Self-Supervised Real-World Denoising by Jointly Learning Visible and Invisible Noise

Shaoyu Wang[†]
wangshapyu@dlmu.edu.cn
Changze Zhou[†]
zcz@dlmu.edu.cn
Bolin Song*
bolin_song@dmu.edu.cn
Yiyang Wang
yywerica@dlmu.edu.cn

College of Artificial Intelligence
Dalian Maritime University
Dalian, China

Abstract

Recently, there has been a rise in the development of self-supervised methodologies that facilitate denoising directly on real-world images without relying on clean image references. Existing self-supervised methods primarily focus on developing techniques for breaking the spatial correlation inherent in real-world noise. However, the inconsistent visibility observed in real-world noisy images, which deviates from that encountered in synthetic noisy images, has yet to be taken into account. In this paper, we propose a new perspective for self-supervised denoising on real-world noisy images, separately and jointly learning both visible and invisible noise using a single blind-spot network. To achieve this objective, a noise visibility map is estimated without relying on any ground truth or reference for the noise level, to direct the network towards focusing on the regions that exhibit similar visual performance using different strategies. Extensive experiments have been conducted to validate the superiority of our method over existing self-supervised denoisers from both quantitative and visual comparisons.

1 Introduction

Image denoising, a crucial task in the field of low-level image processing, aims to recover valuable image structures from noisy observations during the process of noise reduction, resulting in high-quality clear images for downstream tasks such as classification, semantic segmentation, and target identification [25, 41, 42].

With the rapid development of deep learning techniques, learning-based methods have made remarkable progress in comparison to non-learning approaches [5, 11, 13, 36] used in earlier periods. Learning-based approaches initially adopt a supervised manner that require clean-noisy pairs for training [27, 28, 37, 48, 49, 50]. However, this methodology is suitable

[†]The first two authors contributed equally.

* Corresponding author

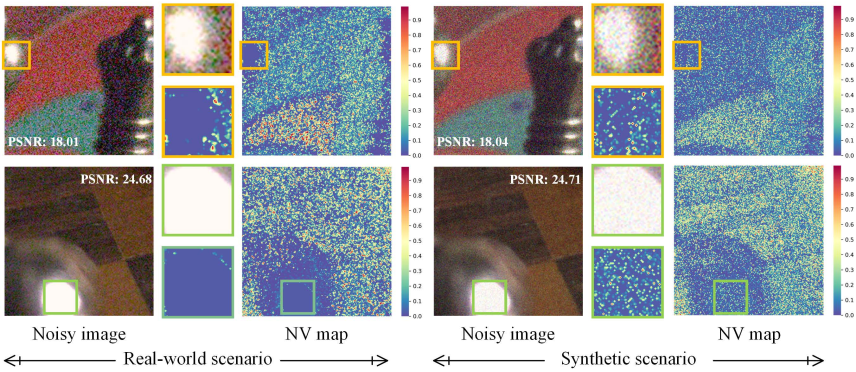


Figure 1: The visibility of noise in the lighter regions of an individual image is reduced in real-world scenarios, whereas this peculiarity does not hold true for synthetic noisy images (generated by adding AWGN to clean images from the SIDD validation dataset).

for synthetic noise removal but falls short in real-world denoising scenarios, due to the necessity of gathering clean-noisy image pairs from strictly controlled research environments, which can be time-consuming and labor-intensive.

To mitigate the need for extensive aligned datasets, various approaches have been developed by leveraging unpaired noisy-clean images to train the denoisers in an unsupervised manner [6, 7, 16, 19, 43]. However, the performance of those methods is still limited due to the intricate nature of modeling real-world noise distributions.

Recently, self-supervised approaches have emerged that enable denoising directly on noisy images without the need for clean image references. The most representative methods rely on blind-spot networks (BSNs), which utilize intricately designed masking schemes enabling the neural network to predict the masked pixels from neighboring noisy ones [17, 29, 32, 44]. However, unlike synthetic denoising, real-world noise removal using BSN typically requires the implementation of techniques aimed at breaking the spatial correlation inherent in real-world noise. The AP-BSN [23], one of the pioneering work that adopts a pixel-shuffle down-sampling (PD) scheme to partially break spatial noise correlation, forms the fundamental basis for diverse subsequent approaches [15, 20, 31, 40]. However, the disparity between real-world and synthetic noisy images extends beyond the spatial noise correlation, which have not been widely concerned.

