

# Zero-Shot Blind-spot Image Denoising via Implicit Neural Sampling

Yuhui Quan<sup>1</sup> Tianxiang Zheng<sup>1</sup> Zhiyuan Ma<sup>2</sup> Hui Ji<sup>2</sup>

<sup>1</sup>School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

<sup>2</sup>Department of Mathematics, National University of Singapore, 119076, Singapore

csyhquan@scut.edu.cn, cszhengt@mail.scut.edu.cn, e0983565@u.nus.edu, matjh@nus.edu.sg

## Abstract

*The blind-spot principle has been a widely used tool in zero-shot image denoising but faces challenges with real-world noise that exhibits strong local correlations. Existing methods focus on reducing noise correlation, which also weaken the pixel correlations needed for accurately estimating missing pixels. In this paper, we first present a rigorous analysis of how noise correlation and pixel correlation impact the statistical risk of a linear blind-spot denoiser. We then propose using an implicit neural representation to resample noisy pixels, effectively reducing noise correlation while preserving the essential pixel correlations for successful blind-spot denoising. Extensive experiments show our method surpasses existing zero-shot denoising techniques on real-world noisy images.*

## 1. Introduction

Image denoising, which restores a clean image from its noisy measurement, is a fundamental task in many low-level vision applications. In recent years, deep learning is the main driving force in this field, with many approaches primarily relying on supervised learning [5, 25, 32, 38, 39, 47–49, 51, 52]. These supervised methods train neural networks (NNs) on paired datasets, which often is synthesized w.r.t. Additive White Gaussian Noise (AWGN). However, Denoising NNs trained on such synthetic noise often fail to perform well (e.g., [2]) on real-world data due to significant differences of AWGN and real-world noise.

Efforts to bridge the gap between synthesized and real-world data have led to the creation of datasets with real-world paired data, such as Smartphone Image Denoising Dataset (SIDD) [1] for camera photographs and Fluorescence Microscopy Denoising (FMD) dataset [54] for fluorescence microscopy images. Note that the creation of such datasets is labor-intensive and often fails to cover the full spectrum of real-world noise. Also, ground-truth (GT) images are difficult to obtain in fields like medical or scientific imaging. The issues caused by the independence of image

pairs remain for supervised denoising NNs. To alleviate the dependence on image pairs with truth images, some methods propose to train denoising NNs using unpaired noisy and clean images [4, 14], or multiple noisy images of the same scene [22, 60]. However, these methods either require clean images or depend on precise image alignment, a challenging task for real-world images.

### 1.1. Learning denoising NN without truth images

To reduce the reliance on truth images in supervised learning methods, considerable efforts have been made to develop truth-free deep learning approaches for denoising.

**Self-supervised denoising using only noisy images:** Self-supervised learning, sometimes also called unsupervised learning in the context of image denoising, aims to train a denoising NN using only noisy images, without requiring clean ones. There is an abundant literature on it. Some generate pseudo-pair data to define self-supervised loss functions (see e.g. [13, 22, 29, 40, 57]). One of the popular and effective techniques is the blind-spot principle [3, 19], which trains the NN to approximate noisy pixel-values of visible pixels by using their neighboring invisible pixels.

**Zero-shot denoising at test time:** While self-supervised methods train denoising models on external datasets with only noisy data, zero-shot methods directly adapt an untrained NN to each noisy image at test time, which make it highly adaptable to specific images or noise types. Also, it suites well in on-line setting, where accessing large datasets is not desired. Such data-independence make it a valuable option in certain cases. Deep Image Prior (DIP) [37] is a pioneering work on zero-shot denoising, leveraging the inductive biases of convolutional neural networks (CNNs) to fit natural images more rapidly than random noise. The concept of the blind-spot principle can also be extended to zero-shot denoising at test time (e.g., [8, 9, 21, 31, 34, 55, 58]), demonstrating promising performance.

### 1.2. Discussions on blind-spot denoising

The practical value of zero-shot denoising and the potential of blind-spot techniques motivate us to study blind-spot

methods for zero-shot real-world image denoising, where a denoising NN is trained for each sample at test time.

Let  $\mathbf{y} = \mathbf{x} + \mathbf{n}$  be a noisy image, where  $\mathbf{x}$  denotes truth image and  $\mathbf{n}$  denotes noise. For an image pixel  $x_i$ , let  $y_i$  denotes its noisy value, and  $\mathcal{N}(i)$  denotes the index set of neighboring pixels of  $x_i$ , excluding  $x_i$ . Then, a denoising NN  $\mathcal{D}_\phi : \mathbf{y} \rightarrow \mathbf{x}$  aims at minimizing the summation of the differences between the noisy value  $y_i$  and truth value  $x_i$  over all pixels. The difference for each pixel  $i$  is then

$$|\mathcal{D}_\phi(\{y_i\} \cup \{y_j\}_{j \in \mathcal{N}(i)}) - x_i|_2^2. \quad (1)$$

In the absence of  $x_i$ , the blind-spot scheme trains the denoising NN to predict  $x_i$ , based solely on the noisy values of neighboring pixels  $\{y_j\}_{j \in \mathcal{N}(i)}$ , explicitly excluding  $y_i$ . If  $y_i$  is included, the NN  $\mathcal{D}_\phi$  will converge to identity map. The resulting blind-spot-based loss is then:

$$|\mathcal{D}_\theta(\{y_j\}_{j \in \mathcal{N}(x_i)}) - y_i|_2^2, \quad (2)$$

The effectiveness of blind-spot scheme depends on how close the loss (2) is to the loss (1). Their expectations are close if the following two conditions are satisfied:

1. The noise  $\mathbf{n}$  is zero-mean and pixel-wise independent.
2. The truth pixel value  $x_i$  is strongly correlated with its neighboring pixels  $\{x_j\}_{j \in \mathcal{N}(i)}$ .

However, real-world noise is likely spatially correlated, which violates the noise assumption of plain blind-spot scheme. The main approach to extend blind-spot techniques to handle spatially correlated noise is to use few neighboring pixels and more pixels far way. For example, pixel-shuffle down-sampling (PD) was proposed in [59] with a small down-sampling stride size, and refined in AP-BSN [21] by using a large/small stride during training/testing. However, small strides fail to break noise correlation, while large strides weaken the pixel-value correlations. PUCA [15] employs patch-unshuffle/shuffle techniques and LG-BPN [41] uses channel self-attention with dilated convolution to relate far-away pixels. MASH [9] also used local shuffling to reduce local noise correlations.

