_s(I_N), c = \mathrm{True}), \quad s = 1, 2, 5 \tag{7}$$

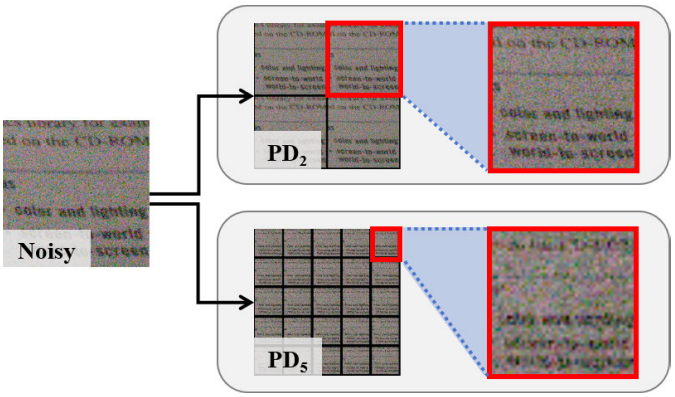


Fig. 9. Visual comparison of PD_2 and PD_5. Illustration of sub-images from PD_2 and PD_5 with red square. Large stride factor can brings more aliasing artifacts to sub-images.

where y_s is the output with PD stride of s, and the hyper-parameters λ_1 and λ_2 imply the contribution weight of each result with different strides.

IV. Experiments

This section begins with a discussion of datasets, metrics and meticulous configurations employed in training our Complementary-BSN framework. Following this, we present extensive experimental results and draw detailed comparisons with the other approaches. Finally, we undertake extensive ablation studies to analyze the efficacy of our framework.

A. Dataset

We assess the performance of our proposed framework using three widely recognized real-world datasets, including Smartphone Image Denoising Dataset (SIDD) [14], Darmstadt Noise Dataset (DND) [57] and Natural Image Noise Dataset (NIND) [58].

SIDD [14] is a comprehensive dataset designed for real-world image denoising. It encompasses paired images captured by five smartphone cameras across 10 diverse scenes. For training, we utilize sRGB images from the SIDD Medium dataset, which offers 320 image pairs of real-world noisy and corresponding clean images. For validation and evaluation, we employ sRGB images from the SIDD validation set and the SIDD benchmark set, respectively. Both sets include 1280 noisy patches (256×256). It is worth noting that the benchmark set does not provide the ground-truth images.

DND [57] consists of only 50 real noisy images from four different cameras. This dataset is typically employed as a test set and encompasses 1000 noisy patches with size of 512×512.

NIND [58] is a set of real photographs with real noise captured by a Fujifilm X-T1 and XF18-55mm. This dataset contains 22, 14, 13, and 79 clean-noisy pairs for ISO levels of 3200, 4000, 5000, and 6400, respectively.

Unless explicitly mentioned otherwise, we utilize the SIDD Medium dataset for training in our subsequent experiments. In addition, as our self-supervised model requires only noisy images to be trained, we also directly train and test

TABLE I

QUANTITATIVE RESULTS OBTAINED FROM VARIOUS DENOISING MODELS ON THE SIDD VALIDATION, SIDD BENCHMARK AND DND BENCHMARK DATASETS. THE PSNR AND SSIM VALUES OF BENCHMARK DATASETS ARE OBTAINED FROM THE OFFICIAL SIDD AND DND WEBSITES. WE EMPLOY † NOTATION TO SIGNIFY THAT THE NETWORK IS TRAINED EXCLUSIVELY IN A FULLY SELF-SUPERVISED APPROACH. ADDITIONALLY, ∗ DENOTES THAT THE APPROACH UTILIZES A SELF-ENSEMBLE STRATEGY. FOR THE SELF-SUPERVISED TECHNIQUES, THE BEST VALUE IS EMPHASIZED IN BOLD

Learning Type	Method	SIDD validation				SIDD benchmark		DND benchmark	
		PSNR↑(dB)	SSIM↑	LPIPS↓	DISTS↓	PSNR↑(dB)	SSIM↑	PSNR↑(dB)	SSIM↑
Non-learning-based	BM3D [3]	25.71	0.576	0.657	-	25.65	0685	34.51	0.851
	WNNM [6]	26.05	0.592	0.635	-	25.78	0.809	34.67	0.865
Supervised	DnCNN [15]	35.25	0.861	0.272	0.218	35.13	0.896	37.89	0.932
	CBDNet [20]	33.07	0.863	0.288	-	33.28	0.868	38.05	0.942
	RIDNet [18]	38.71	0.913	-	-	38.70	0.950	39.24	0.952
	DANet [46]	39.00	0.914	0.263	0.226	38.89	0.955	39.13	0.948
	VDN [45]	39.28	0.909	0.208	-	39.26	0.955	39.38	0.952
Unpaired image-based	C2N* [51] + DIDN [56]	35.39	0.891	0.237	0.199	35.35	0.937	36.38	0.887
Self-supervised	N2V [31]	27.06	0.651	0.468	0.332	26.99	0.652	29.23	0.765
	R2R [55]	35.04	0.844	-	-	34.78	0.898	37.61	0.936
	CVF-SID [10]	34.15	0.871	0.423	0.304	34.71	0.917	36.50	0.924
	AP-BSN [11]	36.73	0.878	0.251	0.218	35.97	0.925	38.09†	0.937†
	LG-BPN [12]	37.31	0.884	0.175	0.172	37.28	0.936	38.43†	0.942†
	SASL [13]	37.39	0.875	0.245	0.204	37.41	0.934	38.18	0.938
	Complementary-BSN (Ours)	**37.51**	**0.885**	**0.173**	**0.171**	**37.43**	**0.936**	**38.24**	**0.940**
	Complementary-BSN† (Ours)	-	-	-	-	-	-	**38.62†**	**0.942†**

