Next-Scale Prediction: A Self-Supervised Approach for Real-World Image Denoising

Yiwen Shan[1] Haiyu Zhao[1] Peng Hu[1] Xi Peng[1,2] Yuanbiao Gou[1,†]
[1]College of Computer Science, Sichuan University, Chengdu, China
[2]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, Chengdu, China
{shan.yiwen.ml, haiyuzhao.gm, penghu.ml, pengx.gm, gouyuanbiao}@gmail.com

Abstract

Self-supervised real-world image denoising remains a fundamental challenge, arising from the antagonistic trade-off between decorrelating spatially structured noise and preserving high-frequency details. Existing blind-spot network (BSN) methods rely on pixel-shuffle downsampling (PD) to decorrelate noise, but aggressive downsampling fragments fine structures, while milder downsampling fails to remove correlated noise. To address this, we introduce Next-Scale Prediction (NSP), a novel self-supervised paradigm that decouples noise decorrelation from detail preservation. NSP constructs cross-scale training pairs, where BSN takes low-resolution, fully decorrelated sub-images as input to predict high-resolution targets that retain fine details. As a by-product, NSP naturally supports super-resolution of noisy images without retraining or modification. Extensive experiments demonstrate that NSP achieves state-of-the-art self-supervised denoising performance on real-world benchmarks, significantly alleviating the long-standing conflict between noise decorrelation and detail preservation. The code is available at https://github.com/XLearning-SCU/2026-CVPR-NSP.

1. Introduction

Self-supervised image denoising [15, 21, 22, 25, 30, 37] aims to estimate the underlying noise-free images from their corrupted observations, without relying on any ground-truth supervision. The ill-posed nature and the absence of clean images make this task challenging and of broad practical applicability.

A common paradigm for self-supervised image denoising is based on the blind-spot network (BSN) [2, 5, 9, 17–19], which predicts the clean value of a pixel from its surrounding pixels. This approach relies on the assumption that the noise is pixel-wise independent. If the noise exhibits spatial correlation, the network inevitably learns this correlation and predicts the noise in the target pixel from the nearby noisy pixels, which deviates from the denoising objective. Consequently, BSN-based methods are limited in handling real-world images, where noise introduced by Image Signal Processing [3] pipeline often exhibits strong spatial correlation. As confirmed by prior studies [20, 23, 28, 38], BSN-based methods generally fail to achieve satisfactory results in real-world image denoising.

To enable BSN to handle real-world noise, researchers use pixel-shuffle downsampling (PD) [44], which splits an image into many smaller sub-images. With a large downsampling factor, PD can effectively dismantle spatially correlated noise, turning it into nearly pixel-wise independent disturbances that BSN can cleanly remove. But that very aggressiveness comes at a steep price: it fragments the high-resolution structures, leaving the network to learn from tiny

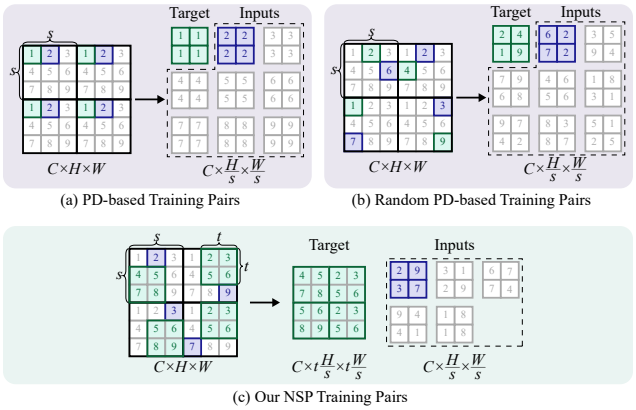


Figure 1. Illustration of different training-pair construction strategies for BSN. Inspired by visual autoregressive modeling, our NSP method constructs lower-scale inputs for noise decorrelation and high-scale targets for detail preservation.

† Corresponding author

patches and crippling its ability to recover fine details. In some cases, BSN even misinterprets high-resolution structures as noise and removes them unintentionally [20, 27]. Reduce the factor and detail returns, yet the noise regains spatial correlation that BSN cannot easily eliminate. In short, PD simultaneously undoes the correlation you must break and erases the resolution you must preserve. This antagonistic trade-off is subtle, fundamental, and central to making real-world denoising truly work.

This inherent contradiction motivates us to seek a more balanced strategy which decorrelates the noise more effectively while preserving the fine details. Ideally, such a strategy should retain the denoising advantages of PD while mitigating its destructive effect on high-resolution structures. Inspired by recent advances in Visual Autoregressive Modeling [35], we propose Next-Scale Prediction (NSP) as a principled solution to this dilemma. Rather than relying on a fixed PD factor, NSP performs denoising through a coarse-to-fine, next-scale prediction process. Specifically, BSN takes sub-images produced with a large PD factor as input, and predicts their higher-resolution counterparts associated with a smaller PD factor. This design enables NSP to effectively decorrelate the spatially-correlated noise at lower scales while preserving the fine details at higher scales. By decoupling noise decorrelation from detail preservation, NSP alleviates the above intrinsic conflict present in PD-based BSN methods, providing a robust paradigm for real-world image denoising.

The contributions are summarized as follows:

- **Next-Scale Prediction**: A new self-supervised paradigm for real-world image denoising that effectively addresses the conflict between noise decorrelation and detail preservation.
- **Data-Pair Construction**: A new strategy for constructing cross-scale/resolution training pairs that blocks access to noise-correlated pixels, coupled with a minimal BSN modification.
- As a by-product of next-scale prediction, our method provides a feasible way to super-resolve real-world noisy images, without retraining or modifying the model.

2. Related Works

In this section, we review related work on self-supervised image denoising and visual autoregressive modeling.

