Prompt-SID: Learning Structural Representation Prompt via Latent Diffusion for Single-Image Denoising

Huaqiu Li*, **Wang Zhang***, **Xiaowan Hu, Tao Jiang, Zikang Chen, Haoqian Wang**†

Tsinghua University
lihq23@mails.tsinghua.edu.cn

Abstract

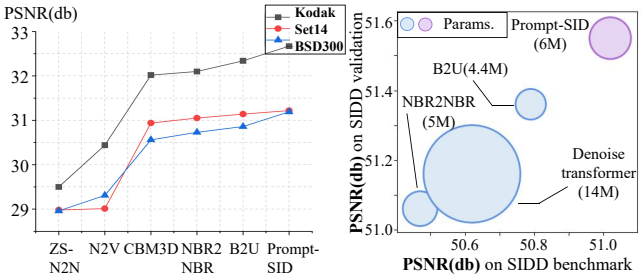
Many studies have concentrated on constructing supervised models utilizing paired datasets for image denoising, which proves to be expensive and time-consuming. Current self-supervised and unsupervised approaches typically rely on blind-spot networks or sub-image pairs sampling, resulting in pixel information loss and destruction of detailed structural information, thereby significantly constraining the efficacy of such methods. In this paper, we introduce Prompt-SID, a **prompt**-learning-based **s**ingle **i**mage **d**enoising framework that emphasizes preserving of structural details. This approach is trained in a self-supervised manner using downsampled image pairs. It captures original-scale image information through structural encoding and integrates this prompt into the denoiser. To achieve this, we propose a structural representation generation model based on the latent diffusion process and design a structural attention module within the transformer-based denoiser architecture to decode the prompt. Additionally, we introduce a scale replay training mechanism, which effectively mitigates the scale gap from images of different resolutions. We conduct comprehensive experiments on synthetic, real-world, and fluorescence imaging datasets, showcasing the remarkable effectiveness of Prompt-SID.

Code — https://github.com/huaqlili/Prompt-SID.

Introduction

Image noise arises from diverse sources, including sensor noise and environmental factors, alongside potential introduction during quantization and image processing procedures, thereby exerting adverse impacts on downstream tasks such as classification (Wang et al. 2017), detection (Shijila, Tom, and George 2019), and segmentation (Liu et al. 2020). Consequently, the quest for efficacious image denoising methodologies assumes critical significance within the domain of computer vision research.

In recent years, there has been a proliferation of learning-based supervised denoising methodologies (Zhang, Zuo, and Zhang 2018; Anwar and Barnes 2019; Menteş et al. 2021; Zhang et al. 2017; Zamir et al. 2022b, 2021; Zhang et al. 2023). Nonetheless, supervised denoising methods are beset by certain limitations, including their reliance on labeled data and their limited adaptability to real-world scenarios.



(a) Performance (b) Parameter size

Figure 1: Comparison of Prompt-SID with other self-supervised image denoising methods in terms of model parameters and experimental results of setting $\sigma \in [5,50]$.

Alternative paradigms such as unsupervised and self-supervised methods (Laine et al. 2019; Wu et al. 2020; Pang et al. 2021; Papkov and Chizhov 2023; Wang et al. 2023; Zhang and Zhou 2023; Lehtinen et al. 2018) have emerged to circumvent these constraints. Traditional self-supervised denoising methods often employ mask strategies (Huang et al. 2021; Wang et al. 2022) to extract downsampled images or introduce blind spots by altering convolutional kernel visibility (Lee, Son, and Lee 2022; Song, Meng, and Ermon 2020; Krull, Buchholz, and Jug 2019). However, these methods, although effective in building image denoising pipelines, suffer from significant pixel information loss. During the process of sampling sub-images, some pixels are selected while others are discarded. Additionally, in the training of blind-spot networks, the central pixel of the convolutional kernel is also invisible. Furthermore, compared to blind-spot networks, downsampled images suffer from more severe structural damage and semantic degradation.

To address the aforementioned issues, we introduce Prompt-SID, a prompt-learning-based single-image denoising framework that primarily addresses the semantic degradation and structural damage caused by the sampling processes of previous self-supervised methods. We design a structural representation generation diffusion (RG-Diff) based on a latent diffusion model, using the degraded struc-

*These authors contributed equally.
†Haoqian Wang is corresponding author.

tural representations as conditional information to guide the recovery of undamaged ones. In this process, we encode information from all pixels, thereby preserving the previously invisible pixels while avoiding identity mapping. We also design a structural attention module (SAM) in the denoiser to integrate the structural representation as a prompt, further decoding them into feature images. Our approach leverages a multi-scale alternating training regime to mitigate the issues of information loss and structural disruption typically encountered during the sub-sampling process. Additionally, through the scale replay mechanism, our method effectively reduces the scale gap and achieves domain adaptation. During inference, our framework generalizes seamlessly to denoising tasks on original scale images, maintaining the integrity of structural details. Notably, just as shown in Fig. 1, our approach has demonstrated impressive results on synthetic, real-world, and fluorescence imaging datasets while maintaining a relatively low parameter count.

The contributions of our work can be summarized as:

- Based on prompt learning, we develop a self-supervised image denoising pipeline, extracting structural representations from the original images to inform and guide the restoration process of the downsampled inputs.
- To bridge the scale gap between the downsampled domain and the original resolution domain, we devised a branch dedicated to processing the original resolution, indirectly contributing to the optimization process to prevent pixel identity mapping.
- Pioneering of applying diffusion models to self-supervised image denoising, we have engineered a novel structural representation generation diffusion, leveraging the powerful capability of the generative model to refine semantic representation prompts within the latent space.
- Our method surpasses existing SOTA approaches across various datasets, including synthetic, real-world, and fluorescence imaging datasets, demonstrating its superiority in image-denoising tasks.