Complementary-BSN† on various datasets to enjoy a fully self-supervised manner.

B. Image Quality Assessment Metric

To measure denoising quality of our Complementary-BSN and the other denoising methods, we employ PSNR, SSIM [42], LPIPS [43] and DISTS [44] metrics. Although most studies only adopt PSNR and SSIM to evaluate denoising methods, they have limitations in capturing perception-related textures because these metrics rely on pixel-level image differences. To accurately measure the restoration of detailed textures, we additionally employ LPIPS and DISTS as deep feature-based metrics for texture and detailed structural similarity. For the SIDD validation and NIND datasets, because the paired noisy-clean images are provided, we can directly test the PSNR/SSIM/LPIPS/DISTS results. For the SIDD benchmark and DND benchmark datasets, the PSNR/SSIM values for the denoising results are acquired through the official online submission system hosted on the SIDD benchmark website[1] and the DND benchmark website.[2] It is worth noting that higher PSNR/SSIM and lower LPIPS/DISTS values indicate better denoising performance.

C. Implementation Details

We adopt the general BSN architecture [38] in our model. Following [11], we empirically set the stride factor s of PD and BRP to 5 for training and set p and T to 0.16 and 8 for the post-refinement processing R^3 in inference. All the experiments are performed on a PC with an Intel i7-7700K CPU,16GB RAM and Nvidia GeForce GTX 1080Ti GPU. In addition, the code is developed on Windows with Cudnn 7.6.4, CUDA SDK 10.1, Pytorch 1.8.1 and Python 3.8. During the training, we adopt Adam optimizer with an initial learning rate of 1e-4. The network is trained with 20 epochs and batch size with 1 until achieving full convergence. To optimize Complementary-BSN, we extract 120×120 noisy patches from training images and augment them through random flipping and 90° rotation. We employed $\lambda_1 = 0.3$ and $\lambda_2 = 0.2$ for the Complementary-BSN architecture. These hyperparameters are determined by our additional experiments described in Section IV-E.

D. Comparison With SOTA Algorithms

We compare our approach with a diverse range of denoising methods, including non-learning, supervised, unpaired, and self-supervised methods. Specifically, the supervised models are trained using paired noisy-clean image pairs from SIDD. The unpaired image-based models create pseudo-paired noisy-clean images by simulating real noise and then train the denoiser using these generated pairs. The self-supervised models are trained on single noisy images. All experimental results are generated by ourselves in the same training scheme using the author's public code. We validate the efficacy of our Complementary-BSN for real-world image denoising through rigorous testing on SIDD, DND and NIND datasets.

The quantitative comparisons are shown in Tables I-II. In Table I, we can observe that our Complementary-BSN

[1]https://www.eecs.yorku.ca/ kamel/sidd/

[2]https://noise.visinf.tu-darmstadt.de/

TABLE II

QUANTITATIVE RESULTS ON THE NIND DATASET. WE EMPLOY † NOTATION TO SIGNIFY THAT THE NETWORK IS TRAINED EXCLUSIVELY IN A FULLY SELF-SUPERVISED APPROACH. FOR THE SELF-SUPERVISED TECHNIQUES, THE BEST VALUE IS EMPHASIZED IN BOLD

Learning Type	Method	NIND ISO3200				NIND ISO5000			
		PSNR↑(dB)	SSIM↑	LPIPS↓	DISTS↓	PSNR↑(dB)	SSIM↑	LPIPS↓	DISTS↓
Supervised	DnCNN [15]	33.82	0.858	0.216	0.131	32.31	0.806	0.269	0.151
	DANet [46]	35.06	0.879	0.267	0.192	33.83	0.857	0.322	0.216
	NAFNet [59]	35.04	0.880	0.251	0.174	34.12	0.864	0.287	0.184
	Restormer [60]	35.05	0.880	0.251	0.172	34.01	0.860	0.290	0.182
Unpaired image-based	C2N* [51] + DIDN [56]	34.86	0.875	0.260	0.174	33.42	0.846	0.296	0.184
Self-supervised	N2V† [31]	28.42	0.766	0.318	0.196	27.04	0.658	0.376	0.217
	NBR2NBR† [54]	29.47	0.770	0.310	0.190	28.20	0.698	0.360	0.221
	AP-BSN† [11]	34.41	0.854	0.329	0.237	33.49	0.847	0.348	0.243
	LG-BPN [12]	33.94	0.840	0.189	0.204	33.33	0.831	0.165	0.173
	Complementary-BSN (Ours)	34.33	**0.855**	**0.161**	**0.176**	33.52	0.839	**0.163**	**0.172**
	Complementary-BSN† (Ours)	**34.57**	0.849	0.187	0.194	**33.79**	**0.848**	0.167	0.178

