Spatially Adaptive Self-Supervised Learning for Real-World Image Denoising

Junyi Li[1], Zhilu Zhang[1], Xiaoyu Liu[1], Chaoyu Feng, Xiaotao Wang, Lei Lei, Wangmeng Zuo[1,2](✉)

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]Peng Cheng Laboratory, China

nagejacob@gmail.com, cszlzhang@outlook.com, liuxiaoyu1104@gmail.com, wmzuo@hit.edu.cn

Abstract

Significant progress has been made in self-supervised image denoising (SSID) in the recent few years. However, most methods focus on dealing with spatially independent noise, and they have little practicality on real-world sRGB images with spatially correlated noise. Although pixel-shuffle downsampling has been suggested for breaking the noise correlation, it breaks the original information of images, which limits the denoising performance. In this paper, we propose a novel perspective to solve this problem, i.e., seeking for spatially adaptive supervision for real-world sRGB image denoising. Specifically, we take into account the respective characteristics of flat and textured regions in noisy images, and construct supervisions for them separately. For flat areas, the supervision can be safely derived from non-adjacent pixels, which are much far from the current pixel for excluding the influence of the noise-correlated ones. And we extend the blind-spot network to a blind-neighborhood network (BNN) for providing supervision on flat areas. For textured regions, the supervision has to be closely related to the content of adjacent pixels. And we present a locally aware network (LAN) to meet the requirement, while LAN itself is selectively supervised with the output of BNN. Combining these two supervisions, a denoising network (e.g., U-Net) can be well-trained. Extensive experiments show that our method performs favorably against state-of-the-art SSID methods on real-world sRGB photographs. The code is available at `https://github.com/nagejacob/SpatiallyAdaptiveSSID`.

1. Introduction

Image denoising aims to restore clean images from noisy observations [5, 11, 14], and it has achieved noticeable improvement with the advances in deep networks [2, 10, 21, 29–31, 33, 38, 40, 41, 46–48, 50, 51]. However, the models trained with synthetic noise usually perform poorly in real-world scenarios in which noise is complex and change-

(a) Noisy Input   (b) CVF-SID [35]

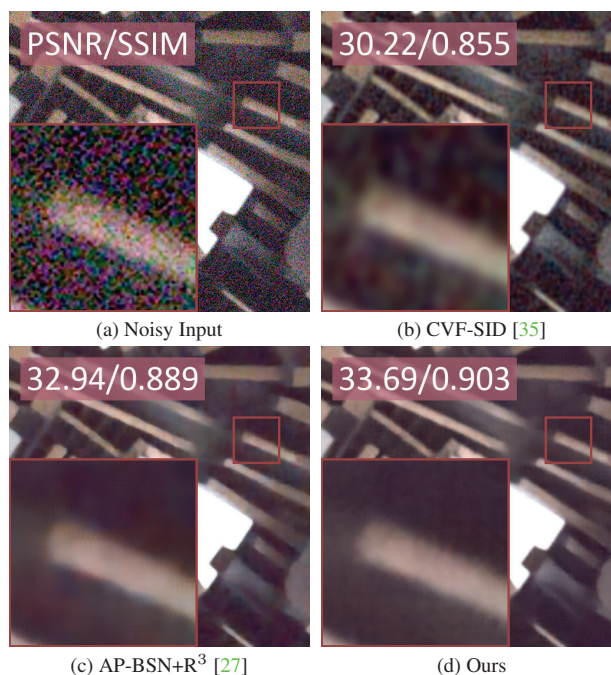(c) AP-BSN+R³ [27]   (d) Ours

Figure 1. Visual comparison between self-supervised denoising methods on the DND dataset [37]. PSNR (dB) and SSIM with respect to the ground-truth are marked on the result for quantitative comparison. Our method performs better in removing spatially correlated noise from real-world sRGB photographs.

able. A feasible solution is to collect real-world clean-noisy image pairs [1, 37] and take them for model training [2, 15, 21, 46, 47]. But building such datasets generally requires strictly controlled environment as well as complicated photographing and post-processing, which is time-consuming and labor-intensive. Moreover, the noise statistics vary under different cameras and illuminating conditions [43, 54], and it is impractical to capture pairs for every device and scenario.

To circumvent the limitations of noisy-clean pairs collection, self-supervised image denoising (SSID) ap-

proaches [3, 17, 19, 20, 24, 26–28, 34–36, 42, 45] have been proposed, which can be trained merely on noisy images. However, the noise model assumptions of a large amount of SSID methods do not match the characteristics of real-world noise in sRGB space. For instance, HQ-SSL [26] improves the denoising performance with posterior inference, but requires explicit noise probability density. Noise2Score [20] and its extension [19] propose a closed-form image denoising schema with score matching followed by noise model and noise level estimation, but the noise is bounded to Tweedie distribution. Although some methods [17,42] are designed for distribution agnostic noise, they can only deal with spatially independent noise.

Recently, a few attempts have been explored to remove spatially correlated noise in a self-supervised manner. CVF-SID [35] disentangles the image and noise components from noisy images, but the difficulty of optimization limits its performance. Some methods [44, 57] break the spatial noise correlation with pixel-shuffle downsampling (PD), then utilize spatially independent denoisers (*e.g.*, blind-spot network [6, 26, 44]) to remove the uncorrelated noise. However, PD breaks the original information of the images and leads to aliasing artifacts, which largely degrade the image quality. AP-BSN [27] applies asymmetric PD factors and post-refinement processing to seek for a better trade-off between noise removal and aliasing artifacts, but it is time-consuming during inference.

