# Iterative Denoiser and Noise Estimator for Self-Supervised Image Denoising

Yunhao Zou[1] Chenggang Yan[2] Ying Fu[1]*
[1]Beijing Institute of Technology [2]Hangzhou Dianzi University

## Abstract

*With the emergence of powerful deep learning tools, more and more effective deep denoisers have advanced the field of image denoising. However, the huge progress made by these learning-based methods severely relies on large-scale and high-quality noisy/clean training pairs, which limits the practicality in real-world scenarios. To overcome this, researchers have been exploring self-supervised approaches that can denoise without paired data. However, the unavailable noise prior and inefficient feature extraction take these methods away from high practicality and precision. In this paper, we propose a Denoise-Corrupt-Denoise pipeline (DCD-Net) for self-supervised image denoising. Specifically, we design an iterative training strategy, which iteratively optimizes the denoiser and noise estimator, and gradually approaches high denoising performances using only single noisy images without any noise prior. The proposed self-supervised image denoising framework provides very competitive results compared with state-of-the-art methods on widely used synthetic and real-world image denoising benchmarks.*

## 1. Introduction

Image denoising is a significant research problem in the field of image processing, which aims to remove noise while preserving the details and structures of the original image. Image denoising benefits not only visual quality perceived by the human eye but also downstream imaging problems, including remote sensing [23, 39], medical imaging [24], microscopy [22], and low-light imaging [6, 40, 44].

Numerous studies have been proposed for image denoising. Traditionally, researchers utilize hand-crafted priors to remove noise. Among them, wavelet [35], total variation [33], and self-similarity [3, 7] based methods are widely used. More recently, with the rapid development of computational resources, deep learning, specially convolutional neural network (CNN) based methods [12, 32, 34, 47, 48] have gained increasing attention due to their strong representation and spatial feature extraction ability. Since DnCNN [47], which is the earliest exploration of deep denoiser, a large number of related works emerged within a few years [12, 32, 34, 48]. Although CNN-based denoisers achieve very promising results compared to traditional methods, the severe dependence on high-quality noisy/clean training pairs limits their application in real scenes.

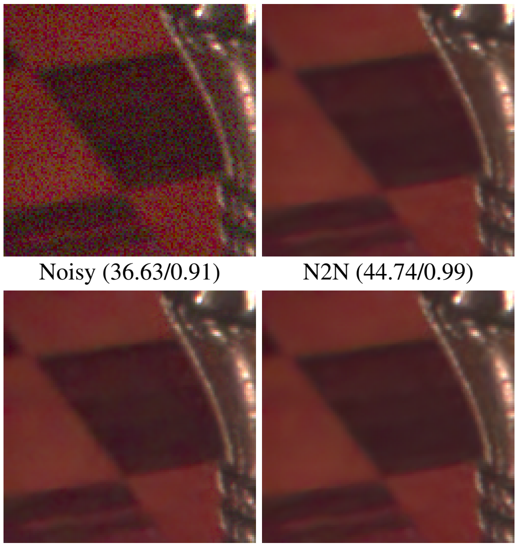

Noisy (36.63/0.91)
N2N (44.74/0.99)
N2V (43.04/0.99)
Ours (45.52/0.99)

Figure 1: Performance overview of our self-supervised denoising method from a single noisy image. Our performance surpasses baseline self-supervised method N2V [18], and has competitive results with a strong baseline N2N [21], which is trained on paired noisy observations. The PSNR/SSIM results are shown in the brackets.

To address the challenge of collecting paired real data, there are two common technical approaches. The first approach involves synthetic noisy/clean image pairs instead of real ones. To achieve this, modeling noise distribution becomes crucial to reduce the domain gap between the synthetic training data and real evaluation data. Early methods model noise distribution using simple statistical models [9, 25, 46]. In recent years, more accurate generative models [1, 5, 15, 43] and sophisticated statistical models [40, 49, 50] are proposed, which can better describe real

[*] Corresponding Author: fuying@bit.edu.cn

sensor noise distribution. Precise noise modeling is verified to be a good alternative for paired real data [40], but such methods still need camera-specific training/calibration data, which limits its application.

The other approach is to design unsupervised and self-supervised training strategies. These methods relieve the network from noisy/clean training pairs. Instead, their denoising model can be trained by unpaired [4] or weakly paired images. For example, Noise2Noise (N2N) [21] needs a clean image's two noisy observations for network learning. To further improve the practicality, in recent years, more methods [16, 18, 20, 30, 37, 38] intend to recover clean images from single noisy images. Among these methods, some of them [16,29,30,42] generate noisy/noisy pairs from a single noisy image, which free the learning of Noise2Noise from two noisy observations, but a prerequisite for these methods is knowing the noise models in advance. Others [18, 20, 38] propose customized blind-spot networks that are designed to avoid identity mapping, which discard important pixel information.

In this work, we propose a novel Denoise-Corrupt-Denoise training pipeline (DCD-Net) for self-supervised image denoising. By carefully analyzing the mainstream self-supervised denoising methods, we devise a solution that can simultaneously make up for the deficiencies of both the impractical and less accurate problems of existing self-supervised methods. Specifically, we present an iterative deep denoising pipeline that takes only single noisy images to approach the performance with additional training information (*e.g*., paired noisy observations for N2N [21]). We repeatedly follow the pipeline of denoising for clean prediction, estimating noise level, corrupting for N2N pairs, and performing N2N learning. Then, by iteratively training the denoiser and noise estimation model, the denoising performance gradually reaches a strong baseline, *i.e*., N2N, even though we have only single noisy observations. The proposed DCD-Net achieves promising results on widely used synthetic and real-world image denoising benchmarks.

The main contributions can be summarized as follows:

1. We carefully analyze the pros and cons of existing denoising methods and propose a novel Denoise-Corrupt-Denoise learning pipeline for self-supervised image denoising.
2. We propose to iteratively train the denoiser with a noise estimation model, which makes the denoising network gradually approach a strong weakly supervised baseline, *i.e*., N2N.
3. We apply our denoising framework on both synthetic and real image denoising benchmarks, and the results verify our superiority over leading self-supervised methods.