Related Works

Self-Supervised Image Denoising

Self-supervised image denoising methods have evolved primarily along two paths. The first path, exemplified by methods like noise2void (N2V) (Krull, Buchholz, and Jug 2019), employs blind spot to introduce invisible pixels within the central region of convolutional kernels, thereby circumventing the issue of identity mapping. Recent advancements such as AP-BSN (Lee, Son, and Lee 2022) extend blind spot networks by introducing a shuffling mechanism to disrupt the spatial continuity of noise in natural images. Additionally, some studies (Wang et al. 2023) modify the blind spot areas within convolutional kernels. The second path, as Fig. 2 shows, represented by noise2noise (N2N) (Lehtinen et al. 2018), positing that training with L2 loss tends to converge towards the mean of observed values. This suggests the feasibility of replacing desired training targets with distributions having similar means. Subsequent endeavors (Huang et al. 2021; Mansour and Heckel 2023; Li et al. 2023) have been dedicated to self-supervised training by downsampling inputs and targets from single noisy images.

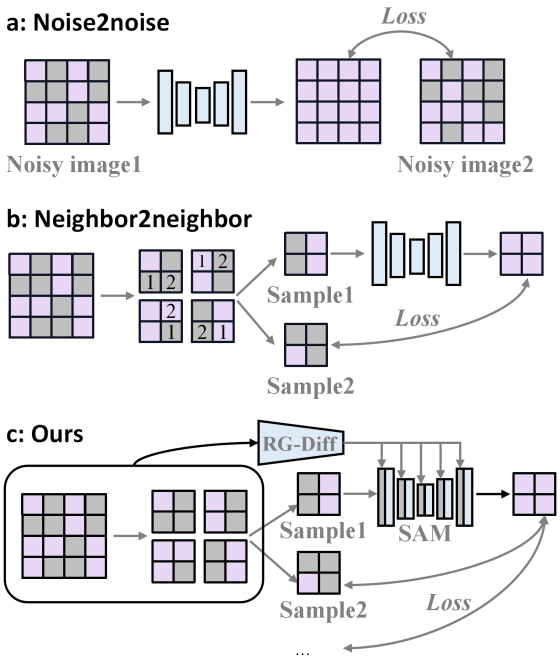


Figure 2: The distinctions between the pipelines of N2N, NBR2NBR, and Prompt-SID.

Diffusion Model in Low-Level Tasks

Diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Kawar et al. 2022) leverage parameterized Markov chains to optimize the lower variational bound on the likelihood function, thereby enabling them to generate more precise target distributions compared to other generative models such as GANs. Over recent years, they have garnered significant attention in image restoration tasks, including super-resolution (Li et al. 2022; Lin et al. 2023), enhancement (Wang et al. 2024), inpainting (Xia et al. 2023), and so forth. These approaches often fine-tune a pre-trained stable diffusion model and directly decode the generated latent space representations to obtain outputs. However, these generation methods inevitably introduce a degree of randomness, resulting in subtle semantic deviations at the image level due to sampling Gaussian noise. Additionally, it fails to meet the requirements for lightweight deployment.

Method

To address the issues of low pixel utilization and structural damage, we made the following improvements in Prompt-SID. Firstly, We applied a spatial redundancy sampling strategy to minimize pixel wastage. Secondly, during the training phase at the downscaled image, we introduced RG-Diff for extracting structural representations via latent diffusion. Leveraging the generative capacity of the diffusion model, We aim for the model to utilize structural information from

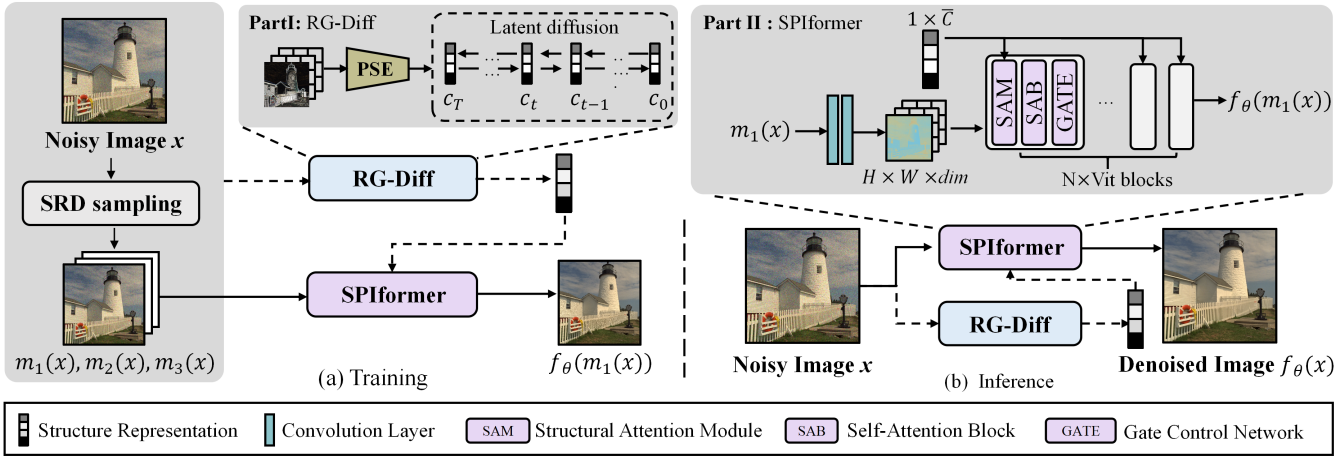


Figure 3: The primary denoising pipeline of Prompt-SID. (a) This method acquires the sub-images for network training through a spatial redundancy sampling strategy. These inputs are denoised using SPIformer, while the original image's structural representation is obtained as a prompt through RG-Diff. Each Transformer block incorporates a SAM to facilitate feature fusion. (b) During inference, Prompt-SID exclusively employs the original scale image through SPIformer and the RG-Diff branch.

