Supplementary Material
Enhanced Latent Space Blind Model for Real Image Denoising via Alternative Optimization

Chao Ren[*] **Yizhong Pan** **Jie Huang**
College of Electronics and Information Engineering, Sichuan University, Chengdu, China
`chaoren@scu.edu.cn, {panyizhong, huangjiechn}@stu.scu.edu.cn`

In this supplementary material, we first illustrate some details about the use of existing assets, and the data of our method. In our main paper and supplementary material, we use existing public datasets for training or testing, including Div2K, SIDD Medium Dataset, RENOIR Dataset, DnD benchmark, SIDD benchmark, SIDD validation dataset, BSD68, Kodak24, and Nam. All these datasets are widely used in the image denoising field and are publicly available. Similar to other image denoising papers, we have cited the original papers that produced these datasets. We also use existing public codes for comparisons, including CBM3D, TNRD, DnCNN, FFDNet, DCDicL, CBDNet, VDN, RIDNet, AINDNet, InvDN, DANet, and DeamNet. All the codes of these methods are publicly available, and we have also cited these original papers. For the data, we have submitted the denoised images of our method on SIDD and DnD benchmarks to their online servers, which further verifies the effectiveness of our method. Note that ScaoedNet is called 'Scaoed_1' and ScaoedNet† is called 'Scaoed_2' on these sites. Please see these sites for details (SIDD: `http://www.cs.yorku.ca/~kamel/sidd/benchmark.php`; DnD: `https://noise.visinf.tu-darmstadt.de/benchmark/#overview`). In addition, we provide more experimental results and analysis about our method. Specifically, the following items are contained in our supplementary material:

S1. ***Algorithm 1****.*

S2. Details of CIG-SA and CIG-CA.

S3. Details of DnD and SIDD.

S4. FSM vs. SCA.

S5. Ablation on FSM and DEM.

S6. Study on Degradation Estimation Update.

S7. The Power of the New Cost Function.

S8. Convergence of ScaoedNet.

S9. Details of the Total Loss.

S10. Computational Complexity and Inference Time.

S11. Real-World Denoising on Nam.

S12. Importance of LS.

S13. Difference from Existing Deep Networks.

S14. Extension to Other Image Restoration Application.

S15. Limitations and Future Works.

---

[*] Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Algorithm 1 SCAOED algorithm

```
Input: Noisy image $\mathbf{y}$.
Initialization: Set iteration number $K$; Initialize the LS estimate $\mathbf{z}^{(0)} = E(\mathbf{y})$; Initialize the noise map $\mathbf{u}^{(0)}$ via $U_{\text{ini}}(\mathbf{y})$.
Main Iteration: Increment $i$ by 1 and perform the following steps:
for $i = 0$ to $K - 1$ do
    Obtain the guidance information by $\mathbf{g}^{(i)} = G(\mathbf{u}^{(i)})$;
    With given $\mathbf{u}^{(i)}$, $\mathbf{g}^{(i)}$, and $\mathbf{z}^{(i)}$, update the LS denoising estimate $\mathbf{z}^{(i+1)}$ by $\arg\min_{\mathbf{z}} \mathcal{H}(\mathbf{z}, \mathbf{u}^{(i)}, \mathbf{y}) + \tau\widetilde{\mathcal{G}}(\mathbf{z}, \mathbf{g}^{(i)}, \mathbf{z}^{(i)})$;
    if $i < K - 1$ then
        With given $\mathbf{u}^{(i)}$ and $\mathbf{z}^{(i+1)}$, update the noise map estimate $\mathbf{u}^{(i+1)}$ via $\arg\min_{\mathbf{u}} \mathcal{H}(\mathbf{z}^{(i+1)}, \mathbf{u}, \mathbf{y}) + \eta\widetilde{\psi}(\mathbf{u}, \mathbf{u}^{(i)})$;
    end if
end for
Obtain the denoising result $\hat{\mathbf{x}}$ via $D(\mathbf{z}^{(K)})$.
Output: Denoised image $\hat{\mathbf{x}}$.
```