## 2. Related Work

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

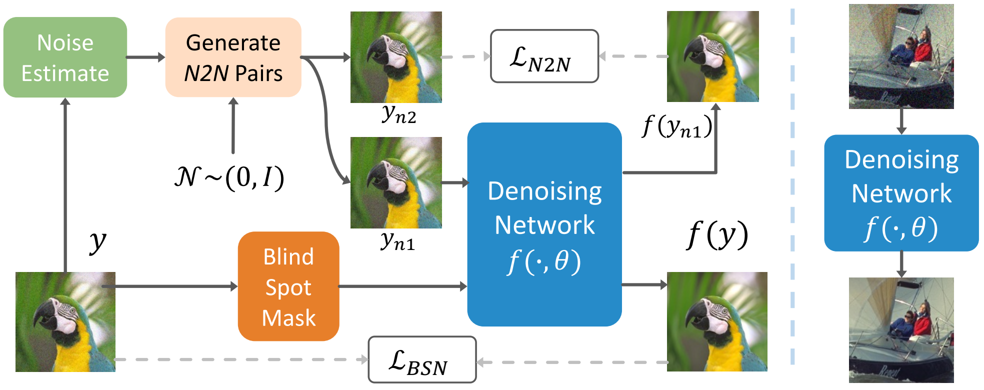
### 2.1. Supervised Image Denoising

In the early years, deep image denoising methods are fed with paired noisy/clean images. In 2017, DnCNN [47] first introduces CNN architecture to the field of image denoising, which consists of multiple convolutional layers with residual connections. Later, RED [27], FFDNet [48] and MemNet [34] further improve the denoising performances by introducing finer network structure, including densely connected layers and cascading architecture. CBDNet [12] further introduces to predict and feed a noise map to deep networks, in order to remove noise for images with unknown noise level. Recently, more and more researchers directly utilize the U-Net [32] architecture that is first introduced in medical image segmentation problem, and verify that it can be effectively used for image denoising problems. On synthetic noise removal, we can synthesize accurate noisy/clean pairs with known noise levels, making deep learning-based methods easily surpass traditional denoisers to a great deal. However, when it comes to denoising tasks in real scenarios [2, 31] where we do not know the noise distribution in advance, such methods suffer from severe domain gap and perform poorly on real scenes. Approaches including paired real data capture and noise modeling can mitigate this problem, but additional labor and data are required and further limit the applications for supervised methods.

### 2.2. Unsupervised and Self-supervised Image Denoising

Although traditional methods, including NLM [3] and BM3D [7] can directly remove noise from a single image, they heavily rely on hand-crafted prior and suffer from unbearable inference time. In recent years, self-supervised deep denoisers have been widely explored, and such methods can be roughly divided into two technical lines.

*Noise2Noise* (N2N) [21] is the pioneer in deep denoising models that attempts to weaken paired training data requirements. It theoretically proves that a CNN can directly estimate the true image by learning from two noisy observations, which have the same underlying clean image. Considering that such methods. Since the two noisy observations required by N2N still cannot meet most real scenarios, some methods [16, 29, 30, 42] attempt to generate noisy/noisy pairs from a single noisy image, and apply them to N2N learning. Noisy-As-Clean (NAC) [42] and Noisier2Noise [29] recorrupt the noisy image and directly train the model on a noisier input. Recorrupted-to-corrupted



(a) Training (b) Inference

Figure 2: The overview of iterative Denoise-Corrupt-Denoise pipeline for self-supervised image denoising.

(R2R) [30] design a data augmentation technique to generate noisy pairs from a single noisy image. Nevertheless, these methods suffer from inconvenience when noise model and level are not available in advance. Neighbor2Neighbor (NBR2NBR) [16] generates noisy pairs from subimages. It is indeed free from noise prior, but the underlying clean images differ.

*Noise2Void* (N2V) [18] is an improvement over N2N since it only needs a single noisy image for training. The main idea behind N2V is to use a part of the noisy image as input and predict the missing part of the same noisy image. The proposed network, *i.e.*, the blind-spot network is trained to predict the missing pixels from the surrounding context. Starting from N2V [18], a lot of methods design more powerful blind-spot network. DBSN [41] introduces a dilated blind-spot networks. AP-BSN [20] devises a blind-spot network based on asymmetric pixel-shuffle downsampling. Blind2Unblind (B2U) [38] proposes a global mask mapper, which is still a blind-spot network. With blind-spot architecture, such methods inevitably suffer from information loss at the pixels which are chosen as blind spots.

In this work, we focus on iteratively learning denoiser with noise estimator from single noisy images, leading to higher practicality than N2N data generation-based methods, and better performance than blind-spot methods.

## 3. Method

In this section, we first revisit the two foundations of recent self-supervised denoising methods. Then, we illustrate our observation and motivation for this work. Next, the details of our Denoise-Corrupt-Denoise self-supervised training pipeline are illustrated. Finally, the learning details are introduced. The overview framework of this work is illustrated in Fig. 2.

### 3.1. Revisit of Previous Works

N2N [21] and N2V [18] represent two mainstream technical lines for recent self-supervised denoising works [16, 18,20,29,30,38,42]. Assuming we have a set of noisy observations $\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_M$, which have the same underlying true image $\boldsymbol{x}$ and are corrupted by noise $\boldsymbol{n}_1, \boldsymbol{n}_2, \cdots, \boldsymbol{n}_M$ distributed from the same zero-mean distribution, *i.e.*,

$$\boldsymbol{y}_i = \boldsymbol{x} + \boldsymbol{n}_i, \quad i \in \{1, 2, \cdots, M\}, \tag{1}$$

N2N [21] theoretically derives that the optimization problem of deep denoiser $f(\cdot;\theta)$, which minimizes the $L_2$ Loss between the prediction and ground truth image, equals directly minimizing between two independently distributed noisy observations

$$\arg\min_{\theta} \mathbb{E}[\|f(\boldsymbol{y};\theta) - \boldsymbol{x}\|_2^2] = \arg\min_{\theta} \mathbb{E}[\|f(\boldsymbol{y}_j;\theta) - \boldsymbol{y}_k\|_2^2], \tag{2}$$

where $j, k \in \{1, 2, \cdots M\}$ and $j \neq k$.