the downscaled images to recover corresponding representations at the original scale. The structural representations generated are then fused into the SPIformer using the SAM mechanism. Additionally, to ensure the trained model generalizes effectively on original-scale images, we incorporated a scale replay mechanism during training: following the processing of downscaled images in each iteration, gradients are frozen, and an additional inference pass is conducted on the original-scale images. The pipeline of our method is illustrated in Fig. 3.

Spatial Redundancy Sampling Strategy

Following the principles of noise2noise, targets that adhere to a zero-mean noise while similar to the ground truth can serve as a supervisory signal. So we sample the input and target for network training within a single noisy image.

By employing spatial redundancy sampling strategy m, we can obtain sub-images $m_1(\mathbf{x})$, $m_2(\mathbf{x})$, $m_3(\mathbf{x})$ from the original-scale noisy image x. First, we divide the image $\mathbf{x}$ into $h/2 \times w/2$ small blocks, with each block containing four pixels. The small block located in the i-th row and j-th column is named $b(\mathbf{i}, \mathbf{j})$. From each block, we randomly sample three adjacent pixels $p_1(b(\mathbf{i}, \mathbf{j}))$, $p_2(b(\mathbf{i}, \mathbf{j}))$, $p_3(b(\mathbf{i}, \mathbf{j}))$, where $p_1(b(\mathbf{i}, \mathbf{j}))$ is adjacent to the other two pixels. The selection of $p_2(b(\mathbf{i}, \mathbf{j}))$ and $p_3(b(\mathbf{i}, \mathbf{j}))$ is random among the remaining two pixels. Subsequently, we obtained three sub-images that are one-fourth the size of the original image.

The process can be written as follows:

$$m_n(\mathbf{x}) = \sum_{\mathbf{i}=1}^{h/2} \sum_{\mathbf{j}=1}^{w/2} p_n(b(\mathbf{i}, \mathbf{j})), n = 1, 2, 3 \quad (1)$$

Structural Representation Generation Diffusion

We propose structural representation generation diffusion (RG-Diff), performing the diffusion process within a $1 \times N$ dimensional vector space. To minimize the randomness in the generation process, we design a joint training framework using the $\mathcal{L}_1$ loss in vector space, and integrating the generated representations into the feature map processing branch, rather than directly decoding them into output results. The operational principle of RG-Diff is illustrated in Fig. 4.

First, we designed a pixel structure encoder (PSE) to compress image information into the implicit space and extract structural representations. The PSE comprises several residual blocks, a global average pooling layer, and two linear layers. We encode the downscaled image $m_1(\mathbf{x})$ and the original scale image $\mathbf{x}$, resulting in the structural representations of the downsampled image $\mathbf{c}_{sub}$ and the original scale image $mathbf{c}_{org(0)}$, respectively. The process can be represented by the following equation:

$$\mathbf{c}_{sub} = PSE(m_1(\mathbf{x})) \quad (2)$$

$$\mathbf{c}_{org(0)} = PSE(\mathbf{x}) \quad (3)$$

Subsequently, we perform the forward diffusion process based on $\mathbf{c}_{org(0)}$. At a sampled time step t, the forward diffusion is carried out using the following equation, where $\mathbf{c}_{org(0)}$ serves as the initial state. We introduce noise to this representation according to the Markov process.

$$q(\mathbf{c}_{org(t)}|\mathbf{c}_{org(0)}) = \mathcal{N}(\mathbf{c}_{org(t)}; \sqrt{\bar{\alpha}_t}\mathbf{c}_{org(0)}; (1-\bar{\alpha}_t)\mathbf{I}) \quad (4)$$

Here, $\mathbf{c}_{org(t)}$ represents the state with noise obtained after $\mathbf{t}$ steps of sampling on $\mathbf{c}_{org(0)}$. $\bar{\alpha}_t$ is a manually designed hyperparameter. β_t is the predefined scale factor, which increases linearly with the time steps. The relationship between $\bar{\alpha}_t$ and β_t satisfies: $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t$.

In the reverse process, we incorporate $\mathbf{c}_{sub}$ as a conditional control input through concatenation during the $\mathbf{t}$-step denoising procedure. Given that the features are one-dimensional vectors, we employ MLP for this task. Unlike the conventional reverse diffusion process, where the inputs

