Zero-Shot Noise2Mean: Gap Minimization for Efficient Denoising from a Single Noisy Image

Duo Liu*, Yiqi Shi*, Guoyin Zhang, Sizhao Li†, Liguo Zhang†

College of Computer Science and Technology, Harbin Engineering University
{liu_duo, shiyiqi, zhangguoyin, sizhao.li, zhangliguo}@hrbeu.edu.cn

Abstract

Acquiring pairwise noisy-clean training data is challenging. Consequently, some self-supervised denoising methods utilize noisy image pairs as both input and target for network training. However, a major issue with these methods is the gap between the clean images of the input and target. In this paper, we achieve high-quality image denoising by reducing or even eliminating this gap. Our method requires no training data or prior knowledge of the noise distribution. It consists of two lightweight networks that can be trained using only a single noisy test image. Specifically, we propose a random mask-based downsampler that generates multiple pairs of downsampled noisy images, which are similar but distinct. These image pairs serve as the input for the first network, with the mean image of each pair used as the target. This initially reduces the gap between the clean images of the input and target. Particularly, in our method, the clean counterpart of the first network's target (i.e., the mean image) can be obtained. We then train a second network using the mean image as input and its clean counterpart as the target. This effectively eliminates the gap and achieves better denoising results. Extensive experiments demonstrate that our method outperforms in both denoising performance and efficiency.

Code — https://github.com/Doyle59217/ZS-N2M

1 Introduction

The primary goal of image denoising is to remove unwanted noise from an input image and restore a clean, noise-free version. The main challenge is distinguishing noise from the original image content without prior knowledge of the noise distribution. Recently, Convolutional Neural Networks have significantly advanced learning-based image denoising methods (Kim, Kim, and Baik 2024; Chen et al. 2021; Fu, Guo, and Wen 2023; Hong et al. 2020; Jang et al. 2021; Ren et al. 2021; Zamir et al. 2022). Most of these methods (Ma et al. 2022; Yue et al. 2020; Xu et al. 2021; Cheng et al. 2021; Liang et al. 2021; Wang et al. 2022a) rely heavily on training with extensive pairs of noisy-clean images, assuming fixed and known noise distributions. However, real-world images often contain unknown noise due to varying environments and different camera models. Thus, collecting a large dataset of paired noisy-clean images covering all possible noise types is impractical.

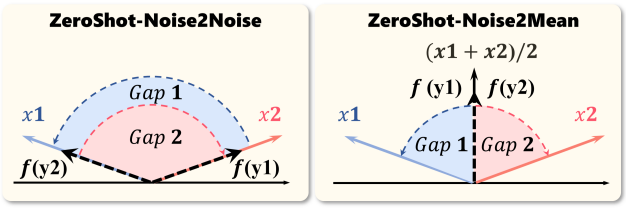

Figure 1: The principle of minimizing the gap in our method.

Self-supervised denoising methods enable model training using only noisy images and allow adaptation to unknown noise in the given images (Lee and Kim 2022; Lee, Son, and Lee 2022; Wang et al. 2023; Lequyer et al. 2022; Wang et al. 2022b; Chihaoui and Favaro 2024; Cheng, Liu, and Tan 2023; Neshatavar et al. 2022; Chen et al. 2024; Li et al. 2023; Vaksman and Elad 2023; Fadnavis et al. 2024). Noise2Noise (Lehtinen et al. 2018) introduces the concept of using different noisy observations of the same scene as both input and target for the network. Assuming the noise in each image is independent and follows a zero-mean distribution, it is equivalent to supervised network training. However, capturing multiple noisy observations of a scene remains highly challenging, especially for dynamic scenes or medical imaging. Neighbor2Neighbor (Huang et al. 2021) and ZS-N2N (Mansour and Heckel 2023) generate a pair of downsampled images from a noisy image to approximate different noisy observations of the same scene, serving as the network's input and target. However, a small but objective gap exists between the clean images of the two downsampled images. This inevitably leads to reduced network performance.

In this work, we propose **Z**ero-**S**hot **N**oise**2M**ean (**ZS-N2M**), a novel zero-shot denoising method that requires no training data or prior knowledge of the noise distribution and needs only a single noisy test image. We extend ZS-N2N (Mansour and Heckel 2023) to achieve higher-quality image denoising by reducing or even eliminating the gap between the clean images of the network's input and target. Specifically, we use two lightweight networks, the mean-based denoising network (MBD-Net) and the gap refinement de-

*These authors contributed equally.
†Corresponding authors.

noising network (GRD-Net), to achieve this goal. A random mask-based downsampler (RM Downsampler) is presented to generate multiple pairs of similar yet distinct downsampled images from a noisy test image, serving as input to the MBD-Net. The mean image of each downsampled pair is used as the target to train the MBD-Net, preliminarily reducing the gap between the clean images, as shown in Figure 1.

Particularly, the mean image of all downsampled noisy image pairs generated by the proposed RM Downsampler is identical and shares the same expected value as the original noisy image. This ensures that the training target constructed for MBD-Net retains the same expected value as the original noisy image. Since the ultimate goal is to denoise the original noisy image, the MBD-Net indirectly satisfies the consistency between the expected values of the network's input and target, as required by Noise2Noise. This allows MBD-Net to achieve improved and more stable denoising performance when processing the original noisy image.

Despite our efforts to bring the constructed target's clean image closer to the input's clean image, a gap still remains. We aim to further address this problem. The most direct way to eliminate the gap is to use the clean image of the network's input as the target. A distinctive feature of our method is that the trained MBD-Net actually outputs the clean counterpart of the mean image (i.e., the constructed target) when processing downsampled images. Therefore, we additionally train the GRD-Net, using the mean image and its clean counterpart as the input and target, respectively. This fundamentally resolves the gap between the clean images and ensures consistency in the expected value. The input and training target of the GRD-Net all share the same expected value as the original noisy image, enabling our method eventually to achieve superior denoising results.

Our contributions can be summarized as follows:

- We propose a novel zero-shot image denoising method, ZS-N2M, which requires no training data or prior knowledge of the noise distribution and uses only a single noisy test image for training.
- By constructing new training targets, we reduce or even eliminate the gap between the clean images of the network's input and target, achieving high-quality image denoising with two lightweight networks.
- Utilizing the presented random mask-based downsampler, multiple pairs of sufficiently similar yet distinct downsampled noisy images are generated to ensure stable denoising performance of the network.

