Exploring Efficient Asymmetric Blind-Spots for Self-Supervised Denoising in Real-World Scenarios

Shiyan Chen[1,2]   Jiyuan Zhang[1,2]   Zhaofei Yu[1,2,3*]   Tiejun Huang[1,2,3]

[1]School of Computer Science, Peking University
[2]National Key Laboratory for Multimedia Information Processing, Peking University
[3]Institute for Artificial Intelligence, Peking University

{strerichia002p, jyzhang}@stu.pku.edu.cn, {yuzf12,tjhuang}@pku.edu.cn

Abstract

Self-supervised denoising has attracted widespread attention due to its ability to train without clean images. However, noise in real-world scenarios is often spatially correlated, which causes many self-supervised algorithms that assume pixel-wise independent noise to perform poorly. Recent works have attempted to break noise correlation with downsampling or neighborhood masking. However, denoising on downsampled subgraphs can lead to aliasing effects and loss of details due to a lower sampling rate. Furthermore, the neighborhood masking methods either come with high computational complexity or do not consider local spatial preservation during inference. Through the analysis of existing methods, we point out that the key to obtaining high-quality and texture-rich results in real-world self-supervised denoising tasks is to train at the original input resolution structure and use asymmetric operations during training and inference. Based on this, we propose Asymmetric Tunable Blind-Spot Network (AT-BSN), where the blind-spot size can be freely adjusted, thus better balancing noise correlation suppression and image local spatial destruction during training and inference. In addition, we regard the pre-trained AT-BSN as a meta-teacher network capable of generating various teacher networks by sampling different blind-spots. We propose a blind-spot based multi-teacher distillation strategy to distill a lightweight network, significantly improving performance. Experimental results on multiple datasets prove that our method achieves state-of-the-art, and is superior to other self-supervised algorithms in terms of computational overhead and visual effects.

1. Introduction

Image denoising is an essential low-level computer vision problem. With the advancements in deep learning, an in-

---

*Corresponding author.

(a) Noisy Input   (b) AP-BSN (R3)   (c) LG-BPN

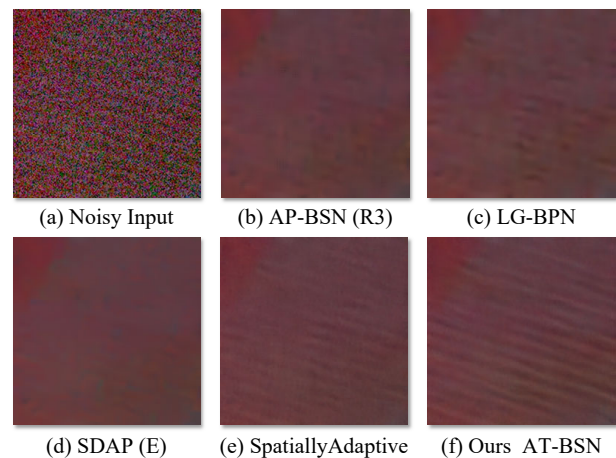(d) SDAP (E)   (e) SpatiallyAdaptive   (f) Ours AT-BSN

Figure 1. Comparisons of our AT-BSN with other methods. Our method recovers more high frequency texture details.

creasing number of studies are focused on supervised learning using clean-noisy pairs [2, 17, 32, 47–49]. Typically, additive white Gaussian noise (AWGN) is introduced into clean datasets to synthesize clean-noisy denoising datasets. However, real-world noise is known to be spatially correlated [7, 23, 37]. Some generative-based methods attempt to synthesize real-world noise from existing clean data [5, 8, 18, 21, 43]. However, synthesizing real-world noise remains challenging, and suffers from generalization issues. To address the issue, some researchers attempt to capture clean-noisy pairs in real-world scenarios [1, 4]. However, in certain scenarios, such as medical imaging and electron microscopy, constructing such datasets can be impractical or even infeasible.

Self-supervised denoising algorithms, represented by Noise2Noise [28], have brought new life to the denoising field. These methods only require noisy observations to train the denoising model. However, in real-world scenarios, noise often exhibits spatial correlation, which contradicts the pixel-wise independent noise assumption [25, 28]

that most self-supervised algorithms [3, 20, 25, 28, 41] rely on. Recent studies have proposed self-supervised denoising algorithms suitable for real-world scenarios [22, 27, 29, 34, 35, 42]. These methods mainly disrupt the noise correlation by downsampling [22, 27, 35] or neighborhood masking [29, 42]. The representative work of the former is AP-BSN [27], which utilized pixel-shuffle downsampling (PD) [27, 51] to disrupt the noise correlation and employed asymmetric PD stride factors for training and inference. However, according to the Nyquist-Shannon sampling theorem, downsampling methods disrupt image spatial structure during inference, leading to lower sampling density and loss of high-frequency details. Conversely, neighborhood masking methods denoise at original resolution structure, retaining more texture information.

In this paper, we carefully analyze the existing related work and point out that **training at the original resolution structure** and **using asymmetric operations during training and inference** are key to producing high-quality, texture-rich clear images in self-supervised real noise removal tasks. Based on these observations, we propose a novel paradigm called **Asymmetric Tunable Blind-Spot Network(AT-BSN)**, where the blind-spot size can be freely adjusted to balance between noise correlation suppression and image local structure destruction. Furthermore, the flexible tunable blind-spot allows us to obtain a potential teacher network distribution, where each sampled teacher has a different blind-spot, making each teacher network's ability to handle flat/texture areas different. We propose a **Blind-Spots Based Multi-Teacher Distillation** strategy, which significantly improves performance and further reduces computational overhead. Experimental results demonstrate the effectiveness of the proposed method.