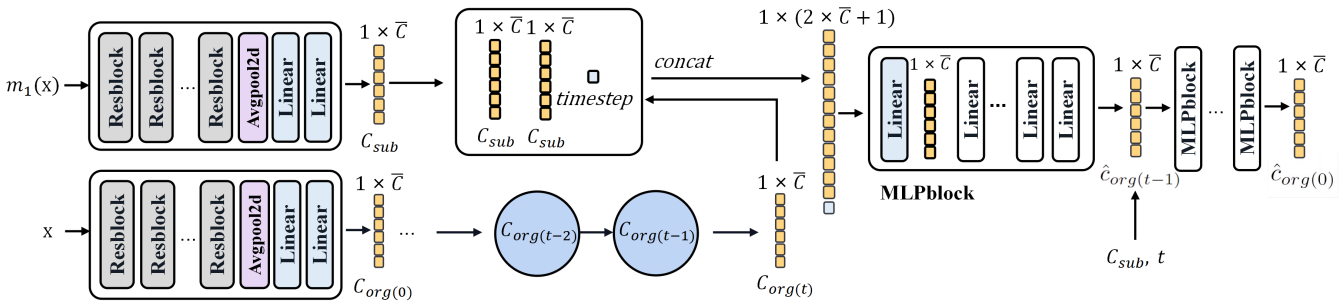


Figure 4: Diagram of the RG-Diff branch. Initially, PSE encodes the image representation into an implicit space, followed by a diffusion process within this space to obtain $\mathbf{c}_{org(t)}$. Utilizing the representation of $m_1(\mathbf{x})$ as a conditioning factor, RG-Diff guides the restoration of the representation of $\mathbf{x}$. This is achieved by merging $\mathbf{c}_{org(t)}$, $\mathbf{c}_{sub}$ and timestep $\mathbf{t}$ in the reverse diffusion stage inputting them into the denoising network.

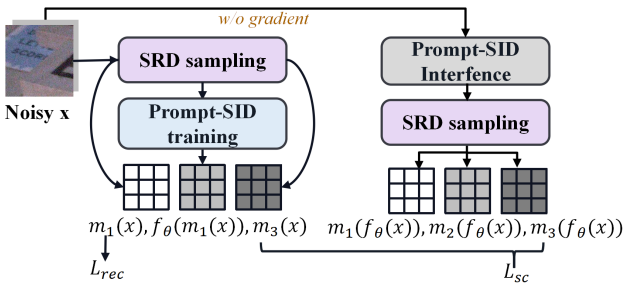


Figure 5: Introducing a scale-replay training branch without gradient backpropagation. We pass the original-scale noisy image $\mathbf{x}$ through Prompt-SID and downsample the denoised result to obtain $m_1(f_\theta(\mathbf{x}))$, $m_2(f_\theta(\mathbf{x}))$, $m_3(f_\theta(\mathbf{x}))$. These downscaled outputs are utilized to enforce regularization constraints on the image restoration loss.

to the denoising network consist of time step $\mathbf{t}$ and the intermediate state $\mathbf{c}_t$, we concatenate $\mathbf{c}_{sub}$ with $\mathbf{c}_{org(t)}$ at the feature level. This joint input is crucial for guiding subsequent generation steps. The reverse diffusion process at each time step can be articulated as follows:

$$\hat{\mathbf{c}}_{org(t-1)} = \frac{1}{\sqrt{\alpha_t}}\{\hat{\mathbf{c}}_{org(t)} - f_\theta(\hat{\mathbf{c}}_{org(t)}, \mathbf{c}_{sub}, \mathbf{t})\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\} \quad (5)$$

In this context, f_θ denotes the parameters of the denoising network. Due to the computational efficiency of this branch, we perform reverse diffusion across all time steps to obtain the final representation $\hat{\mathbf{c}}_{org(0)}$. To control the generation direction, we impose the following constraint using L_1 loss.

$$\mathcal{L}_{diff} = \left\|\hat{\mathbf{c}}_{org(0)} - \mathbf{c}_{org(0)}\right\|_1 \quad (6)$$

The structural representation $\hat{\mathbf{c}}_{org(0)}$ is utilized in the image restoration branch for decoding, thereby directing the generation at the feature map level.

Structural Prompt Integrative Transformer

We employ the vision transformer (ViT) (Dosovitskiy et al. 2020) module as the branch of image reconstruction. Similar to prior research (Zamir et al. 2022a; Zhang et al. 2023; Chen et al. 2022), our transformer module comprises two components: the multi-head self-attention block and the gate control network. Additionally, we introduce a structural attention module (SAM) to incorporate the previously generated structural representation $\hat{\mathbf{c}}_{org(0)}$ into the feature map.

The primary operation principle of SAM can be delineated into two phases: channel attention extraction and computation, and the integration of structural embedding information. We acquire channel attention weights through global average pooling and 1x1 convolution applied to the feature maps, as illustrated by the following equation:

$$\mathbf{c}_{sca} = AvgPool(\hat{\mathbf{F}}) * \mathbf{W}_{l1} + \mathbf{b}_{l1} \quad (7)$$

In this equation, $AvgPool(\hat{\mathbf{F}})$ denotes the global average pooling operation applied to the feature map $\hat{\mathbf{F}}$. The resulting output is subsequently multiplied by the weight $\mathbf{W}_{l1}$ and added to the bias $\mathbf{b}_{l1}$. Subsequently, we merge $\mathbf{c}_{sca}$ and $\hat{\mathbf{c}}_{org(0)}$ to jointly derive channel attention weights that direct the processing of the feature maps. The precise procedure is outlined as follows:

$$\begin{aligned}\mathcal{F} = \mathbf{W}_{s1}\mathbf{c}_{sca} \odot \mathbf{W}_{c1}\hat{\mathbf{c}}_{org(0)} \odot Norm(\hat{\mathbf{F}}) \\ + \mathbf{W}_{s2}\mathbf{c}_{sca} \odot \mathbf{W}_{c2}\hat{\mathbf{c}}_{org(0)}\end{aligned} \quad (8)$$

In the equation mentioned above, $\mathbf{W}_{s1}, \mathbf{W}_{s2}, \mathbf{W}_{c1}, \mathbf{W}_{c2}$ represents the weight matrix of the linear layer. $\mathcal{F}$ represents the feature map processed by the SAM.