The blind spot scheme can be either implemented with specific NN architecture (e.g. [3, 13, 19]), or with the introduction of a mask to exclude the pixel itself in the loss function (e.g. [9, 31]). Let  $M$  denote the binary masks, where  $M_i = 0$  if the pixel  $i$  is masked and 1 otherwise. Let  $\odot$  denote entry-wise multiplication. Then, the blind-spot-based loss function can be expressed as:

$$\|(1 - M) \odot (\mathcal{D}_\phi(M \odot \mathbf{y}) - \mathbf{y})\|. \quad (3)$$

The mask-based loss (3) is a special blind-spot implementation, where the loss is defined as the sum of the differences between noisy intensity values of all invisible pixels  $y_i$  in  $\Omega_{inv}$  and their prediction from the NN using their neighboring visible pixels in  $\Omega_{vis}$ , respectively.

### 1.3. Main idea

In blind-spot techniques, the selection of visible (un-masked) and invisible (masked) pixels is critical for effective denoising, in the presence of spatially correlated noise. The main challenge is how to balance two opposing factors:

- *Over-reliance on visible pixels:* Masking too few pixels leads to excessive use of visible pixels, including their spatially correlated neighbors, introducing noise. Additionally, with the loss computed over a small subset of pixels, the model may overfit by ignoring differences in visible pixels, hindering generalization for denoising.
- *Insufficient visible pixels:* Masking too many pixels reduces available input, forcing the model to rely on distant pixels with weaker correlations for prediction. This will lead to less accurate predictions and poor generalization for denoising invisible pixels.

In short, the selection of visible pixels and invisible pixels plays a critical role in mask-based blind-spot schemes.

In this paper, we propose a novel approach that address this challenges, which leverages the strong intensity correlation of neighboring invisible pixels and the weak noise correlation of distant visible pixels to predict intensity of each invisible pixel.

- *Using distant invisible pixels:* We select a few distant visible pixels to predict the target invisible pixel, as their noise is weakly correlated. However, their intensities also have weaker correlations with the target, making them insufficient on their own for accurate prediction.
- *Leveraging neighboring visible pixels:* To counteract the weak correlation of distant pixels, we incorporate neighboring pixels. However, to avoid introducing correlated noise, instead of directly using their values, we employ an implicit neural representation (INR) trained on visible pixels to estimate the values of these neighboring invisible pixels, ensuring reliable, noise-mitigated information.

In short, our approach balances two pixel sources: breaking noise correlation by using only distant visible pixels and compensating for their weak intensity correlation with the INR’s denoised estimation of neighboring invisible pixels. The INR and denoising NN are jointly trained to effectively handle spatially correlated real-world noise.

### 1.4. Contributions

Our contributions are twofold. First, we provide a theoretical analysis of prediction risk in mask-based blind-spot denoising, examining its relationship with three key factors: noise correlation, pixel-value correlation, and visible set size. While based on a simplified linear model, this analysis provides insights into mask-based blind-spot techniques and motivates the design of our proposed approach.

Second, we propose a mask-based blind-spot denoising

method that leverages both distant visible pixels and INR-based estimates of neighboring invisible pixels. Our approach significantly improves blind-spot denoising for real-world tasks with spatially correlated noise. To conclude, the main contributions of this paper are:

- A quantitative analysis of the statistical risk of a linear blind-spot denoiser, considering noise correlation, pixel-value correlation, and visible set size.
- A blind-spot denoiser that combines distant visible pixels with INR-based estimates of neighboring invisible pixels to effectively handle spatially correlated noise.

Extensive experiments demonstrating the effectiveness of our approach on real-world noisy images.

## 2. Related Work

**Implicit Neural Representation:** INR is a coordinate-based NN that maps spatial or spatiotemporal coordinates to signal values, enabling continuous and memory-efficient representations. INR has been widely used in surface reconstruction tasks (*e.g.*, [26, 35]), as demonstrated by several studies [16]. INR has also been used in many image and video restoration tasks [6, 7, 33, 53], including zero-shot image denoising [16].

**Supervised denoising NNs:** Supervised methods train a NN that maps a noisy image to a clean image. Main architectures include CNN (see *e.g.*, [5, 32, 38, 47, 48, 51, 52]) and transformer [25, 39, 49]. Instead of using noisy-clean image pairs, the weakly supervised methods either use unpaired noisy-clean images [4, 14] or paired noisy images [22, 60]. Zhou *et al.* [59] trained a noise estimator and denoiser on mixed AWGN and impulse noise, using pixel-shuffle down-sampling to adapt for real-world noise.

**Self-supervised denoising NNs:** Self-supervised methods refer to training a denoising NN on an external dataset of only noisy images. The blind-spot scheme, an early and popular approach, ensures the receptive-field center remains unseen during prediction through specific NN design. Lainel *et al.* [20] occludes half of the receptive fields. D-BSN [42] employs a center-masked convolution layer, followed by dilated convolutions with varying step sizes. Neighbour2Neighbour [13] uses 4 down-sampled versions of the noisy images to define visible and invisible pixels. MM-BSN [50] implements multiple convolutional kernels with different mask patterns. SelfFormer [34] proposed a blind-spot implementation for transformers using directional self-attention and Siamese mutual learning.

Another blind-spot approach studies mask-based self-supervised loss. Noise2Void [19] and Noise2Self [3] adopt the loss (3). Noise2Same [43] introduces an additional self-reconstruction loss to utilize center pixels' information. Blind2Unblind [40] trains a masker to better preserve valuable pixels in the masked input. SASL [24] gives separate

treatment to flat and textured regions when defining mask.