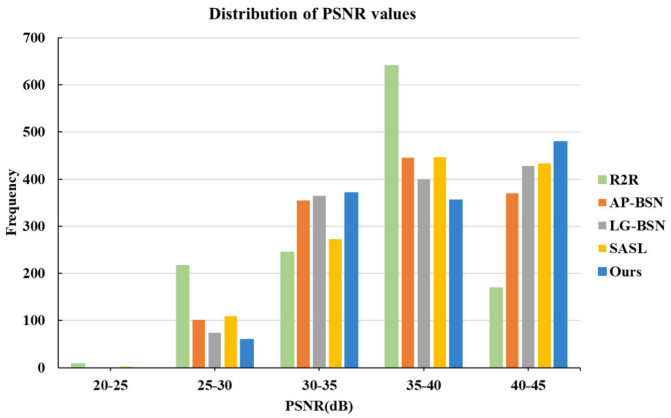


Fig. 10. The distribution of PSNR values on the SIDD validation dataset. We calculate the PSNR results of 1280 images and describe their respective frequencies across different numerical ranges. The results of our algorithm have the highest frequency in the top range.

achieves significantly better results than non-learning methods and even some supervised-based image denoising methods. For instance, on the SIDD benchmark dataset, our method outperforms DnCNN [15] by achieving a 2.30dB higher PSNR and a 0.040 increase in SSIM. Compared with the SOTA SSID approaches, our Complementary-BSN also obtains the best performance. Specifically, our method outperforms AP-BSN [11] by 1.46dB in PSNR on the SIDD benchmark dataset and comparing with the latest BSN methods, LG-BPN [12] and SASL [13], our results of PSNR on SIDD are 0.15dB and 0.02dB higher, respectively. In addition, our method shows the lower LPIPS and DISTS values of 0.078 and 0.047 on the SIDD validation dataset, respectively, when compared to AP-BSN. As shown in Table II, Complementary-BSN is ranked in first place among the SOTA self-supervised denoising methods for all metrics. As for NIND ISO5000, our method performs better than DANet, NAFNet, and Restormer in terms of LPIPS and DISTS with slightly lower value, which means our results are perceptually closer to ground-truth.

TABLE III

MACS AND INF. TIME ARE MEASURED WITH 256 × 256 PATCHES OF THE SIDD VALIDATION DATASET

Methods	Params (M)	MACs (G)	Inf. Time (s)
AP-BSN [11]	3.09	1888	1.937
LG-BPN [12]	2.95	2252	48.375
SASL [13]	1.08	17.4	0.062
Complementary-BSN (Ours)	3.09	2333	2.187

Meanwhile, to demonstrate the specific numerical differences in performance compared to other self-supervised methods, we present the distribution of PSNR values on the SIDD validation dataset in Fig. 10. In this figure, the distribution is grouped with an interval of 5 and our results are represented in blue. It is evident from the graph that when PSNR is below 30, our algorithm has the lowest proportion. Moreover, when the metric reaches 40-45, our algorithm has the highest quantity, indicating that the superior overall performance of our algorithm compared to others.

Figs. 11 and 12 provide visual comparisons among the proposed method and the SOTA self-supervised approaches on the SIDD validation and benchmark datasets. Comparing the red box in these images, we can see that while R2R [55] can preserve some structure to some extent, it cannot directly process sRGB noisy images without additional NLF and ISP functions, harming its performance in real-world situations. CVF-SID [10] overlooks the characteristics and complex of real noise, the outputs still remain noise and show stains in flat regions. AP-BSN exhibits massive grid artifacts, LG-BPN struggles to accurately restore images with complex textures or patterns. Although SASL has a significant texture restoration effect on denoised images, it still reserve the zigzag edges. In contrast, our Complementary-BSN not only removes most spatially correlated noise but also restores texture details more accurately without introducing visual artifacts. The visual comparisons on the DND benchmark and NIND datasets are shown in Figs. 13 and 14, respectively. It can be seen that AP-BSN causes excessive blurring of images and is unable to