Extensive experiments demonstrate that ZS-N2M outperforms state-of-the-art zero-shot denoising methods on both synthetic and real-world noisy images, including microscope images, when comprehensively considering both denoising effectiveness and efficiency.

2 Motivation

2.1 Background

Noise2Noise (Lehtinen et al. 2018) proposes that, assuming zero-mean noise, a neural network can be trained using different noisy observations, y and z, of the same clean image x. It demonstrates that mapping one noisy image y to another noisy image z is equivalent to training the network in a supervised manner to map y to the clean image x:

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}\left[\|f_{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{x}\|_2^2\right] = \arg\min_{\boldsymbol{\theta}} \mathbb{E}\left[\|f_{\boldsymbol{\theta}}(\mathbf{y}) - \mathbf{z}\|_2^2\right] \quad (1)$$

where $f_{\boldsymbol{\theta}}$ is the denoising network parameterized by θ.

Neighbor2Neighbor (Huang et al. 2021) extends Noise2Noise by enabling network training on a set of single noisy images. Assuming noise at different positions in the image is uncorrelated, this method introduces a random neighbor sub-sampling technique to generate independent training image pairs. The method maps one sub-sampled noisy image to another for network training and introduces a regularization term to address the gap between the clean images of the sub-sampled pairs.

ZS-N2N (Mansour and Heckel 2023) is further extended to allow network training with only a single noisy image y. The method introduces a filter that computes the mean of diagonal pixels to generate a pair of downsampled noisy images, y_1 and y_2, and trains the network by mapping them to each other. However, it overlooks a crucial point regarding the gap between the clean images of the two downsampled noisy images. The core principle of Noise2Noise is to map the input y to the clean image of the target z. Since the clean images of y and z are identical, this process is equivalent to mapping y to its own clean image. ZS-N2N, however, actually maps y_1 to the clean image x_2 of y_2, rather than to its own clean image x_1. This inevitably leads to a decline in network performance. Specifically, the gap between the clean images of the network's input and target is defined by $D_{\text{MSE}}(x_1, x_2)$, where D_{MSE} denotes the mean squared error.

2.2 Reducing and Eliminating the Gap

Addressing the aforementioned issue is essential to achieving better denoising results. A new noisy image can be constructed as the network's target, bringing its clean image closer to the input's clean image. For the downsampled noisy image pair y_1 and y_2, the clean image of their mean image $(y_1 + y_2)/2$, which is $(x_1 + x_2)/2$, is simultaneously closer to their individual clean images, x_1 and x_2. Using y_1 as input and $(y_1 + y_2)/2$ as target is equivalent to mapping y_1 to $(x_1 + x_2)/2$. The gap between the clean images of the network's input and target is $D_{\text{MSE}}(x_1, (x_1 + x_2)/2) = \frac{1}{4} D_{\text{MSE}}(x_1, x_2)$. This also holds for y_2. Compared to ZS-N2N, the gap is reduced by three-quarters, as shown in Figure 1. The proof is provided in the Supplementary Material.

Mapping a pair of downsampled noisy images to their mean image already brings the clean images of the network's input and target as close as possible, but the gap still remains. The most effective way to eliminate the gap is to directly use the true clean image of the noisy input as the target. As explained earlier, setting $(y_1 + y_2)/2$ as the target, the network actually learns $(x_1 + x_2)/2$. Specifically, when the network processes y_1 and y_2, it may not output the exact clean images x_1 and x_2, but it can produce $(x_1 + x_2)/2$. Notably, $(x_1 + x_2)/2$ is the clean counterpart of $(y_1 + y_2)/2$. This indicates that a noisy image and its clean counterpart can

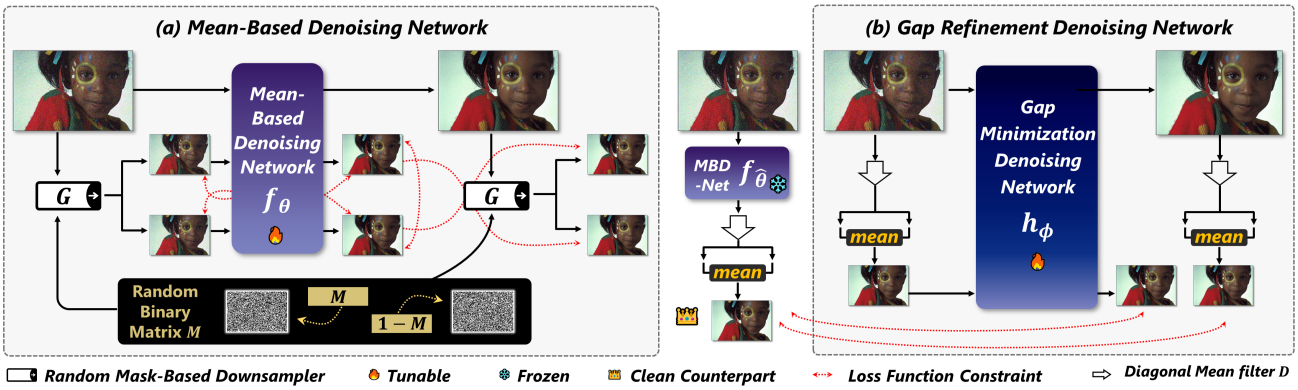

Figure 2: Overview of ZS-N2M framework. (a) Mean-based denoising network. The RM Downsampler $G = (G_{1n}, G_{2n})$ generates N pairs of downsampled images from a single noisy test image y. For the n^{th} pair, $(G_{1n}(y), G_{2n}(y))$ serve as network inputs, with their mean image used as the mapping target. (b) Gap refinement denoising network. The MBD-Net remains frozen to obtain $(D_1(f_{\hat{\theta}}(y)) + D_2(f_{\hat{\theta}}(y)))/2$, serving as the target for the GRD-Net input $(D_1(y) + D_2(y))/2$ during training. $D = (D_1, D_2)$ represents the filter that computes the mean of diagonal pixels. Red lines show our training strategy.