In this paper, we present a novel perspective for SSID by considering the respective characteristics of flat and textured regions in noisy images, resulting in a spatially adaptive SSID method for real-world sRGB images. Instead of utilizing pixel-shuffle downsampling and blind-spot network to learn denoising results directly, we seek for spatially adaptive supervision for a denoising network (*e.g.*, U-Net [39]). Concretely, for flat areas, the supervision can be safely derived from non-adjacent pixels, which are much far from the current pixel for excluding the influence of noise correlation. We achieve it by extending the blind-spot network (BSN) [26] to a blind-neighborhood network (BNN). BNN modifies the architecture of BSN to expand the size of blind region, and takes the same self-supervised training schema as BSN. Note that it is difficult to determine whether an area is flat or not from the noisy images, so we directly apply BNN to the whole image and it has little effect on the handling of flat areas. Moreover, such an operation can give us a chance to detect textured areas from the output of BNN, whose variance is usually higher. For textured areas, neighboring pixels are essential for predicting the details and they can not be ignored. To this end, we present a locally aware network (LAN), which focuses on recovering the texture details solely from adjacent pixels. LAN is supervised by flat areas of BNN output. When training is done, LAN will be applied to textured areas to generate supervision information for these areas.

Combining the learned supervisions for flat and textured areas, a denoising network can be readily trained. During inference, BNN and LAN can be detached, only the ultimate denoising network is used to restore clean images. Extensive experiments are conducted on SIDD [1] and DND [37] datasets. The results demonstrate our method is not only effective but also efficient. In comparison to state-of-the-art self-supervised denoising methods, our method behaves favorably in terms of both quantitative metrics and perceptual quality. The contributions of this paper can be summarized as follows:

- We propose a novel perspective for self-supervised real-world image denoising, *i.e.*, learning spatially adaptive supervision for a denoising network according to the image characteristics.

- For flat areas, we extend the blind-spot network to a blind-neighborhood network (BNN) for providing supervision information. For texture areas, we present a locally aware network (LAN) to learn that from neighboring pixels.

- Extensive experiments show our method has superior performance and inference efficiency against state-of-the-art SSID methods on real-world sRGB noise removal.

2. Related Work

2.1. Deep Image Denoising

The development of convolutional neural networks (CNNs) has led to great improvement on deep image denoising [50, 51]. DnCNN [50] outperforms traditional patch-based methods [5, 11, 14] on Gaussian denoising. FFDNet [51] takes a noise level map as input, which can handle various noise levels with a single model. Other learning based methods, such as RED30 [33], Mem-Net [40], and MWCNN [31], are also developed with advanced architectures. Nevertheless, models trained on AWGN generalize poorly to real scenarios due to the domain discrepancy between synthetic and real noise.

In order to mitigate the gap between synthetic and real noise, one feasible way is to simulate realistic noise as much as possible and introduce it during the network training [15, 49]. For example, CBDNet [15] inverses the demosaicing and gamma correction steps in image signal processing (ISP), then synthesizes signal-dependent Poisson-Gaussian noise [12] in raw space. Zhou *et al.* [57] break spatial-correlated noise into pixel-independent one with pixel-shuffle downsampling, then handle it with AWGN-based denoiser. Another way is to capture paired noisy-clean images for constructing real-world datasets [1, 37].

Taking such datasets for training, models have better potential to generalize to the corresponding real noise [2, 10, 21, 29, 38, 41, 46–48]. However, conducting well-aligned training pairs generally requires a controlled environment and much human labor. Moreover, noise statistics vary under different cameras and different illuminating conditions [43, 54]. It is impractical to collect datasets for every device and scenario.

2.2. Unpaired Image Denoising

Unpaired methods [8, 16, 18, 44] mitigate the data collecting issue by training networks with unpaired noisy and clean datasets. Most of which [8, 16, 18] learn to model the noise statistics with adversarial training. A noise generator is first trained to match the noise distribution of the noisy images, then used to map the clean images to pseudo-noisy ones. Finally, the denoising network is trained with synthetic pseudo-noisy and clean image pairs. In addition, Wu *et al.* [44] learn the noise distribution by jointly training a denoising network and a noise estimator. And the ultimate denoising network is trained with pseudo-noisy and clean pairs as well as noisy and denoised pairs. However, as the noise distribution in sRGB space is very complex and difficult to model [23], the performance of unpaired methods is still limited when facing real-world photographs.

2.3. Self-Supervised Image Denoising

In order to get rid of dependence on clean images, self-supervised methods are proposed, which are trained with noisy images only. Noise2Noise [28] suggests to learn a model from paired noisy images, which remains limited in practice. Subsequently, Noise2Void [24] and Noise2Self [3] separate noisy images into input and target pairs with a mask strategy. Blind-spot networks take a step further by excluding the corresponding input noisy pixel from the receptive field for each output pixel, which can be implemented with multiple network branches [7, 26] or dilated and masked convolutions [6, 44]. Moreover, probabilistic inference [25, 26] and regular loss functions [42] are proposed to alleviate the information loss issue at the blind spot. In addition, Noisier2Noise [34] and NAC [45] take the noisy images as target, and synthesize and add new noise to the noisy images for model training. More recently, Noise2Score [20] and its extension [19] propose a closed-form denoising for Tweedie distributions with score matching and posterior inference. SelfIR [55] introduce blurry images for self-supervised denoising task. Nevertheless, the above self-supervised methods can only handle noisy images with spatially independent noise.

A few attempts have been done to remove the spatially correlated noise in a self-supervised manner. R2R [36] reverses the noisy images to raw space to synthesize training pairs, then renders the synthetic raw images back to sRGB space. But it requires several priors, such as the parameters of the camera ISP and noise model. StructN2V [4] adapts the blind mask from a single pixel [24] to match the structure of the noise. CVF-SID [35] disentangles noisy images into clean images and noise components. Among the self-supervised approaches for real-world sRGB noise, AP-BSN [27] propose asymmetric PD factors and post-refinement processing to make a better trade-off between noise removal and aliasing artifacts, but it is time-consuming during inference.