S1. Algorithm 1

The details about the proposed SCAOED algorithm are given in this section. First, the noisy input $\mathbf{y}$ is transformed to the LS version $\mathbf{z}^{(0)} = E(\mathbf{y})$. In addition, we initially estimate the noise map $\mathbf{u}^{(0)}$ via $U_{\text{ini}}(\mathbf{y})$, and then get the guidance information $\mathbf{g}^{(0)}$ via $G(\mathbf{u}^{(0)})$. After that, we update the LS denoising result $\mathbf{z}^{(i+1)}$ via $\arg\min_{\mathbf{z}} \mathcal{H}(\mathbf{z}, \mathbf{u}^{(i)}, \mathbf{y}) + \tau\widetilde{\mathcal{G}}(\mathbf{z}, \mathbf{g}^{(i)}, \mathbf{z}^{(i)})$. Next, we update the noise map estimate $\mathbf{u}^{(i+1)}$ via $\arg\min_{\mathbf{u}} \mathcal{H}(\mathbf{z}^{(i+1)}, \mathbf{u}, \mathbf{y}) + \eta\widetilde{\psi}(\mathbf{u}, \mathbf{u}^{(i)})$. Then, we repeat the previous steps to iteratively update the guidance information, the estimate of the noise map, and the estimate of the clean image in LS. Finally, we can obtain the denoising result $\hat{\mathbf{x}}$ via $D(\mathbf{z}^{(K)})$.

S2. Details of CIG-SA and CIG-CA

For CIG-SA, the input feature map $\mathbf{a}$ is first processed by average pooling (Apool, between channels), maximum pooling (MPool, between channels), and a Conv layer. Then, 'Concat-Conv' layers are used to adjust the feature map. Next, the complementary feature information in channel dimension is calculated by the APool (within each channel), Conv, and Sigmoid layers. After that, the feature map is adjusted by the element-wise product with the complementary feature information in channel dimension, and followed by a skip connection to get the CIG spatial feature map. Finally, the spatial weights can be obtained by '1×1 Conv-Sigmoid'. For CIG-CA, a spatial information guided feature map is first obtained by the element-wise product of $\mathbf{a}$ and the weights calculated by 'Conv-Sigmoid'. Then, both the guided map and $\mathbf{a}$ are input into APool (within each channel) and MPool (within each channel) followed by a Concat layer to get the CIG channel feature map. After that, 'Conv-ReLU-Conv' layers are used for non-linear mapping. Then, the following 'Reshape-Conv-Reshape-Sigmoid' layers are used to map the previous output map to the channel weighting vector.

S3. Details of DnD and SIDD

DnD consists of 50 pairs of real clean-noisy scenes taken by different cameras with specifical settings, where 20 smaller images of size 512×512 are extracted from each scene. To capture these scenes, the base ISO level is used for taking the clean scenes while higher ISO and appropriately adjusted exposure time are used for taking the noisy scenes. To obtain the ground-truth images for the DnD benchmark, post-processing is further applied to the initial clean images, including small camera shift adjustment, linear intensity scaling, and removal of low-frequency bias. However, the corresponding ground-truth images are not available for users. The PSNR/SSIM results can only be obtained by uploading the denoised images to the DnD website.

SIDD contains about 30,000 real clean-noisy image pairs captured by five representative smartphones from 10 scenes under varying lighting conditions. For each noisy image, the corresponding noise-free

Table 1: PSNRs (dB) and SSIMs on SIDD validation dataset for ScaoedNet with FSM and SCA respectively.

Method	PSNR↑	SSIM↑
ScaoedNet with FSM	39.48	0.9186
ScaoedNet with SCA	39.39	0.9178

Table 2: Ablation on FSM and DEM for SIDD validation dataset.

FSM	✗	✗	✓	✓
DEM	✗	✓	✗	✓
PSNR↑/SSIM↑	39.28/0.9164	39.34/0.9167	39.35/0.9171	39.48/0.9186

image is estimated through some statistical methods [1]. However, only 320 clean-noisy image pairs (SIDD Medium Dataset) are provided for training. For testing, SIDD benchmark provides 40 noisy images without the corresponding clean images, and the PSNR/SSIM results can be obtained through its online server. In addition, SIDD validation also consists of 1280 clean-noisy image pairs to evaluate the denoising performance.

S4. FSM *vs.* SCA

The proposed FSM module can be regarded as an improved SCA module, which fully exploits the spatial and channel information. In fact, by removing the complementary information guidance in FSM, it becomes a traditional SCA module. Different from the existing SCA, complementary information guidance is introduced in FSM. To further prove the effectiveness of FSM, we compare it to SCA. Table 1 shows that higher PSNR/SSIM values can be obtained on the SIDD validation dataset by using FSM. The results indicate that the complementary information guidance can better inform the capture of the feature tensor information, which verifies the superiority of FSM over SCA.