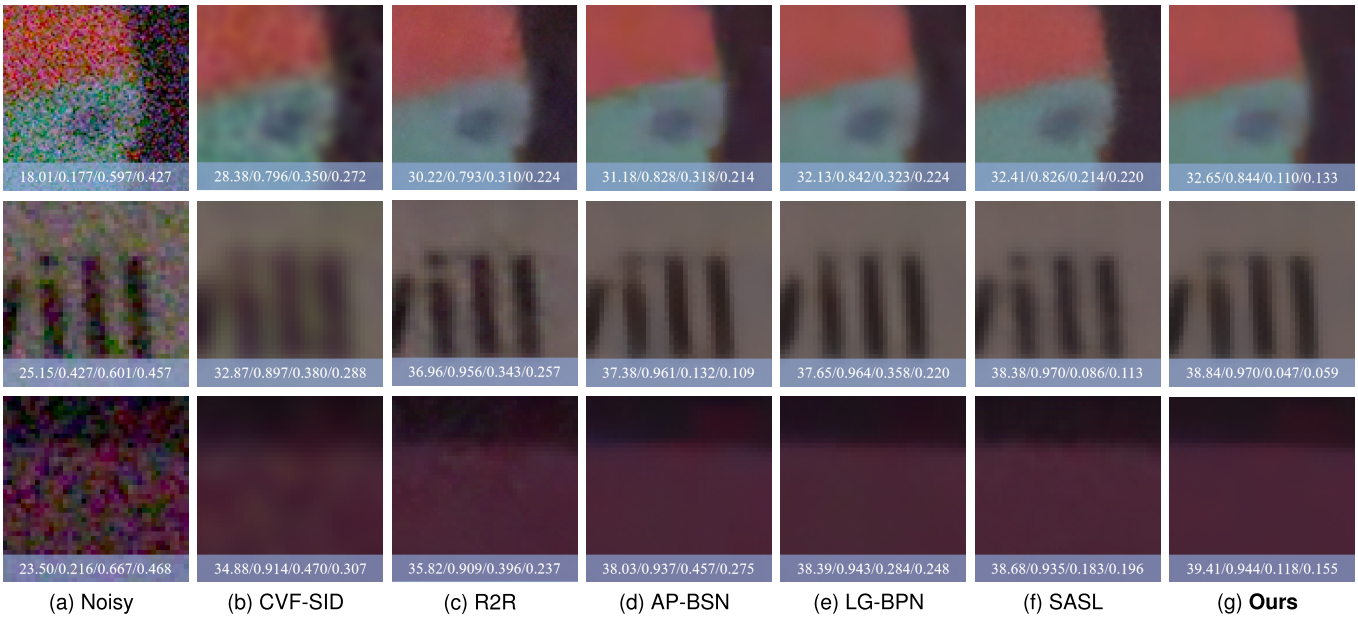


Fig. 11. Visual comparison on the SIDD validation dataset. The result is annotated with PSNR↑(dB) / SSIM↑ / LPIPS↓ / DISTS↓ values, which facilitate quantitative comparisons against the ground-truth.

Fig. 12. Visual comparison on the SIDD benchmark dataset. Quantitative results are inaccessible due to the absence of ground-truth.

restore textural structures, while LG-BPN mistakenly identifies areas of high-intensity noise as image structures, leading to obvious artifacts in denoised results. On the contrary, our method effectively mitigates the impact of dense noise and preserves a greater amount of textures.

In addition to assessing the quality of restored images, we also investigate the number of parameters, multiplier accumulator operations (MACs), and inference time (Inf. Time), which are provided in Table III. The Inf. Time is measured at each 256×256 patch of the SIDD validation dataset. Specifically, Complementary-BSN is approximately 24x faster than recently LG-BPN for denoising because we do not need Transformer block which is computationally intensive. Although the MACs of our method is more expensive than AP-BSN, the number of parameters is same as AP-BSN, which means that we can achieve higher computational complexity with fewer parameters. Furthermore, our Complementary-BSN is featured with longer inference time than SASL, owing to multiple repetitions of the random-replacing strategy. However, compared to the other approaches, our method has a more powerful denoising capability. Thus, these contrasts demonstrate the effectiveness of our Complementary-BSN.

E. Ablation Study

This subsection conducts several ablation studies on the SIDD validation dataset to show the effectiveness of the

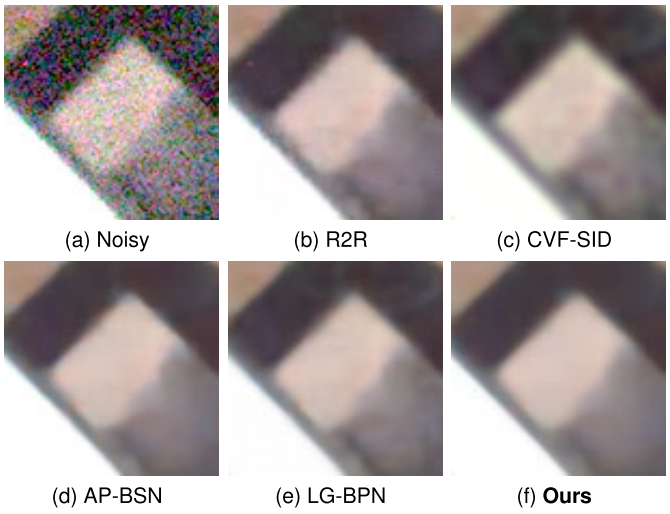


Fig. 13. Visual comparison on the DND benchmark dataset.

