Zero-Shot Blind-spot Image Denoising via Implicit Neural Sampling (Supplementary Materials)

Yuhui Quan[1] Tianxiang Zheng[1] Zhiyuan Ma[2] Hui Ji[2]
[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[2] Department of Mathematics, National University of Singapore, 119076, Singapore
csyhquan@scut.edu.cn, cszhengtx@mail.scut.edu.cn, e0983565@u.nus.edu, matjh@nus.edu.sg

1. Proof of Proposition 1

Let $\mathbf{e} \in \mathbb{R}^M$ denote the column vector of ones, and $\mathbf{E} \in \mathbb{R}^{M\times M}$ the square matrix of ones.

Proposition 1. *Consider a vector* $\mathbf{y} = [y_1, y_2, \ldots, y_M]$*, for each linear denoiser defined by*

$$\mathcal{D}_{\mathbf{a}}(\mathbf{y}) = \mathbf{a}^\top \mathbf{y},$$

where $\mathbf{a} = [a_1, a_2, \ldots, a_M]^\top$ *denotes it weights. its risk is defined by*

$$\mathcal{R}_{\mathbf{a}} = \mathop{\mathbb{E}}_{\boldsymbol{\mu},\mathbf{n},n_0} \left[\|\mathcal{D}_{\mathbf{a}}(\mathbf{y}) - x_0\|_2^2 \right].$$

Let $\mathbf{a}^*$ *be the optimal weights vector learned by minimizing the self-supervised loss:*

$$f(\mathbf{a}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{\mu},\mathbf{n},n_0} \left[\|\mathcal{D}_{\mathbf{a}}(\mathbf{y}) - y_0\|_2^2 \right].$$

That is

$$\mathbf{a}^* \in \mathop{\arg\min}_{\mathbf{a}\in\mathbb{R}^M} f(\mathbf{a}).$$

Then, the risk of the optimal linear denoiser is

$$\begin{aligned}\mathcal{R}_{\mathbf{a}^*} =& -\frac{M(x_0^4 - \lambda_n^2\sigma^4)}{Mx_0^2 + (M-1)\lambda_n^2\sigma^2 + \lambda_v^2\sigma^2 + \sigma^2} + x_0^2 \\ =& -\frac{Mx_0^4 + \frac{M\sigma^2}{M-1}\left(Mx_0^2 + \lambda_v^2\sigma^2 + \sigma^2\right)}{Mx_0^2 + (M-1)\lambda_n^2\sigma^2 + \lambda_v^2\sigma^2 + \sigma^2} + \frac{M\sigma^2}{M-1} + x_0^2.\end{aligned}$$

Lemma 1. *Let* $k_1 \geq 0$ *and* $k_2 > 0$*, then* $k_1\mathbf{E} + k_2\mathbf{I}$ *is a postive definite matrix and its inverse is given by* $(k_1\mathbf{E} + k_2\mathbf{I})^{-1} = -\frac{k_1}{(Mk_1 + k_2)k_2}\mathbf{E} + \frac{1}{k_2}\mathbf{I}$.

Proof of Lemma 1. For all $\mathbf{x} = (x_1, \cdots, x_M)^T \neq 0$, we have

$$\mathbf{x}^T(k_1\mathbf{E} + k_2\mathbf{I})\mathbf{x} = k_1\mathbf{x}^T\mathbf{E}\mathbf{x} + k_2\mathbf{x}^T\mathbf{x} = k_1\left(\sum_{i=1}^M x_i\right)^2 + k_2\|\mathbf{x}\|_2^2 > 0.$$

Therefore, $k_1\mathbf{E}+k_2\mathbf{I}$ is positive definite. Note that $\mathbf{E}^2=M\mathbf{E}$, so

$$\begin{aligned}&(k_1\mathbf{E}+k_2\mathbf{I})\left(-\frac{k_1}{(Mk_1+k_2)k_2}\mathbf{E}+\frac{1}{k_2}\mathbf{I}\right)\\=&-\frac{k_1^2}{(Mk_1+k_2)k_2}\mathbf{E}^2-\frac{k_1}{Mk_1+k_2}\mathbf{E}+\frac{k_1}{k_2}\mathbf{E}+\mathbf{I}=\mathbf{I}.\end{aligned}$$