3. Method

Self-supervised denoising aims to predict the denoised images $\hat{\mathbf{x}}$ from the noisy observations $\mathbf{y}$, without the supervision from clean images $\mathbf{x}$. In this paper, we propose a new perspective for self-supervised denoising. Specifically, we extract appropriate supervisions from noisy images according to the image characters (*i.e.*, flatness) for denoising network training. In this section, we first introduce how to learn the supervision for flat (Sec. 3.1) and textured (Sec. 3.2) areas, respectively. Then we discuss how to utilize the supervisions for the denoising network in Sec. 3.3.

3.1. Supervision for Flat Areas

For a noisy image, denoising of flat areas is generally easier than that of textured areas whose edges and details need to be preserved well [56]. And the same goes for our extraction of supervision information. Thus, we handle these two areas separately with the consideration of their different characteristics. In this subsection, we first learn the supervision for flat areas.

For removing spatially correlated noise in real-world sRGB images, pixel-shuffle downsampling (PD) has been applied to break the noise correlation [27, 44, 57]. Then blind-spot network (BSN) [6, 26, 44] can be applied to the downsampled sub-images to remove the spatially independent noise. However, the high-frequency information is distorted during downsampling [13], leading to aliasing artifacts. Moreover, apart from noise-correlated neighbor pixels, a large amount of non-neighboring pixels are removed, which degrades the denoising performance.

In terms of flat areas in noisy images, a pixel of the underlying clean images is similar to a large number of surrounding pixels, not only the neighboring pixels but also non-neighboring ones. The non-neighboring pixels can be far enough from the current pixel that exceeds the noise correlation range. Thus, instead of PD, we can enlarge the blind-spot of BSN into a blind-neighborhood until covering all noise-correlated pixels. We present a blind-neighborhood network (BNN) to achieve it, as shown in Figure 2(a). Different from PD, BNN can make full use of the uncorrelated noisy pixels and preserve the original information as much as possible. Similar to BSN, the loss
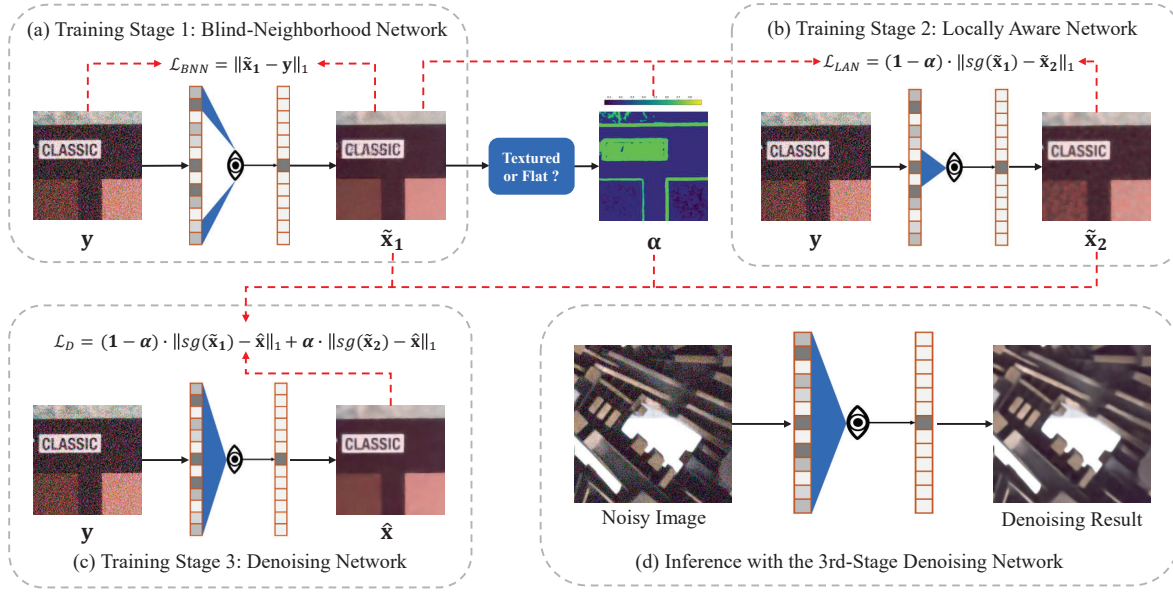
Figure 2. Overview of our self-supervised denoising framework. (a) In training stage one, the blind-neighborhood network (BNN) learns to remove the spatially correlated noise for flat areas in a self-supervised manner. (b) In training stage two, the locally aware network (LAN) is supervised by flat areas of BAN outputs. When training is done, LAN will be applied to texture areas to generate texture details. (c) In training stage three, the denoising network is supervised by the first two stages' results with adaptive coefficients. (d) During inference, the 3rd-stage denoising network can be deployed to denoise real-world photographs.

function of BNN is as follows:

$$\mathcal{L}_{BNN} = \|\tilde{\mathbf{x}}_1 - \mathbf{y}\|_1, \tag{1}$$

where $\hat{\mathbf{x}}_1$ is the output of BNN, $\mathbf{y}$ is the noisy image. As the random noise can not be predicted, the output of BNN will converge to the clean image.