Synthetic noisy images are generated by overlaying the clean images with randomly sampled signals from a predetermined probability distribution, followed by normalizing each pixel value within the standard range. Therefore, the noise is clearly visible across all regions within an individual image, displaying consistent characteristics. However, for real-world images, it may vary within an individual image; e.g., noise is often less visible in the lightest regions [10, 35], where a tonal response curve superimposed on saturation reduces contrast and noisy signals are clipped during the ADC and quantization processes, hence noise [18], as shown in Fig. 1. Besides, such inconsistency can also be observed from the noise visibility (NV) map estimated by our approach. Herein-after, all NV maps are visualized with certain normalization and by calculating the average across three channels.

Therefore, although the skill such as PD is able to partly disrupt the spatial correlation of real-world noise, the visual variability of noise within a real-world noisy image persists even when PD is employed. However, existing approaches employ the strategy of treating these spatially varying noise as equal, which may result in interference during network learning. Existing networks often lack the ability to effectively eliminate such visually changing noise during noise learning, and instead tend to learn to remove them as a compromise; e.g., they tend to eliminate the larger but less visible noise when the more visible noise occupies a small percentage, while leaving behind partly visible noise and creating fragmented artifacts.

In this research, we propose a novel perspective for self-supervised denoising employing BSN with single noisy images, taking into account the spatially inconsistent visibility of noise in real-world scenarios. Concretely, to identify the regions affected by either visible or invisible noise, we develop a multi-path module to estimate a noise visibility (NV) map without depending on any ground truth or reference for the noise level. With the assistance of NV map, we adopt different strategies on areas affected by visible or invisible noise to direct the BSN denoiser toward focusing on regions that exhibit similar visual performance. In addition, rather than training multiple networks, both visible and less visible noise are jointly learnt using a single BSN network, by attaching distinct significance to each of them. Finally, extensive experiments have been conducted to evaluate the proposed framework, which validate that our approach outperforms current self-supervised denoisers in terms of both quantitative metrics and visual comparisons. Our main contributions are outlined.

1. We propose a novel perspective for self-supervised denoising on real-world noisy images, i.e., separately and jointly learning both visible and invisible noise using a single BSN, using different strategies to direct the network towards focusing on each part.

2. Given the region-correlated and structured features of real-world noise, a multi-path module has been developed to serve as a guiding mechanism for effectively separating regions affected by either visible or invisible noise, without relying on any ground truth (GT) or reference indicating the level of noise.

3. We propose different strategies to guide the network in targeting distinct regions for denoising, while simultaneously preserving the details that are prone to being mistakenly identified as noise. From experiments, our method shows state-of-the-art performance in the domain of self-supervised real-world image denoising.

2 Related Work

Denoising techniques have their roots from non-learning approaches, which can be categorized as either filter-based [5, 11, 38] or model-driven methods [2, 13, 36]. Over the past decade, the advent of deep learning technology has brought about significant performance enhancements in comparison with non-learning approaches.

Supervised Image Denoising: Supervised denoising methods were initially developed for additive white Gaussian noise (AWGN) removal, by employing a CNN network with batch normalization and residual learning [48]. Then, many advanced network-based denoisers with intricate architectures have been proposed subsequently [27, 28, 37, 49, 50]. However, the models trained with AWGN exhibit limited generalization capability owing to the inherent domain discrepancy between real-world and synthetic noise.

To mitigate this problem, CBDNet[14] employed a synthesis of Poisson-Gaussian noise and simulated its application through the in-camera ISP model. In addition, for solving real-world denoising problem, several methods train the network directly on real-world noisy-clean pairs in a supervised manner [9, 21, 45, 46]. However, obtaining properly aligned noisy-clean pairs of real images asks for a significant amount of human labor and may not always be feasible in real-world scenarios.

Unpaired Image Denoising: Considering that the ground truth of the noisy image is usually unavailable, several approaches sought to generate unpaired noisy-clean images for training. Most of these approaches aimed to simulate realistic noise [6, 7, 16, 19, 43], for example, the approach named GCBD [7] utilized a generative adversarial network (GAN)[12] for training a generator capable of matching the actual noise distribution found in the plain regions of noisy images. However, their performance remains constrained by the intricate nature of modeling real-world noise distributions.

Self-supervised Denoising for Synthetic Noise: One of the most prominent methods for self-supervised denoising is Noise2Noise [24], which employs a neural network trained on paired images of the same noisy target to predict the clean signal. However, capturing multiple noisy observations per scene continues to pose a significant challenge. Thus, various methodologies have been developed to tackle this issue by generating paired images from a single noisy observation through a series of sampling and transfer techniques [17, 29, 32, 44].

In addition to utilizing paired images for learning purposes, BSNs are employed to acquire self-supervised models from only noisy images with masking techniques[4, 22, 34, 39]. However, the theoretical assurance of those BSNs does not apply to real-world scenarios since that the noise should satisfy the assumption of pixel-wise independence.