As a result, $(k_1\mathbf{E}+k_2\mathbf{I})^{-1}=-\frac{k_1}{(Mk_1+k_2)k_2}\mathbf{E}+\frac{1}{k_2}\mathbf{I}$. □

Proof of Proposition 1. Note that

$$\begin{aligned}f(\mathbf{a})=&\mathop{\mathbb{E}}_{\boldsymbol{\mu},\mathbf{n},n_0}\left[\left\|\mathcal{D}_{\mathbf{a}}(\{y_j\}_{j=1}^M)-y_0\right\|_2^2\right]=\mathop{\mathbb{E}}_{\boldsymbol{\mu},\mathbf{n},n_0}\left[\left(\mathbf{a}^T\mathbf{e}x_0-\mathbf{a}^T\boldsymbol{\mu}x_0+\mathbf{a}^T\mathbf{n}-x_0-n_0\right)^2\right]\\=&\mathop{\mathbb{E}}_{\boldsymbol{\mu},\mathbf{n},n_0}[\mathbf{a}^T\mathbf{e}\mathbf{e}^T\mathbf{a}x_0^2+\mathbf{a}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{a}x_0^2+\mathbf{a}^T\mathbf{n}\mathbf{n}^T\mathbf{a}+x_0^2+n_0^2-2\mathbf{a}^T\mathbf{e}\mathbf{a}^T\boldsymbol{\mu}x_0^2+2\mathbf{a}^T\mathbf{e}\mathbf{a}^T\mathbf{n}x_0\\&-2\mathbf{a}^T\mathbf{e}x_0^2-2\mathbf{a}^T\mathbf{e}x_0n_0-2\mathbf{a}^T\boldsymbol{\mu}\mathbf{a}^T\mathbf{n}x_0+2\mathbf{a}\boldsymbol{\mu}x_0^2+2\mathbf{a}^T\boldsymbol{\mu}x_0n_0-2\mathbf{a}^T\mathbf{n}x_0-2\mathbf{a}^T\mathbf{n}n_0+2x_0n_0]\\=&\mathbf{a}^T\mathbf{E}\mathbf{a}x_0^2+\lambda_v^2\sigma^2\mathbf{a}^T\mathbf{I}\mathbf{a}x_0^2+\mathbf{a}^T\left(\lambda_n^2\sigma^2\mathbf{E}+(1-\lambda_n^2)\sigma^2\mathbf{I}\right)\mathbf{a}+x_0^2+\sigma^2-2\mathbf{a}^T\mathbf{e}x_0^2-2\mathbf{a}^T\mathbf{e}\lambda_n\sigma^2\\=&\mathbf{a}^T\left((x_0^2+\lambda_n^2\sigma^2)\mathbf{E}+(\lambda_v^2+1-\lambda_n^2)\sigma^2\mathbf{I}\right)\mathbf{a}-2\left(x_0^2+\lambda_n\sigma^2\right)\mathbf{a}^T\mathbf{e}+x_0^2+\sigma^2.\end{aligned}$$

If $\sigma=0$, then

$$f(\mathbf{a})=x_0^2\left(\mathbf{a}^T\mathbf{E}\mathbf{a}-2\mathbf{a}^T\mathbf{e}+1\right)=x_0^2\left(\mathbf{a}^T\mathbf{e}-1\right)^2\geq 0,$$

equality holds when $\mathbf{a}^{*T}\mathbf{e}=1$. Note that when $\sigma=0$, we have $x_0=y_0$, so