The main contributions are summarized as follows:

- We carefully analyze existing methods and point out that training at the original resolution structure and using asymmetric operations during training and inference are key to producing high-quality, texture-rich results in self-supervised denoising tasks.
- We propose AT-BSN, which can better balance the suppression of noise correlation and the destruction of image's local spatial structure by applying asymmetric blind-spots during training and inference.
- We propose a Blind-Spots Based Multi-Teacher Distillation strategy, which significantly improves performance by distilling a lightweight student network from teachers with different blind-spots sampled from the teacher network distribution.
- Experimental results on multiple real-world datasets show our method achieves state-of-the-art performance, with clear advantages in computational complexity and preservation of high-frequency texture details.

2. Related Works

Supervised Image Denoising. Deep learning has made remarkable advances in image denoising in recent years. Zhang *et al*. [48] introduced DnCNN, the first CNN-based method for supervised denoising, which significantly outperformed traditional methods [6, 11, 12, 16, 39].The following work aimed to enhance the performance of supervised denoising, such as FFDNet [49], CBDNet [17], RID-Net [2], DANet [47], FADNet [32], and so on. However, supervised-based methods require large amounts of aligned clean-noisy pairs as training data, which are usually difficult and costly to obtain in formal scenarios.

Unpaired Image Denoising. To tackle the challenge in supervised learning, some generative-based [15] approaches synthesize noisy samples from clean images [5, 8, 13, 18, 21, 43]. The simulated clean-noisy pairs can be further used to train a supervised denoising model. However, the performance of unpaired image denoising methods can be limited when the existing clean images do not match the distribution of the current scene.

Self-Supervised Image Denoising. Lehtinen *et al*. [28] proposed Noise2Noise, which demonstrated that a denoising network could be trained with two independent noisy observations of the same scene. However, even if Noise2Noise relaxes the clean image requirement, obtaining two aligned noisy images in real-world scenarios remains difficult. Noise2Void [25] and Noise2Self [3] proposed a blind-spot strategy to learn denoising from only single noisy images. Further works [26, 43] extended the paradigm to blind-spot network (BSN) through shifted convolutions [26] and dilated convolutions [43]. Blind-spot means the network is designed to denoise each pixel from its surrounding spatial neighborhood without itself, thus, the identity mapping to the noisy image itself can be avoided. Noisier2Noise [33], Noisy-As-Clean (NAC) [44], Recorrupted-to-Recorrupted (R2R) [36], and IDR [50] generated noisy training pairs by adding synthetic noise to given noisy inputs. Recently, Neighbor2Neighbor [20] proposed to subsample the noisy input images to obtain noisy pairs for Noise2Noise-like training. Blind2Unblind [41] proposes a global-aware mask mapper and re-visible loss to fully excavate the information in the blind-spot for Noise2Void-like training.

Real-World Image Denoising. Some works [1, 4] attempt to capture clean-noisy pairs in real-world scenarios. Abdelhamed *et al*. [1] carefully took and aligned clean-noisy pairs from different scenes and lighting conditions using five representative smartphone cameras, and proposed the SIDD dataset. These datasets enable supervised methods [10, 19, 24, 31, 46, 47] to train on real-world clean-noisy pairs. However, constructing real datasets requires tremendous human effort and time. Moreover, real-world noise tends to exhibit spatial correlation, which contra-

(a) Downsampling based asymmetric operations

(b) Densely-sampled patch-masked convolution based asymmetric operations

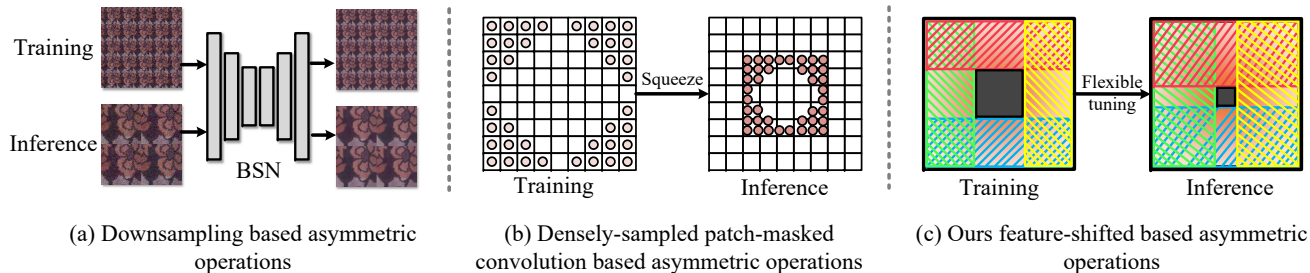(c) Ours feature-shifted based asymmetric operations

Figure 2. Three kinds of asymmetric operations during training and inference. Our scheme can flexibly tune the blind-spot size to meet the requirements of training and inference, achieving a balance between noise correlation suppression and local spatial destruction.

dicts the premise of Noise2Noise [28] that noise follows an independent and identically distributed pattern, rendering it and its subsequent variants unsuitable for direct application to real-world scenarios. In order to apply self-supervised learning to real-world settings, Neshatavar *et al.* [34] introduced a cyclic multi-variate function to disentangle clean images, signal-dependent noise, and signal-independent noise from noisy images. However, the method relies on a simple network without residual connections to avoid learning an identity mapping to the noise signal. Additionally, the simple assumption about real-world noise signals has resulted in its vague denoising results. Lee *et al.* [27] employed pixel-shuffle downsampling (PD) [51] to disrupt the spatial correlation of noise and introduced different PD stride factors for training and inference for better performance. Li *et al.* [29] proposed to use a larger blind-neighborhood to suppress the spatial correlation of noise and present a network to extract the texture within the blind-neighborhood region. However, the method still uses a large blind-spot during testing, which requires a lot of training to extract effective information from a distance to reconstruct the central pixel. LG-BPN [42] proposes to mask the central area of a large convolution kernel to suppress the spatial correlation of noise and proposes a dilated Transformer block to extract global information. However, the introduction of large kernels will bring greater computational overhead. In addition, some methods attempt to improve AP-BSN. Pan *et al.* [35] propose random sub-sampling as data augmentation. Jang *et al.* [22] utilize information from the blind-spot position by proposing conditional masked convolution. Nevertheless, these downsampling-based methods lack texture details.