The blind-spot schemes above work well for independent noise, but struggle with spatially correlated noise. To address this, [59] introduced pixel shuffle downsampling to utilize distant pixels with likely independent noise. AP-BSN [21] refined this by using a larger stride in training and a small stride in testing. LG-BPN [41] proposes a denser sampling kernel for better local texture recovery and a global branch for larger receptive fields. SASL [24] separates flat and textured regions in masking-based self-supervision. PUCA [15] employs patch-unshuffle/shuffle techniques to expand receptive fields for improved global context integration. SS-BSN [12] combines grid SA [36] with a simplified D-BSN [42]. SwinIA [30] masks the diagonal of the attention matrix in a Swin Transformer [25].

Beyond blind-spot, there are also approaches for self-supervised denoising. Recorruption-based methods, such as R2R [29] and its extension [57] simulate supervised losses by deriving pairs from the noisy input. Score-based methods, such as [17, 18, 44], trained the NN via score matching. Disentanglement-based methods (*e.g.* CVF-SID [28]) use cyclic loss to separate clean and noisy components.

**Zero-shot denoising at test time:** Zero-shot does not require any external dataset for training. Instead, it adapts an untrained NN directly to each specific noisy image at test time. The concept of zero-shot denoising was first introduced by DIP [37], which uses the inductive biases of CNNs to fit natural images more rapidly than random noise. Noise2Fast [23] extended the idea of Neighbour2Neighbour to enabling NN training at test time. NoisyAsClean [46] adds different synthetic noise to the given input to make auxiliary training pairs for zero-shot denoising. Based on blind-spot scheme, Self2Self [31] trained a denoising NN with dropout for prediction ensemble.

To handle spatially correlated noise, Zheng *et al.* [55] maps noisy images into a latent space with AGWN, where noise is removed via an AGWN denoiser. ScoreDVI [8] included score priors into Bayesian inference for denoising. MASH [9] introduced an adaptive mask ratio for regions with different noise correlation, and applied local shuffling to reduce local noise correlations.

## 3. Method

In this section, we first establish a theoretical analysis of the statistical risk of a linear blind-spot denoiser. Then we discuss the details of our zero-shot mask-based blind-spot approach for removing spatially-correlated noise.

### 3.1. Risk analysis of linear blind-spot denoiser

Consider an invisible pixel for prediction:

$$y_0 = x_0 + n_0,$$

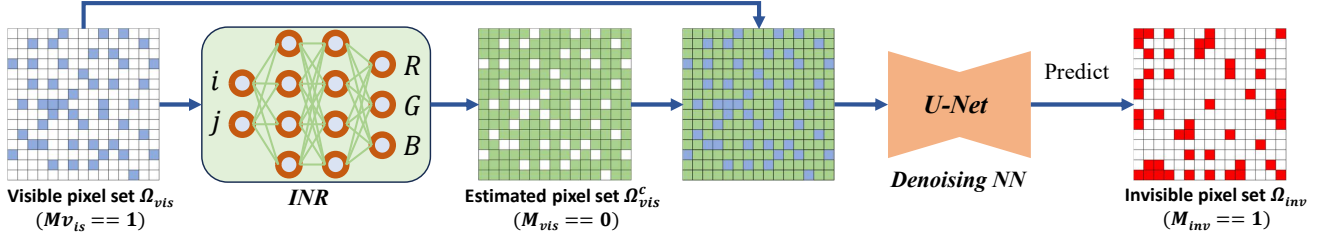


Figure 1. Illustration of the proposed blind-spot scheme, where blue pixels represent visible pixels, green pixels represent all other pixels except the visible ones, and red pixels represent invisible pixels for prediction.

where  $x_0$  denotes the truth intensity value and  $n_0$  random noise with  $\mathbb{E}[n_0] = 0$  and  $\text{Var}[n_0] = \sigma^2$ . Let  $\{y_j = x_j + n_j\}_{j=1}^M$  denote the noisy intensity values of  $M$  visible pixels which are used for predicting  $x_0$ . We consider the following relationship between visible set  $\{y_j\}_j$  and the target  $y_0$ :

$$y_j = (1 - \mu_j)x_0 + n_j, \quad 1 \leq j \leq M, \quad (4)$$

where  $\{\mu_j\}_{j=1}^M$  are i.i.d. random noise, and  $\{n_j\}_{j=1}^M$  are noise which relates to the noise  $n_0$  of the invisible point  $x_0$  by

$$n_j = \lambda_n n_0 + \sqrt{1 - \lambda_n^2} \epsilon_j, \quad 1 \leq j \leq M. \quad (5)$$

The variables  $\{\mu_j\}_j$  and  $\{\epsilon_j\}_j$  satisfy: for  $1 \leq j \leq M$ ,

$$\mathbb{E}[\mu_j] = \mathbb{E}[\epsilon_j] = 0, \quad \text{Var}[\mu_j] = \lambda_v^2 \sigma^2, \quad \text{Var}[\epsilon_j] = \sigma^2. \quad (6)$$

**Remark 1.** Two parameters  $\lambda_n$  and  $\lambda_v$  measure the correlation degree for noise and intensity value, respectively. A smaller  $\lambda_n$  means weaker correlation between any two noises and vice versa. Also, a smaller  $\lambda_v$  implies stronger correlation between the intensity values of two pixels.

Suppose we learn a blind-spot denoiser  $\mathcal{D}_\phi$  that predicts  $x_0$  based on  $\{y_j\}_{j=1}^M$ . Its statistical risk is defined by

$$\mathcal{R}_\phi = \mathbb{E}_{\{n_j\}, \{\mu_j\}, n_0} [\|\mathcal{D}_\phi(\{y_j\}_{j=1}^M) - x_0\|_2^2]. \quad (7)$$

Consider a linear model  $\mathcal{D}_a$  parametrized by  $\mathbf{a} = \{a_j\}_{j=1}^M$ :

$$\mathcal{D}_a(\{y_j\}_{j=1}^M) = \sum_{j=1}^M a_j \cdot y_j. \quad (8)$$

It is then trained by minimizing the self-supervised loss:

$$\mathbf{a}^* := \underset{\mathbf{a} \in \mathbb{R}^M}{\text{argmin}} \left\| \sum_{j=1}^M a_j \cdot y_j - y_0 \right\|_2^2. \quad (9)$$