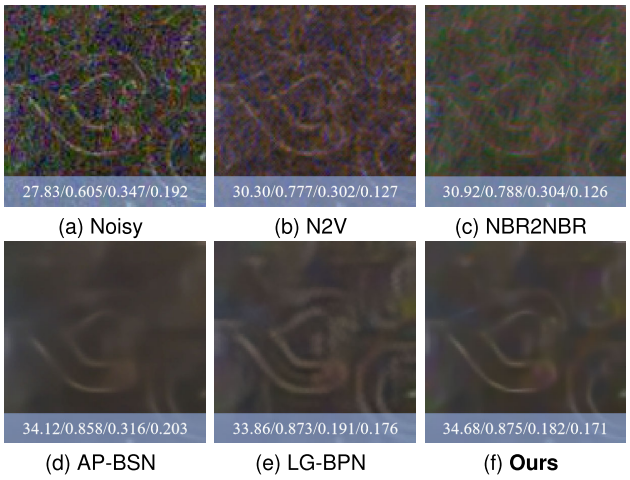


Fig. 14. Visual comparison on the NIND dataset. The result is annotated with PSNR↑(dB) / SSIM↑ / LPIPS↓ / DISTS↓ values, the ground-truth comes from corresponding ISO200 image.

proposed method. First, we investigate three major components of our model (Table IV): a) Mask-Map branch; b) BRP; c) MPD. The presence of a "✓" in the Mask-Map branch column indicates that Mask-Map branch is integrated into the training process, and the network is trained by our proposed L_{total}. Similarly, a "✓" in the BRP column signifies that the BRP strategy is employed during training. If there is a "✓" in the MPD column, it means that the MPD scheme is utilized during inference. Additionally, given that real-world noise often exhibits more intricate patterns, we conduct an ablation study to determine the optimal width for the global mask mapper.

1) Ablation on Mask-Map Branch: Complementary-BSN consists of two branches in parallel. Specifically, we combine Mask-Map branch and Enhanced-PD-BSN branch, aiming at complementing pixels information across them. To investigate the appropriateness of our two-branch configuration, we perform an ablation study on the Mask-Map branch. Table IV shows the results of different combinations. In particular, the first row of Table IV indicates AP-BSN. A comparison

TABLE IV

ABLATION STUDIES OF COMPONENTS OF OUR PROPOSED METHOD, INCLUDING MASK-MAP BRANCH, BRP AND MPD. THE ABLATION STUDIES ARE PERFORMED ON THE SIDD VALIDATION DATASET. THE BEST RESULTS ARE MARKED IN BOLD

Mask-Map branch	BRP	MPD	PSNR↑(dB)/SSIM↑
-	-	-	36.73/0.878
-	✓	-	36.88/0.878
-	✓	✓	36.99/0.879
✓	✓	-	37.36/0.884
-	-	✓	36.82/0.879
✓	-	✓	37.14/0.883
✓	-	-	37.20/0.883
✓	✓	✓	**37.51/0.885**

TABLE V

EXPERIMENTAL RESULT OF W/O MPD, MPD(2, 5) AND MPD(1, 2, 5) FROM OUR FRAMEWORK WITHOUT R^3 ON THE SIDD VALIDATION DATASET. THE BEST RESULTS ARE MARKED IN BOLD

Method	w/o MPD	MPD(2, 5)	MPD(1, 2, 5)
PSNR↑(dB)	36.53	36.58	**36.71**
SSIM↑	0.850	0.854	**0.859**

between the first and seventh rows of this table reveals that simply applying Mask-Map branch to the original AP-BSN model increases the PSNR/SSIM result by 0.47dB/0.005. This significant enhancement underscores the impact of our two-branch structure for the network in enhancing image denoising capabilities.

2) Ablation on BRP: The comparison between the first and second rows of Table IV demonstrates that incorporating the BRP strategy leads to a substantial improvement in performance, with an increase of 0.15dB in PSNR. Similarly, when comparing the fourth and seventh rows, we observe a consistent trend of performance enhancement in both PSNR and SSIM metrics. These findings strongly support the efficacy of our proposed BRP operation in enhancing the denoising capabilities of the network. Moreover, it is crucial to emphasize that the stride of BRP must be set equal to that of PD. This ensures that the structural integrity of the original image is preserved, preventing any distortion or alteration of its fundamental components.