3. Methods

3.1. Revisit of Various Methods to Disrupt Noise Spatial Correlation

Due to the effect of image signal processors (ISP), e.g. image demosaicking [7, 23, 37], real-world noise is generally known to be spatially correlated and pixel-wise dependent.
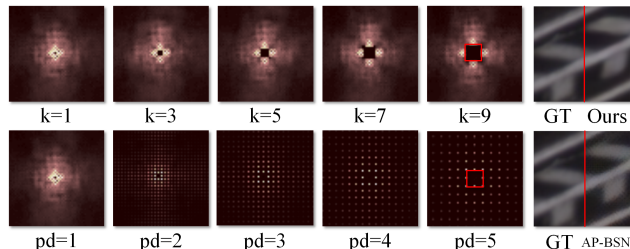

Figure 3. Effective Receptive Field analysis of AP-BSN and our AT-BSN. Each column has the same center blind-spot size. The downsampling operation of AP-BSN can cause aliasing effects.

Lee *et al.* [27] analyzed the spatial correlation of real-world noise and found that different camera devices in the SIDD dataset show similar noise behaviors in terms of spatial correlation. According to Lee *et al.* [27]'s analysis, the correlation of noise presents a Gaussian distribution that decays as distance increases. This correlation of noise violates the pixel-wise independent noise assumption of the BSN, rendering it inadequate for real noise removal.

Recently, some methods have been proposed to break the spatial correlation of noise, so that BSN can be used for real-world noise removal. Basically, these methods can be divided into downsampling based approaches and neighborhood masking based approaches.

Downsampling Based Approaches. AP-BSN [27] first introduced the pixel-shuffle downsampling operation [51] into the self-supervised denoising task to break the noise correlation, and proposed asymmetric PD to balance between noise correlation removal and image structure damage. Jang *et al.* [22] design a conditional blind-spot network, which selectively controls the blindness of the network to use the center pixel information. By retaining some information of the blind-spot part at test time, this method achieved better results. Further, Pan *et al.* [35] propose random sub-sampling to address the data-hungry issue of training BSN with real noisy images, and further propose to use sampling difference as a perturbation to improve per-

formance. Regardless of the exact downsampling method used, these methods aim to optimize the BSN $B_\theta(\cdot)$ mainly by minimizing the following loss functions:

$$\mathcal{L}_{down} = \|D_m^{-1}(B_\theta(D_m(I_{noisy}))) - I_{noisy}\|_1,$$
$$or \ \|B_\theta(D_m(I_{noisy})) - D_m(I_{noisy})\|_1, \quad (1)$$

where $D_m(\cdot)$ denotes a certain downsampling method with a factor of m, $D_m^{-1}(\cdot)$ denotes its inverse operation. Although the downsampling based methods can effectively break the spatial correlation of noise, the results of these methods are often blurry and lack texture details. According to the Nyquist-Shannon sampling theorem, the fidelity of the results is positively correlated with the sampling density. These methods train on degraded low-resolution images, and their receptive fields are present as sparse and diffuse grids, which makes it difficult to learn structural information. Moreover, these methods still test on low-resolution sub-images, which greatly reduces their sampling density and produces aliasing effects, leading to high-frequency detail loss. Additional time-consuming post-processing is also needed to eliminate aliasing effects.

Neighborhood Masking Based Approaches. The neighborhood masking based schemes attempt to train the denoising network on the original resolution structure, the optimization objective is as follows:

$$\mathcal{L}_{neighbor} = \|B_\theta'(I_{noisy}) - I_{noisy}\|_1, \quad (2)$$

where $B_\theta'(\cdot)$ denotes a carefully modified BSN, in which the blind-spot size is enlarged. LG-BPN [42] proposed a densely-sampled patch-masked convolution, which breaks the spatial correlation of noise by masking the center part of a large convolution kernel. LG-BPN also proposed to squeeze the convolution kernel weights during inference to reduce the masked part, to balance between local structure damage and noise correlation removal. Note that this idea is similar to the asymmetric PD in AP-BSN, but the introduction of large convolution kernels brings high computational overhead, and the squeeze of convolution kernel weights is limited to specific convolution kernel sizes, which is inflexible to adjust the masked region. Li*et al.* [29] proposed to use a larger blind-neighborhood to break the noise correlation, and trained another network with a receptive field limited to the blind-neighborhood to fill in the information loss within the large blind-neighborhood position. However, this scheme utilizes the same large blind-neighborhood during training and inference, which lacks the consideration of local spatial structure damage.

Effective Receptive Field Analysis. We further compare the effective receptive fields (ERF) of the two schemes to demonstrate the importance of training at full resolution structure. Noise correlation is generally confined to local regions, according to Lee's statistics [27]. Therefore, it

is unnecessary to disrupt regions beyond the local neighbors. We consider the ERF of the central pixel, and take our method and PD operation in AP-BSN as an example, as shown in Fig. 3. The ERF of AP-BSN is calculated on the subgraph and remapped back to the original resolution structure. One can find that the ERF of AP-BSN in the original image manifests as a sparse grid-like pattern, akin to the effect of simple stacking of dilated convolutions [45]. This characteristic comes with similar drawbacks to the dilated convolutions [9, 40, 45], namely **1)** the loss of local information, posing challenges for the model to learn clues from the grid-like discontinuous sub-image for recovering the clean signal, and **2)** long-ranged information might be not relevant. Moreover, the loss of information continuity outside the center blind-spot can also introduce aliasing artifacts. Nevertheless, it is apparent that our method only loses a portion of the information within the blind-spot area, while the ERF outside the blind-spot remains unaffected. This inspires us to set the size of blind-spot to 9 during training. Due to the higher correlation of the signal compared to the noise, the central pixel can be recovered using the pixels outside the blind-spot that are less correlated with it in the noise domain. It is worth noting that the size of the blind-spot can be minimized during inference to reduce information loss.