Self-supervised Denoising for Real-world Noise: As a pioneer approach in training a self-supervised denoiser directly on real-world noisy images, CVF-SID disentangles the noise components by employing a cyclic multi-variate function, but it assumes spatially-uncorrelated noise, which does not align with the real distribution of noise [30]. The BSN denoisers, which have attracted significant attention for synthetic noise removal, cannot be directly applied to the real-world denoising tasks, primarily due to the violation of the assumption of pixel-wise independence of noise.

Then, AP-BSN, a brilliant approach that proposes pixel-shuffle down-sampling scheme to break spatial noise correlation has received the most attention [23]. Formally, its network output $\mathbf{x}_o$ is calculated by

$$\mathbf{x}_o = \mathcal{B}(\mathbf{y};\omega_n) := \mathcal{P}_s^{-1}(\mathcal{N}_b(\mathcal{P}_s(\mathbf{y});\omega_n)), \tag{1}$$

where $\mathbf{y}$ is the noisy input; $\mathcal{N}_b(\cdot;\omega_n)$ denotes the BSN with well-trained parameter ω_n; $\mathcal{P}_s$ is the pixel-shuffling operator with a stride factor of s, and $\mathcal{P}_s^{-1}$ is its inverse.

A number of approaches have been subsequently developed, building upon the foundation of AP-BSN [15, 20, 31, 40]. SS-BSN[15] directly adopted the asymmetric PD utilized in AP-BSN and incorporated the extraction of valuable information from non-local self-similarity. In addition, SS-BSN also introduced an attention module to enhance the training of BSN, however, it is time-consuming. Given the significant reduction in sampling density caused by the PD, LG-BPN [40] introduced a densely-sampled module to effectively retain more information and improve fine structure recovery. Both C-BSN [20] and SDAP [31] suggested utilizing random sub-sampling schemes to mitigate the artifacts introduced by PD and enhance denoising performance. However, they offered limited performance gains.

In addition to improving the PD technique, SASL[26] is the work that takes into account the diverse intricacies involved in denoising across different regions. It introduced a blind-neighborhood network to offer certain supervision in flat areas and employed a locally aware network for noise elimination of textured regions. Therefore, including the U-Net denoiser, SASL consists of three individually trained networks. The independent learning of flat and textured regions leads to the presence of inconsistent fine structure in the final outcomes.

3 Method

3.1 Blind Estimation of NV Map

In addition to the essential characteristic of spatial correlation, the visibility of real noise also exhibits inconsistency within an individual image, due to variations in luminance range across different regions [10] (not limited to the brightest areas). To comprehensively investigate the noise linked to different visibility levels, we propose constructing a network module of NV map estimation to roughly discern and distinguish them.

To address the issue of inconsistent noise in relation to varying luminance ranges, we in this paper propose a module incorporating a multi-path network for estimating an NV map without relying on any ground truth or reference of the noise level. For every path of the network, we learn from the previous work [43] to build a network that contains five 1×1 convolution layers of 32 channels, while deploying the ReLU nonlinear mapping for all convolution layers except the last one. Although, the 1×1 convolution has been proposed for learning signal-dependent noise in a multivariate heteroscedastic Gaussian model, it faces challenges in capturing the structural texture of real-world noise, particularly when the fluctuations occur over relatively long distances. In other words, using 1×1 convolution only may struggle to handle real-world noise that is generally not fine-grained. Besides, the visibility of real-world noise is generally consistent in regions with comparable luminance and chroma. However, the single-path module lacks the ability to accurately simulate such region-related characteristic. To better simulate the spatially correlated and structured real-world noise, we propose a 3-paths network that integrates the individual estimation of single-paths (see Fig. 2 framed in grey).

It should be emphasized that our proposed module serves as a blind estimator of the NV map, which presents a greater challenge compared to previous work that relies on ground truth [14, 21] or any other noise level reference [6, 7, 19, 43]. Due to the inherent structural texture of real-world noise, it becomes challenging to distinguish them from the underlying image structure. Therefore, accurately estimating real-world noise in a blind manner may pose the risk of underestimating the visible noise while simultaneously misjudging the image structure as noise. Taking this into consideration, we in the following sub-section will carefully construct a single BSN network for removing both visible and less visible (and underestimated) noise, while simultaneously preserving intricate image details.

3.2 Visible and Invisible Denoising

We in this paper desire to optimize a single BSN denoiser by selectively targeting one specific type of noise with a particular visibility level at any given time. This concept draws inspiration from the practice of self-supervised denoising, which involves using multiple synthetic noisy images with diverse levels of noise [24]. Nevertheless, in the context of real-world denoising, the separate learning of distinct noise visibility is not as optimal as it is in synthetic scenarios.