3) Ablation on MPD: Rows 1 and 5 of Table IV shows the impact of our proposed MPD scheme for inference. By incorporating MPD, we observe a significant enhancement in denoising performance, with an improvement of 0.09dB in PSNR, which underscoring the effectiveness of MPD in elevating the quality of the denoised output. In addition to the overall comparison, we further assess the denoising performance with different combinations within MPD and report the results in Table V. The method of w/o MPD represents using the same PD stride with 2 for inference as AP-BSN, MPD(2, 5) represents using PD stride with 2 and 5, MPD(1, 2, 5) represents using PD stride with 1, 2 and 5. It is clear that introducing PD with inference stride factor of 1 contributes

TABLE VI

THE EFFECT OF MPD ON THE SIDD BENCHMARK, DND BENCHMARK, NIND ISO3200 AND NIND ISO5000 DATASETS. TO PREVENT MPD FROM BEING AFFECTED BY POST-PROCESSING, OUR RESULTS ARE OBTAINED WITHOUT R^3. THE BEST RESULTS ARE MARKED IN BOLD

Datasets	SIDD benchmark		DND benchmark		NIND ISO3200				NIND ISO5000			
	PSNR↑(dB)	SSIM↑	PSNR↑(dB)	SSIM↑	PSNR↑(dB)	SSIM↑	LPIPS↓	DISTS↓	PSNR↑(dB)	SSIM↑	LPIPS↓	DISTS↓
w/o MPD	36.48	0.917	38.05	0.930	34.32	0.837	0.173	0.184	33.46	0.834	0.159	0.174
MPD	**36.66**	**0.922**	**38.14**	**0.933**	**34.36**	**0.839**	**0.168**	**0.180**	**33.49**	**0.836**	**0.153**	**0.165**

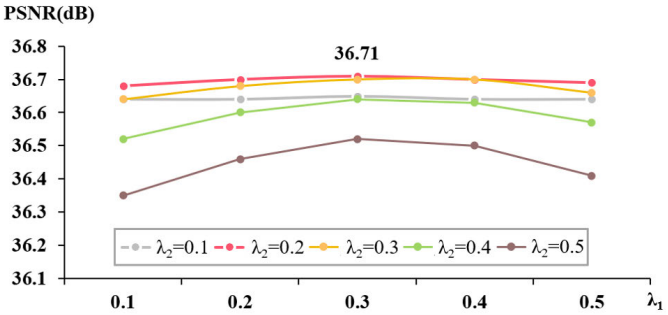


Fig. 15. Ablation study of the hyperparameters of MPD with our Complementary-BSN framework on the SIDD validation dataset. The R^3 refinement is avoided here to illustrate the difference clearly. We investigate the effect of different λ_1 for λ_2 = 0.1, 0.2, 0.3, 0.4, 0.5 and observing that MPD achieves the best performance when λ_1 = 0.3, λ_2 = 0.2.

to the final denoising results. Additionally, irrespective of whether MPD(1, 2, 5) or MPD(2, 5) is used in inference, the denoising performance is superior to framework without MPD strategy and the values of PSNR are 0.05dB and 0.18dB higher respectively, showing the flexibility and effectiveness of MPD. To further validate the performance of our MPD strategy when confronted with various real-world noise types and image structures, we conduct a study about the effect of MPD on the SIDD benchmark, DND benchmark, NIND ISO3200, and NIND ISO5000 datasets in Table VI. This table quantitatively shows that a significant disparity between the results with and without the application of MPD, indicating that MPD is universally applicable to different types of real-world noise.

The ablation of hyperparameters for MPD is illustrated in Fig. 15. We can observe that our design delivers the ultimate capability with $\lambda_1 = 0.3$ and $\lambda_2 = 0.2$. The experiment shows that PD_2 has the greatest contribution to the final restoration, which means that the MPD scheme takes the combined advantages of balanced performance of PD_2 and the complementary information from PD_1 and PD_5.

4) Hyperparameter for Global Mask Mapper: Table VII provides an assessment of the proposed method with varying widths (w) of the global mask mapper. The results indicate that a width of 3 attains the highest performance. It is noteworthy that the width of the global masker determines the frequency of masking pixels, while less width means more masked pixels. Therefore, the model with width of 2 suffers from information loss brought by frequent masking and interpolation and generates sub-optimal denoising performance. On the contrary, although higher width means less masked pixels, the large masking interval results in spatial correlation of the masked images and confuses the network during the training phase.

TABLE VII

ABLATION STUDY OF THE WIDTH SIZE ON THE INTRODUCED GLOBAL MASK MAPPER. THE BEST RESULTS ARE MARKED IN BOLD

Width	2	3	4
PSNR↑(dB)	37.20	**37.51**	37.26
SSIM↑	0.883	**0.885**	0.884

Based on the results presented in Table VII, we determine that setting the width of the global-aware mask mapper to 3 in implementation.