In addition, the network architecture of BNN is modified from the BSN used in HQ-SSL [26]. The BSN applies four network branches whose receptive fields are restricted in different directions. At the end of each branch, a single pixel shift is applied to the features to create the blind-spot. We increase the pixel shift size from 1 to k to create a $(2k-1) \times (2k-1)$ blind-neighborhood. Referring to the analysis of spatial correlation on real-world noise in AP-BSN [27], we set $k=5$ for 9×9 blind-neighborhood. More experimental analysis on the blind-neighborhood size can be seen in Sec. 5.3. The details of network architecture are provided in the supplementary material.

3.2. Supervision for Texture Areas

Note that we apply BNN to the whole image. On the one hand, it is difficult to predetermine whether an area is flat or not from noisy images. Such an operation has little effect on the supervision prediction of flat areas. On the other hand, textured areas can be easily detected from the full output of BNN for their supervision extraction. In this subsection,

we illustrate how to determine and extract supervision for textured areas.

Determination of Textured Areas. In flat areas, BNN sufficiently removes the noise, and the corresponding variance is low. In texture areas, although BNN is ambiguous in reserving details, the output is still not flat and the variance is high. With the above observations, we can infer the flatness from the local variance of BNN output $\tilde{\mathbf{x}}_1$. Specifically, we densely extract image patches for each spatial location (i, j), then calculate the standard deviation map $\boldsymbol{\sigma}$,

$$\boldsymbol{\sigma}(i,j) = std(\tilde{\mathbf{x}}_1(i - \frac{n-1}{2} : i + \frac{n-1}{2}, j - \frac{n-1}{2} : j + \frac{n-1}{2})), \tag{2}$$

where $std(\cdot)$ represents the standard deviation function and is measured on 1-channel patches by averaging RGB values. n denotes the local window size, which we empirically set $n = 7$. Binarization $\boldsymbol{\sigma}(i, j)$ to determine the flatness (textured or flat) is rough and sometimes incorrect. Instead, we convert the standard deviation map $\boldsymbol{\sigma}$ into soft coefficients $\boldsymbol{\alpha}$ which are normalized to $[0, 1]$. In our experiments, $\boldsymbol{\sigma}$ is a reliable indicator that it is usually high in the textured areas (e.g., edges, texts) and is low in flat areas. Io generate the coefficient $\boldsymbol{\alpha}$ that indicating flatness, we convert $\boldsymbol{\sigma}$ to $\boldsymbol{\alpha}$

with a piecewise function,

$$\boldsymbol{\alpha}(i,j) = \begin{cases} S(\boldsymbol{\sigma}(i,j) - 1), & \boldsymbol{\sigma}(i,j) \leq l \\ 0.5, & l < \boldsymbol{\sigma}(i,j) \leq u \\ S(\boldsymbol{\sigma}(i,j) - 5), & \boldsymbol{\sigma}(i,j) > u \end{cases} \quad (3)$$

where $S(\cdot)$ denotes the Sigmoid function. We empirically set $l = 1$ and $u = 5$ respectively. Higher $\boldsymbol{\alpha}(i,j)$ means the local area is more textured.

Locally Aware Network. Blind-neighborhood network performs poorly in texture areas, as it ignores the adjacent pixels which are essential for details prediction. To extract proper supervision for texture areas, we present a locally aware network (LAN). LAN is carefully designed to have a local receptive field, which focuses on recovering the texture details from neighboring pixels. Nonetheless, LAN is a supervised network to be well-trained that requires clean supervision. Fortunately, the flat areas of BNN output give a chance to supervise LAN. On the one hand, the flat areas of BNN are well-denoised to approximate clean signal, and LAN can be trained safely. On the other hand, The small receptive field of LAN makes it able to preserve texture details. The loss function for training LAN is as follows:

$$\mathcal{L}_{LAN} = (\mathbf{1} - \boldsymbol{\alpha}) \cdot \|sg(\tilde{\mathbf{x}}_1) - \tilde{\mathbf{x}}_2\|_1, \quad (4)$$

where $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ is the output of BNN and LAN, respectively. $sg(\cdot)$ denotes stop gradient operation [9]. When training is done, LAN will have a transfer ability to predict the supervision for textured areas.

The network structure of LAN is simple yet delicate. We stack 3×3 convolution layers to create the local receptive field, specifically, k 3×3 layers can make up $(2k+1) \times (2k+1)$ receptive field. In order to further refine the color information, we additionally add several 1×1 convolution blocks with channel attention mechanism [53]. More details are provided in the supplementary material.

3.3. Denoising with Learned Supervisions

The learned images of BNN and LAN can be leveraged in two ways: image-level fusion or use as supervisions to train a denoising network. We note that image-level fusion suffers from performance and efficiency issues. From the performance perspective, although BNN and LAN are designed to denoise flat and texture areas respectively, they still have their weaknesses. BNN may generate oversmooth results in flat areas, while LAN can not completely remove the noise in texture areas. Image-level fusion can not avoid these drawbacks and only provides limited improvement. Instead, utilizing the results as supervision to train a denoising network achieves a better trade-off between detail preserving and noise removal, which yields better performance. From the efficiency perspective, BNN

is computationally complex due to the special network design for blind-neighborhood. Training additional denoising network provides more flexibility on the network complexity. Thus, we tend to train a denoising network with learned spatially adaptive supervisions. More experimental analysis about the two ways can be seen in Sec. 5.2.