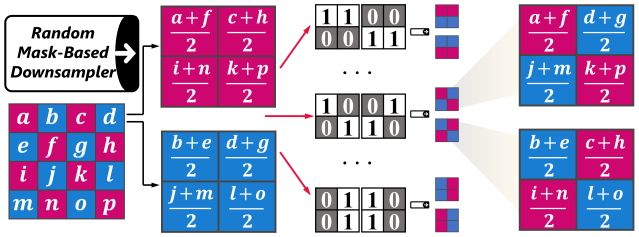

Figure 3: Random mask-based downsampling. The RM Downsampler $G = (G_{1n}, G_{2n})$ generates N pairs of images, each with size $H/2 \times W/2 \times C$, from a noisy image y with size $H \times W \times C$. For the n^{th} pair, the steps are: (1) Decompose y into two images, $D_1(y)$ and $D_2(y)$, each half the size of y, by averaging the diagonal pixels within non-overlapping 2×2 patches. (2) Generate a random binary matrix M_n of size $H/2 \times W/2 \times C$, containing an equal number of 0s and 1s. (3) Compute $G_{1n}(y)$ and $G_{2n}(y)$.

be obtained and subsequently used as the input and target to train a new network for improved denoising performance.

3 Method

Based on our motivation in Section 2, we present ZS-N2M. Figure 2 shows an overview of our framework. Initially, the proposed RM Downsampler is employed to generate N pairs of downsampled noisy images. These image pairs are subsequently utilized to train the MBD-Net. Finally, the frozen MBD-Net is used to obtain the clean counterpart of its mapping target, which is then used to train the GRD-Net. The training framework is detailed in Algorithms 1 and 2 provided in the Supplementary Material.

3.1 Random Mask-Based Downsampler

Since the ultimate goal is to denoise the original noisy image y and obtain its clean image x, the network is needed to stably output the result with an expected value of $\mathbb{E}[x]$. Thus, the constructed training target should share the same expected value as x, implying that the mean image of the downsampled image pair generated by our method should have the expected value of $\mathbb{E}[x]$. However, achieving network stability using only a single pair of downsampled noisy images is challenging. A sufficient number of downsampled image pairs, similar yet distinct from y_1 and y_2, is required. And the targets constructed using these pairs should all share the same expected value as x.

Therefore, as shown in Figure 3, we introduce the RM Downsampler $G = (G_{1n}, G_{2n})$ to generate N pairs of noisy images $(G_{1n}(y), G_{2n}(y))$ from a single noisy test image y, where $n = 1, 2, ..., N$. Specifically, we first apply the filter $D = (D_1, D_2)$ of ZS-N2N (Mansour and Heckel 2023), which calculates the mean of diagonal pixels, to obtain a pair of noisy images $y_1 = D_1(y)$ and $y_2 = D_2(y)$. Next, we generate a random matrix $M_n (n = 1, 2, ..., N)$ of the same size as y_1 and y_2, containing only 0s and 1s, each occupying 50%. Finally, the n^{th} pair is expressed as:

$$\begin{cases} y_{1n} = G_{1n}(y) = D_1(y) \circ M_n + D_2(y) \circ (1 - M_n), \\ y_{2n} = G_{2n}(y) = D_1(y) \circ (1 - M_n) + D_2(y) \circ M_n \end{cases}$$

where $\circ$ denotes element-wise multiplication. This guarantees a sufficient number of downsampled image pairs that are similar yet distinct. A key advantage is that, regardless of changes in M_n, the pixel correspondence between y_{1n} and y_{2n} at corresponding positions remains intact. In other words, the mapping targets constructed for all downsampled image pairs satisfy $(y_{1n} + y_{2n})/2 = (y_1 + y_2)/2$. Similarly, the clean images of these mapping targets satisfy $(x_{1n} + x_{2n})/2 = (x_1 + x_2)/2$. This ensures that the network produces more stable results.

More importantly, RM Downsampler preserves all pixel values of the original noisy image during downsampling, ensuring that all training targets share the same expected value as y, i.e., $\mathbb{E}[(y_{1n} + y_{2n})/2] = \mathbb{E}[(y_1 + y_2)/2] = \mathbb{E}[y]$. Similarly, the clean images of these targets satisfy $\mathbb{E}[(x_{1n} + x_{2n})/2] = \mathbb{E}[(x_1 + x_2)/2] = \mathbb{E}[x]$. Assum-

ing zero-mean noise, the expected value of a noisy image equals that of its clean image, i.e., $\mathbb{E}\left[(y_{1n}+y_{2n})/2\right] = \mathbb{E}\left[(x_{1n}+x_{2n})/2\right] = \mathbb{E}\left[x\right]$. Particularly, $(y_1+y_2)/2$ is also a downsampled version of y and satisfies $\mathbb{E}\left[(y_1+y_2)/2\right] = \mathbb{E}\left[x\right]$. Consequently, training with pairs of downsampled noisy images generated by our RM Downsampler and their mean images would produce a network whose outputs consistently maintain expected values matching $\mathbb{E}\left[x\right]$, contributing to better denoising of the original noisy image.

3.2 Mean-Based Denoising Network

We avoid directly setting MBD-Net's target as $(y_{1n}+y_{2n})/2$ because, while it reduces the gap between the clean images of the input and target, it increases noise similarity. Noise similarity describes how closely the noise at corresponding pixel positions matches between a pair of noisy images. We observed that higher noise similarity impairs the network's learning effectiveness. For instance, when a noisy image serves as both input and target, the network fails to learn effective denoising. Thus, we adopt an indirect approach.

First, to ensure that both noisy images in each downsampled pair yield the same result after network processing, we propose the interactive equality loss, defined as:

$$\begin{aligned}\mathcal{L}_{\text{ie}}(\boldsymbol{\theta}) = \|f_{\boldsymbol{\theta}}(y_{1n}) - G_{2n}(f_{\boldsymbol{\theta}}(y))\|_2^2 \\ + \|f_{\boldsymbol{\theta}}(y_{2n}) - G_{1n}(f_{\boldsymbol{\theta}}(y))\|_2^2\end{aligned} \tag{2}$$

where $f_{\boldsymbol{\theta}}$ is the denoising network parameterized by θ.

$$\begin{cases} G_{1n}(f_{\boldsymbol{\theta}}(y)) = D_1(f_{\boldsymbol{\theta}}(y)) \circ M_n + D_2(f_{\boldsymbol{\theta}}(y)) \circ (1 - M_n), \\ G_{2n}(f_{\boldsymbol{\theta}}(y)) = D_1(f_{\boldsymbol{\theta}}(y)) \circ (1 - M_n) + D_2(f_{\boldsymbol{\theta}}(y)) \circ M_n \end{cases}$$

$G_{1n}(f_{\boldsymbol{\theta}}(y))$ represents first denoising the original noisy image y and then performing RM downsampling on it. It is consistent with $f_{\boldsymbol{\theta}}(G_{1n}(y))$, i.e., $f_{\boldsymbol{\theta}}(y_{1n})$, where RM downsampling is performed on y before denoising. The same applies to $G_{2n}(f_{\boldsymbol{\theta}}(y))$. This enables the network to process downsampled images while also accessing the original noisy image, leading to improved denoising performance.