**Proposition 1.** The risk of the linear denoiser  $\mathcal{D}_{\mathbf{a}^*}$  of the

form (8) learned by minimizing the loss (9) is

$$\begin{aligned} \mathcal{R}^* &= \frac{-[M(x_0^4 - \lambda_n^2 \sigma^4)]}{Mx_0^2 + (M-1)\lambda_n^2 \sigma^2 + \lambda_v^2 \sigma^2 + \sigma^2} + x_0^2 \quad (10) \\ &= \frac{-[Mx_0^4 + \frac{M\sigma^2}{M-1}(Mx_0^2 + \lambda_v^2 \sigma^2 + \sigma^2)]}{Mx_0^2 + (M-1)\lambda_n^2 \sigma^2 + \lambda_v^2 \sigma^2 + \sigma^2} + \frac{M\sigma^2}{M-1} + x_0^2 \quad (11) \end{aligned}$$

*Proof.* See Supplementary Material.  $\square$

The minimal risk in (10) and (11) reveals the interplay between noise correlation and intensity correlation among visible and invisible pixels. From (10), if  $y_0$  is not noise-dominated (i.e.,  $x_0^2 > \lambda_n \sigma^2$ ), the minimal risk decreases as the intensity correlation increases (i.e., lower  $\lambda_v^2$ ). Similarly, from (11), the risk decreases with lower noise correlation (i.e., smaller  $\lambda_n^2$ ). The risk also depends on the visible set size  $M$ , but increasing  $M$  does not always reduce it. Furthermore, (10) and (11) quantify the impact of mask ratio on blind-spot denoising for local correlated noise. It can be seen that the effect of noise correlation  $\lambda_n$  scales with  $(M-1)$ , while intensity correlation  $\lambda_v$  does not. This suggests prioritizing noise correlation reduction over intensity correlation enhancement. This motivated us to advocate aggressively reducing noise correlation by excluding more neighboring pixels and leveraging distant ones, with an INR compensating for pixel correlation.

### 3.2. Zero-shot blind-spot denoising

The basic idea is using noisy values from distant visible pixels and INR-based estimates of neighboring invisible pixels to predict the intensity-value of each target invisible pixel. These pixels are randomly selected in each epoch.

**Defining the visible pixel sets:** Let  $M_{vis}$  denote a binary matrix with entries sampled from a Bernoulli distribution:

$$M_{vis}(i, j) = \begin{cases} 1, & \text{with probability } p_0; \\ 0, & \text{with probability } 1 - p_0. \end{cases}$$

To ensure sufficient separation between visible and invisible pixels in the predictions, as well as adequate spacing among

visible pixels themselves, we select a small value for  $p_0$ . Then, the visible pixel set  $\Omega_{vis}$  is defined as

$$\Omega_{vis} = \{(i, j) : M_{vis}(i, j) = 1\}.$$

**Defining the invisible pixel set for prediction:** In classic mask-based blind-spot methods, the invisible pixel set is defined as the complement of the visible pixel set, which includes neighboring pixels with correlated noise. To address this, we propose using a subset of the complement that excludes any pixels close to visible ones. It is implemented by applying a morphological dilation operation to the visible pixel set  $\Omega_{vis}$ , which expands the visible pixel set by a certain radius. The invisible pixel set  $\Omega_{inv}$  is then defined as the complement of the dilated visible pixel set. That is, let  $k$  denote the structure element for dilation operation:

$$k = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Define  $D = M_{vis} \otimes k$ . Then the mask for invisible pixel set, denoted by  $M_{inv}$  is defined as

$$M_{inv}(i, j) = 1 - \begin{cases} 1, & \text{if } D(i, j) \neq 0; \\ 0, & \text{Otherwise.} \end{cases}$$

Then, the visible pixel set  $\Omega_{inv}$  is defined as

$$\Omega_{inv} = \{(i, j) : M_{inv}(i, j) = 1\}.$$

**Remark 2.** Unlike existing mask-based blind-spot approaches that use the complement of visible pixels for prediction, our approach uses a subset of this complement for prediction that excludes neighboring pixels.

**INR-based estimation of neighboring invisible pixels:**

For each invisible pixel in  $\Omega_{inv}$ , the pixels in  $\Omega_{vis}$  are not in its neighborhood and weakly correlated in noise. However, the intensity correlation between the invisible pixel and these visible pixels is also not strong enough to offer sufficient information of target invisible pixel. To address this, we propose constructing an INR of the image to fit the pixel values of visible pixels. As a continuous representation, the INR can estimate the intensity values of neighboring invisible pixels, providing additional information about the target invisible pixel’s intensity that complements the data from distant visible pixels. Specifically, let  $\mathcal{F}_\phi$  denote the INR that maps the pixel coordinates to the intensity values, which is trained on the visible pixel set  $\Omega_{vis}$ :

$$\mathcal{L}_{inr} = \sum_{(i,j) \in \Omega_{vis}} \|\mathcal{F}_\phi(i, j) - \mathbf{y}(i, j)\|_1. \quad (12)$$

Once trained, the INR can be used to estimate the intensity values of all non-visible pixels, including those in  $\Omega_{inv}$  and excluded neighboring pixels of  $\Omega_{vis}$ .

$$\mathbf{x}_{inr}(i, j) = \mathcal{F}_\phi(i, j), \quad (i, j) \in \Omega_{vis}^c, \quad (13)$$

where  $\Omega_{vis}^c$  denotes the complement of  $\Omega_{vis}$ .

**Loss functions for denoising NN:** Taking the distance visible pixels and the INR-based estimation of neighboring invisible pixels as input, the denoising NN  $\mathcal{D}_\theta$  is trained to predict the intensity values of invisible pixels in  $\Omega_{inv}$ . The loss function for the denoising NN is defined as:

$$\mathcal{L}_{dn} = \|M_{inv}(\mathcal{D}_\theta(M_{vis} \odot \mathbf{y} + (1 - M_{vis}) \cdot \mathbf{x}_{inr}) - \mathbf{y})\|_1. \quad (14)$$

**Sub-pixel consistency between two NNs:** When training the INR to predict the intensity values of visible pixels, not the truth intensity values but the noisy values of visible pixels is used in the loss function (12). Consequently, the INR-based estimation of invisible pixels may still contain some degree of noise. To address the potential impact of this residual noise, we introduce a sub-pixel consistency constraint that leverages INR’s resolution-free property.