Based on the analysis of the existing approaches, we draw the following conclusions. In order to recover clean images with clear textures from noisy images, **the following two points** are crucial:

1) According to the Nyquist-Shannon sampling theorem, both the training and inference stages of the network need to be conducted on original input resolution structure to ensure sampling density.

2) Asymmetric operation during training and inference is crucial to strike a balance between the disruption of noise spatial correlation and the destruction of local spatial structure.

The quantitative analysis of the second point can be found in the supplementary materials. Based on these two observations, we propose AT-BSN, a blind-spot network that can flexibly adjust the blind-spot size during training and inference. Fig. 2 shows three schematic diagrams of asymmetric operations during training and inference. Compared with Fig. 2 (a), AT-BSN operates on original resolution structure to maximize sampling density. Compared with Figure Fig. 2 (b), AT-BSN has a lower computational cost and can adjust the blind-spot size at will, which prompts us to further propose a multi-teacher knowledge distillation strategy based on various blind-spots to further improve performance and reduce network complexity.
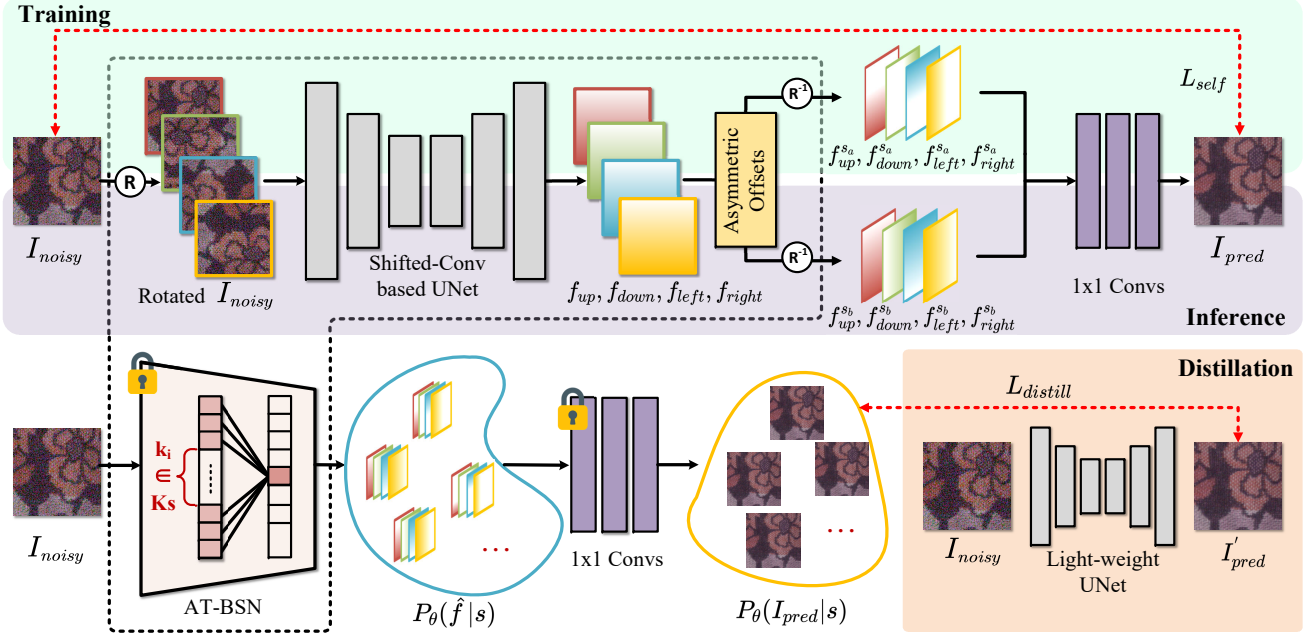
Figure 4. Overview of the proposed AT-BSN framework. We employ asymmetric blind-spots for training and inference to balance the suppression of noise spatial correlation and local information preservation. We regard the trained AT-BSN as a meta-teacher network, generate multiple teacher networks by sampling different blind-spots, and execute multi-teacher distillation on a lightweight network.
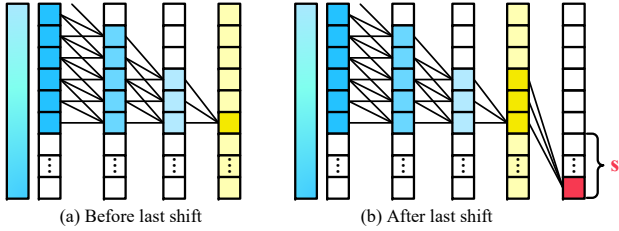

Figure 5. Implementation principle of the tunable blind-spot.

3.2. Tunable Blind-Spot

BSN [25, 27, 43] is designed to denoise each pixel from its surrounding spatial neighborhood without itself. Typically, BSN can be constructed through shifted convolutions [26] or dilated convolution [43].

Restricted Receptive Fields. Our AT-BSN is inspired by Laine's approach [26], which combines four branches with restricted receptive fields, each of which is limited to a half-plane that excludes the central pixel. For a $h \times h$ convolution kernel, we append $d = \lfloor h/2 \rfloor$ rows of zeros at the top of the feature map, apply the convolution, and finally crop the last d rows of the feature map. For the 2×2 pooling layer, we pad the top of the feature map and crop the last row of it before pooling.