Scale Replay Mechanism and Loss

After passing through SPIformer, we derive $f_\theta(m_1(\mathbf{x}))$, where f_θ denotes the network parameters requiring optimization in RG-Diff and SPIformer. The reconstruction loss is computed by evaluating the L_2 loss between $f_\theta(m_1(\mathbf{x}))$ and $m_2(\mathbf{x})$, as well as between $f_\theta(m_1(\mathbf{x}))$ and $m_3(\mathbf{x})$. The specific formula is outlined as follows:

$$\mathcal{L}_{rec} = \|f_\theta(m_1(\mathbf{x})) - m_2(\mathbf{x})\|_2 + \|f_\theta(m_1(\mathbf{x})) - m_3(\mathbf{x})\|_2 \quad (9)$$

In the preceding discussion, we emphasized the necessity of addressing the generalization between downscaled and original-scale images. Our objective is to train a model capable of alleviating the domain gap between them. Therefore, in each iteration, we conduct an additional inference process

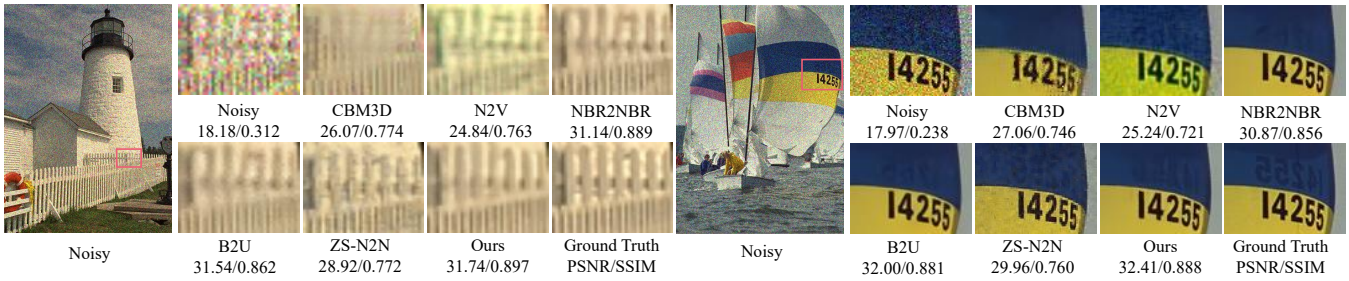


Figure 6: The visual comparison of Prompt-SID with state-of-the-art self-supervised image denoising methods in synthetic noise experiments, demonstrating results for Poisson noise set at level 30 tested on the Kodak dataset.

	Dataset	Baseline,N2C	CBM3D	N2V	NBR2NBR	B2U	ZS-N2N	Ours
σ = 25	Kodak	32.43/0.884	31.87/0.868	30.32/0.821	32.08/0.879	32.27/0.880	29.25/0.779	**32.41/0.883**
	BSD300	31.05/0.879	30.48/0.861	29.34/0.824	30.79/0.873	30.87/0.872	28.56/0.801	**31.16/0.880**
	Set14	31.40/0.869	30.88/0.854	28.84/0.802	31.09/0.864	31.27/0.864	28.29/0.779	**31.45/0.868**
$\sigma \in$ [5,50]	Kodak	32.51/0.875	32.02/0.860	30.44/0.806	32.10/0.870	32.34/0.872	29.50/0.767	**32.67/0.876**
	BSD300	31.07/0.866	30.56/0.847	29.31/0.801	30.73/0.861	30.86/0.861	28.96/0.797	**31.19/0.866**
	Set14	31.41/0.863	30.94/0.849	29.01/0.792	31.05/0.858	31.14/0.857	28.98/0.783	**31.22/0.860**

Table 1: Quantitative results on synthetic datasets in sRGB space for Gaussian noise set. The highest PSNR(dB)/SSIM among unsupervised denoising methods are highlighted in **bold**.

	Dataset	Baseline,N2C	CBM3D	N2V	NBR2NBR	B2U	ZS-N2N	Ours
λ = 30	Kodak	31.78/0.876	30.53/0.856	28.90/0.788	31.44/0.870	31.64/0.871	28.70/0.756	**31.65/0.874**
	BSD300	30.36/0.868	29.18/0.842	28.46/0.798	30.10/0.863	30.25/0.862	28.07/0.787	**30.43/0.869**
	Set14	30.57/0.858	29.44/0.837	27.73/0.774	30.29/0.853	30.46/0.852	27.72/0.758	**30.56/0.858**
$\lambda \in$ [5,50]	Kodak	31.19/0.861	29.40/0.836	28.78/0.758	30.86/0.855	31.07/0.857	28.10/0.725	**31.49/0.864**
	BSD300	29.79/0.848	28.22/0.815	27.92/0.766	29.54/0.843	29.92/0.852	27.68/0.765	**30.01/0.855**
	Set14	30.02/0.842	28.51/0.817	27.43/0.745	29.79/0.838	30.10/0.844	27.51/0.748	**30.34/0.852**

Table 2: Quantitative results on synthetic datasets in sRGB space for Poisson noise set. The highest PSNR(dB)/SSIM among unsupervised denoising methods are highlighted in **bold**.

on the original-scale images. The steps of the model inference process are illustrated in Fig. 3. We encode $\mathbf{x}$ using the PSE and feed the structural representation $\mathbf{c}_x$ into RG-Diff, which introduces random Gaussian noise during inference and performs reverse diffusion guided by $\mathbf{c}_x$. Concurrently, $\mathbf{x}$ undergoes processing through a feature manipulation branch where structural representations are fused.