We choose a common and representative network structure, *i.e.*, U-Net [39], as our denoising network. As shown in Figure 2(c), the U-Net is trained with the following loss,

$$\mathcal{L}_D = (\mathbf{1} - \boldsymbol{\alpha}) \cdot \|sg(\tilde{\mathbf{x}}_1) - \hat{\mathbf{x}}\|_1 + \boldsymbol{\alpha} \cdot \|sg(\tilde{\mathbf{x}}_2) - \hat{\mathbf{x}}\|_1, \quad (5)$$

where $\hat{\mathbf{x}}$ is the output of U-Net, $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ are the output of BNN and LAN, respectively. $sg(\cdot)$ is stop gradient operation [9] and $\boldsymbol{\alpha}$ is the adaptive coefficients mentioned above.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct experiments on SIDD [1] and DND [37], which are widely used datasets for real-world image denoising. SIDD dataset captures noisy-clean pairs with smartphone cameras. Each noisy image is captured multiple times, and the mean image is served as the ground-truth. It provides 320 image pairs (SIDD-Medium) for training, 1,280 patches for validation, and 1280 patches for benchmark testing. DND [37] dataset provides 50 image pairs for testing only. Its noisy-clean pairs are conducted by shooting the same scene twice with different ISO settings. The high-ISO images are taken as noisy inputs, while the corresponding low-ISO images are nearly noise-free and can serve as ground-truth. The benchmark testing results (*i.e.*, PSNR and SSIM) can be achieved by uploading the denoised patches to their official websites. We note that SIDD and DND benchmarks provide evaluations in both the raw space and sRGB space. As our method is developed to remove the spatially correlated noise of sRGB images, we do not compare with the self-supervised denoising methods [6,17,42] for raw images.

Implementation Details. We only leverage the noisy images of SIDD Medium dataset [1] to train our denoising framework. Our BNN, LAN and the denoising network are trained successively with the same training settings. We crop the training images into patches of size 256×256, and augment the image patches with random flipping and rotation. The augmented patches are formed as mini-batches of size 8 to facilitate network training. Each network is trained with Adam optimizer [22] for 400k iterations, for a total of 1200k iterations. The learning rate is initially set to 3×10^{-4}, and decreased to zero with cosine annealing scheduler [32].

Table 1. Quantitative comparison of PSNR (dB), SSIM and LPIPS on SIDD [1] and DND [37] datasets. LPIPS is calculated on SIDD validation dataset only as the ground-truth images on benchmark datasets are not available. Hereinafter, **red** and blue indicate the best and the second best results among unpaired and self-supervised methods, respectively.

	Method	SIDD Validation PSNR↑ / SSIM↑ / LPIPS↓	SIDD Benchmark PSNR↑ / SSIM↑	DND Benchmark PSNR↑ / SSIM↑
Non-learning based	BM3D [11]	25.71 / 0.576 / 0.657	25.65 / 0.685	34.51 / 0.851
	WNNM [14]	26.05 / 0.592 / 0.635	25.78 / 0.809	34.67 / 0.865
Supervised (Synthetic pairs)	DnCNN [50]	26.21 / 0.604 / 0.712	26.25 / 0.599	32.43 / 0.790
	CBDNet [15]	33.07 / 0.863 / 0.288	33.28 / 0.868	38.05 / 0.942
	Zhou *et al.* [57]	33.96 / 0.899 / 0.258	34.00 / 0.898	38.40 / 0.945
Supervised (Real pairs)	DnCNN [50]	37.73 / 0.943 / 0.245	37.61 / 0.941	38.73 / 0.945
	Baseline, N2C [39]	38.98 / 0.954 / 0.201	38.92 / 0.953	39.37 / 0.954
	VDN [46]	39.29 / 0.956 / 0.208	39.26 / 0.955	39.38 / 0.952
	Restormer [48]	39.93 / 0.960 / 0.198	40.02 / 0.960	40.03 / 0.956
Unpaired	GCBD [8]	-	-	35.58 / 0.922
	UIDNet [16]	-	32.48 / 0.897	-
	C2N [18]	35.36 / 0.932 / 0.192	35.35 / **0.937**	37.28 / 0.924
	Wu *et al.* [44]	-	-	37.93 / 0.937
Self-Supervised	Noise2Void [24]	27.48 / 0.664 / 0.592	27.68 / 0.668	-
	Noise2Self [3]	29.94 / 0.782 / 0.556	29.56 / 0.808	-
	NAC [45]	-	-	36.20 / 0.925
	R2R [36]	-	34.78 / 0.898	-
	CVF-SID [35]	34.15 / 0.911 / 0.423	34.71 / 0.917	36.50 / 0.924
	AP-BSN+R³ [27]	36.74 / **0.934** / 0.281	36.91 / 0.931	38.09 / 0.937
	Ours	**37.39** / **0.934** / **0.176**	**37.41** / 0.934	**38.18** / **0.938**

(a) Noisy Input   (b) Baseline, N2C [39]   (c) Noise2Void [24]   (d) CVF-SID [35]   (e) AP-BSN+R³ [27]   (f) Ours
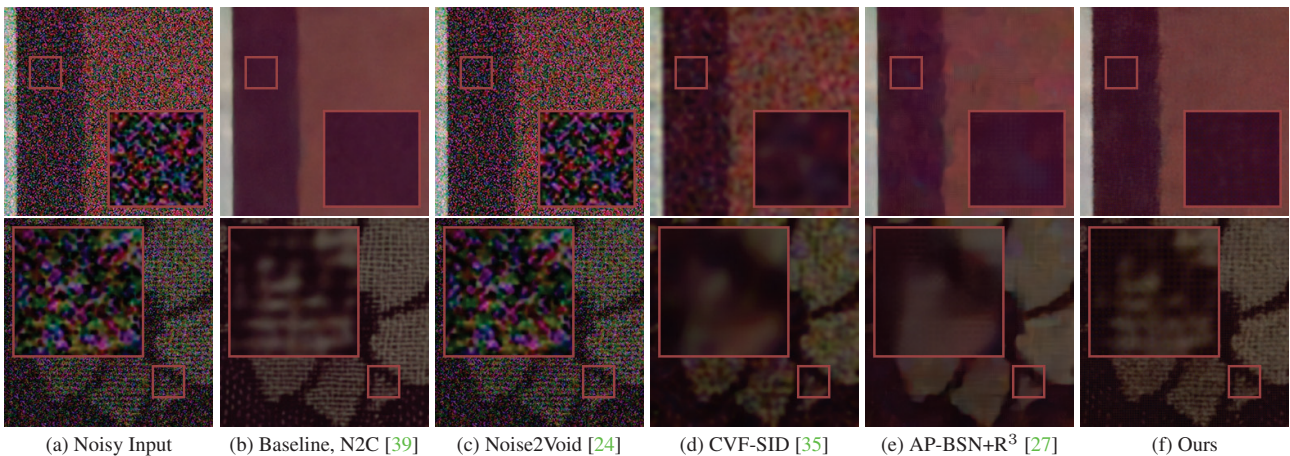
Figure 3. Qualitative comparison on SIDD benchmark dataset [1]. Sub-figure (b) denotes the result of our denoising network trained on SIDD dataset in a supervised manner. Sub-figures (c)-(f) are from self-supervised denoising methods.