Next, we employ the cross mapping loss:

$$\mathcal{L}_{\text{cm}}(\boldsymbol{\theta}) = \|f_{\boldsymbol{\theta}}(y_{1n}) - y_{2n}\|_2^2 + \|f_{\boldsymbol{\theta}}(y_{2n}) - y_{1n}\|_2^2 \tag{3}$$

In simple terms, these two loss terms ensure that the network outputs the same result when processing y_{1n} and y_{2n}, equivalently mapping both to the same image. Additionally, they must map to each other, which is equivalent to mapping y_{1n} and y_{2n} to their mean image $(y_{1n}+y_{2n})/2$.

We also introduce the interactive denoising loss to improve denoising performance:

$$\mathcal{L}_{\text{id}}(\boldsymbol{\theta}) = \sum_{i=1,2} \|f_{\boldsymbol{\theta}}(y_{in}) - (d \circ f_{\boldsymbol{\theta}}(y_{in}) + (1-d) \circ \mu_i)\|_2^2 \tag{4}$$

where μ_i represents the average of all pixel values within a 5×5 window at the corresponding position in $f_{\boldsymbol{\theta}}(y_{in})$. $d = 2\sigma_1\sigma_2(\sigma_1^2+\sigma_2^2+C)^{-1}$. σ_i represents the standard deviation of the pixels within a 5×5 window at the corresponding positions in $f_{\boldsymbol{\theta}}(y_{in})$'s luminance channel. C is a constant.

The total loss function of the MBD-Net is defined as $\mathcal{L}(\boldsymbol{\theta}) = \alpha_{\text{ie}}\mathcal{L}_{\text{ie}}(\boldsymbol{\theta}) + \alpha_{\text{cm}}\mathcal{L}_{\text{cm}}(\boldsymbol{\theta}) + \alpha_{\text{id}}\mathcal{L}_{\text{id}}(\boldsymbol{\theta})$, where the corresponding weights are $\alpha_{\text{cm}} = 5$ and $\alpha_{\text{id}} = 10$, with the specific setting for α_{ie} detailed in the experiments.

3.3 Gap Refinement Denoising Network

To produce results that better approximate the clean image of the input, we propose the GRD-Net. The parameters $\hat{\theta}$ of the trained MBD-Net are frozen to obtain $(D_1(f_{\hat{\boldsymbol{\theta}}}(y)) + D_2(f_{\hat{\boldsymbol{\theta}}}(y)))/2$ to train the GRD-Net.

Specifically, as detailed in Section 2.2, $f_{\hat{\boldsymbol{\theta}}}(y_1)$ serves as the clean counterpart of the noisy image $(y_1+y_2)/2$. Notably, $D_1(f_{\hat{\boldsymbol{\theta}}}(y))$ is consistent with $f_{\hat{\boldsymbol{\theta}}}(D_1(y))$, i.e., $f_{\hat{\boldsymbol{\theta}}}(y_1)$, so $D_1(f_{\hat{\boldsymbol{\theta}}}(y))$ can also serve as $(y_1+y_2)/2$'s clean counterpart. Similarly, $D_2(f_{\hat{\boldsymbol{\theta}}}(y))$ also serves as the clean counterpart of $(y_1+y_2)/2$. To further reduce errors, we calculate the average of these two results and set it as the training target for GRD-Net, with input $(y_1+y_2)/2$. Moreover, GRD-Net's input and target both share the expected value $\mathbb{E}\left[x\right]$, enabling more effective denoising of the original noisy image y.

The reconstruction loss is expressed as:

$$\mathcal{L}_{\text{re}}(\boldsymbol{\phi}) = \|h_{\boldsymbol{\phi}}(\frac{y_1+y_2}{2}) - \frac{D_1(f_{\hat{\boldsymbol{\theta}}}(y)) + D_2(f_{\hat{\boldsymbol{\theta}}}(y))}{2}\|_2^2 \tag{5}$$

where $h_{\boldsymbol{\phi}}$ is the denoising network parameterized by ϕ.

To enable the GRD-Net to access the original noisy image, we apply the consistency loss:

$$\mathcal{L}_{\text{con}}(\boldsymbol{\phi}) = \left\| \frac{1}{2}\sum_{i=1}^{2} D_i(h_{\boldsymbol{\phi}}(y)) - \frac{1}{2}\sum_{i=1}^{2} D_i(f_{\hat{\boldsymbol{\theta}}}(y)) \right\|_2^2 \tag{6}$$

where $\frac{1}{2}\sum_{i=1}^{2} D_i(h_{\boldsymbol{\phi}}(y))$ aligns with $h_{\boldsymbol{\phi}}((y_1+y_2)/2)$, differing only in the order of downsampling and denoising. $\frac{1}{2}\sum_{i=1}^{2} D_i(f_{\hat{\boldsymbol{\theta}}}(y)) = (D_1(f_{\hat{\boldsymbol{\theta}}}(y)) + D_2(f_{\hat{\boldsymbol{\theta}}}(y)))/2$.

The total loss function of GRD-Net is defined as $\mathcal{L}(\boldsymbol{\phi}) = \beta_{\text{re}}\mathcal{L}_{\text{re}}(\boldsymbol{\phi}) + \beta_{\text{con}}\mathcal{L}_{\text{con}}(\boldsymbol{\phi})$, where $\beta_{\text{re}} = 100$ and $\beta_{\text{con}} = 1000$. Notably, our method is zero-shot, the final result $h_{\hat{\boldsymbol{\phi}}}(y)$ is obtained directly during the training on the noisy test image y, eliminating the need for a separate execution phase. $\hat{\phi}$ represents the optimal network parameters.

4 Experiments

4.1 Implementation Details

Datasets for Synthetic Experiments. We use the Kodak24 (Franzen 2010) and Set14 (Zeyde, Elad, and Protter 2012) datasets. To ensure consistency, all test images are center-cropped to 256×256. We consider four types of synthetic noise distributions: (1) Gaussian noise with a fixed level $\sigma = 25$, (2) variable Gaussian noise with levels $\sigma \in [5, 50]$, (3) Poisson noise with a fixed level $\lambda = 25$, and (4) variable Poisson noise with levels $\lambda \in [5, 50]$. Note that the σ values correspond to image intensity levels in the range $[0, 255]$, while λ correspond to that in the range $[0, 1]$.