S5.Ablation on FSM and DEM

The proposed attention module mainly consists of two parts, *i.e.*, FSM and DEM. Table 2 shows that ScaoedNet (*i.e.*, with both FSM and DEM) can achieve the best results compared with the ones without FSM, without DEM, and without FSM&DEM. For instance, by removing FSM, the PSNR/SSIM will decrease 0.14dB/0.0019. Similarly, by removing DEM, the performance will decrease 0.13dB/0.0015. Consequently, the effectiveness of FSM, DEM, and their combination are verified.

S6. Study on Degradation Estimation Update

Degradation estimation is important for the denoising performance of ScaoedNet, and the overall improvements of ScaoedNet partially come from more accurate retrieved noise map. In this section, we show the effectiveness of the update for the noise map estimation, and thus the effects of different settings of K for DE are tested. The results of PSNR and the L_1 distance are provided in Table 3, and we can conclude that with the increasing of DE networks ($\mathbf{u}$-Nets), the estimation result becomes more accurate. That means the update of degradation estimation is an effective scheme in degradation estimation, leading to higher performance. Therefore, such update scheme of the degradation estimation provides a convenient and effective means to progressively obtain the DE results.

Table 3: L_1 distances and PSNRs for verifying the effectiveness of the multiple DE networks. Note that during the calculations of the L_1 distance and PSNR, $\mathbf{u}$ is scaled to lie in the range of [0 255].

K	1	2	3	4	5
L_1 distance↓	5.80	5.01	4.48	4.27	4.20
PSNR↑	33.16	34.18	34.80	35.21	35.40

Table 4: Average PSNR (dB) and SSIM values of the denoised images by each stage in ScaoedNet.

Stage	1	2	3	4	5
PSNR↑	36.52	38.96	39.33	39.44	39.48
SSIM↑	0.8729	0.9140	0.9175	0.9184	0.9186

S7. The Power of the New Cost Function

The cost function mainly introduces LS, noise information, and GC. In Subsection **4.5**, ‘w/o LS’ verifies the power of LS; Noise information is introduced to handle the real denoising. In **S5** of ‘Supplementary Material’, we prove the power of noise information update; The implementation of GC: the reconstruction is guided by DEM with the guidance information $\mathbf{g}$ realized by G-Module. ‘w/o GC’ verifies the power of GC. Consequently, we can conclude that the proposed new cost function is effective in real image denoising.

S8. Convergence of ScaoedNet

In the implementation of ScaoedNet, five stages are used. To analyze the convergence of our ScaoedNet, we report the PSNR/SSIM results of stage 1, stage 2, stage 3, stage 4, and stage 5 in Table 4. Note that, in these stages, the parameters of all the DE networks ($\mathbf{u}$-Nets) (except the initial one $\mathbf{u}_{\text{ini}}$-Net) are shared, and the parameters of all the RE networks ($\mathbf{z}$-Nets) are also shared. Therefore, the LS decoding module D can be directly placed at the end of each RE network ($\mathbf{z}$-Net) to obtain the denoising result. In addition, the visual results are also provided in Fig. 1. We can find that the later iteration can generate better results than the previous iterations step-by-step on SIDD validation dataset, which verifies the convergence of ScaoedNet.

S9. Details of the Total Loss

According to Eq. (6) in the main body of the paper, our total loss contains two main parts: the reconstruction loss and the noise map estimation loss.

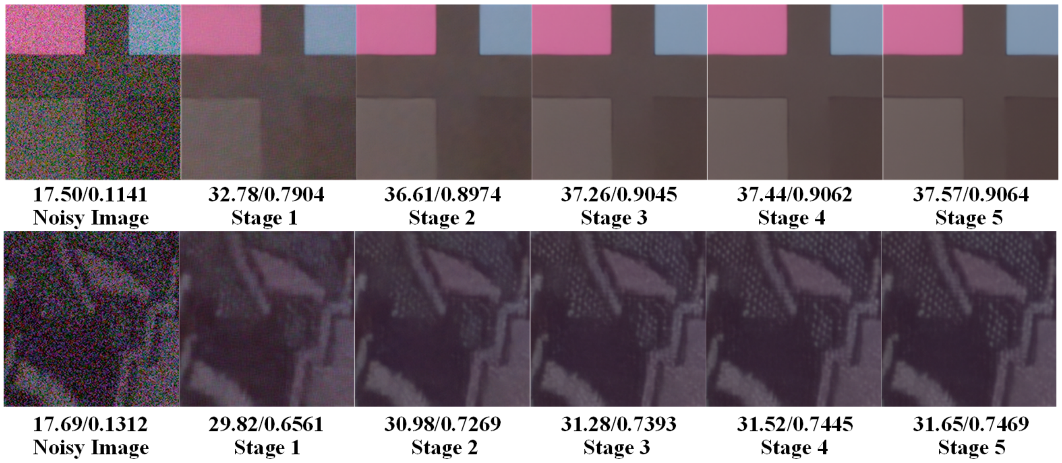
The reconstruction loss, i.e., $\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{g=1}^{N}(\|D(\mathbf{z}_g^{(K)}) - \mathbf{x}_g\|_1^1 + \gamma\mathcal{L}_{\text{S}}(D(\mathbf{z}_g^{(K)}), \mathbf{x}_g))$, further contains two sub-parts: 1) L_1 loss between the output and the ground-truth image to ensure their similarity in pixel-level; 2) the structural similarity loss $\mathcal{L}_{\text{S}}$ to pay attention to image structures. In other words, in the reconstruction loss, both the pixel-level constraint and the structure-level constraint are considered simultaneously to ensure the quality of reconstruction.