The overall training process with the scale replay mechanism is illustrated in Fig. 5. To prevent identity mapping, we compute losses using the downsampled version of the denoised original-scale image, rather than directly supervised by the noisy original image.

$$\begin{aligned}\mathcal{L}_{sc} = &\| f_\theta(m_1(\mathbf{x})) - m_1(f_\theta(\mathbf{x})) - m_2(\mathbf{x}) + m_2(f_\theta(\mathbf{x})) \|_2 \\ &+ \| f_\theta(m_1(\mathbf{x})) - m_1(f_\theta(\mathbf{x})) - m_3(\mathbf{x}) + m_3(f_\theta(\mathbf{x})) \|_2 \end{aligned} \quad (10)$$

The expression for the final loss is as follows:

$$\mathcal{L} = \alpha_{rec}\mathcal{L}_{rec} + \alpha_{sc}\mathcal{L}_{sc} + \alpha_{diff}\mathcal{L}_{diff} \quad (11)$$

α_{rec}, α_{sc} and α_{diff} are adjustable hyperparameters. In our experiments, we set them to 1, 1.5, and 1, respectively.

Experiment

Implementation Details

Training Details. We select supervised method (Ronneberger, Fischer, and Brox 2015), CBM3D (Dabov et al. 2007a), BM3D (Dabov et al. 2007b), anscombe (Makitalo and Foi 2010), noise2void(N2V) (Krull, Buchholz, and Jug 2019), NBR2NBR (Huang et al. 2021), blind2unblind(B2U) (Wang et al. 2022), zero shot noise2noise(ZS-N2N) (Mansour and Heckel 2023) for writing. More comparative experimental results can be found in the supplementary material. We obtain quantitative and qualitative results from other methods by adopting official pre-trained models and running their public codes. For training, we fixed the decay rate for the exponential moving average at 0.999 and initialized the learning rate to 0.0002. Parameter optimization and computation were performed with Adam optimizer, setting β_1 to 0.9 and β_2 to 0.99. All training was executed on one Nvidia RTX3090.

Datasets. In synthetic denoising, we curated a training set comprising 44,328 images from the ILSVRC2012 dataset (Deng et al. 2009), with testsets named kodak, BSD300 (Martin et al. 2001), and set14 (Zeyde, Elad, and

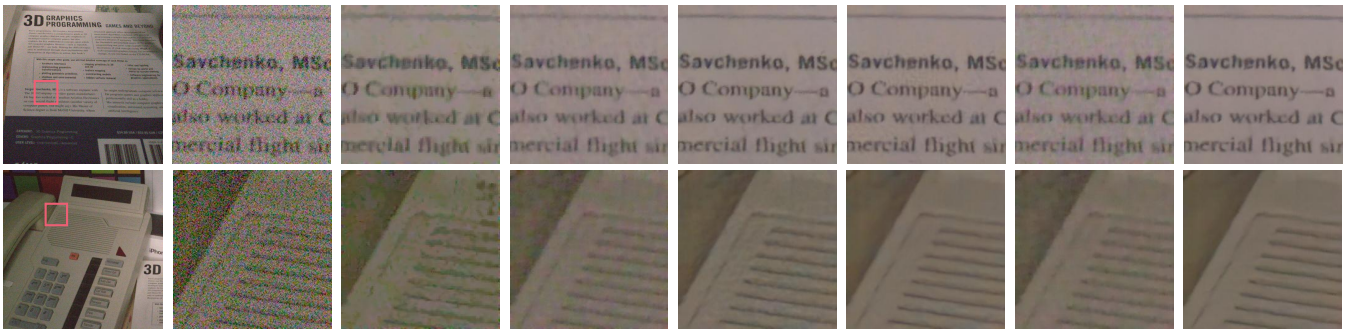


Figure 7: Visual comparison of our method against other methods on SIDD Benchmark. All images are converted from raw RGB space to sRGB space by the ISP provided by SIDD for visualization. Best viewed in color.

SIDD dataset	Baseline,N2C	CBM3D	N2V	NBR2NBR	B2U	ZS-N2N	DT	Ours
Benchmark	50.60/0.991	48.60/0.986	48.01/0.983	50.47/0.990	50.79/0.991	47.68/0.981	47.68/0.981	**51.02/0.991**
Validation	51.19/0.991	48.92/0.986	48.55/0.984	51.06/0.991	51.36/0.992	48.75/0.985	48.75/0.985	**51.55/0.992**

Table 3: Quantitative comparisons (PSNR(dB)/SSIM) on SIDD benchmark and validation datasets in raw-RGB space. The best PSNR(dB)/SSIM results among denoising methods are marked in **bold**.

Protter 2012). In real-world denoising, We utilized SIDD-Medium dataset (Abdelhamed, Lin, and Brown 2018) in the raw-RGB domain as the training set. For testing, we employed two datasets: SIDD validation and SIDD benchmark. Moreover, We employed the two-photon calcium imaging of a large 3D neuronal populations dataset, as proposed by SRDtrans (Li et al. 2023), for training and testing on the fluorescence imaging dataset. More implementation details can be found in the supplementary material.

Benchmarking Results

Synthetic Denoising. For sRGB images, we conducted four sets of experiments under Gaussian and Poisson noise settings. Our results are displayed in Fig. 6, Tab. 1 and 2. Overall, our method attained outstanding results, exhibiting superior performance across the majority of experimental metrics. Specifically, in all experimental trials, our approach surpassed SOTA method B2U (Wang et al. 2022) on all test datasets. Furthermore, our method exhibited a consistent 0.21-0.34dB improvement over another sampling method NBR2NBR (Huang et al. 2021) across various metrics, owing to our structural representation prompt strategy. Notably, Our approach demonstrates measurable enhancements over traditional supervised methods (Ronneberger, Fischer, and Brox 2015). Specifically, in the $\lambda \in [5, 50]$ experiment, we outperformed the baseline by 0.32dB on the Set14 dataset. Moreover, Across twelve experimental settings spanning three datasets, our method outperformed supervised approaches in eight instances.