The basic idea is to regularize the training of INR such that, within the sub-pixel neighborhood of a pixel, the predicted intensity values from INR are close to the denoised intensity value produced by the denoising NN  $\mathcal{D}_\theta$ , which is expected to be closer to the true intensity value than  $\mathbf{y}(i, j)$ . Specifically, the sub-pixel consistency loss is defined as

$$\mathcal{L}_{mc} = \sum_{(i,j) \in \Omega_{vis}} \|\mathcal{F}_\phi(i + \Delta i, j + \Delta j) - \mathbf{x}_{dn}^{\text{detach}}(i, j)\|_1. \quad (15)$$

where  $0 < \Delta i, \Delta j < 1$  denote random sub-pixel shift, and

$$\mathbf{x}_{dn}^{\text{detach}}(i, j) = \text{sg}(\mathcal{D}_\theta(M_{vis} \odot \mathbf{y} + (1 - M_{vis}) \cdot \mathbf{x}_{inr}))(i, j)$$

denotes the output from denoising NN, same as (14), but with stopping gradient (denoted by  $\text{sg}(\cdot)$ ). That is, the denoising NN  $\mathcal{D}_\theta$  is detached for this loss, when updating the parameters of the INR  $\mathcal{F}_\phi$  during training.

### 3.3. Summary

We jointly trains two NNs: an INR  $\mathcal{F}_\phi$  and a denoising NN  $\mathcal{D}_\theta$ , where the later take the output of the former as part of its input. The overall loss function consists of a self-supervised loss  $\mathcal{L}_{dn}$ , an INR-based loss  $\mathcal{L}_{inr}$  and a sub-pixel consistency loss  $\mathcal{L}_{mc}$ :

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{dn} + \mathcal{L}_{inr} + \mathcal{L}_{mc}. \quad (16)$$

## 4. Experiment

### 4.1. NN architecture for denoising NN and INR

Our approach does not reply on specific design of NN architecture. Thus, we adopt the off-the-shelf NN architectures to evaluate the performance gain from the proposed blind-spot scheme. The denoising NN  $\mathcal{D}_\theta$  follows the one used in Noise2Noise [22]. The design of  $\mathcal{F}_\phi$  follows SIREN [35],

Table 1. Quantitative comparisons (PSNR(dB)/SSIM) of different denoising methods. The best of each category are **bolded**.

Category	Method	SIDD Validation [1]	SIDD Benchmark [1]	FMDD [54]	PolyU [45]	CC [27]	
Supervised	DnCNN [51]	<b>37.73/0.943</b>	<b>37.61/0.941</b>	-	-	-	
Self-supervised & un-paired learning	Noise2Void [19]	27.06/0.651	26.99/0.652	-	-	-	
	NBR2NBR [13]	27.94/0.604	27.90/0.679	-	-	-	
	CVF-SID [28]	34.81/ <b>0.944</b>	34.71/0.917	<b>32.73/0.843</b>	35.86/0.937	33.29/0.913	
	LUD-VAE [56]	34.91/ <b>0.944</b>	34.82/0.926	-	<b>36.99/0.955</b>	<b>35.48/0.941</b>	
	R2R [29]	35.04/0.844	34.78/0.898	-	-	-	
	APBSN [21]	<b>36.73/0.878</b>	36.91/0.931	31.99/0.836	37.03/0.951	34.88/0.925	
	PUCA [15]	-	37.54/0.936	-	-	-	
	SelfFormer [34]	-	<b>37.69/0.937</b>	-	-	-	
Zero-shot	BM3D [10]	25.65/0.475	25.65/0.685	30.06/0.771	37.40/0.953	35.19/0.858	
	WNNM [11]	26.05/0.592	25.78/0.809	-	-	-	
	DIP [37]	32.11/0.740	-	32.90/0.854	37.17/0.912	35.61/0.912	
	Self2Self [31]	29.46/0.595	29.51/0.651	30.76/0.695	37.52/0.926	36.63/0.932	
	PD-denoising [58]	33.97/0.820	33.61/0.894	33.01/0.856	37.04/0.940	35.85/0.923	
	NN+denoising [55]	-	33.18/0.895	32.21/0.831	37.66/0.956	36.52/0.943	
	APBSN-single [21]	30.90/0.818	30.71/0.869	28.43/0.804	29.61/0.897	27.72/0.891	
	ScoreDVI [8]	34.75/0.856	34.60/0.920	33.10/0.865	37.77/ <b>0.959</b>	37.09/0.945	
	MASH [9]	35.06/0.851	34.78/0.900	33.71/0.882	37.62/0.932	36.87/0.935	
		Ours	<b>35.31/0.868</b>	<b>35.05/0.922</b>	<b>33.95/0.885</b>	<b>37.88/0.959</b>	<b>37.20/0.948</b>

with the same number of hidden neurons across all layers of the MLP: two input neurons and three output neurons. Sinusoidal functions are used as activation functions in all layers, except the last one. In addition, dropout is enabled when training  $\mathcal{F}_\phi$  to further alleviate possible overfitting to noise. More details can be found in complementary file.

## 4.2. Experiment configurations

Four datasets are used for benchmarking. (1) *Smartphone Image Denoising Dataset (SIDD)* [1], a dataset of the images captured by five different smartphones. SIDD validation set and benchmark set are used for validation and evaluation, respectively. Both sets contain 1,280 noisy patches of size  $256 \times 256$ , where only validate set have clean image patches. (2) *Fluorescence Microscopy Denoising Dataset (FMDD)* [54] with real grayscale fluorescence images acquired using commercial confocal, two-photon, and wide-field microscopes. These images include biological samples such as cells, zebrafish, and mouse brain tissues. Following [8], we denoise the raw test set images. The size of all images is  $512 \times 512$ . (3) *PolyU* [45] dataset with high-resolution images from various scenes, captured by five cameras. To keep evaluation consistency, we follow [8] to center-crop the PolyU images into patches of size  $256 \times 256$ . (4) *CC* [27] dataset with 15 real noisy images captured with different ISOs (1,600, 3,200 and 6,400). The size of all images is  $512 \times 512$ .