4.2. Results for Real-World Denoising

Quantitative Comparison. The quantitative comparison results of SIDD [1] and DND [37] datasets can be seen in Table 1. Our method outperforms all unpaired and self-supervised approaches. In terms of PSNR and SSIM metrics, our method is comparable to DnCNN [50] trained with

real-world pairs. It demonstrates our method further mitigates the performance gap between self-supervised methods and supervised ones. Among self-supervised methods, Noise2Void [24] and Noise2Self [3] fail to handle the noise in sRGB images due to their spatially independent noise assumption. NAC [45] and R2R [36] depend on the parameters of the camera ISP to simulate training pairs, which

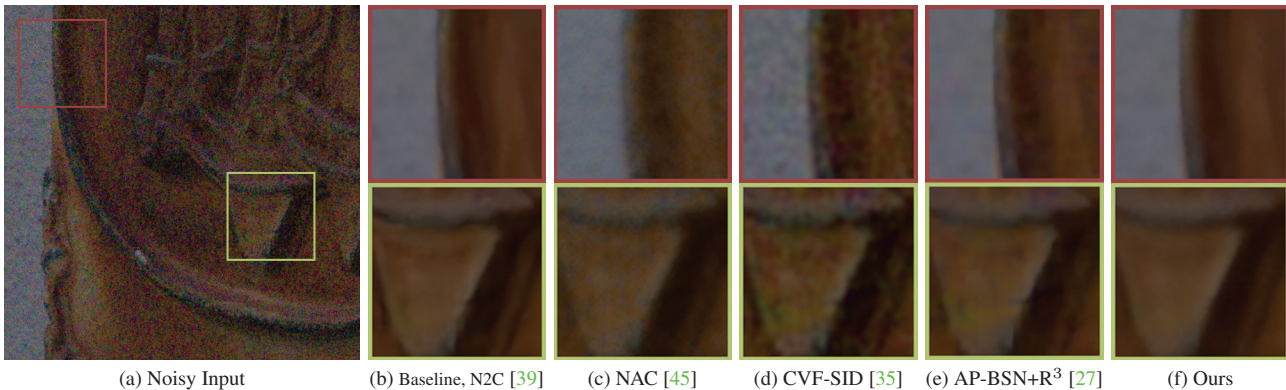| (a) Noisy Input | (b) Baseline, N2C [39] | (c) NAC [45] | (d) CVF-SID [35] | (e) AP-BSN+R^3 [27] | (f) Ours |

Figure 4. Qualitative comparison on DND benchmark dataset [37]. Sub-figures (c)-(f) are from self-supervised denoising methods.

are usually unavailable. CVF-SID [35], AP-BSN [27] and our method are directly trained on noisy images, while our method achieves the best performance. We also provide LPIPS [52] results on SIDD validation dataset to measure the perceptual quality. Our method shows the lowest LPIPS score in all competing denoising methods. It shows that our denoised images are most perceptually similar to the ground-truth.

Qualitative Comparison. The visual comparison of state-of-the-art self-supervised methods on the benchmark datasets is shown in Figure 3 and Figure 4. As the clean ground-truth image is not provided, we train a supervised counterpart of our denoising network for reference, which is called baseline N2C. From Figure 3, our method removes most spatially correlated noise and recovers the texture details better. Other methods hardly generate visually plea-sure results. Noise2Void [24] has little denoising effect and seems to converge into the noisy image. CVF-SID [35] can not decouple the noisy and clean components well from the noisy observations, leading to partially denoised results. AP-BSN [27] suffers from aliasing artifacts due to pixel-shuffle downsampling, e.g., the texture details on the fabric are missed. The qualitative comparison on DND benchmark testing dataset [37] is shown in Figure 4. In comparison with other competitive methods, our result is cleaner and more photo-realistic.

5. Ablation Study

We conduct extensive ablation studies on SIDD valida-tion dataset [1] to analyze the effectiveness of the proposed framework.

5.1. Effects of Supervision Components

Table 2 shows that each component of the proposed model is essential for the denoising network training. From the table, $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ are not ideal supervisions and have their pros and cons. Supervision by $\tilde{\mathbf{x}}_1$ clearly removes the noise but at the risk of smoothing-out fine details. Supervi-sion by $\tilde{\mathbf{x}}_2$ preserves the texture details but exhibits inferior denoising effect. The network trained with both supervi-sions makes a tradeoff between noise removal and texture preserving, thus showing better results than separate ones. Each one of them alone is not sufficient enough to facilitate the denoising network training. Specifically, using BNN output as supervision alone leads to 0.46dB performance drop, and using LAN output as supervision alone widens the gap to 1.35dB. This demonstrates that both BNN and LAN provide indispensable supervisions for training the denois-ing network. Adaptive coefficients $\boldsymbol{\alpha}$ is used for balancing the supervision strength of BNN and LAN outputs accord-ing to the local flatness. We apply $\boldsymbol{\alpha}$ to the loss function (i.e., Eqn. (5)) to emphasize the BNN output in flat areas, as well as the LAN output in texture areas. It provides 0.09dB performance boost, which demonstrates the effectiveness of learning spatially adaptive supervision for self-supervised denoising.