Different from N2N that needs paired noisy images, N2V [18] proposes to train the network on a single noisy image $\boldsymbol{y}$

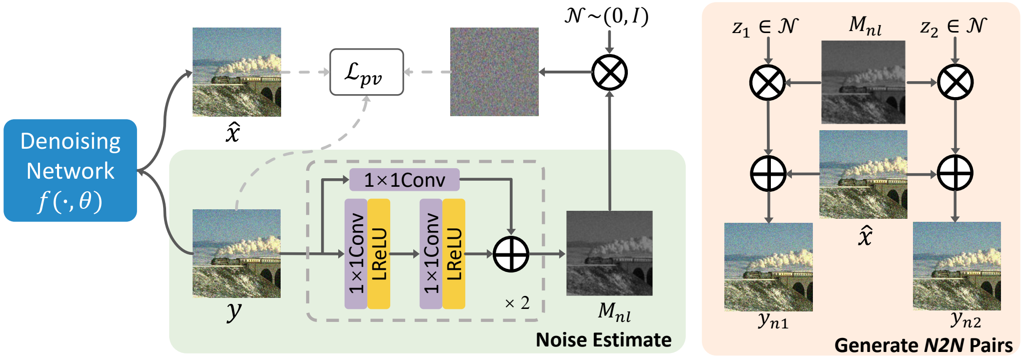
$$\arg\min_{\theta} \mathbb{E}[\|f(\boldsymbol{y}_{RF(p)};\theta) - \boldsymbol{y}(p)\|_2^2], \tag{3}$$

where $p$ denotes the pixel coordinate, $\boldsymbol{y}_{RF(p)}$ represents a specific receptive field centered around pixel $p$, excluding the pixel itself. Such network design is termed as blind-spot network, and can well avoid learning identity mapping.

Based on the above two self-supervised denoising foundations, successive works either dedicate in generating N2N pairs from a single noisy image, or designing more elaborate blind-spot networks.

### 3.2. Observations and Motivations

Implied by the evaluation of existing works [40] and our empirical study (Section 4), the denoising performance of



(a) Noise level estimation (b) N2N pair generation

Figure 3: The overview of our noise level estimation and N2N pair generation model.

N2N in real datasets like SIDD [2] and ELD [40] still lags behind that of supervised learning (Noise2Clean, N2C). However, in other synthetic denoising situations, plenty of works [16, 38] indicate that when N2N is fed with sufficient training samples, it can reach the same precision as N2C, and better than state-of-the-art self-supervised denoising methods. This phenomenon can be attributed to the fact that N2N's training in the latter case is supplemented with a wide variety of random noise samples, which is not possible in real-world settings where the same noisy samples are presented in every training epoch. In addition, according to the results of existing works [16, 38], we observe that N2N data generation based methods [16, 29, 30, 42] basically depend on inaccessible real noise distribution, or affected by unpaired ground truth. Although blind-spot networks [18, 20, 38, 41] relax the training conditions for N2N, the image denoising accuracy is compromised due to the ignorance of information at blind spots. As a result, such methods cannot compete with sufficiently trained N2N models. The above observations can be concluded as follows: Given sufficient training samples,

1. N2N can achieve nearly the same performance as supervised learning (N2C)
2. N2N surpasses existing self-supervised methods, including N2N data generation based methods and blind-spot networks.

In a word, a sufficiently trained N2N model can serve as a strong baseline for self-supervised learning. Hence, we are inspired to optimize denoising networks to gradually approach the performance of N2N.

### 3.3. Denoise-Corrupt-Denoise Pipeline

To explore the potential of N2N, obtaining *sufficient & accurate* N2N training pairs is the critical issue that needs to be addressed. Although several previous works [29, 30] attempt to generate noisy pairs directly from single noisy observations, they are affected by accumulated optimization errors. Hence, we still desire clean images produced by perfect denoiser. In addition, to create noisy observation pairs, we need to know the noise model and noise level. Luckily, numerous previous works [10, 13, 26] have verified that the noise produced by real image sensors follows Poisson-Gaussian model, and can be well approximated by the heteroscedastic Gaussian model. In other words, real noise $\boldsymbol{n}$ can be modeled by

$$\boldsymbol{n} = \mathcal{N} \sim (0, \sigma_d^2\boldsymbol{x} + \sigma_i^2), \tag{4}$$

where $\sigma_d^2$ and $\sigma_i^2$ are the signal-dependent and signal-independent variance. Though Eq. (4) specifies noise distribution, noise level still remains unknown for self-supervised methods. Thus we introduce a noise estimation network to learn it.

To conclude, there are two prerequisites to achieve close results as N2N, *i.e.*, a powerful denoising network $f(\cdot,\theta)$, and an accurate noise estimator $\mathcal{E}(\cdot)$. Here, we introduce a Denoise-Corrupt-Denoise pipeline for self-supervised image denoising, which iteratively optimizes the denoising network and noise estimator. Our pipeline follows the following steps.

**Pre-Denoise.** First, we perform deep denoising network $f(\cdot,\theta)$ on the given single noisy image $\boldsymbol{y}$ to obtain a prediction $\hat{\boldsymbol{x}}$

$$\hat{\boldsymbol{x}} = f(\boldsymbol{y},\theta). \tag{5}$$

**Corrupt.** Then, we predict the noise level from $\boldsymbol{y}$ and the predicted clean image $\hat{\boldsymbol{x}}$, and obtain the pixel-wise noise level map $\boldsymbol{M}_{nl}$

$$\boldsymbol{M}_{nl} = \mathcal{E}(\boldsymbol{y},\hat{\boldsymbol{x}}). \tag{6}$$

The noise level map $\boldsymbol{M}_{nl}$ is used to corrupt the predicted clean image and obtain two different noisy images that have the same noise level as the original image $\boldsymbol{y}$, as shown in Fig. 3(b). In our experiment, though real noise produced by image sensor are spatially uncorrelated, we find that incorporating spatial feature further facilitate the learning of noise levels. Thus, we use resnet blocks [14] with $3 \times 3$ convolution kernels, as shown in Fig. 3(a). Directly performing Gaussian distribution using $\boldsymbol{M}_{nl}$ would bring a non-differential problem. Therefore, we use the reparameterization trick [17] to make sure that the noise sampling process can be backward. The process is illustrated in Fig. 3(b), and can be formulated as

$$\boldsymbol{y}_{ni} = \hat{\boldsymbol{x}} + \boldsymbol{z}_i * \boldsymbol{M}_{nl}, \quad \boldsymbol{z}_i \in \mathcal{N}(0, I), i \in \{1, 2\} \tag{7}$$

We propose a patch variance loss $\mathcal{L}_{pv}$ to constrain the noise estimator $\mathcal{E}$, which can be denoted as

$$\mathcal{L}_{pv} = \sum_i \|Var(\mathcal{P}(\boldsymbol{y}_{ni}, p)) - Var(\mathcal{P}(\boldsymbol{y}, p))\|, \tag{8}$$

where $\mathcal{P}(\cdot, p)$ means extracting $p \times p$ patches from an image.