The noise map estimation loss is $\frac{\eta}{N}\sum_{i=1}^{K-1}\sum_{g=1}^{N}\|\boldsymbol{\kappa}_{i,g}\otimes(\hat{\mathbf{u}}_g^{(i)} - \mathbf{u}_g)\|_1^1$. That means the estimated noise maps of the first to $(k-1)$th degradation estimation (DE) networks are constrained to be close to the ground-truth one.

The weight of each stage is determined by $\boldsymbol{\kappa}_{i,g} = \nu_i \cdot \boldsymbol{\alpha}_g$, where ν_i is the weight for the ith DE network. Considering that in multiple DE stages, the later DE network will produce more accurate estimation, and thus a larger weight should be assigned. Therefore, the geometric sequence with a common ratio ι (greater than 1) and sum 1 can be used for ν_i-s, i.e., $\nu_i = (\iota - 1)\iota^{i-1}/(\iota^{K-1} - 1)$. In addition, the parameter $\boldsymbol{\alpha}_g$ is the gth element of the indicator vector $\boldsymbol{\alpha}$ for the noise constraint.

Table 5: The average PSNR(dB)/SSIM results of different denoising methods on Nam dataset.

DnCNN[21]	VDN[19]	CBDNet[6]	RIDNet[2]	DANet[20]	ScaoedNet	ScaoedNet†
35.54 0.8785	38.72 0.9628	39.17 0.9667	39.12 0.9615	39.83 0.9686	41.41 0.9762	41.54 0.9786


17.50/0.1141 Noisy Image **32.78/0.7904 Stage 1** **36.61/0.8974 Stage 2** **37.26/0.9045 Stage 3** **37.44/0.9062 Stage 4** **37.57/0.9064 Stage 5**

17.69/0.1312 Noisy Image **29.82/0.6561 Stage 1** **30.98/0.7269 Stage 2** **31.28/0.7393 Stage 3** **31.52/0.7445 Stage 4** **31.65/0.7469 Stage 5**

Figure 1: PSNR(dB)/SSIM and visual results of ScaoedNet in each stage.

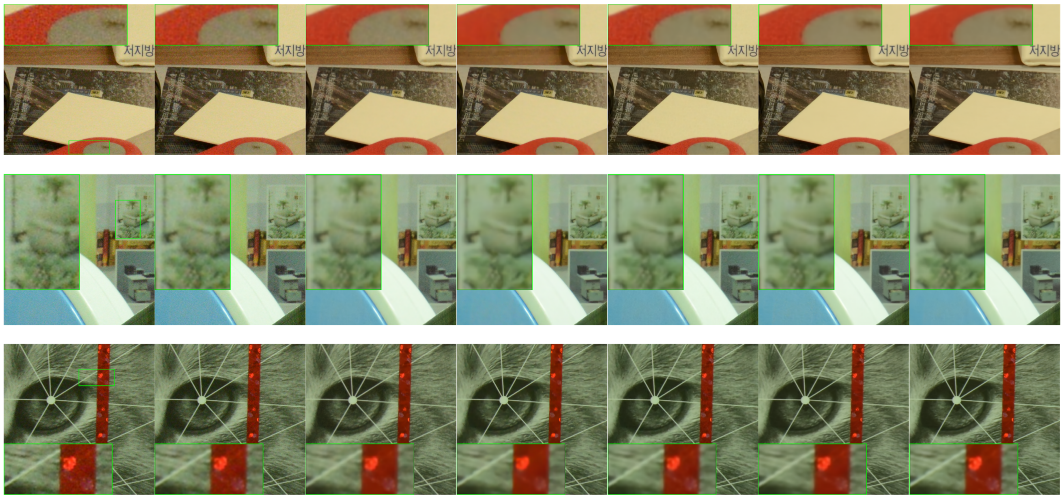
The reason for introducing $\boldsymbol{\alpha}$ is given in the following. For the commonly used training datasets like SIDD, RENOIR, etc., in real denoising, the specific noise maps $\mathbf{u}_g$-s are unknown, and thus cannot be used for training the DE network. In this case, the noise map estimation loss should be invalidated by setting its weight to 0; to update the parameters of the DE networks, we need to synthesize the dataset according to the real noise model established in Eq. (2) for training. Therefore, when the training data is synthetic with a known noise map, the weight for the noise map estimation loss should be 1. This is why we call $\boldsymbol{\alpha}_g$ the gth element of the indicator vector $\boldsymbol{\alpha}$ for the noise constraint.