2.1. Self-Supervised Image Denoising

Different from traditional image denoising [10, 11, 14, 24, 34, 43], which relies on noisy-clean pairs for training, self-supervised image denoising seeks to recover clean images directly from noisy observations. It leverages intrinsic image statistics rather than requiring paired supervision.

A foundational work of self-supervised image denoising is Noise2Noise [21], which demonstrates that a denoiser can be trained using only corrupted image pairs. By replacing clean targets with another noisy observation of the same image, Noise2Noise achieves comparable performance to the models trained with clean targets. Furthermore, Noise2Void [18] and Noise2Self [2] proposed a blind-spot masking paradigm to generate pseudo noisy-noisy pairs from a single noise image. Specifically, these approaches mask a pixel and predict its clean value using surrounding noisy pixels. Since the masked pixels are excluded from the receptive field, they could be used as targets to supervise the training of the blind-spot network (BSN). However, this paradigm is inefficient since it trains the network using only masked pixels rather than the entire image at once. To address this, Laine et al. [19] and DBSN [39] modified the network architecture by masking feature maps within the convolutions, enabling all pixels in the input image to serve as supervision signals for the BSN, thereby significantly enhancing training efficiency.

However, in real-world scenarios, the noise often exhibits strong spatial correlations, which pose challenges for these methods that rely on local correlations for restoration. To address this, several studies have explored strategies to break the spatial dependencies in the noise, enabling BSN to perform effective denoising in real-world scenarios. For example, AP-BSN [20] introduces asymmetric pixel-shuffle downsampling for training and inference to break spatial noise correlation. SDAP [28] proposes random subsample generation strategy to disrupt noise dependencies during training. TBSN [23] leverages transformer-based attention with masked spatial and grouped channel mechanisms to maintain the blind-spot constraint while capturing long-range dependencies.

Although these methods effectively reduce spatial noise correlations to enable real-world denoising, they also disrupt the natural correlations between image pixels, which could compromise image details. In contrast, we propose a novel paradigm for self-supervised image denoising that preserves fine-grained details. Specifically, it constructs training pairs using a downsampled low-resolution image to disrupt noise correlations and a high-resolution target to allow the network to learn fine-grained structures. With this design, our paradigm achieves effective denoising while minimally compromising image fidelity.

2.2. Visual Autoregressive Modeling

Visual Autoregressive Modeling (VAR) [35], awarded the NeurIPS 2024 Best Paper, is a recently proposed and highly promising paradigm for image generation. Inspired by the human tendency to perceive and create images from coarse global structures to fine local details, VAR enables a transformer [36] to predict the next higher-scale token map conditioned on all token maps from previous scales. Compared to traditional next-token prediction, this next-scale predic-

tion (NSP) is more natural for images, which expand in two-dimensional space rather than one-dimensional sequence. By avoiding the flattening of 2D into 1D, VAR preserves spatial structure and addresses the fundamental limitation of image autoregressive models. In practice, VAR significantly outperforms previous autoregressive baselines and surpasses strong diffusion models on multiple metrics.

Motivated by these observations, we propose NSP to disentangle noise decorrelation from detail preservation. The BSN first operates on a downsampled scale where noise is effectively decorrelated for removal, and subsequently predicts its high-resolution counterpart to restore fine details. This coarse-to-fine hierarchy ensures that denoising and structural reconstruction are addressed independently at their optimal scales.

3. Proposed Method

In this section, we start with a brief review of the foundational denoising models, followed by a detailed explanation of our proposed method.

3.1. PD-based BSN Methods

BSN [18] is an effective architecture for self-supervised image denoising. By assuming zero-mean, pixel-wise independent noise, it predicts the clean value of each pixel from its surrounding noisy pixels. Formally, the training of BSN $f_\theta : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{C\times H\times W}$ can be formulated as

$$\min_\theta \mathcal{L}\big(f_\theta(\mathbf{I}) - \mathbf{I}\big), \tag{1}$$

where θ is the parameter set, $\mathbf{I}$ is the noisy image, and $\mathcal{L}$ is certain loss function.

A fundamental limitation of BSN is that it can only remove pixel-wise independent noise, making it less effective for real-world images where noise often exhibits strong spatial correlation. To address this, PD [23, 44] decomposes a real-world noisy image into multiple sub-images (Fig. 1(a)), effectively transforming spatially correlated noise into approximately pixel-wise independent noise. Building on this, later work [28] introduces randomness into the pixel-shuffle process. As shown in Fig. 1(b), a single pixel is randomly selected from each $s \times s$ patch to form a sub-image. This substantially increases the number of possible sub-images, which can be calculated as $N = \sum_{i=0}^{s^2-1}(s^2 - i)^{\frac{HW}{s^2}}$, far exceeding s^2 sub-images generated by standard PD.

Despite notable advances, a fundamental bottleneck remains. To completely decorrelate spatially structured noise in real-world images, PD must use a large downsampling factor s, so that formerly neighboring, noise-correlated pixels are redistributed into different sub-images. That redistribution indeed weakens noise correlations, but it simultaneously destroys the high-frequency cues that encode fine structure. BSN trained on these sub-images lacks the statistical signal needed to recover subtle edges and textures. Reducing s preserves those cues, yet it also preserves spatial noise correlation, leaving BSN with residual, signal-like noise that cannot distinguish from true detail.

In essence, PD-based BSN methods face an identifiability problem: the operations that decorrelate spatially structured noise inevitably remove the high-frequency cues that distinguish true image details from noise. This is not a mere tuning issue but an intrinsic trade-off: ***noise decorrelation and detail preservation act on the same spatial dependencies in opposite directions, so any solution must explicitly decouple the two objectives rather than hope to meet them simultaneously***.