$$\mathcal{R}_{\mathbf{a}^*}=f(\mathbf{a}^*)=0,$$

which satifies the theorem.
If $\lambda_v=0$ and $\lambda_n=1$, then

$$f(\mathbf{a})=(x_0^2+\sigma^2)(\mathbf{a}^T\mathbf{E}\mathbf{a}-2\mathbf{a}^T\mathbf{e}+1)=(x_0^2+\sigma^2)\left(\mathbf{a}^T\mathbf{e}-1\right)^2\geq 0,$$

equality holds when $\mathbf{a}^{*T}\mathbf{e}=1$. So, we have

$$\mathcal{R}_{\mathbf{a}^*}=(x_0^2+\sigma^2)\mathbf{a}^{*T}\mathbf{E}\mathbf{a}^*-2x_0^2\mathbf{a}^{*T}\mathbf{e}+x_0^2=\sigma^2,$$

which satifies the theorem.
If $\lambda_v=0$, $\lambda_n=-1$ and $x_0^2+\sigma^2\neq 0$, then

$$\begin{aligned}f(\mathbf{a})=&(x_0^2+\sigma^2)(\mathbf{a}^T\mathbf{E}\mathbf{a})-(x_0^2-\sigma^2)(\mathbf{a}^T\mathbf{e})+x_0^2+\sigma^2\\=&(x_0^2+\sigma^2)\left(\mathbf{a}^T\mathbf{e}-\frac{x_0^2-\sigma^2}{x_0^2+\sigma^2}\right)^2-\frac{(x_0^2-\sigma^2)^2}{x_0^2+\sigma^2}+x_0^2+\sigma^2\geq\frac{4x_0^2\sigma^2}{x_0^2+\sigma^2},\end{aligned}$$

equality holds when $\mathbf{a}^{*T}\mathbf{e}=\frac{x_0^2-\sigma^2}{x_0^2+\sigma^2}$. So, we have

$$\mathcal{R}_{\mathbf{a}^*}=(x_0^2+\sigma^2)\mathbf{a}^{*T}\mathbf{E}\mathbf{a}^*-2x_0^2\mathbf{a}^{*T}\mathbf{e}+x_0^2=\sigma^2,$$

which satifies the theorem.
Next, we assume $(\lambda_v^2+1-\lambda_n^2)\sigma^2\neq 0$. Since $|\lambda_n|\leq 1$, then $(\lambda_v^2+1-\lambda_n^2)\sigma^2>0$. Note that

$$f''(\mathbf{a})=2\left((x_0^2+\lambda_n^2\sigma^2)\mathbf{E}+(\lambda_v^2+1-\lambda_n^2)\sigma^2\mathbf{I}\right).$$

From the lemma, we derive $f''(\mathbf{a})$ is positive definite, so $f(\mathbf{a})$ is convex. Therefore, $f(\mathbf{a})$ has a minimum, and the minimum archieved when $f'(\mathbf{a}^*)=0$.

$$f'(\mathbf{a}^*)=2\left((x_0^2+\lambda_n^2\sigma^2)\mathbf{E}+(\lambda_v^2+1-\lambda_n^2)\sigma^2\mathbf{I}\right)\mathbf{a}^*-2\left(x_0^2+\lambda_n\sigma^2\right)\mathbf{e}=0.$$

Then, from the lemma, we have

$$\begin{aligned}\mathbf{a}^* &= \left((x_0^2+\lambda_n^2\sigma^2)\mathbf{E} + (\lambda_v^2+1-\lambda_n^2)\sigma^2\mathbf{I}\right)^{-1}(x_0^2+\lambda_n\sigma^2)\mathbf{e} \\ &= -\frac{\left(x_0^2+\lambda_n\sigma^2\right)\left(x_0^2+\lambda_n^2\sigma^2\right)}{\left(Mx_0^2+M\lambda_n^2\sigma^2+\left(\lambda_v^2+1-\lambda_n^2\right)\sigma^2\right)\left(\lambda_v^2+1-\lambda_n^2\right)\sigma^2}\mathbf{E}\mathbf{e} + \frac{x_0^2+\lambda_n\sigma^2}{\left(\lambda_v^2+1-\lambda_n^2\right)\sigma^2}\mathbf{I}\mathbf{e} \\ &= \frac{x_0^2+\lambda_n\sigma^2}{Mx_0^2+(M-1)\lambda_n^2\sigma^2+(\lambda_v^2+1)\sigma^2}\mathbf{e}.\end{aligned}$$