Datasets for Real-world Experiments. We consider four types of real-world noise conditions. (1) For camera noise, we randomly select 20 images, each sized 512×512, from

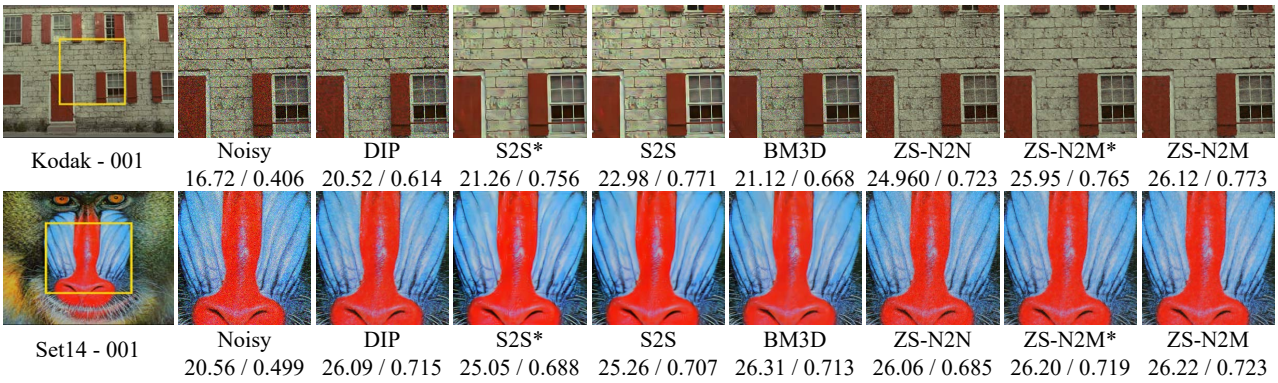

Figure 4: Visual comparison of different dataset-free denoising methods for synthetic noise. The first row shows visual comparison in the setting of Poisson noise with $\lambda = 19$ on the Kodak24 dataset. The second row shows visual comparison in the setting of Gaussian noise with $\sigma = 25$ on the Set14 dataset. The quantitative PSNR(dB)/SSIM results are listed below the images.

	Method	Gaussian Noise				Poisson Noise			
		$\sigma = 25$		$\sigma \in [5, 50]$		$\lambda = 25$		$\lambda \in [5, 50]$	
		Kodak24	Set14	Kodak24	Set14	Kodak24	Set14	Kodak24	Set14
data-based	N2C	28.01 / 0.778	27.44 / 0.751	27.46 / 0.759	27.62 / 0.729	27.13 / 0.744	26.77 / 0.758	25.55 / 0.701	25.88 / 0.705
	N2N	27.92 / 0.776	27.31 / 0.749	27.49 / 0.750	27.62 / 0.728	27.11 / 0.725	26.85 / 0.743	25.47 / 0.700	25.86 / 0.704
	N2V	27.72 / 0.765	27.22 / 0.730	27.58 / 0.742	27.52 / 0.724	25.28 / 0.702	27.05 / 0.770	25.38 / 0.698	25.28 / 0.709
	NB2NB	27.90 / 0.773	27.23 / 0.728	27.92 / 0.746	27.58 / 0.721	27.01 / 0.760	27.08 / 0.769	25.41 / 0.701	25.33 / 0.702
dataset-free	DIP	26.47 / 0.707	26.47 / 0.707	26.97 / 0.713	27.07 / 0.723	25.83 / 0.716	26.11 / 0.699	26.56 / 0.710	25.78 / 0.685
	S2S	27.38 / 0.758	27.82 / 0.752	28.58 / 0.756	26.52 / 0.699	**28.02** / 0.737	**28.13 / 0.795**	24.05 / 0.693	**27.42** / 0.773
	S2S*	27.21 / 0.752	27.75 / 0.699	28.28 / 0.750	26.43 / 0.695	27.12 / 0.702	**27.98** / 0.787	25.05 / 0.674	26.53 / 0.733
	BM3D	**30.95 / 0.866**	**30.94 / 0.865**	27.94 / 0.730	27.94 / 0.730	25.63 / 0.693	25.63 / 0.691	23.94 / 0.653	24.32 / 0.623
	ZS-N2N	28.27 / 0.776	27.88 / 0.753	28.28 / 0.761	27.88 / 0.753	27.03 / 0.714	25.99 / 0.720	26.45 / 0.714	26.46 / 0.731
	ZS-N2M*	29.39 / 0.812	28.76 / 0.814	**29.28 / 0.793**	**28.85 / 0.801**	27.82 / **0.766**	27.22 / 0.784	**26.97 / 0.734**	27.27 / **0.777**
	ZS-N2M	**29.56 / 0.827**	**29.05 / 0.823**	**29.90 / 0.811**	**29.23 / 0.817**	**27.93 / 0.790**	27.43 / **0.792**	**27.17 / 0.753**	**27.55 / 0.785**

Table 1: Quantitative comparison (PSNR (dB) / SSIM) of different denoising methods for synthetic noise. The best and second-best results are both highlighted in **bold**.

the PolyU (Xu et al. 2018) dataset, which contains high-resolution images of various scenes captured by five different cameras. (2) We also use the SIDD (Abdelhamed, Lin, and Brown 2018) dataset, comprising images captured by various smartphone cameras under different lighting conditions and noise patterns. 20 images are randomly selected and cropped to 256×256. (3) We select 20 images, each 512×512 in size, from the Fluorescence Microscopy (Zhang et al. 2019) dataset, containing real grayscale fluorescence images obtained with commercial confocal, two-photon, and wide-field microscopes, including representative biological samples. (4) The FastMRI (Zbontar et al. 2018) dataset is used to extend testing to the domain of medical imaging.