S10. Computational Complexity and Inference Time

According to subsection **4.7**, we can see that ScaoedNet has a moderate parameter number (about 3M), which is significantly lower than CBDNet, VDN, AINDNet, and DANet. We have included FLOPs and inference time so that the readers could better understand the network cost. The FLOPS and inference time of ScaoedNet with 1 stage, 3 stages and 5 stages for output resolution 512×512 are 212G/0.07s, 640G/0.15s, and 1071G/0.28s, respectively. For the second-best method DeamNet, it is 589G/0.18s. Moreover, for 256×256 images, the FLOPs and inference times of ScaoedNet with 1 stage, 3 stages and 5 stages are 53G/0.02s, 160G/0.05s, and 268G/0.08s, respectively. For the second-best method DeamNet, it is 146G/0.05s. We can further reduce the complexity by using the multi-scale encoder-decoder based UNet architecture similar to DeamNet with four scales, where each encoder or decoder can consist of two FM^{2}ARBs. Then, the FLOPs in this case for 256×256 images will reduce to about 130G. We should note that the second-best method DeamNet is also an iterative network. The results show that our method can achieve higher PSNR/SSIM values than the iterative network DeamNet only with slightly longer but reasonable inference time.

S11. Real-World Denoising on Nam

The real-world noise dataset Nam [15] includes various noisy images from 11 static scenes, whose ground-truth versions are captured by the mean of 500 noisy images of the same scene. We crop the images into 512×512 sub-images and randomly select 100 sub-images from those for testing. Different from SIDD and DnD, the images from Nam are JPEG compressed. During the generation



Noisy Image **DnCNN** **VDN** **CBDNet** **RIDNet** **DANet** **ScaoedNet**

Figure 2: Real denoising results on Nam dataset.

of the synthetic clean-noisy image pairs, the synthetic noisy images are further JPEG compressed with the quality factor sampled from the range of [60, 100]. In addition, we add Poly dataset [18] for training according to DANet [20]. Both the PSNR/SSIM and the visual results are provided in Table 5 and Fig. 2. Note that many works are not designed for the noisy images with JPEG compression, and only several methods (*e.g.*, CBDNet [6], RIDNet [2], and DANet [20]) have taken the JPEG compression into consideration. Following DANet [20], the deep network based methods DnCNN [21], VDN [19], CBDNet [6], RIDNet [2], DANet [20] are tested on Nam. We can observe that our performance in terms of PSNR/SSIM is higher than any of the previous algorithms. For the visual results, the noise in the denoised images of DnCNN [21] are not well removed. Other methods, *i.e.*, VDN [19], CBDNet [6] and RIDNet [2], although can achieve better results than DnCNN, there still exist certain artifacts. Among all the competing methods, DANet [20] can achieve the second best denoising performance. In comparison, the noise in the recovered images by ScaoedNet is well suppressed. Consequently, the effectiveness of the proposed method for real-world image denoising is further confirmed.

S12. Importance of LS

LS has the following advantages: 1) better representation ability than the low-dimensional space; 2) allows better information flow in the unfolding network.

Theoretically, without using LS, Eq. (5) in the main body of the paper becomes:

$$\begin{cases} \mathbf{g}^{(i)} = G(\mathbf{u}^{(i)}), \\ \mathbf{x}^{(i+1)} = \arg\min_{\mathbf{x}} \mathcal{H}(\mathbf{x}, \mathbf{u}^{(i)}, \mathbf{y}) + \tau \widetilde{\mathcal{G}}(\mathbf{x}, \mathbf{g}^{(i)}, \mathbf{x}^{(i)}), \\ \mathbf{u}^{(i+1)} = \arg\min_{\mathbf{u}} \mathcal{H}(\mathbf{x}^{(i+1)}, \mathbf{u}, \mathbf{y}) + \eta \widetilde{\psi}(\mathbf{u}, \mathbf{u}^{(i)}). \end{cases} \tag{1}$$

It is obvious that, denoising is performed in the pixel domain to reconstruct $\mathbf{x} \in \mathbb{R}^{n \cdot c}$. It will lead to the $\mathbf{z}$-Net with a low input/output channel dimension c, reducing the information flow in the network.