Real-world Denoising. The quantitative results on real-world datasets are presented in Tab. 3. On the SIDD dataset in the raw-RGB domain, we achieved a 0.55 dB and 0.49 dB advantage on the SIDD validation and SIDD benchmark datasets respectively, relative to the original architecture of our method NBR2NBR (Huang et al. 2021). In comparison to the previous state-of-the-art method, B2U (Wang et al. 2022), we demonstrated improvements of 0.23 dB and 0.19 dB. This can be attributed to the more efficient attention mechanism of the transformer compared to traditional convolutions, and it also underscores the effectiveness of the integrated re-visualization pixel strategy from the network structure. Furthermore, we surpassed the Denoise Transformer(DT) (Zhang and Zhou 2023) method that utilizes a transformer as the backbone. This further validates the effectiveness of diffusion in generating prompts that fuse multi-scale information. By visualizing the results in Fig. 7, we observe that Prompt-SID outperforms in preserving details, minimizing edge blurring and color imbalance.

Sampling rate	Baseline,N2C	NBR2NBR	B2U	Ours
1 Hz	15.65	15.18	15.46	**15.89**
3 Hz	16.28	15.58	15.65	**15.98**
10 Hz	16.14	15.79	15.98	**16.06**
30 Hz	20.89	20.21	**21.12**	21.10

Table 4: The fluorescence imaging denoising experiment's quantitative results were assessed using SNR(db). The best results among denoising methods are marked in **bold**.

Fluorescence Imaging Denoising. Our results on fluorescence imaging denoising are outlined in Tab. 4. For comparative analysis, we selected baseline methods N2C (Ronneberger, Fischer, and Brox 2015), NBR2NBR (Huang et al. 2021), and B2U (Wang et al. 2022). Prompt-SID outperforms other self-supervised methods and achieves results comparable to supervised approaches. Notably, our results surpass the supervised baseline performance at both 1Hz and 30Hz scanning speeds. Upon visualizing the results in Fig. 8, we observed that our method exhibits strong generalization to fluorescence imaging data distribution and achieves

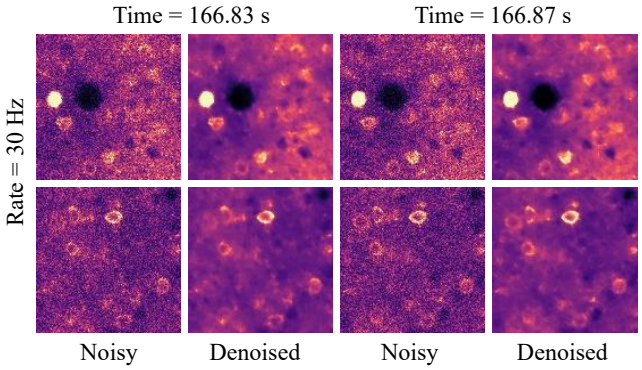


Figure 8: Visual results of fluorescence imaging datasets.

Dataset	SIDD	Kodak	Set14
w/o RG-Diff	51.32/0.991	31.97/0.874	31.06/0.862
w/o RG condition	51.40/0.991	32.25/0.881	31.32/0.867
w/o $\mathcal{L}_{sc}$	50.97/0.990	32.01/0.879	30.89/0.861
w/o $\mathcal{L}_{diff}$	51.03/0.990	32.32/0.883	31.38/0.868
ours	**51.55/0.992**	**32.41/0.883**	**31.45/0.868**

Table 5: Ablation studies on the effect of different modules in real-world and Gaussian σ = 25 datasets.

remarkable image restoration even with significant noise.

Ablation Study

The Ablation of Several Modules. We conducted the following module ablation experiments and tested them on the SIDD benchmark, as well as Kodak and Set14.

The settings for the four sets of experiments are as follows: 1) Ablation experiment on the Structural representation Prompt. We conducted an ablation on RD-Diff within Prompt-SID, removing the Structural representation Prompt, and simultaneously eliminating the fusion mechanism of SAM in the denoiser. 2)Within the RD-Diff branch, we omitted the mechanism that uses the structural representation of downsampled images as a conditioning factor for generation, substituting it with an equally shaped Gaussian noise. Consequently, the diffusion model branch transformed into a traditional unconstrained generative branch. 3)We removed the scale replay mechanism, thereby excluding its influence on model training in the loss function $\mathcal{L}_{sc}$. During training, the denoiser solely processes downsampled images and structural representation prompts. 4)We removed $\mathcal{L}_{diff}$, imposing no loss constraints on the generation of structural representations.

The experimental results are presented in Tab. 5 and Fig. 9. The full version of Prompt-SID exhibits performance improvements compared to the other ablation experiments. In sets 1 and 2, there is a degradation in image semantic details (such as the stacking of petals at the top of the rose). After removing the scale replay mechanism, the denoised images are blurrier than Prompt-SID, as the model did not encounter higher resolution information during training.

How Does the Prompt Work? To validate the effectiveness of the structural representation mechanism, we designed an

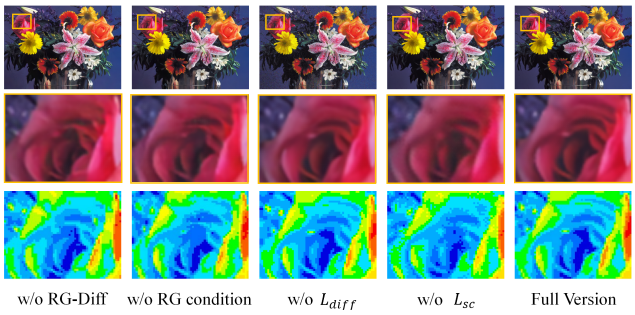


Figure 9: Visual results of ablation experiments.