Then, we calculate the risk of the optimal linear denoiser and derive

$$\begin{aligned}\mathcal{R}_{\mathbf{a}^*} &= \mathop{\mathbb{E}}_{\boldsymbol{\mu},\mathbf{n},n_0}\left[\left\|\mathcal{D}_{\mathbf{a}}(\{y_j\}_{j=1}^M)-x_0\right\|_2^2\right] = \mathbb{E}\left[\left(\mathbf{a}^{*T}\mathbf{e}x_0 - \mathbf{a}^{*T}\boldsymbol{\mu}x_0 + \mathbf{a}^{*T}\epsilon - x_0\right)^2\right] \\ &= \mathbf{a}^{*T}\left((x_0^2+\lambda_n^2\sigma^2)\mathbf{E} + (\lambda_v^2+1-\lambda_n^2)\sigma^2\mathbf{I}\right)\mathbf{a}^* - 2x_0^2\mathbf{a}^{*T}\mathbf{e} + x_0^2 \\ &= -\frac{M(x_0^4-\lambda_n^2\sigma^4)}{Mx_0^2+(M-1)\lambda_n^2\sigma^2+\lambda_v^2\sigma^2+\sigma^2} + x_0^2 \\ &= -\frac{Mx_0^4+\frac{M\sigma^2}{M-1}\left(Mx_0^2+\lambda_v^2\sigma^2+\sigma^2\right)}{Mx_0^2+(M-1)\lambda_n^2\sigma^2+\lambda_v^2\sigma^2+\sigma^2} + \frac{M\sigma^2}{M-1} + x_0^2.\end{aligned}$$

Thus, we finish the proof. We can observe that $\mathcal{R}_{\mathbf{a}^*}$ decreases when $|\lambda_n|$ approaches 0. Additionally, if $x_0^2 \geq \lambda_n\sigma^2$, then $\mathcal{R}_{\mathbf{a}^*}$ decreases when $|\lambda_v|$ approaches 0. □

2. Details of NN architecture

NN architecture of $\mathcal{F}_\phi$: The neural network $\mathcal{F}_\phi$ is structured as a six-layer Multilayer Perceptron (MLP) with 32 feature nodes, which processes input coordinates normalized to $[-1, 1]$ and employs sinusoidal activation functions in each layer, aligning with the principles of Sinusoidal Representation Networks (SIRENs) [10]. To mitigate the risk of overfitting, dropout layers with incrementally increasing dropout rates $(0.35, 0.45, 0.55, 0.65, 0.75)$ are strategically integrated between the MLP layers. The final output layer utilizes a linear transformation to project the network's output into RGB space. To efficiently generate high-frequencies components, the input spatial coordinates are subjected to a higher dimensional space via a high-frequency transformation function $\gamma(\cdot)$, which is sinusoidal in nature:

$$\gamma(\boldsymbol{p}) = \left(\boldsymbol{p}, \sin\left(2^0\pi\boldsymbol{p}\right), \cdots \sin\left(2^{L-1}\pi\boldsymbol{p}\right), \cos\left(2^{L-1}\pi\boldsymbol{p}\right)\right)$$

Here, $\boldsymbol{p}$ denotes the normalized coordinate pair (i, j) within the range $[-1, 1]$, and L is an positive integer.