With LS in Eq. (5), denoising is performed in LS to reconstruct $\mathbf{z} \in \mathbb{R}^{n \cdot m}$. Since $m > c$, $\mathbf{z}$-Net has better representation ability. Moreover, the channel number between two consecutive $\mathbf{z}$-Nets becomes larger, and thus the network information flow is enhanced. All of these will benefit the performance.

We can also observe from Sec. 4.5 that LS is important. Moreover, by visualizing the features after E, we can find that the embeddings are the hierarchical high-dimensional features of the noisy image, where the noise and image components can be easily decoupled, which can effectively promote the performance.

S13. Difference from Existing Deep Networks

S13.1. Differences from Existing Deep Denoising Networks

A large number of deep denoising methods have been proposed in recent years, such as TNRD [3], UNLNet [10], FFDNet [22], FOCNet [7], MemNet [17], DnCNN [21], RED [13], DPDNN [5], CBDNet [6], VDN [19], RIDNet [2], AINDNet [8], InvDN [11], DANet [20], BCDNet [9], GainTuning [14], and DeamNet [16], *etc.* However, the network design of our ScaoedNet is significantly different from those existing methods.

ScaoedNet belongs to the unfolding-based methods. With respect to the theoretical novelty, we first define a novel enhanced model-based denoising cost function to improve the traditional denoising method by introducing the concepts of LS, noise information introduction, and GC. Then, we split the previous problem into three alternative optimization subproblems (*i.e.*, GR, DE and RE subproblems) via the SC scheme. Finally, the iterative optimization steps of the model-based denoising method are utilized to inform the network design, leading to an end-to-end trainable and interpretable denoising network (ScaoedNet). The proposed ScaoedNet consists of LS encoding module (E), LS decoding module (D), G-module, RE network, and DE network.

For the network novelty, a novel NFAE layer is proposed to further enhance noise feature with parameter-free scheme, which leads to better noise estimation results. To the best of our knowledge, we are the first to propose the NFAE layer for noise map estimation. In the RE network, a novel feature multi-modulation attention residual block called FM^{2}ARB (consists of FSM and DEM) is proposed to synchronously and dynamically fuse the encoded degradation representation and the image feature representation, leading to higher denoising performance.

In contrast, the designs of many deep denoising networks are not informed by the traditional methods and are non-iterative, including FFDNet [22], MemNet [17], DnCNN [21], RED [13], CBDNet [6], RIDNet [2], AINDNet [8], InvDN [11], BCDNet [9], and DANet [20]. Although UNLNet [10], TNRD [3], DPDNN [5], VDN [19], FOCNet [7], and DeamNet [16] also use some traditional methods to inform their network designs, their mathematical foundations are quite different. Specifically, UNLNet [10] is based on the non-local variational operator, TNRD [3] is based on the iterative nonlinear reaction diffusion, DPDNN [5] is based on the plug-and-play framework, VDN [19] is based on the variational inference method, FOCNet [7] is based on the fractional optimal control theory, and DeamNet [16] is based on the adaptive consistency prior. Some other works focus on how to improve network generalization for various data distributions, such as [14] which proposes a 'GainTuning' methodology for adaptive denoising. In summary, the mathematical foundation and the network architecture design of the proposed ScaoedNet are significantly different from the existing deep networks.

S13.2. Differences from Deep Alternating Network (DAN) [12]

Although both DAN and our method adopt alternative optimization (different tasks, i.e., super-resolution (SR) and denoising), their theoretical novelties and the network novelties are much different.

Theoretical Differences. 1) DAN performs SR in low-dimensional pixel space, while ours performs denoising in high-dimensional LS; 2) DAN only considers degradation estimation and reconstruction. In addition to degradation estimation ($\mathbf{u}$-Net) and reconstruction ($\mathbf{z}$-Net), we introduce guidance representation (GR, G-Module) to improve performance; 3) Traditional alternating optimization is used in DAN. But we propose a novel self-correction (SC) alternating optimization method.