V. CONCLUSION

In this paper, we propose Complementary-BSN, a novel self-supervised real-world image denoising framework, aiming to address the details lost by the limit on the masked pixels for BSN, and the lack of consideration for the respective characteristics of image regions in inference. First, we design a two-branch structure with the global mask mapper and a novel revisible loss to achieve lossless denoising through selectively controlling blind-spot. Second, we propose BRP, injecting randomly rearranges into the previously PD process, and can further break the pixel-wise spatial correlation. Finally, we propose an efficient inference scheme (MPD) to maintain more useful information by fusing restored images from different stride of PD. Our method does not require any prior knowledge about clean data or noise distribution. The experimental results on real-world noisy datasets show the effectiveness of the proposed model over recent self-supervised denoisers for enhanced structure recovery. In our future work, we aim to learn the correlation parameter in a self-supervised manner. This approach holds the potential to enhance the adaptability and robustness of our model, enabling it to better handle a wide range of scenarios and data variations. Furthermore, we consider combining self-supervised denoising methods based on mask reconstruction with image classification algorithms to devise a joint training framework. This framework will integrate both tasks with corresponding loss functions, realizing the migration of denoising information between different data domains.

REFERENCES

[1] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 479–486.

[2] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, 2005, pp. 60–65.

[3] K. Dabov, A. Foi, and K. Egiazarian, "Image denoising with block-matching and 3D filtering," *Proc. SPIE*, vol. 6064, pp. 354–365, Feb. 2006.

[4] L. Fan, H. Li, M. Shi, Z. Hua, and C. Zhang, "Two-stage image denoising via an enhanced low-rank prior," *J. Sci. Comput.*, vol. 90, no. 1, p. 57, Jan. 2022.

[5] X. Zhang, X. Feng, and W. Wang, "Two-direction nonlocal model for image denoising," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 408–412, Jan. 2013.

[6] 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.

[7] L. Fan, X. Li, H. Fan, and C. Zhang, "An adaptive boosting procedure for low-rank based image denoising," *Signal Process.*, vol. 164, pp. 110–124, Nov. 2019.

[8] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[9] L. Fan, X. Li, H. Fan, Y. Feng, and C. Zhang, "Adaptive texture-preserving denoising method using gradient histogram and nonlocal self-similarity priors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 11, pp. 3222–3235, Nov. 2019.

[10] 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.

[11] 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. 17725–17734.

[12] Z. Wang, Y. Fu, J. Liu, and Y. Zhang, "LG-BPN: Local and global blind-patch network for self-supervised real-world denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18156–18165.

[13] J. Li et al., "Spatially adaptive self-supervised learning for real-world image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9914–9924.

[14] 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.

[15] 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.

[16] C. Tian, Y. Xu, and W. Zuo, "Image denoising using deep CNN with batch renormalization," *Neural Netw.*, vol. 121, pp. 461–473, Jan. 2020.

[17] 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.

[18] A. Saeed and B. Nick, "Real image denoising with feature attention," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3155–3164.

[19] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[20] 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.

[21] G. Singh, A. Mittal, and N. Aggarwal, "ResDNN: Deep residual learning for natural image denoising," *IET Image Process.*, vol. 14, no. 11, pp. 2425–2434, Sep. 2020.

[22] A. Lahiri, S. Bairagya, S. Bera, S. Haldar, and P. K. Biswas, "Lightweight modules for efficient deep learning based image restoration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1395–1410, Apr. 2021.

[23] H. Yue, J. Liu, J. Yang, X. Sun, T. Q. Nguyen, and F. Wu, "IENet: Internal and external patch matching ConvNet for web image guided denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3928–3942, Nov. 2020.

[24] L. Sun, Y. Wang, F. Wu, X. Li, W. Dong, and G. Shi, "Deep unfolding network for efficient mixed video noise removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4715–4727, Sep. 2023.

[25] Z. Zhou, Y. Chen, and Y. Zhou, "Deep dynamic memory augmented attentional dictionary learning for image denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4784–4797, Sep. 2023.

[26] B. Jiang, Y. Lu, J. Wang, G. Lu, and D. Zhang, "Deep image denoising with adaptive priors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5124–5136, Oct. 2022.

[27] L. Guo, S. Huang, H. Liu, and B. Wen, "Towards robust image denoising via flow-based joint image and noise model," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 25, 2023, doi: 10.1109/TCSVT.2023.3345667.

[28] Y. Pan, C. Ren, X. Wu, J. Huang, and X. He, "Real image denoising via guided residual estimation and noise correction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1994–2000, Apr. 2023.

[29] 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.

[30] K. Kim, T. Kwon, and J. C. Ye, "Noise distribution adaptive self-supervised image denoising using tweedie distribution and score matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1998–2006.

[31] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2 void–learning denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2124–2132.

[32] J. Lehtinen et al., "Noise2Noise: Learning image restoration without clean data," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, Jul. 2018, pp. 2965–2974.

[33] N. Moran, D. Schmidt, Y. Zhong, and P. Coady, "Noisier2Noise: Learning to denoise from unpaired noisy data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12061–12069.