Compared Methods and Metrics. We compare our ZS-N2M with one supervised method (N2C), three self-supervised methods (Noise2Noise (N2N) (Lehtinen et al. 2018), Neighbor2Neighbor (NB2NB) (Huang et al. 2021), and Noise2Void (N2V) (Krull, Buchholz, and Jug 2019)), three zero-shot methods (ZS-N2N (Mansour and Heckel 2023), DIP (Ulyanov, Vedaldi, and Lempitsky 2018), and Self2Self (S2S) (Quan et al. 2020)), and one traditional method (BM3D (Dabov et al. 2007)). For real-world noise, we also compare to Noise2Fast (N2F) (Lequyer et al. 2022) and MASH (Chihaoui and Favaro 2024). Notably, we denote the results of S2S without any ensemble as S2S* and the results using only the MBD-Net in our method as ZS-N2M*. For the dataset-based methods (N2C, N2N, N2V, and NB2NB), we train the networks on 500 color images from the ImageNet (Deng et al. 2009) dataset, covering noise level ranges of $\sigma, \lambda \in [5, 50]$. BM3D requires the noise level as the input. For Gaussian noise, we use $\sigma = 25$, while for Poisson noise, the noise level is estimated using the method in (Chen, Zhu, and Ann Heng 2015). We employ two commonly used metrics, PSNR and SSIM, which measure the similarity between the reconstructed image and the original image, where higher values generally indicate better results.

4.2 Benchmark Evaluations

Synthetic Experiments. As shown in Table 1, the performance of dataset-based methods is slightly lower than that of dataset-free methods, primarily due to the limited training on only 500 images. However, with an increase in training data, the performance of these methods improves significantly. This indicates a heavy dependence on training data,

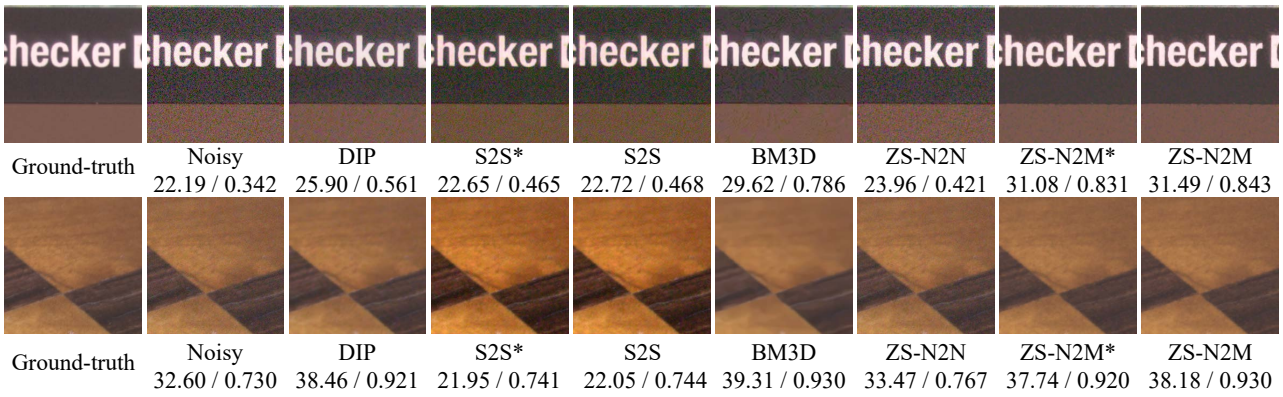

Figure 5: Visual comparison of different dataset-free denoising methods for real-world noise on the SIDD dataset. The quantitative PSNR (dB) / SSIM results are listed below the images.

	Method	SIDD	PolyU	Microscopy
data-based	N2C	33.58 / 0.907	34.28 / 0.909	34.02 / 0.890
	N2N	33.49 / 0.905	34.25 / 0.907	33.98 / 0.886
	N2V	33.21 / 0.898	33.93 / 0.895	33.83 / 0.886
	NB2NB	33.12 / 0.902	33.95 / 0.902	33.92 / 0.892
dataset-free	DIP	33.28 / 0.921	**37.78 / 0.938**	33.56 / 0.901
	S2S	33.48 / 0.920	35.88 / 0.915	34.09 / 0.900
	S2S*	32.62 / 0.892	34.61 / 0.895	33.97 / 0.893
	BM3D	**37.23 / 0.932**	**38.13 / 0.957**	34.03 / 0.891
	MASH	36.27 / 0.902	37.58 / 0.928	34.27 / 0.908
	N2F	-	-	**34.58 / 0.921**
	ZS-N2N	33.28 / 0.879	34.95 / 0.897	33.48 / 0.897
	ZS-N2M*	36.73 / 0.910	36.38 / 0.926	34.25 / 0.907
	ZS-N2M	**36.89 / 0.927**	37.65 / 0.933	**34.53 / 0.915**

Table 2: Quantitative comparison (PSNR (dB) / SSIM) of different denoising methods for real-world noise. The best and second-best results are both highlighted in **bold**.

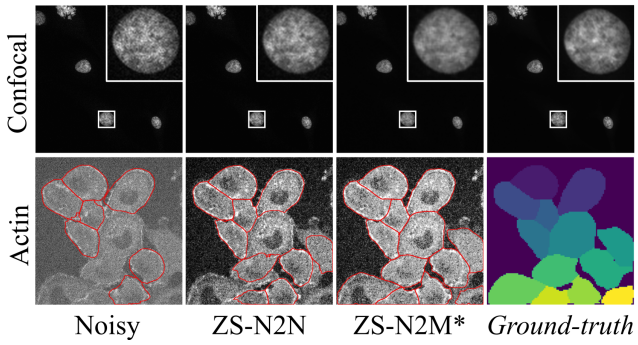

Figure 6: Experimental results of our method on biological cell images (Lequyer et al. 2022).

potentially limiting their effectiveness on unfamiliar images. BM3D, a traditional method specifically designed for Gaussian noise, performs well when the noise variance is known but degrades when the noise characteristics are unclear.

Among the zero-shot methods, DIP heavily depends on the number of training iterations, which is difficult to predetermine in practice. S2S shows robust performance under Poisson noise. However, a comparison with S2S* reveals that its superior results largely rely on ensembling, which incurs high computational costs. ZS-N2N shows relatively stable performance under unknown noise levels and types. In comparison, our method achieves superior metrics, being a well-performing zero-shot denoising method.

Figure 4 presents the visual effects of different methods on synthetic noise. BM3D produces excellent results for Gaussian noise but struggles with Poisson. S2S introduces some color deviation compared to others. In contrast, our method performs consistently well on both types of noise, effectively removing noise while preserving image details.

Real-World Experiments. In real-world photography, explicit noise modeling is challenging, and simply modeling noise with Gaussian and Poisson distributions is insufficient. As a result, dataset-based methods often fail to generalize well to real-world denoising, as shown in Table 2. Among the zero-shot methods, our method achieves satisfactory results on real-world noisy images captured by smartphones and cameras, demonstrating its practical applicability. Figure 5 illustrates the visual effects of different methods on real-world noise. It is evident that S2S exhibits more severe color deviations when removing real-world noise. Our method is the only one that effectively removes noise without over-smoothing the image.