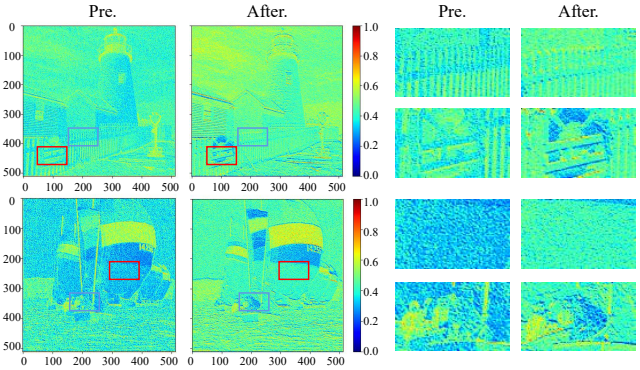


Figure 10: The visualization of the feature map, where pre. and after. represent before and after the prompt fusion.

ablation experiment to visualize the feature maps before and after prompt integration. We averaged the images along the channel dimension during the first SAM operation on the feature maps. The experimental results are shown in Fig. 10. The results demonstrate that the structural representation possesses the implicit ability to restore semantic structures and high-frequency edges. In feature map computation, prompt fusion emphasizes channels with richer structural details and semantic representations while attenuating the influence of noisy channels for high-frequency filtering.

Conclusion

We present Prompt-SID, a prompt-learning-based self-supervised image denoising framework that primarily addresses the semantic degradation and structural damage caused by the sampling processes of previous self-supervised methods. Our approach demonstrates the immense potential of the diffusion model and prompt-learning in image denoising tasks. We design a structural representation generation diffusion(RG-Diff) based on a latent diffusion model, using the degraded structural representations as conditional information to guide the recovery of undamaged ones. Additionally, through the scale replay mechanism, our method effectively reduces the scale gap between subsampled and original scale images. Extensive experiments demonstrate that our method consistently achieves state-of-the-art performance across synthetic, real-world, and fluorescence imaging datasets.

Acknowledgements

This work is supported by the Shenzhen Science and Technology Project under Grant (JCYJ20220818101001004).

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.

Anwar, S.; and Barnes, N. 2019. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3155–3164.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *European conference on computer vision*, 17–33. Springer.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007a. Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space. In *2007 IEEE international conference on image processing*, volume 1, I–313. IEEE.

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007b. 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.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35: 23593–23606.

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.

Laine, S.; Karras, T.; Lehtinen, J.; and Aila, T. 2019. High-quality self-supervised deep image denoising. *Advances in Neural Information Processing Systems*, 32.

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*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Li, X.; Hu, X.; Chen, X.; Fan, J.; Zhao, Z.; Wu, J.; Wang, H.; and Dai, Q. 2023. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nature Computational Science*, 3(12): 1067–1080.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Fei, B.; Dai, B.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*.

Liu, D.; Wen, B.; Jiao, J.; Liu, X.; Wang, Z.; and Huang, T. S. 2020. Connecting image denoising and high-level vision tasks via deep learning. *IEEE Transactions on Image Processing*, 29: 3695–3706.

Makitalo, M.; and Foi, A. 2010. Optimal inversion of the Anscombe transformation in low-count Poisson image denoising. *IEEE transactions on Image Processing*, 20(1): 99–109.

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.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, 416–423. IEEE.

Menteş, S.; Kınlı, F.; Özcan, B.; and Kıraç, F. 2021. [Re] Spatial-Adaptive Network for Single Image Denoising. In *ML Reproducibility Challenge 2020*.

Pang, T.; Zheng, H.; Quan, Y.; and Ji, H. 2021. Recorrupted-to-recorrupted: unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2043–2052.

Papkov, M.; and Chizhov, P. 2023. SwinIA: Self-Supervised Blind-Spot Image Denoising with Zero Convolutions. *arXiv preprint arXiv:2305.05651*.

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.

Shijila, B.; Tom, A. J.; and George, S. N. 2019. Simultaneous denoising and moving object detection using low rank approximation. *Future Generation Computer Systems*, 90: 198–210.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, W.; Yang, H.; Fu, J.; and Liu, J. 2024. Zero-Reference Low-Light Enhancement via Physical Quadruple Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26057–26066.

Wang, Z.; Fu, Y.; Liu, J.; and Zhang, Y. 2023. 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*, 18156–18165.

Wang, Z.; Liu, J.; Li, G.; and Han, H. 2022. Blind2unblind: Self-supervised image denoising with visible blind spots. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2027–2036.

Wu, X.; Liu, M.; Cao, Y.; Ren, D.; and Zuo, W. 2020. Unpaired learning of deep image denoising. In *European conference on computer vision*, 352–368. Springer.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13095–13105.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022a. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2022b. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 1934–1948.

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, D.; and Zhou, F. 2023. Self-supervised image denoising for real-world images with context-aware transformer. *IEEE Access*, 11: 14340–14349.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155.

Zhang, K.; Zuo, W.; and Zhang, L. 2018. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9): 4608–4622.

Zhang, Y.; Li, D.; Shi, X.; He, D.; Song, K.; Wang, X.; Qin, H.; and Li, H. 2023. Kbnet: Kernel basis network for image restoration. *arXiv preprint arXiv:2303.02881*.