[34] J. Xu et al., "Noisy-as-clean: Learning self-supervised denoising from corrupted image," *IEEE Trans. Image Process.*, vol. 29, pp. 9316–9329, 2020.

[35] J. Batson and L. Royer, "Noise2Self: Blind denoising by self-supervision," in *Proc. Int. Conf. Mach. Learning. (ICML)*, vol. 97, Jun. 2019, pp. 524–533.

[36] Y. Zhang, D. Li, K. L. Law, X. Wang, H. Qin, and H. Li, "IDR: Self-supervised image denoising via iterative data refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2088–2097.

[37] M. Yao, D. He, X. Li, F. Li, and Z. Xiong, "Toward interactive self-supervised denoising," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5360–5374, Oct. 2023.

[38] X. Wu, M. Liu, Y. Cao, D. Ren, and W. Zuo, "Unpaired learning of deep image denoising," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*, Oct. 2020, pp. 352–368.

[39] S. Laine, T. Karras, J. Lehtinen, and T. Aila, "High-quality self-supervised deep image denoising," in *Proc. Neural Inf. Process. Systems. (NIPS)*, Dec. 2019, vol. 32, pp. 6970–6980.

[40] Y. Zhou et al., "When AWGN-based denoiser meets real noises," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13074–13081.

[41] D. Zhang, F. Zhou, Y. Jiang, and Z. Fu, "MM-BSN: Self-supervised image denoising for real-world with multi-mask based on blind-spot network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 4188–4197.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[44] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[45] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, "Variational denoising network: Toward blind noise modeling and removal," in *Proc. Neural Inf. Process. Systems. (NIPS)*, Dec. 2019, pp. 1690–1701.

[46] Z. Yue, Q. Zhao, L. Zhang, and D. Meng, "Dual adversarial network: Toward real-world noise removal and noise generation," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*, Nov. 2020, pp. 41–58.

[47] S. W. Zamir et al., "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12370, 2020, pp. 492–511.

[48] Z. Tu et al., "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5769–5780.

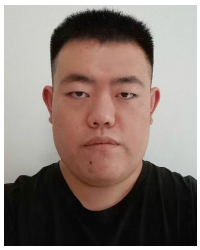
[49] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho, "Transfer learning from synthetic to real-noise denoising with adaptive instance normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3482–3492.

[50] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3155–3164.
[51] G. Jang, W. Lee, S. Son, and K. Lee, "C2N: Practical generative noise modeling for real-world denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2330–2339.
[52] 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.
[53] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Neural Inf. Process. systems. (NIPS)*, vol. 27, Dec. 2014, pp. 2672–2680.
[54] 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.
[55] 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.
[56] 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.
[57] T. Plotz and S. Roth, "Benchmarking denoising algorithms with real photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1586–1595.
[58] B. Brummer and C. De Vleeschouwer, "Natural image noise dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1777–1784.
[59] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vision. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 17–33.
[60] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

Linwei Fan received the Ph.D. degree from Shandong University, Jinan, China, in 2019. She is currently an Associate Professor with the School of Computer Science and Technology, Shandong University of Finance and Economics, and a member of Shandong Provincial Key Laboratory of Digital Media Technology. Her research interests include computer graphics and image processing.

Jin Cui received the B.E. degree from Yanshan College, Shandong University of Finance and Economics, Jinan, China, in 2022. She is currently pursuing the M.E. degree with the School of Computer Science and Technology, Shandong University of Finance and Economics. Her research interests include deep learning and image restoration.

Huiyu Li received the Ph.D. degree from Shandong University, Jinan, China, in 2021. He is currently a Lecturer with the School of Management Science and Engineering, Shandong University of Finance and Economics. His research interests include virtual reality, human–computer interaction, and computer graphics.



Xiaoyu Yan received the B.S. degree from Shandong University of Finance and Economics, Jinan, China, in 2017, where he is currently pursuing the M.E. degree with the School of Computer Science and Technology. His research interests include deep learning theory and image restoration.

Hui Liu received the B.S., M.S., and Ph.D. degrees in computer science from Shandong University, Jinan, China, in 2001, 2004, and 2008, respectively. From 2017 to 2018, she was a Visiting Scholar with the Medical Physics Division, Department of Radiation Oncology, Stanford University, USA. She is currently a Professor with the School of Computer Science and Technology, Shandong University of Finance and Economics, China. Her current research interests include computer aided diagnosis, medical image processing, and machine learning.

Caiming Zhang received the B.S. and M.S. degrees in computer science from Shandong University, Jinan, China, in 1982 and 1984, respectively, and the Ph.D. degree in computer science from Tokyo Institute of Technology, Tokyo, Japan, in 1994. From 1998 to 1999, he was a Post-Doctoral Fellow with the University of Kentucky, Lexington, USA. He is currently a Professor with the School of Software, Shandong University. His research interests include CAGD, information visualization, and medical image processing.