Tunable Blind-Spot. After applying the shifted-conv based UNet to I_{noisy}, we obtain the resulting feature map denoted as f_{up}, whose receptive field is fully contained within an upward half-plane, including the center row. See Fig. 5 for

more details.

We further shift the feature map f_{up} downward by s pixels, resulting in a shifted feature map f_{up}^s.

$$f_{up}^s = M(f_{up}; s), \tag{3}$$

where $M(\cdot; s)$ denotes the shift operation of a offset s. At this point, the receptive field of the central pixel only includes the positions beyond s rows above the current location. Moreover, **the use of $M(\cdot; s)$ on the feature domain is decoupled from feature extraction, which grants it more flexibility**. To expand the receptive field of a pixel to all directions around it, we rotate the input image by multiples of $90°$ and feed them into the network. This results in feature maps f_{up}^s, f_{down}^s, f_{left}^s, and f_{right}^s. Finally, the four feature maps are rotated to the correct orientation and linearly combined through several 1×1 convolutions to produce the final output I_{pred}.

$$I_{pred} = Conv_{1 \times 1}([\hat{f}_{up}^s, \hat{f}_{down}^s, \hat{f}_{left}^s, \hat{f}_{right}^s]), \tag{4}$$

where $[,]$ denotes feature concatenation, $\hat{f}^s$ denotes the corresponding f^s in the correct orientation.

Now a blind-spot area with a length of $k = 2s - 1$ is established in the center around each pixel. We can freely tune the size k of the blind-spot by adjusting the shift factor s of the feature map f^s before the 1×1 convolutions. So far, we have achieved a BSN with a tunable blind-spot size.

We denote the main network parameterized by θ as $F_\theta(\cdot; s)$, the entire process can be formulated as:

$$I_{pred} = Conv_{1\times1}(F_\theta(I_{noisy}; s)). \quad (5)$$

3.3. Asymmetric Blind-Spots

Based on the fact that the noise correlation is less than the signal correlation, we can use appropriate blind-spot k to suppress the noise correlation while minimizing the impact on signal correlation. The central pixel within a large blind-spot can be inferred from pixels outside the blind-spot that are less correlated with it in the noise domain. We minimize the following loss to train the network:

$$\mathcal{L}_{self} = \|Conv_{1\times1}(F_\theta(I_{noisy}; s)) - I_{noisy}\|_1$$
$$= \|I_{pred} - I_{noisy}\|_1. \quad (6)$$

Following previous works, we use L^1 norm for better generalization [27]. In practice, we choose $k = 9$, that is, $s = 4$, during training.

We propose to achieve a balance between training and inference by employing asymmetric blind-spots, so that the trade-off between noise correlation removal and image structure damage can be achieved. Since larger blind-spots have already been utilized during training to enable the BSN to learn to denoise, we can select smaller blind-spots during inference to minimize information loss. We denote the k used in training and inference as k_a and k_b, respectively. Similar notation rule is also used for s. In Sec. 4.3, we will demonstrate the robustness of our approach to different blind-spot combinations.

3.4. Blind-Spots Based Multi-Teacher Distillation

While larger blind-spots can more effectively suppress spatial correlations between neighboring noise signals, they also result in more loss of information. Conversely, smaller blind-spots exhibit an opposite trend.

To better integrate the advantages of our tunable blind-spot nature, we propose a blind-spot based multi-teacher distillation strategy. Under different blind-spot sizes (or different $M(\cdot; s_i)$), the features extracted by AT-BSN satisfy the distribution $P_\theta(\hat{f}|s)$, and the restored clear images follow $P_\theta(I_{pred}|s)$. We consider the trained AT-BSN as a meta-teacher network that can generate many potential teacher networks almost **cost-free** by adjusting the size of blind-spots. Therefore, we obtain multiple potential teacher networks, where different teachers can provide different knowledge, i.e., different teachers handle smooth/texture areas differently (detailed analysis can be found in the supplementary materials). This characteristic is a key strength of our approach, allowing the student network to learn from various teachers.

Specifically, we pass I_{noisy} through the trained network to get the feature f. Subsequently, we sample multiple k

from $Ks \in \{0, 1, 3, ..., 2S - 1\}$ (where S denotes the preset maximum offset), apply $M(\cdot; s_i)$ to f multiple times, and obtain the features $\hat{f}^{s_i}$ under different blind-spot sizes. Then we apply trained 1×1 convolutions to get the clear images $I_{pred}^{s_i}$. Finally, we use $I_{pred}^{s_i}$ to distill a lightweight non-blind-spot network $N_\theta(\cdot)$. In order for the student network to learn fairly from $P_\theta(I_{pred}|s)$, we do not distinguish between different teacher signals explicitly. We set the same weight $\alpha_i = 1$ for each teacher. The student network is distilled by optimizing the following objective:

$$\mathcal{L}_{distill} = \sum_{s_i \in K_s} \alpha_i \|N_\theta(I_{noisy}) - sg(I_{pred}^{s_i})\|_1, \quad (7)$$

where $sg(\cdot)$ denotes stop gradient operation. Under the multi-teacher distillation scheme, our student network can be lightweight and avoid the additional computational cost brought by the rotation operation of the BSN.

Moreover, the distillation itself is also computationally efficient, as multiple teachers share the same feature f. The specific complexity analysis can be found in the supplementary materials. The overall scheme of our methods can be found in Fig. 4.

4. Experiments

4.1. Experimental Configurations

Real-World Datasets. We conduct experiments on two real-world image denoising datasets, SIDD [1] and DND [38]. SIDD-Medium training dataset consists of 320 clean-noisy pairs captured under various scenes and illuminations. The SIDD validation dataset contains 1280 noisy patches with a size of 256×256 for performance evaluation. DND benchmark consists of 50 noisy images captured with consumer-grade cameras of various sensor sizes and does not provide clean images. DND dataset is captured under normal lighting conditions compared to the SIDD dataset, and therefore presenting less noise. We adopt PSNR and SSIM metrics to evaluate our method. We set $k_a = 9$ and $k_b = 3$ for training and inference, respectively. For multi teacher distillation, we sample blind-spots from $Ks \in \{0, 1, 3, 5, 7, 9, 11\}$. More implementation details can be found in supplementary materials.