**N2N Denoise.** Finally, we obtain two different noisy samples $\boldsymbol{y}_{n1}$ and $\boldsymbol{y}_{n2}$ under the same truth image, which meets the requirement for N2N learning. The denoising network $f(\cdot, \theta)$ is trained by the N2N learning described in Eq. (2), and is constrained by minimizing $L_2$ loss

$$\mathcal{L}_{N2N} = \|f(\boldsymbol{y}_{n1}) - \boldsymbol{y}_{n2}\|_2^2. \tag{9}$$

By iteratively repeating Eqs. (5)-(9), the denoising performance is approaching the performance of N2N, while the noise estimation network progressively learns to accurately reveal the noise level of a given single noisy image.

**Blind-Spot Network for Initial Denoising.** While the original Denoise-Corrupt-Denoise pipeline exhibits potential in progressively reaching available paired noisy observations, there remains a requisite for an initial coarse denoised image estimation. Without this, the denoising network might default to an identity mapping, while the noise estimator might revert to a zero mapping. To address this, we employ blind-spot networks that operate independently of any noise priors. For example, following N2V [18], 10% random blind-spots shall be masked and predicted using the rest of the image. Other blind-spot network [16, 38, 41] can also seamlessly integrate into the pipeline, serving as preliminary denoisers. The loss for the blind-spot network is expressed as

$$\mathcal{L}_{BSN} = \|f(\boldsymbol{y}_{RF}; \theta) - \boldsymbol{y}\|_2^2 \tag{10}$$

### 3.4. Iterative Denoising Strategy

During the training of our Denoise-Corrupt-Denoise pipeline, we iteratively and separately step the optimization of the denoising network and noise estimator to avoid learning identity mapping for the denoising network, and zero mapping for the noise estimator.

The total training loss of our Denoise-Corrupt-Denoise pipeline are

$$\mathcal{L} = \mu \mathcal{L}_{pv} + \gamma \mathcal{L}_{N2N} + \lambda \mathcal{L}_{BSN} \tag{11}$$

In the early training phase, we set small $\gamma$ and large $\lambda$ since the blind-spot network is trained to provide a coarse denoiser. As training processes, $\gamma$ is getting larger to dominate the training and approach N2N learning. In addition, $\mu$ decreases quickly to make the network training more stable.

## 4. Experiment

In this section, we first introduce the experimental settings, including the metrics, datasets and learning details we use. Then, we conduct denoising experiments on both synthetic sRGB and real raw-RGB datasets. Finally, ablation studies are performed to verify our training strategy.

### 4.1. Experimental Details

**Dataset.** Following previous self-supervised image denoising methods [16, 38], we assess denoising performance under two noise conditions, *i.e.*, synthetic sRGB and real raw-RGB. For the synthetic sRGB case, to exploit and evaluate the performances of denoising methods, we utilize the ImageNet [8] validation dataset for training, comprising 50k clean sRGB images. Then, the images between resolution of $256 \times 256$ and $512 \times 512$ are used for training. The evaluation is conducted on several established denoising benchmarks, such as Kodak [11], BSD300 [28], and Set14 [45]. During training, all images are cropped to a resolution of $256 \times 256$. In accordance with [16,38], we explore four typical noise distributions for the synthetic scenario: (1) Gaussian noise with a fixed level $\sigma = 25$, (2) Gaussian noise with varied noise levels $\sigma \in [5, 50]$, (3) Poisson noise with a fixed level $\lambda = 30$, and (4) Poisson noise with varied noise levels $\lambda \in [5, 50]$. For the real raw-RGB experiments, we train all methods on the SIDD raw-RGB Medium dataset [2] and evaluate them on the SIDD raw-RGB validation set. Given that SIDD is collected using five diverse smartphone cameras—including Samsung Galaxy S6 Edge, iPhone 7, Google Pixel, Motorola Nexus 6, and LG G4—this setup allows a comprehensive assessment of applicability in real-world denoising scenarios.

**Metrics.** To evaluate the quality of denoised images, we utilize two widely used image-quality metrics, including Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [36] . PSNR measures pixel-wise fidelity, and SSIM can be used the evaluate the 2D spatial similarity. Larger PSNR and SSIM suggest better results.

**Implementation.** Considering that this work focuses on self-supervised image denoising strategies instead of network architecture, we fix the denoising model as U-Net [38] for all methods, which guarantees fair comparisons for all training strategies. During training, the batch size is set to 4, and we use the Adam optimizer initiated with a learning rate of $10^{-4}$, and is halved per 20 epochs. The total number of epochs is 100. As for the hyper-parameter, $\gamma$, $\mu$ and $\lambda$ are set to 0, 1 and 1 at the beginning of the training and evenly change to 1, 0 and 0 as training processes. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

## 4.2. Comparison with State-of-The-Arts

**Compared Methods.** To validate the effectiveness of our training strategy, we compare our method with several supervised baselines and state-of-the-art self-supervised methods. First, as analyzed in Section 3.2, supervised learning (N2C) and N2N [21] can be served as strong baselines for this task, which utilize information more than single noisy images in the training stage. Therefore, if sufficiently trained, N2C and N2N are expected to perform better than other self-supervised methods that compromise of them from single noisy observations. In addition, a traditional denoising method BM3D [7] are evaluated, which is built based on block matching and self-similarity. We also compare with state-of-the-art self-supervised image denoising methods. R2R [30], NAC [42], Ni2N [29] and NBR2NBR [16] are representative methods that generate N2N [21] pairs from a single noisy image. For blind-spot based networks, we make comparisons with N2V [18], SSDN [19] and B2U [38]. We carefully re-implement N2C, N2N, BM3D, N2V, NAC, Ni2N, using the same code base and training iterations for a fair comparison. We directly use the reported results of SSDN and R2R by previous works [16, 38], and we use the pretrained models provided by NBR2NBR and B2U for evaluation.