Network Differences. 1) By using SC, our degradation estimation and reconstruction networks can better exploit the last estimates $\mathbf{z}^{(i)}$ and $\mathbf{u}^{(i)}$ for higher performance than DAN that is without SC; 2) We introduce G-Module to guide denoising, which is not considered in DAN; 3) We propose the FM^{2}A module with FSM and DEM, which is different from DAN; 4) For $\mathbf{u}$-Net, we propose a novel parameter-free NFAE layer, which is not used in DAN.

S13.3. Differences from Noise Basis Network (NBNet) [4]

Mapping noise image to space that better distinguish noise has also been proposed in NBNet. The differences between our method and NBNet are given in the following:

Difference Between Using Distinguishable Space. 1) The space projection in NBNet is achieved by the subspace attention (SSA) module, which needs to construct the basis vector for obtaining the projection matrix. It is essentially an attention module. Our method directly using E to project the input image to the distinguishable space without using attention mechanism; 2) NBNet uses SSA to project the output of each decoder module separately in a UNet-based architecture, and thus multiple SSA modules are exploited. However, our method only needs to introduce the latent space (LS) at the beginning of the whole network to ensure the whole reconstruction process is carried out in the high-dimensional LS; 3) SSA in NBNet requires two inputs, i.e., the low-level feature from skip-connection and the upsampled high-level feature. Specifically, the low-level feature from skip-connection is projected into the signal subspace guided by the upsampled high-level feature. However, our method only needs one single input, i.e., the input noisy image.

Superiority of Our Usage. 1) To obtain the projection in NBNet, a lot of matrix operations are required. But the projection of our method can be more easily obtained by the E module, avoiding matrix operations; 2) NBNet requires two inputs, where one is used as the guidance for the other one. Thus, it cannot be directly used in our work since only one noisy input is available at the beginning of our network. However, since LS only needs a single input, it can directly perform projection on the noisy input image at the beginning of the network.

S14. Extension to Other Image Restoration Application

Our method can be potentially applied to other image restoration (IR) problems. We take image deblurring problem as an example. Image deblurring can be expressed as:

$$\mathbf{y} = \mathbf{x} * \mathbf{k} + \mathbf{n} \tag{2}$$

where $\mathbf{y}$ is the degraded image, $\mathbf{x}$ is the ground-truth image, $\mathbf{k}$ is the blur kernel, $\mathbf{n}$ is the noise, and $*$ denotes the convolutional operation. Then, Eq. (3) in the main body of the paper becomes:

$$\left\{\hat{\mathbf{z}}, \hat{\mathbf{u}}, \hat{\mathbf{k}}\right\} = \arg\min_{\mathbf{z},\mathbf{u},\mathbf{k}} \mathcal{H}(\mathbf{z}, \mathbf{u}, \mathbf{k}, \mathbf{y}) + \tau\mathcal{G}(\mathbf{z}, \mathbf{g}) + \eta_1\psi(\mathbf{u}) + \eta_2\phi(\mathbf{k}), s.t., \hat{\mathbf{x}} = D(\hat{\mathbf{z}}) \tag{3}$$

where τ, η_1, η_2 are the weights for the regularizers $\mathcal{G}(\mathbf{z}, \mathbf{g})$ (guidance constraint, GC), $\psi(\mathbf{u})$ (noise map prior), and $\phi(\mathbf{k})$ (blur kernel prior). By using our self-correction (SC) alternative optimization, we can obtain

$$\begin{cases} \mathbf{g}^{(i)} & = G(\mathbf{u}^{(i)}, \mathbf{k}^{(i)}), \\ \mathbf{z}^{(i+1)} & = \arg\min_{\mathbf{z}} \mathcal{H}(\mathbf{z}, \mathbf{u}^{(i)}, \mathbf{k}^{(i)}, \mathbf{y}) + \tau\widetilde{\mathcal{G}}(\mathbf{z}, \mathbf{g}^{(i)}, \mathbf{z}^{(i)}), \\ \mathbf{u}^{(i+1)} & = \arg\min_{\mathbf{u}} \mathcal{H}(\mathbf{z}^{(i+1)}, \mathbf{u}, \mathbf{k}^{(i)}, \mathbf{y}) + \eta_1\widetilde{\psi}(\mathbf{u}, \mathbf{u}^{(i)}), \\ \mathbf{k}^{(i+1)} & = \arg\min_{\mathbf{k}} \mathcal{H}(\mathbf{z}^{(i+1)}, \mathbf{u}^{(i+1)}, \mathbf{k}, \mathbf{y}) + \eta_2\widetilde{\phi}(\mathbf{k}, \mathbf{k}^{(i)}), \end{cases} \tag{4}$$

where $G(\cdot)$ is the guidance information generator, $\widetilde{\mathcal{G}}(\cdot)$ becomes the joint constraint of GC and SC for $\mathbf{z}$, $\widetilde{\psi}(\cdot)$ becomes the joint constraint of noise information and SC for $\mathbf{u}$, and $\widetilde{\phi}(\cdot)$ becomes the joint constraint of blur kernel and SC for $\mathbf{k}$.