Additionally, we demonstrate the potential of our method in biological imaging applications. The results in Table 2 on the Fluorescence Microscopy (Zhang et al. 2019) dataset show that our method achieves scores comparable to N2F (which provides code for grayscale denoising) and outperforms other methods. Figure 6 shows the visual effects of our method on the Fluorescence Microscopy dataset. It is evident that, compared to ZS-N2N, the image denoised by our ZS-N2M* alone is more similar to the ground truth. Figure 6 also demonstrates our effectiveness in downstream cell image segmentation tasks. The segmentation accuracy significantly improves after denoising images with our ZS-N2M*, surpassing the results of ZS-N2N.

	DIP	S2S	ZS-N2N	ZS-N2M*	ZS-N2M
GPU time	3 min	35 min	**20 sec**	**20 sec**	2 min
CPU time	45 min	4.5 hr	**80 sec**	**80 sec**	40 min
Size	2.2M	1M	**22K**	**22K**	2.7M

Table 3: GPU/CPU time and network size of different zero-shot denoising methods. The best and second-best results are both highlighted in **bold**.

Noise Type	ZS-N2N	$+ G$	w/o G	w/o $\mathcal{L}_{\text{id}}$	ZS-N2M*
Gaussian $\sigma = 25$	28.27	28.57	28.65	29.20	**29.39**
Gaussian $\sigma \in [5, 50]$	28.28	28.59	28.75	29.05	**29.28**
Poisson $\lambda = 25$	27.03	27.32	27.38	27.64	**27.82**
Poisson $\lambda \in [5, 50]$	26.45	26.51	26.56	26.78	**26.97**

Table 4: Ablation study on the RM Downsampler $G = (G_{1n}, G_{2n})$ and the loss terms $\mathcal{L}_{\text{ie}}(\boldsymbol{\theta})$ and $\mathcal{L}_{\text{id}}(\boldsymbol{\theta})$. PSNR (dB) results are evaluated on the Kodak24 dataset.

Computational Efficiency. Dataset-based methods require extensive training time. However, once trained, the network parameters are fixed, and inference becomes nearly instantaneous, as it involves only a forward pass through the model. Consequently, the denoising time for these methods is negligible compared to zero-shot methods, which optimize parameters for each test image individually. Table 3 presents the time required to denoise a 256×256 color image using different zero-shot methods, along with the total number of trainable parameters for each model.

Specifically, among zero-shot methods, ZS-N2M ranks second in efficiency, surpassed only by ZS-N2N. This is due to our GRD-Net utilizing a relatively large UNet (Ronneberger, Fischer, and Brox 2015) architecture. However, our MBD-Net uses the same simple network structure as ZS-N2N, consisting of two convolutional layers with kernel sizes of 3×3 and one layer with a kernel size of 1×1. With approximately 20K parameters, ZS-N2M* achieves efficiency comparable to ZS-N2N. Previous experiments have shown that ZS-N2M* produces better denoising results than ZS-N2N. Therefore, if higher denoising efficiency is desired, ZS-N2M* is a suitable choice, while ZS-N2M is preferred for superior denoising performance.

4.3 Ablation Study

Since our MBD-Net is based on ZS-N2N (Mansour and Heckel 2023), we include the N pairs of downsampled images generated by our RM Downsampler in the comparison with this method. As shown in Table 4, training ZS-N2N with our N pairs of noisy images improves metrics compared to training with a single pair, confirming the effectiveness of our proposed RM Downsampler. This also validates the effectiveness of our $\mathcal{L}_{\text{ie}}(\boldsymbol{\theta})$ loss term. When the $\mathcal{L}_{\text{id}}(\boldsymbol{\theta})$ loss term is excluded and the model is trained on the same N pairs of downsampled noisy images, our method differs from ZS-N2N only by the $\mathcal{L}_{\text{ie}}(\boldsymbol{\theta})$ loss term. Nevertheless, our metrics are higher. This confirms that the inclusion of the $\mathcal{L}_{\text{ie}}(\boldsymbol{\theta})$ loss term allows our method to effectively reduce the gap between the clean images of the network's input and target. We achieve this by indirectly setting the mean image as the mapping target, whereas ZS-N2N directly uses the downsampled image.

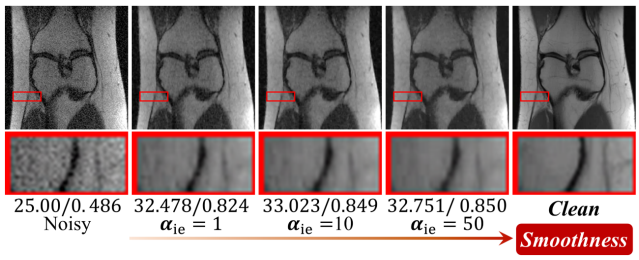

Figure 7: Visual comparison of different α_{ie} values for denoising Gaussian noise on a Knee MRI from the fastMRI (Zbontar et al. 2018) dataset. The quantitative PSNR (dB) / SSIM results are listed below the images.

Additionally, compared to ZS-N2M*, removing the RM Downsampler and training with a single downsampled pair leads to decreased metrics. This suggests that using a sufficient number of similar downsampled pairs as input, with their mean image as the target, leads to a better match. It effectively enhances denoising performance and network stability. The decrease in metrics resulting from removing the $\mathcal{L}_{\text{id}}(\boldsymbol{\theta})$ loss term further demonstrates its effectiveness.

Figure 7 illustrates the effect of different weight values α_{ie} of the $\mathcal{L}_{\text{ie}}(\boldsymbol{\theta})$ loss term on denoising performance for the same noisy image. As α_{ie} increases, the denoised image becomes progressively smoother, indicating that an excessively high α_{ie} value may cause loss of details. Therefore, choosing an appropriate α_{ie} value is crucial for achieving clean and clear results, depending on the noise level and denoising requirements. In this paper, we use $\alpha_{\text{ie}} = 1$ for synthetic experiments and $\alpha_{\text{ie}} = 10$ for real-world experiments.