3.2. Framework Overview

Motivated by above observation, we introduce Next-Scale Prediction (NSP), a self-supervised framework designed to explicitly decouple noise decorrelation and detail preservation. To be specific, NSP builds upon the PD-based BSN paradigm but reformulates denoising as a next-scale prediction problem. BSN first operates on sub-images produced with a large PD factor and then learns to predict their high-scale counterparts corresponding to a smaller PD factor. In this way, ***noise removal is performed at the lower scale, where the noise has been largely decorrelated, while detail recovery occurs at the higher scale, where the detail has been largely preserved.***

The overall framework of NSP is illustrated in Fig. 2. Given a real-world noisy image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$, NSP first constructs cross-scale training pairs using our Data-Pair Construction strategy (Sec. 3.3):

$$(\mathbf{I}_s, \mathbf{I}_t) \;=\; \mathcal{G}(\mathbf{I}; s, t), \tag{2}$$

where $\mathcal{G}$ denotes the proposed pair-construction operator. s is the PD factor to generate the sub-image $\mathbf{I}_s \in \mathbb{R}^{C\times \frac{H}{s}\times \frac{H}{s}}$, and t represents the relative scale of the target image $\mathbf{I}_t \in \mathbb{R}^{C\times \frac{H}{s}t\times \frac{H}{s}t}$ w.r.t. $\mathbf{I}_s$. In this formulation, $\mathbf{I}_s$ serves as the coarse, noise-decorrelated input, whereas $\mathbf{I}_t$ provides the high-scale reference that retains more spatial structure and detail.

With the cross-scale training pairs $(\mathbf{I}_s, \mathbf{I}_t)$, BSN $\mathcal{F}_\theta$ is trained to predict the high-scale target $\mathbf{I}_t$ from the coarse-scale input $\mathbf{I}_s$:

$$\hat{\mathbf{I}}_t = \mathcal{F}_\theta(\mathbf{I}_s), \tag{3}$$

where θ denotes the parameter set, which is optimized by minimizing the reconstruction loss between the prediction and target:

$$\mathcal{L}(\theta) = \big\|\hat{\mathbf{I}}_t - \mathbf{I}_t\big\|_1. \tag{4}$$

Since both $\mathbf{I}_s$ and $\mathbf{I}_t$ are derived from the same noisy image $\mathbf{I}$, this training process forms a self-supervised loop:

$$\mathbf{I} \xrightarrow{\mathcal{G}(\cdot;s,t)} (\mathbf{I}_s, \mathbf{I}_t) \xrightarrow{\mathcal{F}_\theta} (\hat{\mathbf{I}}_t, \mathbf{I}_t) \xrightarrow{\mathcal{L}(\theta)} \text{update } \theta. \tag{5}$$

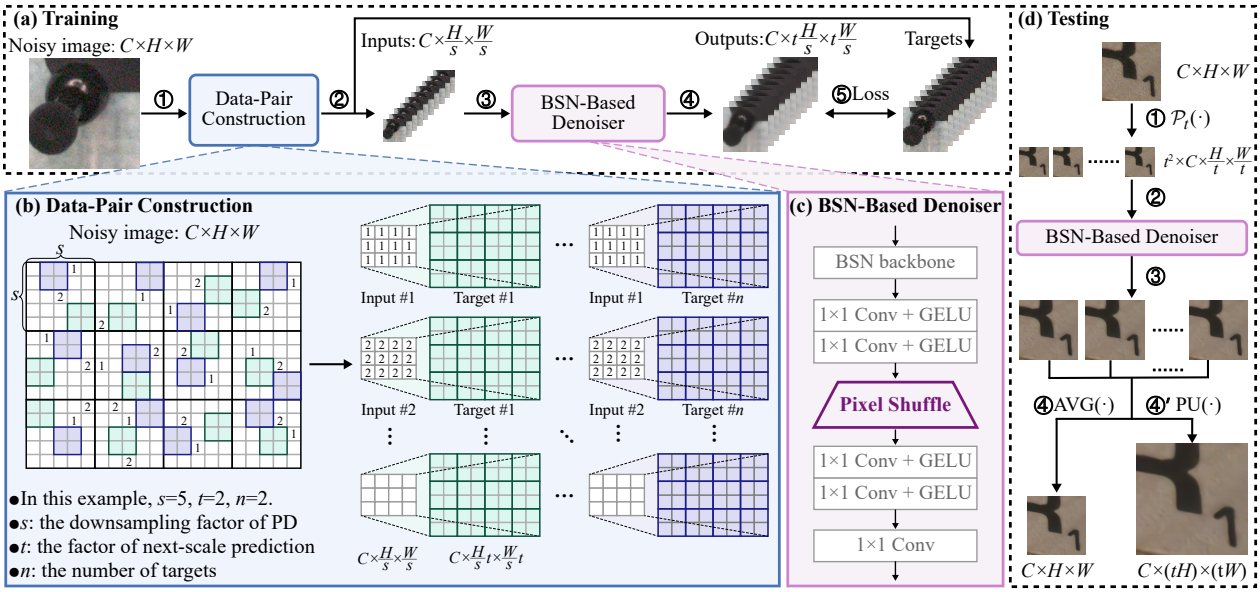


Figure 2. Framework of the proposed NSP. It is built on the PD-based BSN paradigm but reformulates denoising as a next-scale prediction task. The BSN first processes sub-images generated with a large PD factor and then learns to predict their higher-resolution counterparts corresponding to a smaller PD factor. In this way, NSP explicitly decouples the objectives of noise decorrelation and detail preservation.

Through this next-scale prediction formulation, BSN learns to perform noise removal at the coarse scale, where noise is largely decorrelated, while progressively refining spatial details guided by the high-scale targets.