After estimating the NV map (denoted as $\mathbf{m}_{nv}$) by our multi-path module (denoted as $\mathcal{M}(\mathbf{y};\mu_m)$ with the parameter μ_m to be learned), i.e.,

$$\mathbf{m}_{nv} = \mathcal{M}(\mathbf{y};\mu_m), \tag{2}$$

it is inevitable that $\mathcal{M}(\mathbf{y};\mu_m)$ encompasses a combination of partly visible noise and image structures. Thus, we firstly generate a network target $\mathbf{y}_i^-$ to separate the estimated visible

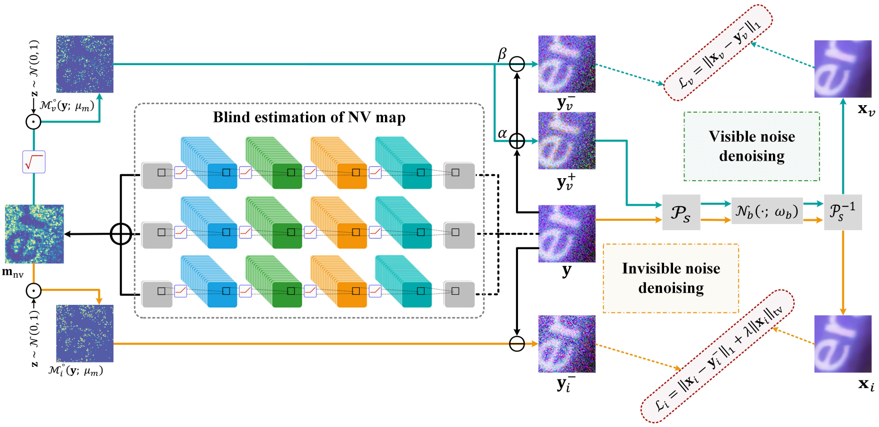

$\mathbf{z} \sim \mathcal{N}(0,1)$
$\mathcal{M}_v^\circ(\mathbf{y};\mu_m)$
Blind estimation of NV map
β
α
$\mathbf{y}_v^-$
$\mathbf{y}_v^+$
$\mathbf{y}$
$\mathbf{y}_i^-$
$\mathcal{L}_v = \|\mathbf{x}_v - \mathbf{y}_v^-\|_1$
Visible noise denoising
$\mathbf{x}_v$
$\mathcal{P}_s$
$\mathcal{N}_b(\cdot;\ \omega_b)$
$\mathcal{P}_s^{-1}$
Invisible noise denoising
$\mathcal{L}_i = \|\mathbf{x}_i - \mathbf{y}_i^-\|_1 + \lambda\|\mathbf{x}_i\|_{tv}$
$\mathbf{x}_i$
$\mathbf{m}_{\mathrm{nv}}$
$\mathbf{z} \sim \mathcal{N}(0,1)$
$\mathcal{M}_i^\circ(\mathbf{y};\mu_m)$

Figure 2: The integrated flow diagram of our proposed framework.

noise from the remaining noise, i.e.,

$$\mathbf{y}_i^- = \mathbf{y} - \mathcal{M}_i^\circ(\mathbf{y};\mu_m), \tag{3}$$

where $\mathcal{M}_i^\circ(\mathbf{y};\cdot) := \mathcal{M}(\mathbf{y};\cdot) \circ \mathbf{z}$, with the Hadamard product $\circ$; besides, $\mathbf{z} \backsim \mathcal{N}(0,1)$. Then, we desire to train a BSN denoiser by minimizing the loss between $\mathbf{y}_i^-$ and the BSN output of $\mathbf{y}$ to pay more focus on less visible and underestimated noise learning.

Given that the estimated NV map inevitably contain image structures, we desire to employ an ℓ_1-norm-based term to account for the sparsity of the differences caused by details removal. Moreover, the total variation (TV) regularization [36] (denoted as $\|\cdot\|_{tv}$) is also used to enhance the edge-preserving capability of the output. Thus, the loss function for learning less visible noise is set as

$$\mathcal{L}_i = \|\mathbf{x}_i - \mathbf{y}_i^-\|_1 + \lambda\|\mathbf{x}_i\|_{\mathrm{tv}}, \tag{4}$$

where $\mathbf{x}_i$ is the output of the BSN denoiser $\mathcal{B}(\mathbf{y};\omega_b)$ with to-be-learned ω_b; while λ is a positive constant to balance the two objective terms.

Taking $\mathbf{y}_i^-$ as the target is also beneficial to the PD process of $\mathcal{B}(\cdot;\cdot)$. Since the NV map is likely to contain image structures, removing them will remain relatively flat regions for denoiser learning, thereby mitigating artifacts caused by the shuffling procedure while emphasizing the network's focus on denoising itself. However, a negative side is that the denoiser trained exclusively on invisible noise may possess inadequate efficacy in eliminating much more visible noise and yields too-smoothed outcomes.