**Experiments on Synthetic Data.** First, we evaluate the denoising performances on synthetic datasets under four noise cases, as described in Section 4.1. The numeric evaluation results are provided in Table 1. It can be inferred that our Denoise-Corrupt-Denoise training strategy works well in Gaussian and Poisson distribution, with both fixed and varied noise levels. Table 1 shows that our method achieves state-of-the-art results in PSNR and SSIM, especially for the setting of Gaussian noise. We notice that R2R provides very promising results in fixed noise level settings. However, due to the requirement for noise model and noise level as prior, it inevitably suffered from performance degradation in blind noise removal. In addition, though B2U reaches similar results with our training pipeline, the training cost of B2U is 16 times larger than direct U-Net training (compared to 3 times of ours), which reduces the practicality of B2U. The visual results are shown in Figs. 4 and 5, from which we can conclude that our DCD-Net pipeline preserves more scene details while maximizing the performance of image denoising.

| Noise Type | Method | KODAK | BSD300 | SET14 |
|---|---|---|---|---|
| Gaussian $\sigma = 25$ | N2C | 32.46/0.884 | 31.20/0.881 | 31.43/0.868 |
| | N2N [21] | 32.48/0.885 | 31.22/0.882 | 31.45/0.869 |
| | BM3D [7] | 29.97/0.808 | 28.48/0.788 | 29.63/0.818 |
| | N2V [18] | 31.81/0.875 | 30.52/0.870 | 30.53/0.853 |
| | SSDN [19] | 30.62/0.840 | 28.62/0.803 | 29.93/0.830 |
| | R2R [30] | 32.25/0.880 | 30.91/0.872 | 31.32/**0.865** |
| | NAC [42] | 25.69/0.521 | 25.51/0.583 | 25.67/0.586 |
| | Ni2N [29] | 30.45/0.811 | 29.34/0.803 | 29.75/0.815 |
| | NBR2NBR [16] | 32.08/0.879 | 30.79/0.873 | 31.09/0.864 |
| | B2U [38] | **32.27**/0.880 | 30.87/0.872 | 31.27/0.864 |
| | Ours | **32.27/0.881** | **31.01/0.876** | **31.29**/ 0.862 |
| Gaussian $\sigma \in [5, 50]$ | N2C | 32.58/0.876 | 31.27/0.870 | 31.50/0.864 |
| | N2N [21] | 32.57/0.876 | 31.26/0.870 | 31.46/0.863 |
| | BM3D [7] | 29.38/0.781 | 28.83/0.795 | 30.74/0.834 |
| | N2V [18] | 31.72/0.863 | 30.39/0.855 | 30.24/0.843 |
| | SSDN [19] | 30.52/0.833 | 28.43/0.794 | 29.71/0.822 |
| | R2R [30] | 31.50/0.850 | 30.56/0.855 | 30.84/0.850 |
| | NAC [42] | 25.40/0.516 | 24.98/0.560 | 25.44/0.575 |
| | Ni2N [29] | 32.17/0.868 | 30.93/0.862 | 30.87/0.852 |
| | NBR2NBR [16] | 32.10/0.870 | 30.73/0.861 | 31.05/**0.858** |
| | B2U [38] | 32.34/**0.872** | 30.86/0.861 | **31.14**/0.857 |
| | Ours | **32.35/0.872** | **31.09/0.866** | 31.09/0.855 |
| Poisson $\lambda = 30$ | N2C | 31.84/0.877 | 30.54/0.872 | 30.63/0.859 |
| | N2N [21] | 31.84/0.877 | 30.54/0.872 | 30.63/0.858 |
| | BM3D [7] | 27.89/0.738 | 26.58/0.717 | 27.11/0.744 |
| | N2V [18] | 31.18/0.864 | 29.88/0.858 | 29.79/0.841 |
| | SSDN [19] | 30.19/0.833 | 28.25/0.794 | 29.35/0.820 |
| | R2R [30] | 30.50/0.801 | 29.47/0.811 | 29.53/0.801 |
| | NAC [42] | 24.36/0.486 | 24.33/0.559 | 23.93/0.541 |
| | Ni2N [29] | 29.43/0.775 | 28.29/0.764 | 28.63/0.778 |
| | NBR2NBR [16] | 31.44/0.870 | 30.10/0.863 | 30.29/0.853 |
| | B2U [38] | **31.64/0.871** | **30.25**/0.862 | **30.46**/0.850 |
| | Ours | 31.60/0.870 | 30.22/**0.865** | 30.41/**0.855** |
| Poisson $\lambda \in [5, 50]$ | N2C | 31.25/0.862 | 30.17/0.859 | 30.28/0.848 |
| | N2N [21] | 31.17/0.861 | 30.10/0.859 | 30.19/0.847 |
| | BM3D [7] | 27.08/0.702 | 25.85/0.688 | 26.44/0.724 |
| | N2V [18] | 30.55/0.844 | 29.46/0.844 | 29.44/0.831 |
| | SSDN [19] | 29.76/0.820 | 27.89/0.778 | 28.94/0.808 |
| | R2R [30] | 29.14/0.732 | 28.68/0.771 | 28.77/0.765 |
| | NAC [42] | 23.12/0.447 | 23.47/0.534 | 23.14/0.516 |
| | Ni2N [29] | 30.31/0.812 | 29.45/0.821 | 29.40/0.812 |
| | NBR2NBR [16] | 30.86/0.855 | 29.54/0.843 | 29.79/0.838 |
| | B2U [38] | **31.07/0.857** | 29.92/0.852 | **30.10/0.844** |
| | Ours | 31.00/**0.857** | **29.99/0.855** | 29.99/0.843 |

Table 1: Quantitative comparison on synthetic dataset.