5 Conclusion

We propose ZS-N2M, a novel zero-shot image denoising method. By constructing new mapping targets for the network, we reduce or even eliminate the gap between the clean images of the input and target, achieving high-quality denoising results. We introduce a random mask-based downsampler that generates multiple pairs of downsampled images from a single noisy image to facilitate training. This ensures stable denoising performance of the network. Additionally, our method consists of two lightweight networks, providing significant advantages in efficiency. This allows for "test-as-train", where no training data or prior knowledge of the noise distribution is required. Denoising can be done with just a single noisy test image. This is particularly critical for processing real-time images acquired in dynamic environments. Extensive experiments demonstrate the superiority of our method in handling both synthetic noise and real-world noise from cameras and microscopes.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2022YFB4400703, 2021YFC3320302).

The authors would also like to acknowledge anonymous reviewers and chairs for providing insightful comments to help improve this work.

References

Abdelhamed, A.; Lin, S.; and Brown, M. S. 2018. A High-Quality Denoising Dataset for Smartphone Cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1692–1700.

Chen, G.; Zhu, F.; and Ann Heng, P. 2015. An Efficient Statistical Method for Image Noise Level Estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 477–485.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021. Hinet: Half Instance Normalization Network for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 182–192.

Chen, S.; Zhang, J.; Yu, Z.; and Huang, T. 2024. Exploring Efficient Asymmetric Blind-Spots for Self-Supervised Denoising in Real-World Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2814–2823.

Cheng, J.; Liu, T.; and Tan, S. 2023. Score Priors Guided Deep Variational Inference for Unsupervised Real-World Single Image Denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12937–12948.

Cheng, S.; Wang, Y.; Huang, H.; Liu, D.; Fan, H.; and Liu, S. 2021. Nbnet: Noise Basis Learning for Image Denoising with Subspace Projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4896–4906.

Chihaoui, H.; and Favaro, P. 2024. Masked and Shuffled Blind Spot Denoising for Real-World Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3025–3034.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Transactions on Image Processing*, 16(8): 2080–2095.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. IEEE.

Fadnavis, S.; Chowdhury, A.; Batson, J.; Drineas, P.; and Garyfallidis, E. 2024. Patch2Self2: Self-supervised Denoising on Coresets via Matrix Sketching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27641–27651.

Franzen, R. 2010. Kodak Lossless True Color Image Suite. https://r0k.us/graphics/kodak/. Accessed: 2024-05-27.

Fu, Z.; Guo, L.; and Wen, B. 2023. sRGB Real Noise Synthesizing with Neighboring Correlation-Aware Noise Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1683–1691.

Hong, Z.; Fan, X.; Jiang, T.; and Feng, J. 2020. End-to-end Unpaired Image Denoising with Conditional Adversarial Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4140–4149.

Huang, T.; Li, S.; Jia, X.; Lu, H.; and Liu, J. 2021. Neighbor2neighbor: Self-Supervised Denoising from Single Noisy Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14781–14790.

Jang, G.; Lee, W.; Son, S.; and Lee, K. M. 2021. C2n: Practical Generative Noise Modeling for Real-World Denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2350–2359.

Kim, C.; Kim, T. H.; and Baik, S. 2024. LAN: Learning to Adapt Noise for Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25193–25202.

Krull, A.; Buchholz, T.-O.; and Jug, F. 2019. Noise2Void-Learning Denoising from Single Noisy Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2129–2137.

Lee, S.; and Kim, T. H. 2022. Noisetransfer: Image Noise Generation with Contrastive Embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 3569–3585.

Lee, W.; Son, S.; and Lee, K. M. 2022. 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*, 17725–17734.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning Image Restoration without Clean Data. *arXiv preprint arXiv:1803.04189*.

Lequyer, J.; Philip, R.; Sharma, A.; Hsu, W.-H.; and Pelletier, L. 2022. A Fast Blind Zero-Shot Denoiser. *Nature Machine Intelligence*, 4(11): 953–963.

Li, J.; Zhang, Z.; Liu, X.; Feng, C.; Wang, X.; Lei, L.; and Zuo, W. 2023. Spatially Adaptive Self-Supervised Learning for Real-World Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9914–9924.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image Restoration using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Ma, R.; Li, S.; Zhang, B.; and Li, Z. 2022. Generative Adaptive Convolutions for Real-World Noisy Image Denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1935–1943.

Mansour, Y.; and Heckel, R. 2023. Zero-Shot Noise2Noise: Efficient Image Denoising without any Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14018–14027.

Neshatavar, R.; Yavartanoo, M.; Son, S.; and Lee, K. M. 2022. 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*, 17583–17591.

Quan, Y.; Chen, M.; Pang, T.; and Ji, H. 2020. Self2Self with Dropout: Learning Self-Supervised Denoising from Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1890–1898.

Ren, C.; He, X.; Wang, C.; and Zhao, Z. 2021. Adaptive Consistency Prior Based Deep Network for Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8596–8606.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep Image Prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Vaksman, G.; and Elad, M. 2023. Patch-Craft Self-Supervised Training for Correlated Image Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5795–5804.

Wang, J.; Di, S.; Chen, L.; and Ng, C. W. W. 2023. Noise2info: Noisy Image to Information of Noise for Self-Supervised Image Denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16034–16043.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022a. Uformer: A General U-Shaped Transformer for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

Wang, Z.; Liu, J.; Li, G.; and Han, H. 2022b. Blind2unblind: Self-supervised Image Denoising with Visible Blind Spots. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2027–2036.

Xu, J.; Li, H.; Liang, Z.; Zhang, D.; and Zhang, L. 2018. Real-World Noisy Image Denoising: A New Benchmark. *arXiv preprint arXiv:1804.02603*.

Xu, L.; Zhang, J.; Cheng, X.; Zhang, F.; Wei, X.; and Ren, J. 2021. Efficient Deep Image Denoising via Class Specific Convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3039–3046.

Yue, Z.; Zhao, Q.; Zhang, L.; and Meng, D. 2020. Dual Adversarial Network: Toward Real-World Noise Removal and Noise Generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, 41–58. Springer.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zbontar, J.; Knoll, F.; Sriram, A.; Murrell, T.; Huang, Z.; Muckley, M. J.; Defazio, A.; Stern, R.; Johnson, P.; Bruno, M.; et al. 2018. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *arXiv preprint arXiv:1811.08839*.

Zeyde, R.; Elad, M.; and Protter, M. 2012. On Single Image Scale-up using Sparse-Representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, 711–730. Springer.

Zhang, Y.; Zhu, Y.; Nichols, E.; Wang, Q.; Zhang, S.; Smith, C.; and Howard, S. 2019. A Poisson-Gaussian Denoising Dataset with Real Fluorescence Microscopy Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11710–11718.