Given a test noisy image $\mathbf{I}' \in \mathbb{R}^{C\times H\times W}$, let $\mathcal{P}_t$ denote PD operator that first decomposes $\mathbf{I}'$ into t^2 sub-images:

$$\{\mathbf{I}'^{(i)}\}_{i=1}^{t^2} = \mathcal{P}_t(\mathbf{I}'), \quad \mathbf{I}'^{(i)} \in \mathbb{R}^{C\times \frac{H}{t}\times \frac{W}{t}}. \tag{6}$$

Then, each sub-image is independently processed by the trained BSN $\mathcal{F}_\theta$:

$$\hat{\mathbf{I}}'^{(i)} = \mathcal{F}_\theta\big(\mathbf{I}'^{(i)}\big), \quad i = 1, \ldots, t^2, \tag{7}$$

where $\hat{\mathbf{I}}'^{(i)} \in \mathbb{R}^{C\times H\times W}$ denotes the denoised output corresponding to the i-th sub-image at the high scale. Finally, the denoised image is obtained by averaging these t^2 reconstructed results:

$$\hat{\mathbf{I}}' = \mathrm{AVG}(\{\hat{\mathbf{I}}'^{(i)}\}_{i=1}^{t^2}) \in \mathbb{R}^{C\times H\times W}. \tag{8}$$

As a by-product, our pipeline also yields a feasible way to produce a $t\times$ super-resolved image of the noisy input without retraining. Let PU denote the pixel-shuffle upsampling that reassembles t^2 denoised sub-images into a single image of size $tH \times tW$. The resulting super-resolved output is then given by

$$\hat{\mathbf{I}}^{\uparrow t} = \mathrm{PU}\big(\{\hat{\mathbf{I}}'^{(i)}\}_{i=1}^{t^2}\big) \in \mathbb{R}^{C\times (tH)\times (tW)}. \tag{9}$$

3.3. Data-Pair Construction

To construct effective cross-scale training pairs, we adhere to three key principles:

- **Blocking noise-correlated pixels across scales** to guarantee BSN learns true noise removal rather than exploiting residual noise correlations.
- **Maintaining structural consistency** to preserve the pixels spatial arrangement in the high-scale targets, ensuring better reconstruction of details.
- **Leveraging random sampling** to generate a wide variety of training pairs, covering diverse noise realizations and structural patterns.

Following these principles, our cross-scale pair construction process is illustrated in Fig. 2(b).

Given a real-world noisy image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$, we first divide it into $s \times s$ non-overlapping patches:

$$\{\mathbf{P}_{i,j}\} = \mathrm{Patch}(\mathbf{I}; s), \quad i/j = 1, \ldots, \frac{H}{s}/\frac{W}{s}, \tag{10}$$

where s denotes both the patch size and the PD factor from the original image to the sub-images.

Next, for each patch $\mathbf{P}_{i,j} \in \mathbb{R}^{C\times s\times s}$, we perform random sampling to select $t \times t$ pixels that construct the high-scale target:

$$\mathbf{L}_{i,j} = \mathrm{Sample}(\mathbf{P}_{i,j}; t) \in \mathbb{R}^{C\times t\times t}, \tag{11}$$

where $t \in (1, s)$ is the scaling factor of the cross-scale pair.

The remaining $(s^2 - t^2)$ pixels in $\mathbf{P}_{i,j}$ are then distributed

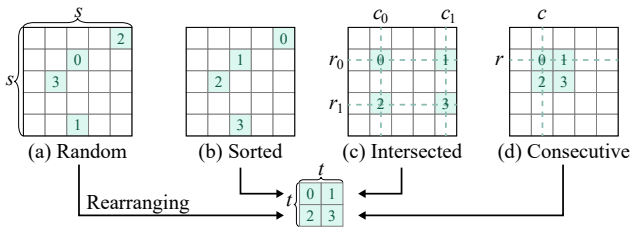


Figure 3. Alternative strategies for high-scale target construction. (a) Randomly select t^2 pixels. (b) Randomly select t^2 pixels and sort in a row-major order. (c) Select the pixels at the intersections of t random rows and columns. (d) Select a consecutive $t \times t$ patch. The comparison between those strategies can be found in Tab. 3.

across $(s^2 - t^2)$ sub-images:

$$\{\mathbf{I}_{i,j}^{(k)}\}_{k=1}^{s^2-t^2} = \text{Distribute}(\mathbf{P}_{i,j} \setminus \mathbf{L}_{i,j}), \tag{12}$$

where each $\mathbf{I}_{i,j}^{(k)} \in \mathbb{R}^{C\times 1\times 1}$ corresponds to a pixel position.

Finally, each sub-image is paired with the same high-scale target, producing the final cross-scale training set:

$$\{(\mathbf{I}_{i,j}^{(k)}, \mathbf{L}_{i,j})\}_{k=1}^{s^2-t^2}. \tag{13}$$

For the Eq. (11), we design several alternative sampling strategies, as illustrated in Fig. 3, and empirically find that those preserving the relative spatial arrangement of pixels within each patch perform the best. Consequently, we employ the consecutive patch strategy as our default choice, which can produce $(s - t + 1)^2$ possible targets.

For the Eq. (12), this one-to-one random distribution assigns noise-correlated pixels to different sub-images, which can generate $\sum_{i=0}^{s^2-t^2-1}(s^2 - t^2 - i)^{\frac{HW}{s^2}}$ sub-images. Consequently, BSN takes these decorrelated sub-images as inputs to predict high-scale targets composed of distinct pixels, thereby effectively blocking noise correlations across scales.

For the Eq. (13), while the formulation constructs a single high-scale target, the strategy naturally generalizes to the multi-target case. Multiple targets are generated by repeatedly sampling $t \times t$ pixels not occupied by previous targets, while the remaining pixels are assigned to sub-images. Let the number of targets be $n \in [1, \lfloor s^2/t^2 \rfloor]$, so that the number of inputs is $(s^2 - n \cdot t^2)$. As a result, the training pairs are formed as the Cartesian product of the inputs and targets, yielding a total of $n \cdot (s^2 - n \cdot t^2)$ pairs per image.