**Experiments on Real Data.** To evaluate how these denoising methods perform in real scenarios, we further conduct experiments on a widely used real-world dataset, *i.e.*, SIDD [2]. For baseline methods N2C and N2N, we directly feed the network with noisy/clean or noisy/noisy pairs provided by SIDD training set. Since SSDN requires the specific noise model as a prior, two representative noise distributions (Gaussian and Poisson) are used to evaluate their method. The quantitative and qualitative results are provided in Table 2 and Fig. 6. As shown in Table 2, our method achieves competitive results with the state-of-the-art method B2U which is trained with a large computational cost. Noting that R2R performs well in synthetic setting (Table 1), but it fails to stay high-level performance in real noise. This is caused by the unknown noise prior in real scenarios. Moreover, Ni2N suffers from the unknown noise prior even more severely, and can hardly eliminate noise when the noise level is absent. The overall quantitative and

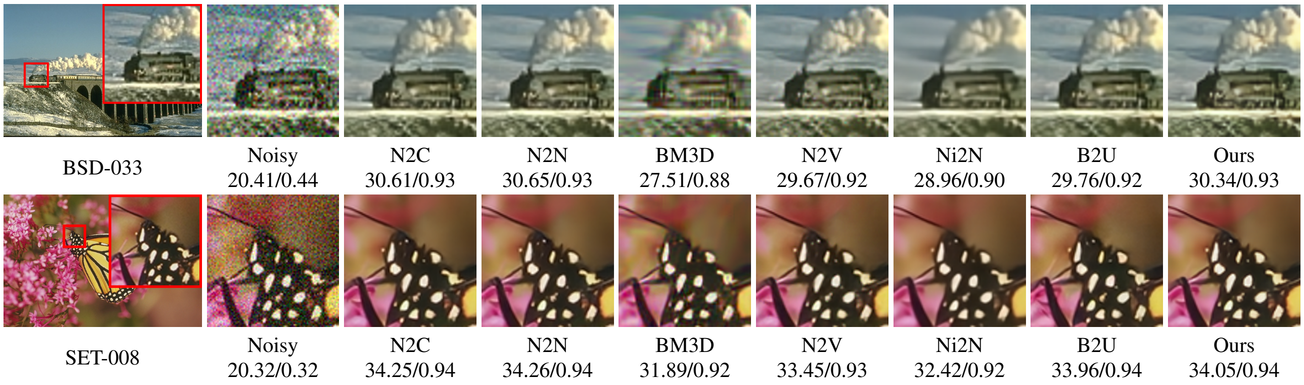


Figure 4: The denoising results on sRGB dataset, under Gaussian noise with $\sigma = 25$. The PSNR/SSIM results are shown below the figure.

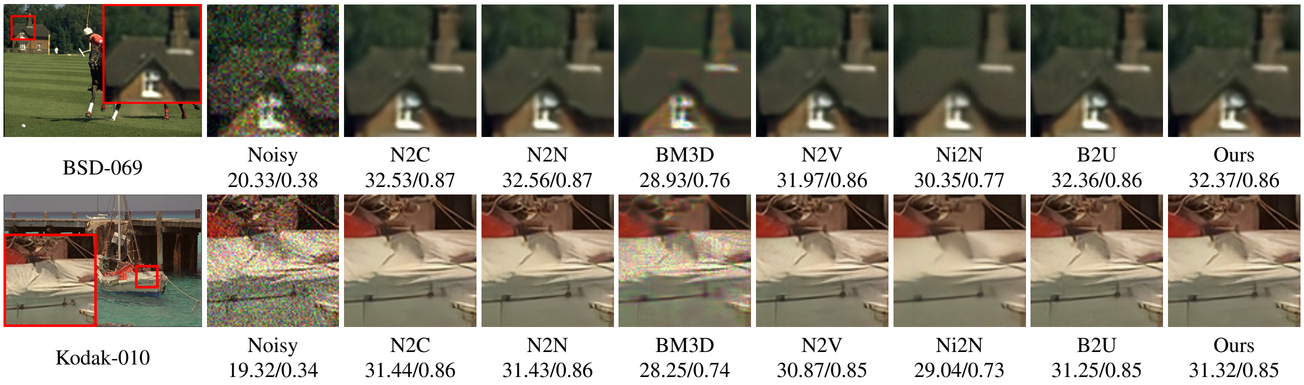


Figure 5: The denoising results on sRGB dataset, under Poisson noise with $\lambda = 30$. The PSNR/SSIM results are shown below the figure.

qualitative results shown in Table 2 and Fig. 6 verify that our training strategy well facilitates the exploitation of self-similar features, leading to better application in real image denoising.

### 4.3. Ablation Studies

To verify the effectiveness of our training strategy, we conduct extensive experiments on the major optimizing components of our pipeline.

**Patch Variance Loss for Noise Estimator.** First, we verify the patch size for the proposed patch variance loss. Five different receptive fields for each patch are used, including $4 \times 4$, $8 \times 8$, $32 \times 32$, $64 \times 64$ and a global receptive field. The denoising results on Kodak dataset with Gaussian noise at $\sigma = 25$ are evaluated, and the results are provided in Table 4. We can see that the patch size of $8 \times 8$ is the most suitable for our patch variance loss. This phenomenon is intuitive, since the variance calculation of smaller patch size is unstable, while larger patch size includes more high-frequency scene information that affects the computation for variance loss. In addition, fewer gradients are backward for larger patch sizes. Table 3 explores the weight for the noise estimation model, which indicates that as the training goes on, the noise estimator can be reduced at early steps

| Methods | Network | Train Cost | PSNR | SSIM |
|---|---|---|---|---|
| N2C | U-Net | ×1 | 51.27 | 0.983 |
| N2N [21] | U-Net | ×1 | 51.29 | 0.991 |
| BM3D [7] | - | - | 48.13 | 0.983 |
| N2V [18] | U-Net | ×1 | 50.46 | 0.990 |
| SSDN [19] (Gaussian) | U-Net | ×4 | 50.44 | 0.990 |
| SSDN [19] (Poisson) | U-Net | ×4 | 50.89 | 0.990 |
| R2R [30] | U-Net | ×1 | 47.20 | 0.980 |
| NAC [42] | U-Net | ×1 | 43.24 | 0.961 |
| Ni2N [29] | U-Net | ×1 | 33.74 | 0.752 |
| NBR2NBR [16] | U-Net | ×2 | 51.06 | 0.991 |
| B2U [38] | U-Net | ×17 | 51.36 | **0.992** |
| Ours | U-Net | ×3 | **51.40** | **0.992** |

Table 2: Quantitative comparisons on SIDD dataset. The PSNR, SSIM and relative training computational cost for each method are provided. The computational cost for a N2C learning is served as the base (×1).