By comparing this equation with Eq. (5) in the main body of the paper, we can find the difference between the denoising and deblurring using our method is: in addition to the estimation of $\mathbf{z}$, $\mathbf{u}$, we have to estimate the blur kernel information $\mathbf{k}$ and construct the new guidance information $\mathbf{g}$ for deblurring. How to construct the $\mathbf{k}$ estimation module and the guidance information generator for deblurring is very important. Other image restoration problems can be solved similarly.

S15. Limitations and Future Works

In this section, we analyze the limitations and the future works of the proposed method. First, the self-attention module is very promising in image restoration, but the parameter number and complexity will be much larger than those of our current method. How to use self-attention more efficiently in ScaoedNet is one of our future works. Since the proposed ScaoedNet is derived from the enhanced model-based real image denoising method, it is limited to the real image denoising task. In fact, this work can be naturally transferred to other image restoration tasks, such as real image deblurring, real image super-resolution, and so on. In these cases, a key and challenging problem is how to design a degradation estimator that can estimate the blurring degradation and noise degradation of the low-quality input image synchronously, and this is one of our future works. In addition, the real noise model used in the proposed method is based on CBDNet. Other powerful real noise models can be introduced for further improving the generalization and performance of ScaoedNet. Therefore, another future work is to extend the generalization by more powerful real noise model, so as to increase its future application potential.

References

[1] A. Abdelhamed, S. Lin, and M. S. Brown. A high-quality denoising dataset for smartphone cameras. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1692–1700, Jun. 2018.

[2] S. Anwar and N. Barnes. Real image denoising with feature attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3155–3164, Oct. 2019.

[3] Y. Chen and T. 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, 2017.

[4] S. Cheng, Y. Wang, H. Huang, D. Liu, H. Fan, and S. Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4896–4906, Jun. 2021.

[5] W. Dong, P. Wang, W. Yin, and G. Shi. Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10): 2305–2318, 2019.

[6] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang. Toward convolutional blind denoising of real photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1712–1722, Jun. 2019.

[7] X. Jia, S. Liu, X. Feng, and Z. Lei. Focnet: A fractional optimal control network for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, Jun. 2019.

[8] Y. Kim, J. W. Soh, G. Y. Park, and N. I. Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3482–3492, Jun. 2020.

[9] K. Ko, Y. J. Koh, and C.-S. Kim. Blind and compact denoising network based on noise order learning. *IEEE Transactions on Image Processing*, 31:1657–1670, 2022.

[10] S. Lefkimmiatis. Universal denoising networks : A novel cnn architecture for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, Jun. 2018.

[11] Y. Liu, Z. Qin, S. Anwar, P. Ji, D. Kim, S. Caldwell, and T. Gedeon. Invertible denoising network: A light solution for real noise removal. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13365–13374, Jun. 2021.

[12] Z. Luo, Y. Huang, S. Li, L. Wang, and T. Tan. Unfolding the alternating optimization for blind super resolution. In *Advances in Neural Information Processing Systems*, volume 33, pages 5632–5643, 2020.

[13] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. *Advances in neural information processing systems (NeurIPS)*, 29:2802–2810, 2016.

[14] S. Mohan, J. L. Vincent, R. Manzorro, P. Crozier, C. Fernandez-Granda, and E. Simoncelli. Adaptive denoising via gaintuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 23727–23740, 2021.

[15] S. Nam, Y. Hwang, Y. Matsushita, and S. J. Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1683–1691, Jun. 2016.

[16] C. Ren, X. He, C. Wang, and Z. Zhao. Adaptive consistency prior based deep network for image denoising. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8596–8606, Jun. 2021.

[17] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *IEEE international conference on computer vision (ICCV)*, pages 4539–4547, Oct. 2017.

[18] J. Xu, H. Li, Z. Liang, D. Zhang, and L. Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018.

[19] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang. Variational denoising network: Toward blind noise modeling and removal. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1690–1701, 2019.

[20] Z. Yue, Q. Zhao, L. Zhang, and D. Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *European Conference on Computer Vision (ECCV)*, pages 41–58, Sep. 2020.

[21] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7): 3142–3155, 2016.

[22] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.