4.2. Comparisons for Real-World Denoising

Quantitative Measure. Tab. 1 presents quantitative comparisons with other methods. We denote the distilled network as AT-BSN (D). Note that AT-BSN (D) is actually a Non-BSN despite its name. As a self-supervised algorithm, our method outperforms all existing unpaired and self-supervised methods, achieving state-of-the-art performance. Our results without † marks indicate we employ the model trained on SIDD-Medium directly on the benchmark. These results show the generalization ability of our method.

	Methods	SIDD Benchmark		SIDD Validation		DND Benchmark	
		PSNR$^\uparrow$ (dB)	SSIM$^\uparrow$	PSNR$^\uparrow$ (dB)	SSIM$^\uparrow$	PSNR$^\uparrow$ (dB)	SSIM$^\uparrow$
Non-Learning	BM3D[11]	25.65	0.685	31.75	0.706	34.51	0.851
	WNNM[16]	25.78	0.809	26.31	0.524	34.67	0.865
Supervised	DnCNN[48]	37.61	0.941	37.73	0.943	37.90	0.943
	CBDNet[17]	33.28	0.868	30.83	0.754	38.05	0.942
	RIDNet[2]	37.87	0.943	38.76	0.913	39.25	0.952
	VDN[46]	39.26	0.955	39.29	0.911	39.38	0.952
	AINDNet(R)[24]	38.84	0.951	38.81	-	39.34	0.952
	DANet[47]	39.25	0.955	39.47	0.918	39.58	0.955
	InvDN[31]	39.28	0.955	38.88	-	39.57	0.952
Unpaired	GCBD[8]	-	-	-	-	35.58	0.922
	D-BSN[43] + MWCNN[30]	-	-	-	-	37.93	0.937
	C2N[21]	35.35	0.937	35.36	0.932	37.28	0.924
Self-Supervised	Noise2Void[25]	27.68	0.668	29.35	0.651	-	-
	Laine-BSN[26]	-	-	23.80$^\diamond$	0.493$^\diamond$	-	-
	Noise2Self[3]	29.56	0.808	30.72	0.787	-	-
	NAC[44]	-	-	-	-	36.20	0.925
	R2R[36]	34.78	0.898	35.04	0.844	-	-
	CVF-SID[34]	34.43 / 34.71†	0.912 / 0.917†	34.51	**0.941**	36.31 / 36.50†	0.923 / 0.924†
	AP-BSN + R^3[27]	35.97 / 36.91†	0.925 / 0.931†	35.76	-	- / 38.09†	- / 0.937†
	C-BSN[22]	36.82 / -	0.934 / -	36.22	0.935	**38.45** / **38.60**†	**0.939** / **0.941**†
	SDAP (E)[35]	37.24 / **37.53**†	**0.936** / **0.936**†	37.30	0.894	37.86 / 38.56†	0.937 / 0.940†
	LG-BPN[42]	37.28 / -	**0.936** / -	37.32	0.886	- / 38.43†	- / **0.942**†
	Spatially-Adaptive (UNet)[29]	**37.41** / 37.37†	0.934 / 0.929†	**37.39**	0.934	38.18 / 38.58†	**0.938** / 0.936†
	AT-BSN (Ours)	36.73 / 36.74†	0.924 / 0.925†	36.80	0.934	37.76 / 38.19†	0.934 / 0.939†
	AT-BSN (D) (Ours)	**37.77** / **37.78**†	**0.942** / **0.944**†	**37.88**	**0.946**	38.29 / **38.68**†	**0.939** / **0.942**†

Table 1. Comparison among different denoising methods on real-world datasets. We report the official results from the benchmark website or related paper. The † marks indicate the method is trained directly on the corresponding benchmark dataset in a fully self-supervised manner. The ◇ marks indicate the result is measured by ourselves.

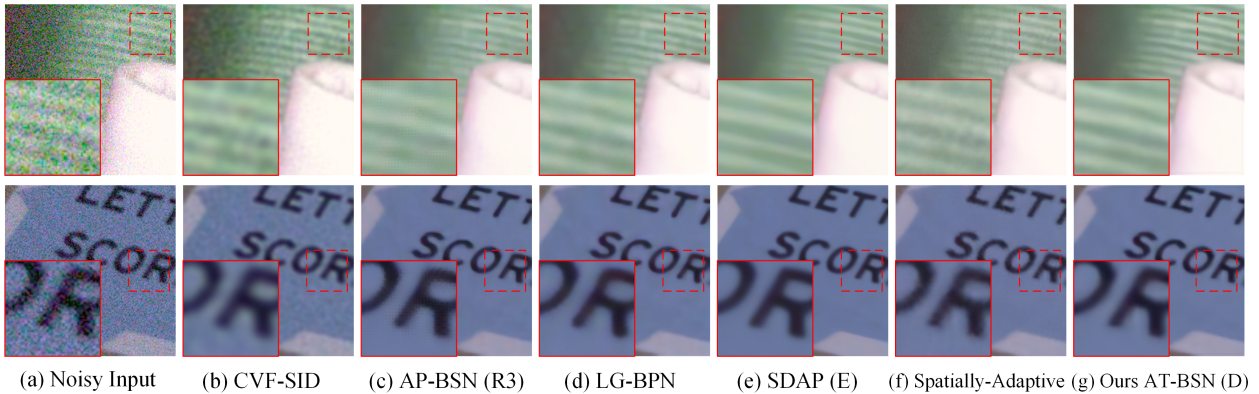| (a) Noisy Input | (b) CVF-SID | (c) AP-BSN (R3) | (d) LG-BPN | (e) SDAP (E) | (f) Spatially-Adaptive | (g) Ours AT-BSN (D) |

Figure 6. Quantitative comparisons on SIDD validation dataset.