5.2. Comparison with Image-Level Fusion

Apart from as supervisions, the outputs of BNN and LAN can also serve as denoised images directly, and fur-ther fused to get better results. However, as illustrated in Sec. 3.3, the way of image-level fusion have performance and efficiency limits. We try spatially adaptive fusion strat-egy that the outputs of BNN and LAN are weighted av-eraged according to the adaptive coefficients $\boldsymbol{\alpha}$, the result is shown in Table 3. Although the fusion strategy has im-provement over individual images, it still falls 0.55dB from our proposed method. Besides, due to the complexity of BNN, the inference time of the fusion strategy is much higher. Figure 5 provides visual comparison between the fused image and our denoising network output $\hat{\mathbf{x}}$. $\hat{\mathbf{x}}$ shows better denoising effects. It further demonstrates the effec-tiveness of taking BNN and LAN outputs as supervisions.

Table 2. Ablation study of supervision components. The notation of $\tilde{\mathbf{x}}_1$, $\tilde{\mathbf{x}}_2$ and $\boldsymbol{\alpha}$ follows Eqn. (5).

Supervision of $\tilde{\mathbf{x}}_1$	✓		✓	✓
Supervision of $\tilde{\mathbf{x}}_2$		✓	✓	✓
Adaptive Coefficients $\boldsymbol{\alpha}$				✓
PSNR of $\hat{\mathbf{x}}$	36.84	35.95	37.30	37.39

Table 3. Comparison between image-level fusion strategy and our method.

Image	$\tilde{\mathbf{x}}_1$	$\tilde{\mathbf{x}}_2$	$(1-\boldsymbol{\alpha}) \cdot \tilde{\mathbf{x}}_1 + \boldsymbol{\alpha} \cdot \tilde{\mathbf{x}}_2$	$\hat{\mathbf{x}}$
PSNR	36.37	35.00	36.84	37.39
Time (ms)	16.7	5.9	22.9	4.8

(a) $\tilde{\mathbf{x}}_1$ (b) $\tilde{\mathbf{x}}_2$ (c) $\boldsymbol{\alpha}$ (d) Fused (e) $\hat{\mathbf{x}}$
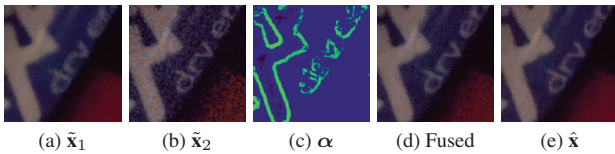
Figure 5. Visual comparison with image-level fusion strategy. (d) denotes the spatially adaptive fusion result $(1-\boldsymbol{\alpha}) \cdot \tilde{\mathbf{x}}_1 + \boldsymbol{\alpha} \cdot \tilde{\mathbf{x}}_2$.

5.3. Blind-Neighborhood Size

As illustrated in Sec. 3.1, the blind-neighborhood of BNN should cover the spatially correlated noisy pixels accurately. Figure 6(a) analyses the effects of blind-neightborhood size of BNN. For 1×1 blind-neighborhood size, BNN degrades to BSN which does not work for spatially correlated noise. When gradually increasing the blind-neighborhood size, BNN shows better denoising effect as more noise-correlated pixels are excluded from the receptive field. The max performance of BNN is achieved at 9×9 blind-neighborhood size, which means all the noise correlation pixels have lain in this range. It is consistent with the observation in AP-BSN [27]. Further increasing the blind-neighborhood size excludes noise-independent pixels from the receptive field, which is harmful to denoising performance. The visualization of denoised images in Figure 7(a) also supports the 9×9 blind-neighborhood size. In addition, the calculation of adaptive coefficients $\boldsymbol{\alpha}$ and the training of LAN are highly depend on the denoising quality of BNN on flat areas. Thus, the performance of the denoising network shows the same trend as BNN.

5.4. Local Receptive Size

Figure 6(b) and 7(b) show the effects of local receptive size of LAN. For 1×1 local receptive size, LAN has little denoising effect due to no neighboring pixels are leveraged. For larger local receptive sizes, LAN can effectively predict the favorable signal from neighbor pixels with good details.
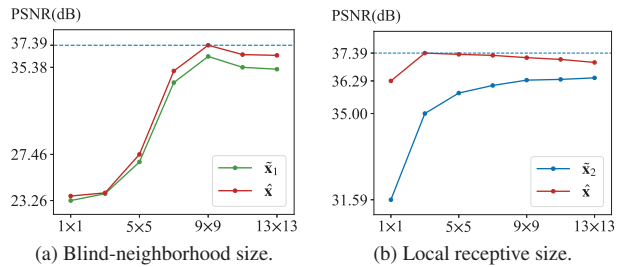
(a) Blind-neighborhood size. (b) Local receptive size.

Figure 6. Effects of the blind-neighborhood size and local receptive size. (a) BNN achieves best performance with blind-neighborhood size 9×9. (b) LAN achieves better performance as the local receptive size increases, but the final denoising network performs best with local receptive size 3×3.