NN architecture of $\mathcal{D}_\theta$: The NN architecture of $\mathcal{D}_\theta$ used in this paper is based on Noise2Noise [7], which features an encoder-decoder structure with skip connections, enhanced by dilated convolutions ($dilation = 2$) to increase the receptive field efficiently. The encoder consists of five blocks (en_block1 to en_block5), each starting with a Conv2d layer using 48 filters of size 3×3, followed by LeakyReLU activations. Max pooling and dropout are employed for dimensionality reduction and regularization. The decoder (de_block1 to de_block5) upsamples feature maps, with each block increasing filter counts (96 in de_block1, 144 in de_block2 to de_block4) and applying similar Conv2d and LeakyReLU operations. The final block (de_block5) reduces filters to 3, corresponding to RGB output channels.

3. More Ablation Study

Comparison to MASH with similar inference time: We conducted the experiments with our method trained for 50 fewer iterations, resulting in an inference time of 24.89 seconds (vs. MASH's 25.11 seconds). As shown in Table 1, our method still outperforms MASH across all datasets.

4. Evaluation on AWGN denoising

In this study, we examine the effectiveness of proposed method on removing synthesized i.i.d. AWGN noise from images. Two datasets are used for the performance evaluation in the case of, including Set9 [11] with 9 color images and BSD68 [5] with 68 gray-scale images. The noise level is set to $\sigma = 25$. The quantitative results of synthetic image denoising are shown in Table 2.

Table 1. Quantitative comparison with fair inference time.

Method	SIDD Validation	SIDD Benchmark	FMDD	PolyU	CC
MASH	35.06/0.851	34.78/0.900	33.71/0.882	37.62/0.932	36.87/0.93
Ours (faster)	**35.29/0.868**	**35.02/0.921**	**33.91/0.885**	**37.86/0.957**	**37.20/0.946**

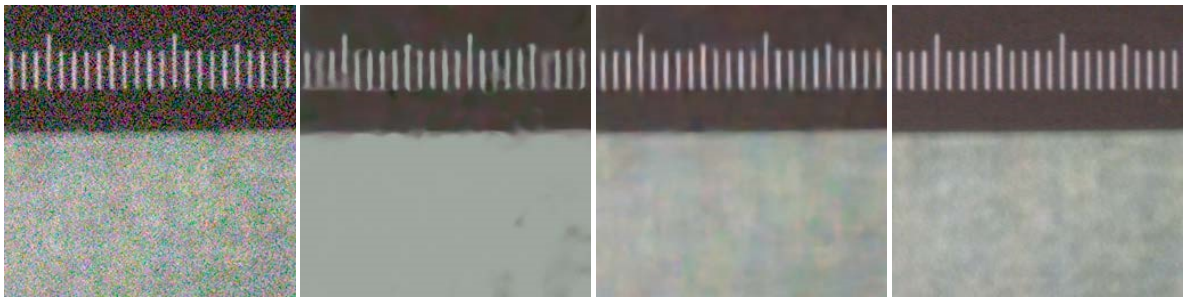
Table 2. Quantitative comparisons of different denoising methods on synthetic denoising with $\sigma = 25$.

Method	Set9		BSD68	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
BM3D [4]	31.67	0.955	28.56	0.801
DIP [11]	30.77	0.942	27.96	0.774
Self2Self [9]	31.74	0.956	28.70	0.803
APBSN-single [6]	26.41	0.862	24.14	0.572
ScoreDVI [2]	29.33	0.925	26.78	0.671
MASH [3]	29.54	0.921	26.53	0.666
Ours	29.73	0.926	26.81	0.677

It can be seen that the original mask-based blind-spot method, Self2Self [9], achieves the best performance on AWGN denoising. In contrast, its extensions designed for real-world denoising, including MASH [3] and our method, perform, less effectively in this context. This decrease in performance is expected because these extensions are tailored to address the spatial correlation of noise in real-world data. While they effectively reduce noise correlation, they inevitably also diminish the intensity value correlation among pixels used to predict invisible pixels. In the absence of noise correlation, the plain blind-spot scheme inherently maintains the highest intensity value correlation, explaining why the original mask-based methods outperform the extensions on removing synthesized AWGN from images.