The proposed model is implemented using PyTorch 1.10, and the experiments are conducted on an NVIDIA 3090 GPU. The model is trained by the Adam optimizer, with

an initial learning rate of  $10^{-4}$  for the INR and  $4 \cdot 10^{-4}$  for the U-Net. The model is trained for 1200 iterations and  $p_0 = 0.1$ . Two metrics, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), are used for quantitative evaluation. Note that the results for SIDD benchmark results were obtained from online submission.

## 4.3. Evaluation on denoising real-world images

As a zero-shot self-supervised approach, our proposed method is benchmarked against a comprehensive selection of existing zero-shot techniques, including DIP [37], Self2Self [31], PD-denoising [58], NN+denoiser [55], APBSN-single [21], ScoreDVI [8], and MASH [9]. For APBSN-single, we adapted the one for dataset [21] to directly denoising images at test-time. The strides for PD-denoising are 5 (training) and 2 (test). For NN+denoiser, we use its optimized version, NN+BM3D, for benchmarking. The included dataset-based self-supervised methods are CVF-SID [28], LUD-VAE [56], and APBSN [21]. For reference, we also include DNCNN [51], a representative supervised method, and BM3D [10], a representative classic method. For each method, we utilize the authors’ original code or, where available, cite the published results directly.

**Quantitative evaluation:** Table 1 presents quantitative results. It can be seen that our method achieves the best PSNR and SSIM in this category. PD-denoising, with a limited stride of 2, struggles to effectively handle complex noise with long-range dependencies. APBSN-single, lacking dataset training, shows detail loss and color arti-

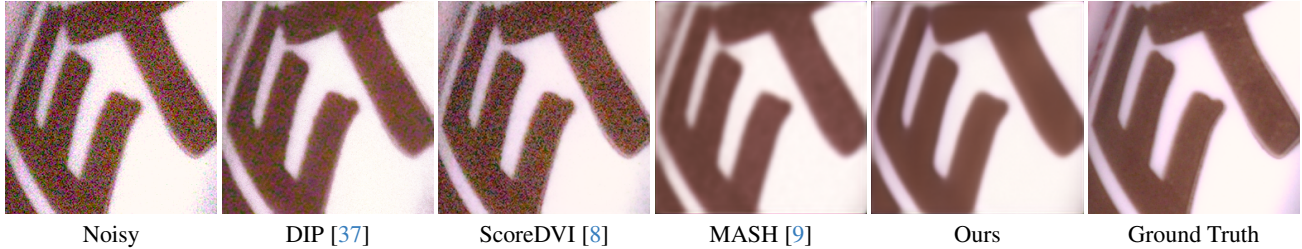


Figure 2. Visual comparison of results from different methods on samples from SIDD-Validation. Zoom in for detailed inspection.

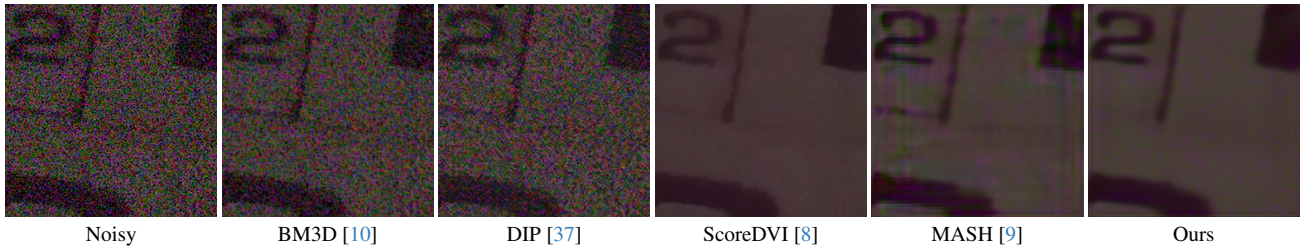


Figure 3. Visual comparison of results from different methods on samples from SIDD-Benchmark. Zoom in for detailed inspection.

facts due to the large PD stride, as it weakens intensity-value correlations of pixels used for prediction. BM3D and NN+denoiser, optimized for specific noise types, underperform when noise distribution is unknown. Noise2Void and Self2Self depend on spatial noise independence, resulting in poor outcomes on real-world images with locally correlated noise. Utilizing INR-based estimation of intensity values of nearby-pixels and distant pixels with noise independence, our method outperforms these zero-shot methods, including MASH and ScoreDVI.

**Visual comparison:** Refer to Fig 2,3,4,5 for visual comparisons of the results from different zero-shot methods on sample images from different datasets. Consistent with quantitative results, the results from the proposed method are better than that from the other methods in terms of visual quality, with fewer artifacts and better image preservation of image details.

**Complexity comparison:** Table 2 shows the inference time, model parameters, and FLOPs for several zero-shot methods, when processing an image of size  $256 \times 256$ . Together with the results from Table 1, these results demonstrate that our method consistently outperforms most existing zero-shot approaches, without sacrificing computational efficiency, making it a highly competitive and practical solution for real-world image denoising.

#### 4.4. Ablation Study

**Impact of masking probability  $p_0$ :** This study examines the impact of masking ratios  $p_0$  to denoising performance. Table 3 shows that real-world denoising performance improves as  $p_0$  decreases until it becomes too small,

Table 2. Complexity comparison of different zero-shot methods.

Method	Infer. time(s)	Params(M)	FLOPs(G)
DIP [37]	48.7	13.4	31.06
Self2Self [31]	1182	1.0	9.55
PD-denoising [58]	0.12	0.7	46.94
NN+denoising [55]	299.2	13.4	31.06
APBSN-single [21]	40.47	3.66	234.63
ScoreDVI [8]	27.07	13.5	37.87
MASH [9]	25.11	0.99	11.44
Ours	25.96	1.0	11.82

while AWGN denoising performance worsen with smaller  $p_0$ . This contrast arises from the stronger noise correlation in real-world images. Thus, using a lower masking ratio, which favors more distant visible pixels for prediction, is more effective for denoising correlated real-world noise.

Table 3. Comparison PSNR/SSIM of different masking ratio  $p_0$  on SIDD-Validation and AWGN with  $\sigma = 25$  on BSD68.