On the contrary, in addition to the less visible and underestimated noise, we also develop a different line to enhance the focus of the BSN denoiser on effectively learning high visibility noise. Specifically, we use a noisier input $\mathbf{y}_v^+$ and a target $\mathbf{y}_v^-$ with lower noise, i.e.,

$$\mathbf{y}_v^+ = \mathbf{y} + \alpha\mathcal{M}_v^\circ(\mathbf{y};\mu_m), \quad \mathbf{y}_v^- = \mathbf{y} - \beta\mathcal{M}_v^\circ(\mathbf{y};\mu_m), \tag{5}$$

where $\alpha > 1$ and $\beta \in (0,1)$, focusing primarily on regions with high visible noise by amplifying the discrepancies between the input and the target in such areas while keeping other regions unchanged.

Dataset	Self-supervised learning						
	CVF-SID[30]	AP-BSN[23]	SDAP[31]	SS-BSN[15]	C-BSN[20]	SASL[26]	Ours
SIDD	34.43/0.912	35.97/0.925	36.54/0.919	36.73/0.923	36.82/0.934	37.41/0.934	37.16/0.936
DND	36.31/0.923	38.09/0.937	37.71/0.928	37.72/0.928	38.45/0.939	38.58/0.936	38.74/0.943

Table 1: Quantitative comparison of PSNR and SSIM on SIDD and DND datasets is conducted, Red and blue colors are employed to represent the best and second-best outcomes among the self-supervised methods.

Moreover, since the sharp image structures are also learnt in the NV map, we use a square root operation on the estimated NV map to enhance the significance of details within it, i.e., $\mathcal{M}_v^{\circ}(\mathbf{y};\cdot) = \mathrm{sqrt}(\mathcal{M}(\mathbf{y};\cdot)) \circ \mathbf{z}$. By minimizing the loss for learning visible noise, i.e.,

$$\mathcal{L}_v = \|\mathbf{x}_v - \mathbf{y}_v^-\|_1, \tag{6}$$

where $\mathbf{x}_v$ represents the output of $\mathcal{B}(\mathbf{y}_v^+;\omega_b)$, we expect the BSN denoiser to remove visible noise while simultaneously preserving meaningful details of the image.

Finally, the well-trained parameters of both the blind estimation module of NV map and the BSN-based denoising module, i.e., μ_m and ω_b are obtained by jointly optimizing:

$$\min_{\mu_m,\omega_b} \mathcal{L}_i + \gamma\mathcal{L}_v, \tag{7}$$

with a balance parameter $\gamma > 0$. Furthermore, the architecture of our BSN denoiser $\mathcal{B}(\cdot;\omega_b)$ is the same with the one utilized in [23], employing PD_5 for training while PD_2 for testing. It is worth noting that the effectiveness of our approach lies in the separately and jointly learning of visible and invisible noise within a single real-world noisy image, which will be verified through the subsequent experiments. For clarity, we provide the flow chart of our proposed method in Fig. 2.

4 Experiments

4.1 Experimental Details

Real-world datasets. To evaluate the effectiveness of our proposed approach, all experiments are conducted on two publicly concerned datasets, i.e., the Smartphone Image Denoising Dataset (SIDD)[1] and Darmstadt Noise Dataset (DND)[33] for training and testing real-world sRGB camera noise removal. We utilize the SIDD medium dataset for model training, which comprises 320 pairs of aligned real noisy-clean images. Each noisy image is captured multiple times, and the average image is served as the reference for GT. However, being a self-supervised method, only noisy images are used as the training samples. Furthermore, the validation and benchmark datasets of SIDD are respectively adopted for validation and testing which both contain 1280 image blocks of size 256×256.

The DND dataset comprises image patches of size 512×512 which is cropped from 50 high-resolution test images. Since DND does not provide specific datasets for training and validation, we use the whole image patches for training in a fully self-supervised manner. As the benchmark dataset does not provide GT, the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) of the test results can be obtained through the online submission system available on the SIDD and DND benchmark websites.

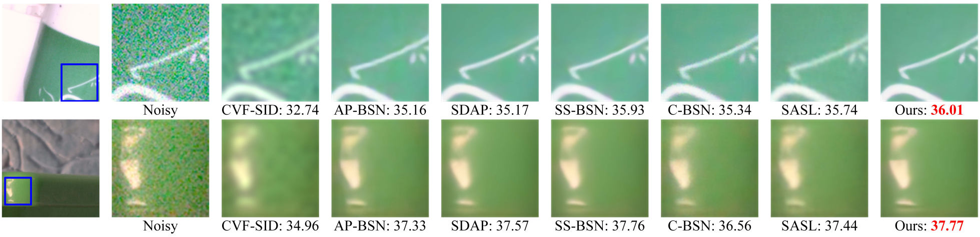


Figure 3: Visual comparisons of our method with other self-supervised SOTAs for denoising sRGB images on the SIDD Validation dataset.