3.4. BSN-Based Denoiser

Within the NSP framework, a BSN is required to predict the next higher-scale sub-images rather than operating within the same scale as in its original design. To achieve this, a pixel shuffle layer is integrated into BSN to facilitate scale transformation. To keep the blind-spot property, the layer is inserted between the 1×1 convolutional blocks near the network tail, as shown in Fig. 2(c).

4. Experiments

In this section, we conduct experiments to evaluate the proposed method. In the following, we first introduce the experimental settings, followed by quantitative and qualitative results on image denoising and super-resolution tasks. Finally, we perform analysis experiments to validate the effectiveness of our key designs.

4.1. Experimental Settings

To comprehensively evaluate the proposed NSP framework, we adopt two representative BSNs as backbones, namely DBSN and TBSN, which are CNN-based and Transformer-based architectures, respectively. In experiments, we denote these two variants as NSP(DBSN) and NSP(TBSN). The PD factor s, which determines the downsampling from the original image to the sub-images, is empirically set to 5, while the scaling factor t from the sub-images to the high-scale targets is set to 2. Following prior works [20, 23, 28], we set the patch size to 160×160 and the batch size to 16. Training is conducted for 750 epochs, with each epoch consisting of 400 iterations. The learning rate is fixed at 1e-4, and the optimizer is Adam with default parameters. All experiments are implemented in PyTorch on NVIDIA GeForce RTX 3090 and A800 GPUs.

4.2. Experimental Results

The denoising methods based on the proposed paradigm are trained on the noisy subset of SIDD Medium, which consists of 320 high-resolution real-world noisy images. For fairness, all comparison methods are evaluated on SIDD Validation, SIDD Benchmark and DND dataset. Specifically, both SIDD Validation and SIDD Benchmark have 1280 noisy images of size 256×256, while DND contains 1000 noisy images of size 512×512.

The quantitative results of all comparison methods are presented in Table 1. As can be seen, NSP(DBSN) and NSP(TBSN) almost achieve the best performance among all self-supervised methods, except for the PSNR on DND dataset being slightly lower than TBSN by 0.03dB. Specifically, both NSP(DBSN) and SDAP use DBSN as the backbone with similar parameter amounts (3.75M and 3.66M). However, NSP(DBSN) outperforms SDAP by 0.44dB/0.0235 on SIDD Validation, according to the PSNR/SSIM metric. A similar result can be found between NSP(TBSN) and TBSN, both of which use the TBSN as the backbone. Such improvement achieved by the NSP paradigm is attributed to the better prediction of the high-scale details, as shown in the highlighted windows of Fig. 4(h)∼(j). Moreover, the performance of NSP(DBSN) and NSP(TBSN) are competitive with that of multiple supervised and pseudo-supervised methods. Among them, NSP(DBSN) outperforms DnCNN, TNRD, CBNDet, PDD, GCBD and C2N with at least 1.66dB margin in PSNR.

Table 1. Quantitative comparisons on real-world noisy image datasets. †denotes results from the previous SIDD evaluation server, which is no longer accessible.

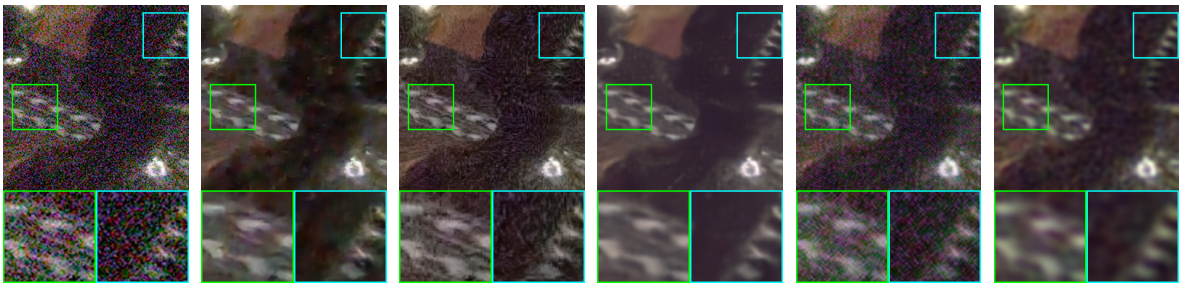
Type	Train Data	Method	#Param	SIDD Validation PSNR/SSIM	SIDD Benchmark PSNR/SSIM	DND PSNR/SSIM
Non-Learning	None	BM3D [7]	-	31.75/0.7061	34.26/0.6950†	34.51/0.8507
		WNNM [12]	-	26.31/0.5240	30.52/0.4498†	34.67/0.8646
Supervised	Paired Noisy-Clean	DnCNN [42]	0.66M	26.20/0.4414	30.31/0.4371†	32.43/0.7900
		TNRD [6]	-	26.99/0.7440	-	33.65/0.8306
		CBDNet [13]	6.79M	33.07/0.8655	34.51/0.8402	38.00/0.9400
		PDD [44]	0.71M	33.96/0.8195	35.22/0.8221	38.40/0.9434
		RIDNet [1]	1.50M	38.71/0.9511	-	39.25/0.9528
		VDNet [41]	7.82M	39.29/0.9109	39.49/0.9117†	39.38/0.9518
		DeamNet [31]	2.23M	39.40/0.9169	39.58/0.9118†	39.63/0.9531
Pseudo-Supervised	Unpaired Noisy-Clean	GCBD [4]	-	-	-	35.58/0.9217
		D-BSN [39]	6.62M	-	-	37.93/0.9373
		C2N [16]	217.26M	35.36/0.8901	36.06/0.8825†	37.28/0.9237
	Paired Noisy-Noisy	R2R [29]	0.56M	35.04/0.8440	35.50/0.8550†	37.61/0.9368
Self -Supervised	Only Noisy	N2V [18]	2.58M	29.07/0.5915	31.77/0.5979	33.37/0.8412
		N2S [2]	2.58M	30.72/0.7870	32.57/0.6768	33.63/0.8564
		NEI2NEI [15]	1.26M	28.00/0.5890	-	31.40/0.7880
		CVF-SID [26]	1.19M	34.14/0.8550	35.03/0.8561	36.50/0.9233
		AP-BSN [20]	3.66M	34.46/0.8296	36.09/0.8310	37.46/0.9244
		SDAP [28]	3.66M	36.58/0.8630	36.80/0.8529	37.71/0.9278
		TBSN [23]	12.74M	36.59/0.8574	37.08/0.8519	**37.90**/0.9288
		NSP(DBSN)	3.75M	**37.02/0.8865**	**37.42/0.8748**	37.80/**0.9319**
		NSP(TBSN)	12.77M	**37.12/0.8853**	**37.46/0.8744**	37.87/**0.9342**