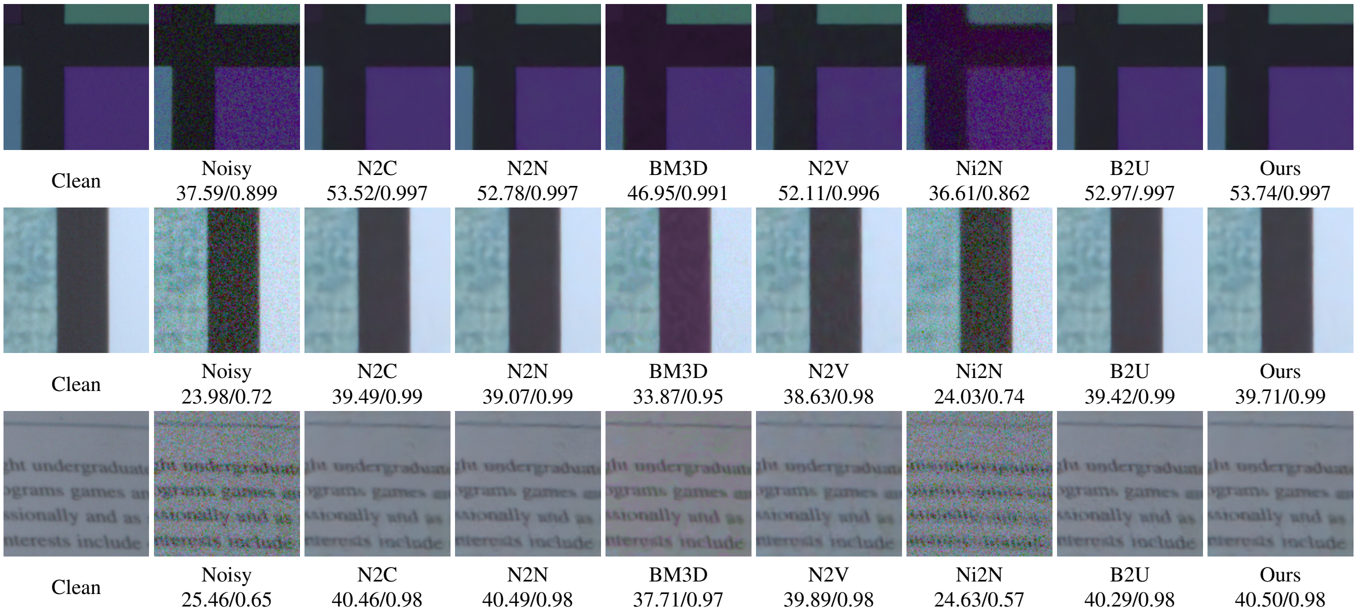


Figure 6: The result on our dataset on SIDD validation dataset. The PSNR/SSIM results are shown below the figure.

Table 3: The ablation study for noise estimation network on Kodark dataset.

| | $\mu = 0$ | $\mu = 1$ | $\mu = 1 \rightarrow 0$ |
|---|---|---|---|
| PSNR | 24.21 | 31.86 | **32.27** |
| SSIM | 0.675 | 0.872 | **0.881** |

Table 4: The ablation study in Kodark dataset. We conduct ablation study on patch size for patch variance loss.

| patch size | 4 | 8 | 32 | 64 | Global |
|---|---|---|---|---|---|
| PSNR | 32.21 | **32.27** | 32.23 | 32.24 | 32.18 |
| SSIM | 0.877 | **0.881** | 0.880 | 0.879 | 0.875 |

Table 5: The ablation study for blind-spot network on Kodark dataset.

| | $\gamma = 0$ | $\gamma = 1$ | $\gamma = 0 \rightarrow 1$ |
|---|---|---|---|
| PSNR | 31.81 | 32.14 | **32.27** |
| SSIM | 0.875 | 0.873 | **0.881** |

Table 6: The ablation study for the regularization term on Kodark dataset.

| | $\lambda = 0$ | $\lambda = 1$ | $\lambda = 1 \rightarrow 0$ |
|---|---|---|---|
| PSNR | 14.46 | 31.75 | **32.27** |
| SSIM | 0.5312 | 0.875 | **0.881** |

since it is more easily to be trained than the denoiser.

**N2N Training.** The ablation study results for N2N training are presented in Table 5. We consider three values for the weight of $\mathcal{L}_{N2N}$. It can be inferred from Table 5 that by removing $\mathcal{L}_{N2N}$ ($\gamma = 0$), the denoising capability degrades since the pipeline reduces to a simple blind-spot network.

**Blind-Spot Training.** Here, we assess the impact of the blind-spot network. As illustrated in Table 6, omitting the blind-spot training results in significant convergence challenges during the training phase. This issue arises because the noise estimator struggles to determine accurate noise levels when relying on flawed denoisers. Consequently, our pipeline leverages the blind-spot network for the initial learning of a coarse denoiser, a process that can be conducted using only single noisy images. We believe that our approach offers possibilities for collaboration with other meticulously crafted blind-spot networks [16, 38].

## 5. Conclusion

In this paper, we propose an effective Denoise-Corrupt-Denoise pipeline (DCD-Net) for self-supervised image denoising. Based on careful analysis and observations that N2N can be served as a strong baseline when training samples are sufficient, the proposed pipeline is trained in an iterative manner and gradually reaches the performance of a sufficiently trained N2N model. The proposed pipeline repeats the following training steps: 1) use a deep denoising network to obtain predicted clean images; 2) predict the noise level to generate new N2N pairs, and optimize the noise estimator; 3) training through N2N strategy and optimize the deep denoiser. Our pipeline is verified to have state-of-the-art performance compared to other self-supervised image denoising methods under a wide variety of synthetic and real noise conditions.