$p_0$	SIDD (correlated noise)	BSD68 (i.i.d AWGN)
0.50	34.84/0.849	27.51/0.712
0.30	35.02/0.857	27.13/0.683
0.10	<b>35.31/0.868</b>	<b>26.81/0.677</b>
0.05	35.25/0.866	26.74/0.676

**Performance contribution of three losses:** In this study, we remove each term from the overall loss function and evaluate its performance. When training the NN using only  $\mathcal{L}_{\text{dn}}$ , the invisible set for prediction is  $\Omega_{\text{vis}}^c$ , same as

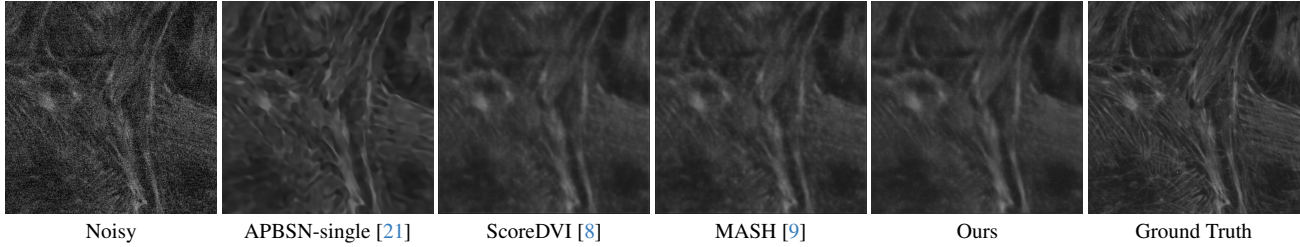


Figure 4. Visual comparison of results from different methods on one sample from FMDD. Zoom in for detailed inspection.

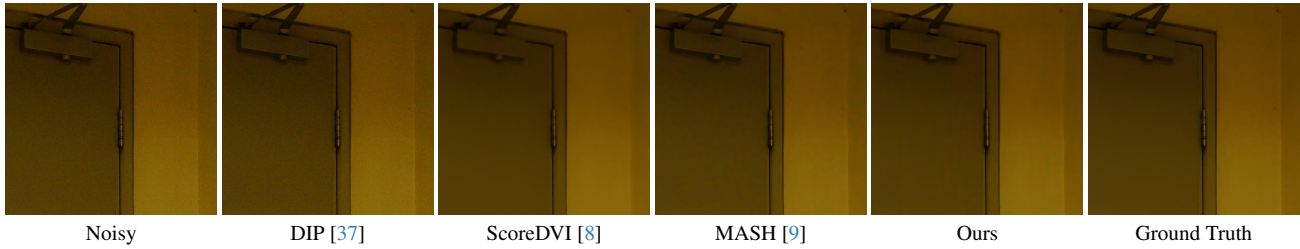


Figure 5. Visual comparison of results from different methods on one sample from PolyU. Zoom in for detailed inspection.

self2self [31]. For  $\mathcal{L}_{dn} + \mathcal{L}_{mc}$ , the INR is retained and trained solely with  $\mathcal{L}_{mc}$ . See Table 4 for the results. The results show that both proposed losses,  $\mathcal{L}_{inr}$  and  $\mathcal{L}_{mc}$ , made noticeable contributions to the performance gain.

Table 4. Contribution of different loss terms on SIDD-Validation.

	$\mathcal{L}_{dn}$	$\mathcal{L}_{dn} + \mathcal{L}_{inr}$	$\mathcal{L}_{dn} + \mathcal{L}_{mc}$	$\mathcal{L}_{dn} + \mathcal{L}_{inr} + \mathcal{L}_{mc}$
PSNR	29.13	34.75	35.14	35.31
SSIM	0.567	0.849	0.857	0.868

**Impact of dilation:** In this study, we evaluate the impact of dilation on the denoising performance. The results show that Dilation has a negligible impact on performance, as shown in Table 5 shows that dilation has a negligible impact on performance. This is because the INR effectively captures the intensity correlation among neighboring pixels, regardless of the dilation factor.

Table 5. Ablation of dilation on SIDD datasets

SIDD	w/o dilation	dilation=1	dilation=2	dilation=3
Validation	35.27/0.866	35.29/0.866	<b>35.31/0.868</b>	35.25/0.865
Benchmark	35.00/0.918	35.02/0.921	<b>35.05/0.922</b>	34.96/0.915

**Impact of iteration numbers:** This study evaluates how the iteration number will impact the performance. As shown in Figure 6, the performance of the proposed method consistently improves with more iterations, without requiring additional early stopping, unlike MASH.

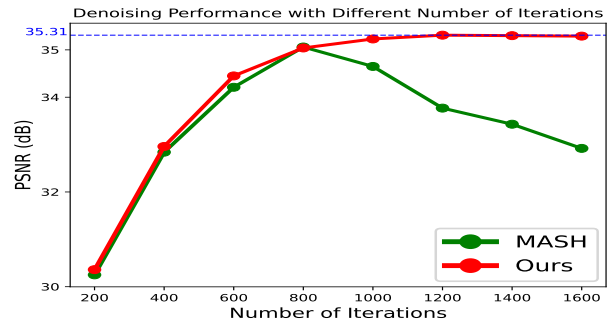


Figure 6. Performance vs. iteration number of our method and MASH on SIDD-Validation dataset with default settings.

**More analysis and visual comparisons:** Refer to supplementary file for more analysis and visual comparisons of the results from different methods.

## 5. Conclusion

This paper addressed the limitations of existing blind-spot methods for removing real-world correlated noise. Our work rigorously quantifies the impact of noise and pixel correlation on the statistical risk of a self-supervised linear blind-spot denoiser. Then, we introduced a novel approach using INP for resampling noisy invisible pixels. This method effectively reduces noise correlation while maintaining important intensity correlations among neighboring pixels, leading to improved denoising performance. Extensive experiments validate the effectiveness of our approach, as it consistently outperforms existing zero-shot denoising techniques on real-world noisy images.

## Acknowledgments

Yuhui Quan would like to acknowledge the support from National Natural Science Foundation of China under Grant 62072188. Hui Ji would like to acknowledge the support from Singapore MOE Academic Research Fund (AcRF) Tier 1 Research Grant A-8000981-00-00.