The qualitative results are shown in Fig. 4. In Fig. 4(e), Noise2Void (N2V) leaves too much noise in the denoised result. This verifies the BSN cannot remove the spatially-correlated noise effectively, since it would learn the spatial correlation inadvertently and predict the noise in each pixel. In Fig. 4(g)∼(i), the "PD+BSN"-based counterparts, *i.e.*, AP-BSN, SDAP, and TBSN, generate chessboard artifacts, which is due to no module can help the BSN learn to predict the higher-scale details from the lower-scale sub-images output by PD. By contrast, the proposed NSP paradigm compensates this by supervising the BSN to predict the higher-scale counterparts where more fine details are preserved. Hence, as shown in Fig. 4(j)∼(k), more details are recovered by NSP(DBSN) and NSP(TBSN).

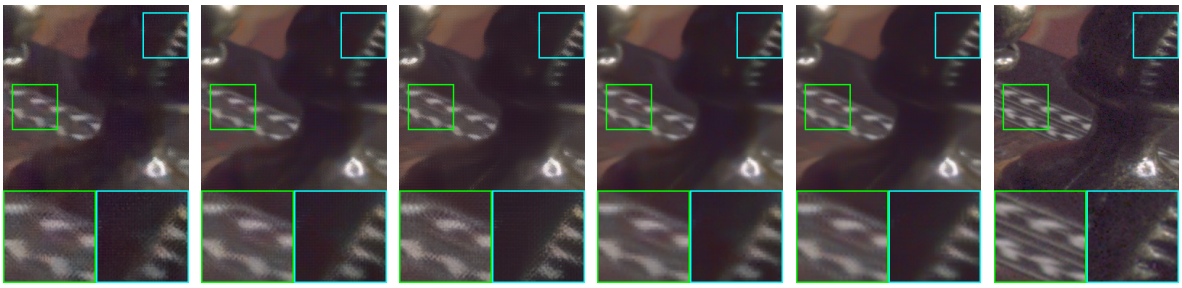
An interesting by-product of the proposed paradigm is noisy image super-resolution (SR), since the BSN predicts a higher-resolution version of the input. Different from most existing SR methods which focus solely on SR, the proposed paradigm performs the image denoising and SR simultaneously. To conduct the noisy image SR experiment, we first downsample the noisy subset of the SIDD Medium and SIDD Validation datasets using a factor of 2 and bicubic interpolation. Then, the proposed NSP(DBSN) is trained and tested on the corresponding downsampled datasets. All the settings are kept the same as those in the denoising experiment, except that the patch size is changed to 180×180. The high-resolution results are obtained by the procedure along the right branch in Fig. 2(d).

For a fair comparison, all competing methods are trained on the downsampled noisy subset of SIDD Medium. To provide denoising capability for the SR methods, we first train the standard SDAP and then use its denoised outputs as the training inputs for the SR methods. Four zero-shot SR methods [8, 32, 33, 40] are selected and re-trained on the dataset to create comparable dataset-trained versions for evaluation.

The quantitative results are listed in Table 2. In the first group, the zero-shot SR methods do not obtain satisfactory results, since they have no capacity of denoising and can only output the noisy high-resolution images, as shown in Fig. 5(b). In the following two groups, the two-stage "SDAP+SR" methods are able to output the denoised high-resolution results. Hence their PSNR and SSIM are significantly higher than those in the first group. In spite of



(a) Noisy (b) CBDNet (c) PDD (d) RIDNet (e) N2V (f) CVF-SID



(g) AP-BSN (h) SDAP (i) TBSN (j) NSP(DBSN) (k) NSP(TBSN) (l) GT

Figure 4. Qualitative comparisons of image denoising on SIDD Validation.

this, the proposed NSP(DBSN) still achieves the highest results with a competitive amount of parameters. The qualitative results are shown in Fig. 5. Due to the difficulty of noisy image SR problem, the comparison methods generate the artifacts to some extent. In spite of this, the proposed NSP(DBSN) still recovers the edges more clearly and generates fewer artifacts than the two-stage methods.

4.3. Analysis Experiments

In this section, we test the influence of three important settings, *i.e.*, the strategy of constructing the targets, the number of targets constructed from single image, and the upsampling factor from the sub-images to the targets.