Our results with † show the potential of fully self supervised learning. Furthermore, the results of AT-BSN (D) demonstrate the potential to enhance performance by integrating the advantages of different blind-spot sizes through multi teacher distillation.

Qualitative Measure. Fig. 6 presents the qualitative comparisons. Our method is capable of preserving the most texture details. In addition, results of downsampling-based methods [27, 35] tend to transition smoothly. LG-BPN [42], Spatially Adaptive [29], and our AT-BSN, as training on the original resolution structure, can preserve better results. Specifically, although AP-BSN uses R^3 post-processing, its visualization still exhibits aliasing effects. More SIDD and DND benchmark results are in the supplementary materials.

Blind-Spots	{1,3, 5,7}	{0,1,3, 5,7}	{0,1,3, 5,7,9}	{0,1,3 5,7,9,11}	{0,1,3,5 7,9,11,13}
PSNR	37.31	37.32	37.41	**37.47**	37.40
SSIM	0.943	0.943	0.944	**0.945**	0.944

Table 2. Ablation study of the ensemble of teacher networks.

	Mean Teacher	Multi Teacher	Param
Student A	36.79 / 0.942	36.98 / 0.943	**0.12 M**
Student B	37.60 / 0.945	37.76 / 0.946	0.86 M
Student C	**37.72 / 0.945**	**37.88 / 0.946**	1.02 M

Table 3. Ablation study of different distillation methods on students with different parameters.

Methods	Params$^\downarrow$ (M)	MACs$^\downarrow$ (G)	PSNR$^\uparrow$ (dB)
CVF-SID [34]	1.19	311.44	34.81
AP-BSN + R^3[27]	3.66	7653.97	36.48
SDAP (E)[35]	3.66	1628.57	37.30
LGBPN[42]	4.56	12168.22	37.32
Spatially-Adaptive (UNet)[29]	1.08	70.11	37.39
AT-BSN (Ours)	1.27	330.51	36.80
AT-BSN (D) (Ours)	**1.02**	**48.92**	37.88

Table 4. Complexity Analysis. The multiplier-accumulator operations (MACs) are measured on 512×512 patches.

4.3. Ablation Study

Analysis of Asymmetric Blind-Spots. We perform experiments on the combinations of training blind-spot sizes $k_a \in \{7,9,11\}$ and inference blind-spot sizes $k_b \in \{0,1,3,5,7,9,11,13\}$, testing on the SIDD validation dataset. From Fig. 7, Our method could achieve the best performance at $k_a = 9$ and $k_b = 3$. This is due to the noise correlation is fully suppressed during training, the network has learned well denoising ability. During testing, only small blind-spot is needed to suppress the areas with the highest noise correlation, while maximizing the preservation of local information. Performance degradation is observed in the $k_a = 7$ setting. To address this, we introduce early stopping in this setting ($k_a = 7*$), resulting in relatively better results. This suggests that a blind-spot with $k_a = 7$ can partially suppress spatial noise but not entirely. Increasing epochs leads the network to learn from outside the blind-spot to reconstruct central noise.

Ensemble of Teacher Networks. To investigate how the student network benefits from learning various teachers, we conducted additional ablation experiments. A straightforward approach is to ensemble multiple teacher networks with different blind-spots by averaging their outputs for the final denoising result. Tab. 2 displays the performance achieved by averaging results from teacher networks across different sets Ks. Performance tends to increase with larger
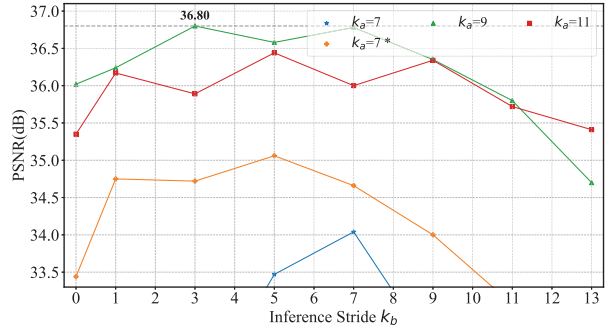

Figure 7. Ablation experiments on the combinations of different blind-spots between training and inference.

Ks, indicating that teacher networks with distinct blind-spots perform differently across image areas, and weighting them effectively enhances performance. However, excessively large Ks can lead to overly smooth average images, resulting in decreased performance.

Different Approaches of Distillation. We compare two methods of distillation: mean teacher distillation, utilizing the average outputs of various teachers for student training, and multi-teacher distillation, our chosen method in this study. Tab. 3 presents the outcomes of distilling three UNets with varying parameters using these approaches, with the latter consistently yielding superior performance. We omit a prior for distinguishing image areas, allowing the student to adaptively capture complementary information from different teachers, learning from multiple perspectives and thus enhancing generalization [14]. Conversely, mean teacher distillation employs pixel averaging as a prior, reducing student network robustness and performance.

Complexity Analysis. Tab. 4 compares the complexity of different methods. Our method attains superior performance with the lowest computational overhead. It's worth noting that both AP-BSN and LG-BPN utilize R^3 [27] operations, substantially increasing computational overhead.

5. Conclusions

In this paper, we first analyze existing self-supervised denoising techniques, highlighting the importance of training at the original resolution structure and using asymmetric operations. We then introduce a new approach using asymmetric blind-spots to balance noise suppression and spatial structure preservation, and present a blind-spots based multi-teacher distillation strategy. Experimental results show that our method achieves state-of-the-art and is superior in computational complexity and visual effect.