5. Visualization of the denoising process

We visualize the output of the INR to further illustrate the trade-off between noise correlation and intensity value correlation. As shown in Fig. 1, the INR output provides complementary information to the original noisy images, exhibiting a stronger intensity value correlation with the ground truth.



Input Noisy Image
18.68 dB/0.239

Output of INR
27.68 dB/0.844

Output of Denoising NN
32.31 dB/0.875

Ground Truth
∞ dB/1

Figure 1. Output visualization of INR.

6. Limitations

Local-shuffle-based MASH [3] struggles with long-range noise correlations Our method leverages more distant noisy pixels, whose resulting loss of pixel dependencies is handled by INR estimates. The sub-pixel consistency loss from INR further boosts performance. However, in extremely low-brightness regions, our approach may lose some details, where MASH did slightly better, as shown in Fig 2.

Figure 2. Unsatisfactory cases of our method

7. Additional qualitative comparison

The additional visual comparison on PolyU [12], CC [8] SIDD-Validation [1] and SIDD-Benchmark [1], are provided in Fig. 3, 4, 5, 6.

Figure 3. Qualitative comparison of the results from different methods on some samples from PolyU.

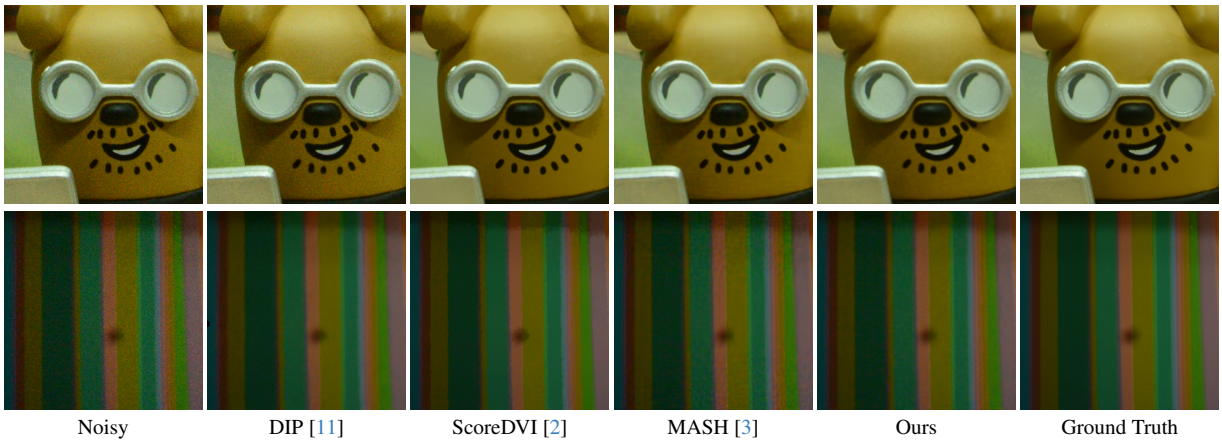
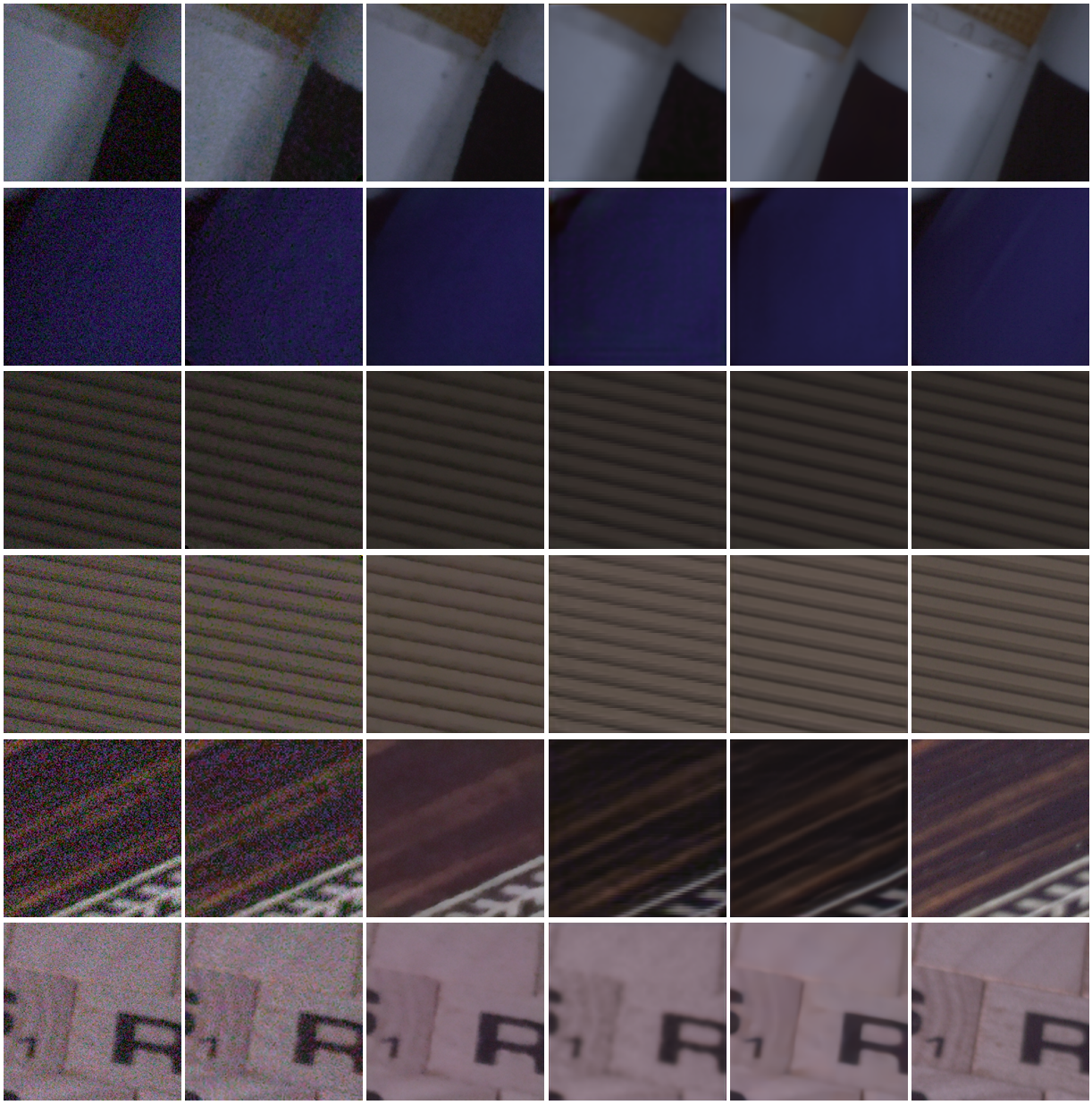


Figure 4. Qualitative comparison of the results from different methods on some samples from CC.



Noisy DIP [11] ScoreDVI [2] MASH [3] Ours Ground Truth

Figure 5. Qualitative comparison of the results from different methods on some samples from SIDD-Validation.