1×1 3×3 5×5 7×7 9×9 11×11

(a) BNN output with different blind-neighborhood sizes.
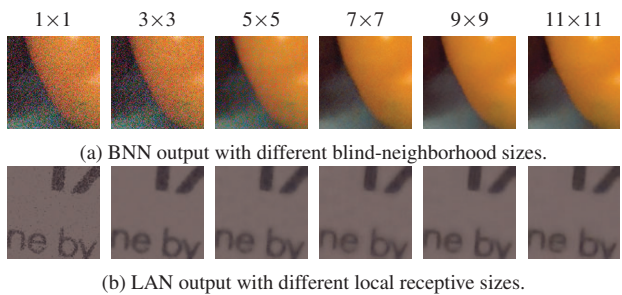
(b) LAN output with different local receptive sizes.

Figure 7. Visual comparison of different blind-neighborhood sizes and local receptive sizes.

But at the same time, the output of LAN tends to be more blurry. The outputs of LAN should serve as good supervision of texture areas for the denoising network training. The best performance of our denoising framework is achieved when LAN has 3×3 local receptive size.

6. Conclusion

In this work, we propose a novel perspective to solve real-world image self-supervised denoising, *i.e.*, seeking for spatially adaptive supervision for a denoising network according to the image characteristics. The clean signal in flat areas can be inferred from non-neighboring noise-independent pixels, while the texture details should come from neighboring ones. Thus, blind-neighborhood network (BNN) is proposed to learn the supervision for flat areas. Adaptive coefficients and locally aware network (LAN) are proposed to determine the flatness and learn the supervision for texture areas, respectively. The denoising network trained by learned supervisions outperforms state-of-the-art self-supervised and unpaired methods with better efficiency.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 1, 2, 5, 6, 7

[2] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3155–3164, 2019. 1, 3

[3] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 2, 3, 6

[4] Coleman Broaddus, Alexander Krull, Martin Weigert, Uwe Schmidt, and Gene Myers. Removing structured noise with self-supervised blind-spot networks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 159–163. IEEE, 2020. 3

[5] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005. 1, 2

[6] Jaeseok Byun, Sungmin Cha, and Taesup Moon. Fbi-denoiser: Fast blind image denoiser for poisson-gaussian noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5768–5777, 2021. 2, 3, 5

[7] Sungmin Cha and Taesup Moon. Fully convolutional pixel adaptive image denoiser. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4160–4169, 2019. 3

[8] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3155–3164, 2018. 3, 6

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 5

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4896–4906, 2021. 1, 3

[11] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 1, 2, 6

[12] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, 17(10):1737–1754, 2008. 2

[13] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009. 3

[14] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2869, 2014. 1, 2, 6

[15] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019. 1, 2, 6

[16] Zhiwei Hong, Xiaocheng Fan, Tao Jiang, and Jianxing Feng. End-to-end unpaired image denoising with conditional adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4140–4149, 2020. 3, 6

[17] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14781–14790, 2021. 2, 5

[18] Geonwoon Jang, Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. C2n: Practical generative noise modeling for real-world denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2350–2359, 2021. 3, 6

[19] Kwanyoung Kim, Taesung Kwon, and Jong Chul Ye. Noise distribution adaptive self-supervised image denoising using tweedie distribution and score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2016, 2022. 2, 3

[20] Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021. 2, 3

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3482–3492, 2020. 1, 3

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Shayan Kousha, Ali Maleky, Michael S Brown, and Marcus A Brubaker. Modeling srgb camera noise with normalizing flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17463–17471, 2022. 3

[24] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019. 2, 3, 6, 7

[25] Alexander Krull, Tomáš Vičar, Mangal Prakash, Manan Lalit, and Florian Jug. Probabilistic noise2void: Unsupervised content-aware denoising. *Frontiers in Computer Science*, 2:5, 2020. 3

[26] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. *Ad-*

vances in Neural Information Processing Systems, 32, 2019. 2, 3, 4

[27] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Apbsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022. 1, 2, 3, 4, 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 *International Conference on Machine Learning*, pages 4620–4631, 2018. 2, 3

[29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 1, 3

[30] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *Advances in neural information processing systems*, 31, 2018. 1

[31] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018. 1, 2

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[33] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems*, 29, 2016. 1, 2

[34] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12064–12072, 2020. 2, 3

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17583–17591, 2022. 1, 2, 3, 6, 7

[36] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2043–2052, 2021. 2, 3, 6

[37] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1586–1595, 2017. 1, 2, 5, 6, 7

[38] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8596–8606, 2021. 1, 3

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 5, 6, 7

[40] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017. 1, 2

[41] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1, 3

[42] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2027–2036, 2022. 2, 3, 5

[43] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020. 1, 3

[44] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *European conference on computer vision*, pages 352–368. Springer, 2020. 2, 3, 6

[45] 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 Transactions on Image Processing*, 29:9316–9329, 2020. 2, 3, 6, 7

[46] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32, 2019. 1, 3, 6

[47] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *European Conference on Computer Vision*, pages 41–58. Springer, 2020. 1, 3

[48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 3, 6

[49] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2696–2705, 2020. 2

[50] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1, 2, 6

[51] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 1, 2

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[53] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 5

[54] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4593–4601, 2021. 1, 3

[55] Zhilu Zhang, RongJian Xu, Ming Liu, Zifei Yan, and Wangmeng Zuo. Self-supervised image restoration with blurry and noisy pairs. In *Advances in Neural Information Processing Systems*, 2022. 3

[56] Wenzhao Zhao, Qiegen Liu, Yisong Lv, and Binjie Qin. Texture variation adaptive image denoising with nonlocal pca. *IEEE Transactions on Image Processing*, 28(11):5537–5551, 2019. 3

[57] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. When awgn-based denoiser meets real noises. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13074–13081, 2020. 2, 3, 6