## References

- [1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018. 1, 6
- [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] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *Proceedings of the International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 1, 2, 3
- [4] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2018. 1, 3
- [5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European Conference on Computer Vision*, pages 17–33. Springer, 2022. 1, 3
- [6] Xiang Chen, Jinshan Pan, and Jiangxin Dong. Bidirectional multi-scale implicit neural representations for image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25627–25636, 2024. 3
- [7] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, and Xiaolong Wang. Videoir: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2047–2057, 2022. 3
- [8] Jun Cheng, Tao Liu, and Shan Tan. Score priors guided deep variational inference for unsupervised real-world single image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12937–12948, 2023. 1, 3, 6, 7, 8
- [9] Hamadi Chihaoui and Paolo Favaro. Masked and shuffled blind spot denoising for real-world images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3025–3034, 2024. 1, 2, 3, 6, 7, 8
- [10] 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. 6, 7
- [11] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xianguo Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2862–2869, 2014. 6
- [12] Young-Joo Han and Ha-Jin Yu. SS-BSN: Attentive blind-spot network for self-supervised denoising with nonlocal self-similarity. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 800–809, 2023. 3
- [13] 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. 1, 2, 3, 6
- [14] 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. 1, 3
- [15] Hyemi Jang, Junsung Park, Dahuin Jung, Jaihyun Lew, Ho Bae, and Sungho Yoon. PUCA: patch-unshuffle and channel attention for enhanced self-supervised image denoising. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 6
- [16] Chaewon Kim, Jaeho Lee, and Jinwoo Shin. Zero-shot blind image denoising via implicit neural representations. *arXiv preprint arXiv:2204.02405*, 2022. 3
- [17] 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. 3
- [18] 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. 3
- [19] 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. 1, 2, 3, 6
- [20] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image

- denoising. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022. 1, 2, 3, 6, 7, 8
- [22] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *Proceedings of the International Conference on Machine Learning*, pages 2965–2974, 2018. 1, 3, 5
- [23] Jason Lequyer, Reuben Philip, Amit Sharma, Wen-Hsin Hsu, and Laurence Pelletier. A fast blind zero-shot denoiser. *Nature Machine Intelligence*, 4(11): 953–963, 2022. 3
- [24] Junyi Li, Zhilu Zhang, Xiaoyu Liu, Chaoyu Feng, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Spatially adaptive self-supervised learning for real-world image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9914–9924, 2023. 3
- [25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 3
- [26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021. 3
- [27] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1683–1691, 2016. 6
- [28] 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. 3, 6
- [29] 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. 1, 3, 6
- [30] Mikhail Papkov and Pavel Chizhov. SwinIA: Self-supervised blind-spot image denoising with zero convolutions. *arXiv preprint arXiv:2305.05651*, 2023. 3
- [31] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1890–1898, 2020. 1, 2, 3, 6, 7, 8
- [32] Yuhui Quan, Yixin Chen, Yizhen Shao, Huan Teng, Yong Xu, and Hui Ji. Image denoising using complex-valued deep cnn. *Pattern Recognition*, 111:107639, 2021. 1, 3
- [33] Yuhui Quan, Xin Yao, and Hui Ji. Single image defocus deblurring via implicit neural inverse kernels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12600–12610, 2023. 3
- [34] Yuhui Quan, Tianxiang Zheng, and Hui Ji. Pseudo-siamese blind-spot transformers for self-supervised real-world denoising. *Advances in Neural Information Processing Systems*, 37:13794–13817, 2024. 1, 3, 6
- [35] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33: 7462–7473, 2020. 3, 5
- [36] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *Proceedings of the European Conference on Computer Vision*, pages 459–479, 2022. 3
- [37] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempit-sky. Deep image prior. In *Proceedings of the IEEE/Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018. 1, 3, 6, 7, 8
- [38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 1, 3
- [39] 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
- [40] 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. 1, 3

- [41] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18156–18165, 2023. 2, 3
- [42] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *Proceedings of the European Conference on Computer Vision*, pages 352–368, 2020. 3
- [43] Yaochen Xie, Zhengyang Wang, and Shuiwang Ji. Noise2same: Optimizing a self-supervised bound for image denoising. *Advances in Neural Information Processing Systems*, 33:20320–20330, 2020. 3
- [44] Yutong Xie, Mingze Yuan, Bin Dong, and Quanzheng Li. Unsupervised image denoising with score function. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [45] Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018. 6
- [46] 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. 3
- [47] 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
- [48] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 3
- [49] 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
- [50] Dan Zhang, Fangfang Zhou, Yuwen Jiang, and Zhengming Fu. Mm-bsn: Self-supervised image denoising for real-world with multi-mask based on blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4189–4198. IEEE, 2023. 3
- [51] 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, 3, 6
- [52] 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, 3
- [53] Tianjing Zhang, Yuhui Quan, and Hui Ji. Cross-scale self-supervised blind image deblurring via implicit neural representation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3
- [54] Yide Zhang, Yinhao Zhu, Evan Nichols, Qingfei Wang, Siyuan Zhang, Cody Smith, and Scott Howard. A poisson-gaussian denoising dataset with real fluorescence microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11710–11718, 2019. 1, 6
- [55] Dihan Zheng, Sia Huat Tan, Xiaowen Zhang, Zuoqiang Shi, Kaisheng Ma, and Chenglong Bao. An unsupervised deep learning approach for real-world image denoising. In *Proceedings of the International Conference on Learning Representations*, 2021. 1, 3, 6, 7
- [56] Dihan Zheng, Xiaowen Zhang, Kaisheng Ma, and Chenglong Bao. Learn from unpaired data for image restoration: A variational bayes approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5889–5903, 2022. 6
- [57] Huan Zheng, Tongyao Pang, and Hui Ji. Unsupervised deep video denoising with untrained network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3651–3659, 2023. 1, 3
- [58] 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*, pages 13074–13081, 2020. 1, 6, 7
- [59] 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*, pages 13074–13081, 2020. 2, 3
- [60] Magauiya Zhussip, Shakarim Soltanayev, and Se Young Chun. Extending stein’s unbiased risk estimator to train deep denoisers with correlated pairs of noisy images. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3