Implementation Details. We employ a batch size of 32 and a patch size of 120 × 120 to train our proposed approach. The Adam method is utilized as our optimizer with an initial learning rate of $3e^{-4}$. Then, the learning rate is reduced by a factor of 10 for every 9 epochs during our training for a total of 21 epochs. We set λ to 0.1 for the TV regularization of our module of learning invisible noise. Additionally, we have chosen $\alpha = 1.8$ and $\beta = 0.6$ as the parameters for our visible denoising module, while setting the balance parameter γ to 1. In addition, all the experiments are conducted on the NVIDIA RTX 3090 in PyTorch 1.9.

4.2 Comparison with State-of-the-arts

We compare our approach against state-of-the-art methods, including traditional non-learning based methods (i.e., BM3D [11] and WNNM [13]), supervised learning methods with synthetic noisy-clean pairs (i.e., DnCNN [48] and Zhou et al. [49]) and with real-world noisy-clean pairs (i.e., TNRD [8], CBDNet [14], RIDNet [3], AINDNet [21], VDN [45], DANet [46] and MIRNet[47]), unpaired learning approaches for real-world image denoising (i.e., GCBD [7], C2N [19] and D-BSN [43]), with quantitative comparisons detailed in the supplementary materials. Furthermore, we place special emphasis on the comparisons with self-supervised learning approaches for real-world image denoising, including CVF-SID[30], AP-BSN[23], SDAP[31], SS-BSN[15], C-BSN[20] and SASL[26], as shown in Table 1.

In Table 1, we present the PSNR/SSIM results of various state-of-the-art (SOTA) denoising methods on both the SIDD Benchmark and DND Benchmark. Specifically, the values of all the SOTAs are posted from the corresponding papers that can be cross-verified with the benchmark websites. These values are obtained by respectively using SIDD medium dataset and DND benchmark for training purposes.

when compared to the self-supervised approaches, our method outperforms other SOTA methods in all aspects, except for the slightly lower PSNR of SIDD in comparison to SASL. However, the results in relation to the SSIM index clearly demonstrate the effectiveness of our proposed method on the SIDD benchmark, The superiority of our method can also be noticed from the visual comparisons in Fig. 3. Our approach clearly demonstrates its capability in generating denoised images with enhanced sharpness, outperforming both SASL and C-BSD which are considered as more competitive methods. However, it is worth noting that these alternative approaches exhibit certain drawbacks such as the presence of irregular artifacts and a loss of detail clarity.

Furthermore, both SDAP and SS-BSN exhibit better results than other approaches, which is on par with the visual quality of our results. However, the split technique of SDAP is more likely to produce excessively smoothed results. On the other hand, SS-BSN exhibits elevated model complexity due to the incorporation of an attention module, whereas its training time

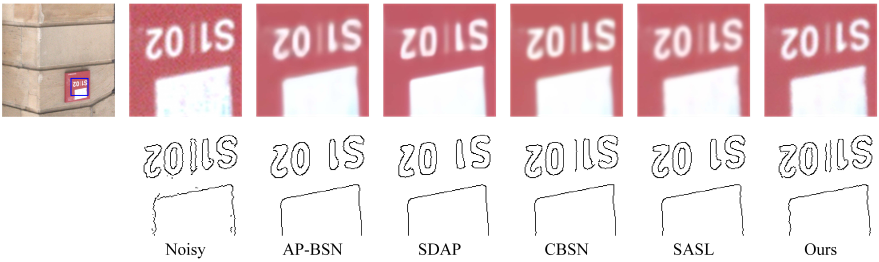


Figure 4: The top row shows the visual comparisons with other self-supervised SOTAs for denoising sRGB images on the DND benchmark dataset, while the bottom row represents the corresponding edge maps computed using the same way.

is prolonged. Specifically, the training time required for a single epoch on the SIDD medium dataset is about 0.7 hours in our framework. However, SS-BSN commends approximately 1.2 hours, surpassing the time taken by our method considerably. In addition, the total number of epochs for our approach and SS-BSN are comparable (i.e., 21 vs 20). Therefore, to achieve the results presented in Tab. 1, SS-BSN requires significantly longer cost of training in comparison with our approach. The qualitative comparison of the DND benchmark testing dataset is illustrated in Fig. 4. Our results exhibit superior cleanliness and enhanced preservation of details when compared to other state-of-the-art approaches, which also verifies the superiority of our approach.