Acknowledgment

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 1, 2, 6

[2] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *ICCV*, 2019. 1, 2, 7

[3] Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *ICML*, 2019. 2, 7

[4] Benoit Brummer and Christophe De Vleeschouwer. Natural image noise dataset. In *CVPR Workshops*, 2019. 1, 2

[5] Sungmin Cha, Taeeon Park, and Taesup Moon. GAN2GAN: Generative noise learning for blind image denoising with single noisy images. In *ICLR*, 2021. 1, 2

[6] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20:89–97, 2004. 2

[7] Priyam Chatterjee, Neel Joshi, Sing Bing Kang, and Yasuyuki Matsushita. Noise suppression in low-light images through joint denoising and demosaicing. In *CVPR*, 2011. 1, 3

[8] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *CVPR*, 2018. 1, 2, 7

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017. 4

[10] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. NBNet: Noise basis learning for image denoising with subspace projection. In *CVPR*, 2021. 2

[11] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE TIP*, 16(8):2080–2095, 2007. 2, 7

[12] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *TIP*, 15: 3736–3745, 2006. 2

[13] Zixuan Fu, Lanqing Guo, and Bihan Wen. srgb real noise synthesizing with neighboring correlation-aware noise model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1683–1691, 2023. 2

[14] Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701, 2017. 8

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2

[16] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *CVPR*, 2014. 2, 7

[17] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, 2019. 1, 2, 7

[18] Zhiwei Hong, Xiaocheng Fan, Tao Jiang, and Jianxing Feng. End-to-end unpaired image denoising with conditional adversarial networks. In *AAAI*, 2020. 1, 2

[19] Xiaowan Hu, Ruijun Ma, Zhihong Liu, Yuanhao Cai, Xiaole Zhao, Yulun Zhang, and Haoqian Wang. Pseudo 3D auto-correlation network for real image denoising. In *CVPR*, 2021. 2

[20] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2Neighbor: Self-supervised denoising from single noisy images. In *CVPR*, 2021. 2

[21] Geonwoon Jang, Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. C2N: Practical generative noise modeling for real-world denoising. In *ICCV*, 2021. 1, 2, 7

[22] Yeong Il Jang, Keuntek Lee, Gu Yong Park, Seyun Kim, and Nam Ik Cho. Self-supervised image denoising with down-sampled invariance loss and conditional blind-spot network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12196–12205, 2023. 2, 3, 7

[23] Qiyu Jin, Gabriele Facciolo, and Jean-Michel Morel. A review of an old dilemma: Demosaicking first, or denoising first? In *CVPR Workshops*, 2020. 1, 3

[24] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *CVPR*, 2020. 2, 7

[25] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void-learning denoising from single noisy images. In *CVPR*, 2019. 1, 2, 5, 7

[26] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. In *NeurIPS*, 2019. 2, 5, 7

[27] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *CVPR*, pages 17725–17734, 2022. 2, 3, 4, 5, 6, 7, 8

[28] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *ICML*, 2018. 1, 2, 3

[29] Junyi Li, Zhilu Zhang, Xiaoyu Liu, Chaoyu Feng, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Spatially adaptive self-supervised learning for real-world image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9924, 2023. 2, 3, 4, 7, 8

[30] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level Wavelet-CNN for image restoration. In *CVPR Workshops*, 2018. 7

[31] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *CVPR*, pages 13365–13374, 2021. 2, 7

[32] Ruijun Ma, Shuyi Li, Bob Zhang, and Zhengming Li. Generative adaptive convolutions for real-world noisy image denoising. In *AAAI*, pages 1935–1943, 2022. 1, 2

[33] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2Noise: Learning to denoise from unpaired noisy data. In *CVPR*, 2020. 2

[34] Reyhaneh Neshatavar, Mohsen Yavartanoo, Sanghyun Son, and Kyoung Mu Lee. Cvf-sid: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image. In *CVPR*, pages 17583–17591, 2022. 2, 3, 7, 8

[35] Yizhong Pan, Xiao Liu, Xiangyu Liao, Yuanzhouhan Cao, and Chao Ren. Random sub-samples generation for self-supervised real image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12150–12159, 2023. 2, 3, 7, 8

[36] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-Recorrupted: Unsupervised deep learning for image denoising. In *CVPR*, 2021. 2, 7

[37] Sung Hee Park, Hyung Suk Kim, Steven Lansel, Manu Parmar, and Brian A Wandell. A case for denoising before demosaicking color filter array data. In *Asilomar Conference on Signals, Systems and Computers*, 2009. 1, 3

[38] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 6

[39] Luminita A. Vese and S. Osher. Modeling textures with total variation minimization and oscillating patterns in image processing. *Journal of Scientific Computing*, 19:553–572, 2003. 2

[40] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *WACV*, pages 1451–1460. Ieee, 2018. 4

[41] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *CVPR*, pages 2027–2036, 2022. 2

[42] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. 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*, pages 18156–18165, 2023. 2, 3, 4, 7, 8

[43] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *ECCV*, 2020. 1, 2, 5, 7

[44] Jun Xu, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao. Noisy-As-Clean: Learning self-supervised denoising from corrupted image. *IEEE TIP*, 29: 9316–9329, 2020. 2, 7

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015. 4

[46] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: Toward blind noise modeling and removal. In *NeurIPS*, 2019. 2, 7

[47] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *ECCV*, 2020. 1, 2, 7

[48] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 2, 7

[49] Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE TIP*, 27(9):4608–4622, 2018. 1, 2

[50] Yi Zhang, Dasong Li, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Idr: Self-supervised image denoising via iterative data refinement. In *CVPR*, pages 2098–2107, 2022. 2

[51] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. When AWGN-based denoiser meets real noises. In *AAAI*, 2020. 2, 3