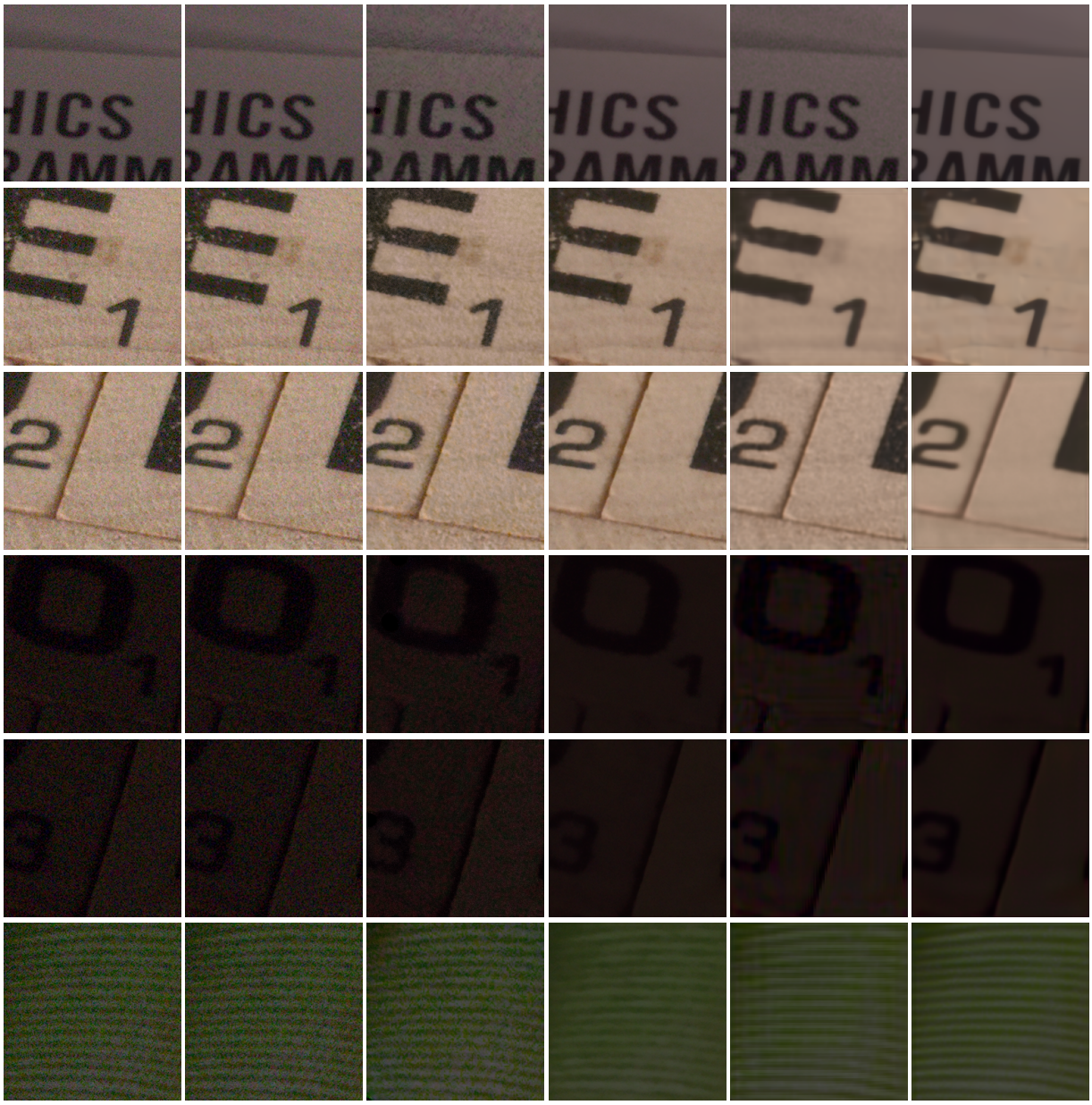


Figure 6. Qualitative comparison of the results from different methods on some samples from SIDD-Benchmark.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018.

[2] Jun Cheng, Tao Liu, and Shan Tan. Score priors guided deep variational inference for unsupervised real-world single image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12937–12948, 2023.

[3] Hamadi Chihaoui and Paolo Favaro. Masked and shuffled blind spot denoising for real-world images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3025–3034, 2024.

[4] 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.

[5] 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.

[6] 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.

[7] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. In *Proceedings of the International Conference on Machine Learning*, pages 2965–2974, 2018.

[8] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1683–1691, 2016.

[9] 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.

[10] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[11] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE/ Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[12] Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018.