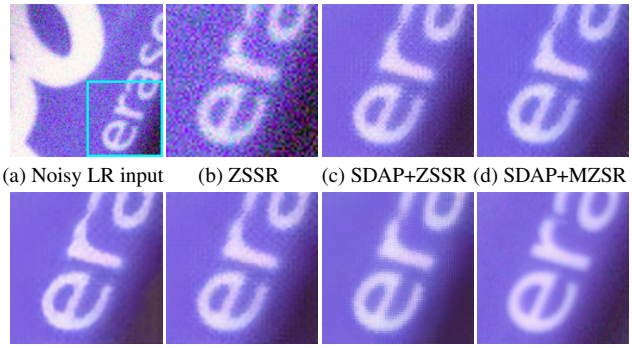
Target Construction Strategies. To construct a high-scale target, t^2 pixels are first selected from each $s \times s$ patch and then rearranged into a $t \times t$ patch, as illustrated in Fig. 2(b). During this rearrangement, it is important to preserve the relative positions of the selected pixels, as these positions encode structural information of the image. The degree of structural preservation depends on the selection strategy: purely random selection, such as in Fig. 3(a), often fails to maintain the original relative positions. Therefore, the choice of pixel selection strategy is critical. In this work, we design four alternative strategies shown in Fig. 3, where the purely random strategy serves as a baseline to evaluate the effectiveness of the other structured approaches.

Table 3 compares the performance of different target construction strategies. As shown in the first column, the purely random selection performs the worst, since the relative positions of the selected pixels are mostly lost. The second strategy, which sorts the randomly selected pixels in row-major order (Fig. 3(b)), improves performance by partially recovering the relative positions between some pixels.

Table 2. Quantitative results on noisy image super-resolution over the SIDD Validation dataset. "SDAP+" denotes a pipeline that first applies SDAP for denoising and then performs super-resolution. "U-" denotes the variant trained on the dataset, as opposed to zero-shot training.

Method	PSNR/SSIM	#Param
ZSSR [32]	25.38/0.4183	0.22M
RZSR [40]	25.00/0.3982	3.04M
MZSR [33]	25.93/0.4426	0.46M
dualSR [8]	24.98/0.4473	0.41M
SDAP+ZSSR	34.39/0.8165	3.88M
SDAP+RZSR	34.43/0.8189	6.70M
SDAP+MZSR	34.17/0.8284	4.12M
SDAP+dualSR	34.67/0.8380	4.06M
SDAP+U-ZSSR	35.10/0.8373	3.88M
SDAP+U-RZSR[1]	-	6.70M
SDAP+U-MZSR[1]	-	4.12M
SDAP+U-dualSR[2]	23.05/0.6123	4.06M
NSP(DBSN)	35.19/0.8490	3.75M
NSP(DBSN)[3]	**35.53/0.8771**	

[1] The design of the method precludes it from being adapted from a zero-shot setting to a standard training regime.
[2] Despite extensive attempts, training on a dataset consistently fails.
[3] At test time, directly input noisy LR image and output SR version without PD/PU.



(a) Noisy LR input (b) ZSSR (c) SDAP+ZSSR (d) SDAP+MZSR

(e) SDAP+dualSR (f) SDAP+U-ZSSR (g) NSP(DBSN) (h) GT

Figure 5. Qualitative comparison of image denoising and super-resolution on SIDD Validation.

Table 3. The comparison of different strategies of target construction and different number of targets n. The four strategies in Fig. 3 are denoted as "Random", "Sorted", "Intersected" and "Consecutive", respectively.

n	Pairs	Random PSNR/SSIM	Sorted PSNR/SSIM	Intersected PSNR/SSIM	Consecutive PSNR/SSIM
1	21	36.77/0.8835	36.91/0.8835	37.01/0.8854	**37.02/0.8865**
2	34	36.87/0.8849	36.97/0.8842	37.02/0.8864	**37.11/0.8884**
3	39	36.83/0.8837	37.05/0.8848	37.07/0.8855	**37.08/0.8861**
4	36	36.89/0.8841	37.01/0.8848	**37.25/0.8865**	37.04/0.8855
5	25	36.88/0.8841	36.88/0.8827	-	-

For instance, the relative order of pixels 1 and 2 is corrected compared to Fig. 3(a). However, this strategy still fails to preserve all relative positions; for example, the horizontal order between pixel 0 and 3 remains reversed.

To address this limitation, two additional strategies are proposed, as illustrated in Fig. 3(c) and (d). The third strategy randomly selects t rows and t columns and chooses the pixels at their intersections, while the fourth strategy selects a consecutive $t \times t$ patch. Both strategies perfectly preserve the relative positions of the selected pixels, leading to superior performance compared with the sorting-based strategy, as shown in the last two columns of Table 3. Overall, these results highlight the critical importance of maintaining relative pixel positions in constructing high-scale targets.

Number of Targets. The number of targets n serves as a hyper-parameter that controls how many training pairs can be constructed from a single noisy image. As discussed in Section 3.3, the total number of pairs is given by a quadratic function of n: $-t^2 \cdot n^2 + s^2 \cdot n$. While increasing n can provide more training pairs and potentially improve the performance, it also incurs higher GPU memory consumption.

Given $s = 5$ and $t = 2$, the impact of the number of targets n on denoising performance is summarized in the first few rows of Table 3. When n increases from 1 to 2, the total number of training pairs grows from 21 to 34, leading to consistent performance improvements across all target construction strategies. Moreover, the highest results for each strategy are obtained when the number of pairs exceeds 30, confirming that a larger number of pairs benefits the training of the proposed paradigm. Although the best performance is achieved with $n > 1$, we report the results for $n = 1$ in Table 1 to ensure compatibility with widely available GPUs having 24GB of memory.