**Acknowledgments** This work was supported by the National Natural Science Foundation of China (62171038, 61931008, 62088101, and U21B2024), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), and the Fundamental Research Funds for the Central Universities.

## References

[1] Abdelrahman Abdelhamed, Marcus A Brubaker, and Michael S Brown. Noise flow: Noise modeling with conditional normalizing flows. In *IEEE Int. Conf. Comput. Vis.*, pages 3165–3173, 2019. 1

[2] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1692–1700, 2018. 2, 4, 5, 6

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 60–65, 2005. 1, 2

[4] Yue Cao, Xiaohe Wu, Shuran Qi, Xiao Liu, Zhongqin Wu, and Wangmeng Zuo. Pseudo-isp: Learning pseudo in-camera signal processing pipeline from a color image denoiser. *arXiv preprint arXiv:2103.10234*, 2021. 2

[5] Ke-Chi Chang, Ren Wang, Hung-Jin Lin, Yu-Lun Liu, Chia-Ping Chen, Yu-Lin Chang, and Hwann-Tzong Chen. Learning camera-aware noise models. In *Eur. Conf. Comput. Vis.*, pages 343–358. Springer, 2020. 1

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3291–3300, 2018. 1

[7] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Proc. SPIE*, volume 6064, page 606414, 2006. 1, 2, 6, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–255. Ieee, 2009. 5

[9] Alessandro Foi, Sakari Alenius, Vladimir Katkovnik, and Karen Egiazarian. Noise measurement for raw-data of digital imaging sensors by automatic segmentation of uniform targets. *IEEE Sens. J.*, 7(10):1456–1461, 2007. 1

[10] Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE Trans. Image Process.*, 17(10):1737–1754, 2008. 4

[11] Rich Franzen. Kodak lossless true color image suite. https://r0k.us/graphics/kodak/. Accessed: 2023-02-10. 5

[12] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1712–1722, 2019. 1, 2

[13] Samuel W Hasinoff. Photon, poisson noise., 2014. 4

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 5

[15] Bernardo Henz, Eduardo SL Gastal, and Manuel M Oliveira. Synthesizing camera noise using generative adversarial networks. *IEEE Trans. Vis. Comput. Graph.*, 27(3):2123–2135, 2020. 1

[16] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14781–14790, 2021. 2, 3, 4, 5, 6, 7, 8

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[18] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2129–2137, 2019. 1, 2, 3, 4, 5, 6, 7

[19] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. *Adv. Neural Inform. Process. Syst.* 6, 7

[20] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17725–17734, 2022. 2, 3, 4

[21] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *Int. Conf. Mach. Learn.*, pages 2965–2974. PMLR, 2018. 1, 2, 3, 6, 7

[22] Jason Lequyer, Reuben Philip, Amit Sharma, Wen-Hsin Hsu, and Laurence Pelletier. A fast blind zero-shot denoiser. *Nat. Mach. Intell.*, pages 1–11, 2022. 1

[23] Miaoyu Li, Ying Fu, and Yulun Zhang. Spatial-spectral transformer for hyperspectral image denoising. In *AAAI Conf. Artif. Intell.*, volume 37, pages 1368–1376, 2023. 1

[24] Xinyang Li, Yixin Li, Yiliang Zhou, Jiamin Wu, Zhifeng Zhao, Jiaqi Fan, Fei Deng, Zhaofa Wu, Guihua Xiao, Jing He, et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nat. Biotechnol.*, pages 1–11, 2022. 1

[25] Wei Liu and Weisi Lin. Additive white gaussian noise level estimation in svd domain for images. *IEEE Trans. Image Process.*, 22(3):872–883, 2012. 1

[26] Markku Mäkitalo and Alessandro Foi. Noise parameter mismatch in variance stabilization, with an application to poisson–gaussian noise estimation. *IEEE Trans. Image Process.*, 23(12):5348–5359, 2014. 4

[27] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Adv. Neural Inform. Process. Syst.*, 29, 2016. 2

[28] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE Int. Conf. Comput. Vis.*, volume 2, pages 416–423. IEEE, 2001. 5

[29] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12064–12072, 2020. 2, 3, 4, 6, 7

[30] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: unsupervised deep learning for image denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2043–2052, 2021. 2, 3, 4, 6, 7

[31] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1586–1595, 2017. 2

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Med. Image Comput. Comput. Assist. Interv.*, pages 234–241, 2015. 1, 2

[33] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60(1-4):259–268, 1992. 1

[34] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *IEEE Int. Conf. Comput. Vis.*, pages 4539–4547, 2017. 1, 2

[35] Yang Wang and Haomin Zhou. Total variation wavelet-based medical image denoising. *Int. J. Biomed. Imaging*, 2006, 2006. 1

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 5

[37] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18156–18165, 2023. 2

[38] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2unblind: Self-supervised image denoising with visible blind spots. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2027–2036, 2022. 2, 3, 4, 5, 6, 7, 8

[39] Kaixuan Wei, Ying Fu, and Hua Huang. 3-d quasi-recurrent neural network for hyperspectral image denoising. *IEEE Trans. Neural Netw. Learn. Syst.*, 32(1):363–375, 2020. 1

[40] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 1, 2, 3, 4

[41] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *Eur. Conf. Comput. Vis.*, pages 352–368, 2020. 3, 4, 5

[42] 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 Trans. Image Process.*, 29:9316–9329, 2020. 2, 3, 4, 6, 7

[43] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *Eur. Conf. Comput. Vis.*, pages 41–58. Springer, 2020. 1

[44] Yuhang Zeng, Yunhao Zou, and Ying Fu. 3D$^2$unet: 3d deformable unet for low-light video enhancement. In *Chinese Conf. Pattern Recog. Comput. Vis.*, pages 66–77. Springer, 2021. 1

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces 2010*, pages 711–730. Springer, 2012. 5

[46] Jiachao Zhang and Keigo Hirakawa. Improved denoising via poisson mixture modeling of image sensor noise. *IEEE Trans. Image Process.*, 26(4):1565–1578, 2017. 1

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

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

[49] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4593–4601, 2021. 1

[50] Yunhao Zou and Ying Fu. Estimating fine-grained noise model via contrastive learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12682–12691, 2022. 1