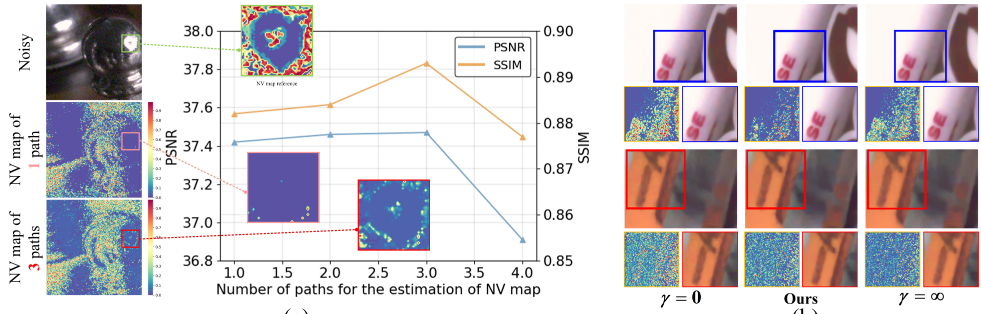
5 Ablation Study

We perform extensive ablation studies on the SIDD validation and benchmark datasets to analyze the effectiveness of our proposed method. This includes investigating the impact of the path number for NV map estimation, examining the effectiveness of the visible and invisible learning modules, evaluating the TV regularization parameter λ, and assessing the sensitivity of parameters α and β in the learning module of visible noise Besides, details of the ablation experiments on hyperparameters λ, α, and β are in the supplementary materials.

5.1 Path Number of NV Map Estimation

As depicted in Fig. 5 (a), we used a range of one to four paths for investigate the impact of the path number of our NV map estimation module on the SIDD validation dataset. When utilizing one to three noise estimation paths, PSNR demonstrates a gradual improvement, while SSIM exhibits a more pronounced increase, reaching its peak with the utilization of three paths of network. Nevertheless, the continuous increase in the number of paths brings about a significant decline in denoising effectiveness, which is attributed to the over-estimation of the noise level.

The utilization of one-path noise estimation network is more prevalent in advance [14, 21, 43]. However, its effectiveness is limited in the blind estimation scenarios encountered in real-world image denoising. Furthermore, we provide a visual illustration to present the disparity in the NV map between the one-path network and our three-paths estimation. It is clearly observed that our NV map is more consistent with the visual perception of the noisy



(a) (b)

Figure 5: (a)Ablation study of the path number of $\mathcal{M}(\mathbf{y};\mu_m)$ for the blind estimation of NV map. (b)Visual comparisons of the visible/invisible learning flow.

image, exhibiting greater consistency with GT reference that shows the disparity between the noisy image and the clean reference provided in the SIDD validation dataset.

5.2 Effects of Visible/Invisible Learning

We conduct ablation studies to examine the impacts of visible and invisible learning flows in our method. By setting $\gamma = 0$, that is, the BSN denoiser is exclusively trained by utilizing invisible noise. Under this scenario, the well-trained BSN denoiser will lose its ability to effectively eliminate visible noise, bringing about the presence of uncleaned artifacts in the denoised outcomes, as presented in the first column of Fig. 5(b). On the contrary, when the invisible noise is not taken into account during training, that is, $\gamma = \infty$, the BSN denoiser $\mathcal{B}(\cdot;\omega_b)$ solely relies on visible noise, which causes an excessive smoothing effect and blurriness in the resulting denoised images, as shown in the last column of Fig. 5(b).

From the quantitative comparisons, the PSNR metric indicates a relatively small difference between the cases when $\gamma = 0$ (PSNR = 36.95) and $\gamma = \infty$ (PSNR = 36.96). However, the SSIM metric reveals a more pronounced disparity, with SSIM values of 0.931 for $\gamma = \infty$ and 0.934 for $\gamma = \infty$, which better aligns with the observed visual effects. Whether through qualitative or quantitative comparisons, it becomes evident that the significance lies in separately addressing both visible and invisible noise within a single BSN denoiser.

6 Conclusion

In this research, we propose a novel perspective for self-supervised denoising by utilizing a BSN with single noisy images, taking into consideration the spatially inconsistent visibility of noise in real-world scenarios. In order to identify the regions affected by either visible or invisible noise, we develop a multi-path module to blindly estimate an NV map, which directs the BSN network towards focusing on the regions of similar level of noise. Extensive experiments have validated that our proposed method performs favorably against SOTA methods on real-world cases. It should be noted that all current methods yield results with certain artifacts (e.g., as depicted in the second row of Fig. 3) caused by color mixture with real-world noise, which will be addressed in our future research.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Natural Science Foundation of Liaoning Province (No. 2023-MS-126) and the Central Universities (No. 3132023523).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018.

[2] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[3] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *ICCV*, 2019.