Table 4. The comparison of different upsampling factor t. The four strategies in Fig. 3 are denoted as "Random", "Sorted", "Intersected" and "Consecutive", respectively.

t	Random PSNR/SSIM	Sorted PSNR/SSIM	Intersected PSNR/SSIM	Consecutive PSNR/SSIM
1	34.48/0.8415	3415/0.8440	33.97/0.8353	34.12/0.8346
2	**36.77/0.8835**	**36.91/0.8835**	**37.01/0.8854**	**37.02/0.8865**
4	35.87/0.8615	36.29/0.8682	36.56/0.8716	36.46/0.8694

Influence of Scaling Factor. The factor t is a hyper-parameter controlling the prediction scale. When t is too large, the mapping from low-resolution inputs to high-resolution targets becomes overly complex, as the targets are t times larger than the inputs, making it harder for the denoiser to learn and leading to degraded performance. Conversely, if t is too small, e.g., $t = 1$, the paradigm degenerates to the conventional PD-based BSN method, which may fail to accurately recover fine details. Furthermore, at test time, t determines the input size of the denoiser (Fig. 2(d)). For $t = 1$, the input is not downsampled, so the spatial correlation of noise remains, preventing the BSN from effectively removing it.

Those observations are confirmed by the results in Table 4. The NSP paradigm achieves the best performance when $t = 2$, while the performance degrades noticeably for both smaller and larger values of t. And, this trend holds consistently across different target construction strategies.

5. Conclusion

In this paper, we propose Next-Scale Prediction (NSP), a self-supervised paradigm for real-world image denoising. NSP enables a BSN to denoise low-resolution sub-images, where noise correlations are largely broken, while simultaneously predicting their high-resolution counterparts to preserve the fine details. This coarse-to-fine approach explicitly decouples the objectives of noise decorrelation and detail preservation, resolving the intrinsic conflict inherent in conventional PD-based BSN methods. To facilitate training, we introduce a data-pair construction strategy that generates a diverse set of cross-scale pairs from a single noisy image. Extensive experiments on real-world benchmarks validate the effectiveness of NSP, and comprehensive analyses investigate the influence of key hyper-parameters and design choices.

Acknowledgments

This work was supported in part by the Postdoctoral Fellowship Program (Grade C) of China Postdoctoral Science Foundation under Grant GZC20251052; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0038; in part by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM610.

References

[1] 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. 6

[2] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International conference on machine learning*, pages 524–533, 2019. 1, 2, 6

[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019. 1

[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 conference on computer vision and pattern recognition*, pages 3155–3164, 2018. 6

[5] Shiyan Chen, Jiyuan Zhang, Zhaofei Yu, and Tiejun Huang. 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*, pages 2814–2823, 2024. 1

[6] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016. 6

[7] 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

[8] Mohammad Emad, Maurice Peemen, and Henk Corporaal. Dualsr: Zero-shot dual learning for real-world super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1630–1639, 2021. 6, 7

[9] Lianru Gao, Degang Wang, Lina Zhuang, Xu Sun, Min Huang, and Antonio Plaza. Bs 3 lnet: A new blind-spot self-supervised learning network for hyperspectral anomaly detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–18, 2023. 1

[10] Yuanbiao Gou, Boyun Li, Zitao Liu, Songfan Yang, and Xi Peng. Clearer: Multi-scale neural architecture search for image restoration. *Advances in neural information processing systems*, 33:17129–17140, 2020. 2

[11] Yuanbiao Gou, Peng Hu, Jiancheng Lv, Joey Tianyi Zhou, and Xi Peng. Multi-scale adaptive network for single image denoising. *Advances in Neural Information Processing Systems*, 35:14099–14112, 2022. 2

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

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

[14] Jingwen He, Chao Dong, and Yu Qiao. Interactive multi-dimension modulation with dynamic controllable residual learning for image restoration. In *European conference on computer vision*, pages 53–68. Springer, 2020. 2

[15] 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, 6

[16] 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. 6

[17] Yeong Il Jang, Keuntek Lee, Gu Yong Park, Seyun Kim, and Nam Ik Cho. Self-supervised image denoising with downsampled invariance loss and conditional blind-spot network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12196–12205, 2023. 1

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

[19] Samuli Laine, Tero Karras, Jaakko Lehtinen, and Timo Aila. High-quality self-supervised deep image denoising. *Advances in neural information processing systems*, 32, 2019. 1, 2

[20] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022. 1, 2, 5, 6

[21] 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 35th International Conference on Machine Learning*, pages 2965–2974. PMLR, 2018. 1, 2

[22] 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. 1

[23] Junyi Li, Zhilu Zhang, and Wangmeng Zuo. Rethinking transformer-based blind-spot network for self-supervised image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4788–4796, 2025. 1, 2, 3, 5, 6

[24] Miaoyu Li, Ying Fu, and Yulun Zhang. Spatial-spectral transformer for hyperspectral image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1368–1376, 2023. 2

[25] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2noise: Learning to denoise from unpaired noisy data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12064–12072, 2020. 1

[26] 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. 6

[27] Alan V Oppenheim, Alan S Willsky, and Syed Hamid Nawab. *Signals & systems*. Pearson Educación, 1997. 2

[28] Yizhong Pan, Xiao Liu, Xiangyu Liao, Yuanzhouhan Cao, and Chao Ren. Random sub-samples generation for self-supervised real image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12150–12159, 2023. 1, 2, 3, 5, 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. 6

[30] 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

[31] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8596–8606, 2021. 6

[32] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 6, 7

[33] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3516–3525, 2020. 6, 7

[34] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: An overview. *Neural Networks*, 131:251–275, 2020. 2

[35] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[37] 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

[38] 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. 1

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

[40] Jun-Sang Yoo, Dong-Wook Kim, Yucheng Lu, and Seung-Won Jung. Rzsr: Reference-based zero-shot super-resolution with depth guided self-exemplars. *IEEE Transactions on Multimedia*, 25:5972–5983, 2022. 6, 7

[41] 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. 6

[42] 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. 6

[43] 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. 2

[44] 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, 3, 6