[4] Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *ICML*, 2019.

[5] Antoni Buades, Bartomeu Coll, and Jean Michel Morel. A non-local algorithm for image denoising. In *CVPR*, 2005.

[6] Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, and Donglai Wei. Learning to generate realistic noisy images via pixel-level noise-aware adversarial training. In *NeurIPS*, 2021.

[7] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *CVPR*, 2018.

[8] 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.

[9] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. NBNet: Noise basis learning for image denoising with subspace projection. In *CVPR*, 2021.

[10] Yiheng Chi, Xingguang Zhang, and H. Stanley Chan. Hdr imaging with spatially varying signal-to-noise ratios. In *CVPR*, 2023.

[11] 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.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[13] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *CVPR*, 2014.

[14] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, 2019.

[15] Young-Joo Han and Ha-Jin Yu. SS-BSN: Attentive blind-spot network for self-supervised denoising with nonlocal self-similarity. In *IJCAI*, 2023.

[16] Zhiwei Hong, Xiaocheng Fan, Tao Jiang, and Jianxing Feng. End-to-End unpaired image denoising with conditional adversarial networks. In *AAAI*, 2020.

[17] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2Neighbor: Self-supervised denoising from single noisy images. In *CVPR*, 2021.

[18] Imatest. Noise in photographic images. https://www.imatest.com/support/docs/23-1/noise/.

[19] Geonwoon Jang, Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. C2N: Practical generative noise modeling for real-world denoising. In *ICCV*, 2021.

[20] 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 *ICCV*, 2023.

[21] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *CVPR*, 2020.

[22] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void-learning denoising from single noisy images. In *CVPR*, 2019.

[23] 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 *CVPR*, 2022.

[24] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *ICML*, 2018.

[25] You Lei, Wang Yuan, Hongpeng Wang, You Wenhu, and Wu Bo. A skin segmentation algorithm based on stacked autoencoders. *IEEE Transactions on Multimedia*, 19(4): 740–749, 2016.

[26] 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 *CVPR*, 2023.

[27] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S. Huang. Non-local recurrent network for image restoration. In *NeurIPS*, 2018.

[28] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NeurIPS*, 2016.

[29] Nick Moran, Dan Schmidt, Yu Zhong, and Patrick Coady. Noisier2Noise: Learning to denoise from unpaired noisy data. In *CVPR*, 2020.

[30] 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 *CVPR*, 2022.

[31] Yizhong Pan, Xiao Liu, Xiangyu Liao, Yuanzhouhan Cao, and Chao Ren. Random sub-samples generation for self-supervised real image denoising. In *ICCV*, 2023.

[32] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-Recorrupted: Unsupervised deep learning for image denoising. In *CVPR*, 2021.

[33] Tobias Plötz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017.

[34] Yuhui Quan, Mingqin Chen, Tongyao Pang, and Hui Ji. Self2Self With Dropout: Learning self-supervised denoising from single image. In *CVPR*, 2020.

[35] Giovanni Ramponi, Norbert Strobel, Sanjit K. Mitra, and Tian Hu Yu. Nonlinear unsharp masking methods for image contrast enhancement. *Journal of Electronic Imaging*, 5(3):353–366, 1996.

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

[37] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. MemNet: A persistent memory network for image restoration. In *ICCV*, 2017.

[38] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *ICCV*, 1998.

[39] Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2Unblind: Self-supervised image denoising with visible blind spots. In *CVPR*, 2022.

[40] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. LG-BPN: Local and global blind-patch network for self-supervised real-world denoising. In *CVPR*, 2023.

[41] Jie Wen, Yong Xu, and Hong Liu. Incomplete multiview spectral clustering with adaptive graph learning. *IEEE Transactions on Cybernetics*, 50(4):1418–1429, 2018.

[42] Jie Wen, Zheng Zhang, Zhao Zhang, Lunke Fei, and Meng Wang. Generalized incomplete multiview clustering with flexible locality structure diffusion. *IEEE Transactions on Cybernetics*, 51(1):101–114, 2020.

[43] Xiaohe Wu, Ming Liu, Yue Cao, Dongwei Ren, and Wangmeng Zuo. Unpaired learning of deep image denoising. In *ECCV*, 2020.

[44] Jun Xu, Yuan Huang, Ming-Ming Cheng, Li Liu, Fan Zhu, Zhou Xu, and Ling Shao. Noisy-As-Clean: Learning self-supervised denoising from corrupted image. *IEEE Transactions on Image Processing*, 29:9316–9329, 2020.

[45] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational Denoising Network: Toward blind noise modeling and removal. In *NeurIPS*, 2019.

[46] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *ECCV*, 2020.

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020.

[48] 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.

[49] 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.

[50] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019.