Multi-Scale Selective Feedback Network with Dual Loss for Real Image Denoising

Xiaowan Hu¹,², Yuanhao Cai¹, Zhihong Liu¹, Haoqian Wang¹,²* and Yulun Zhang³

¹The Shenzhen International Graduate School, Tsinghua University, China
²The Shenzhen Institute of Future Media Technology, Shenzhen 518071, China
³Northeastern University, MA 02115, US

{hu-xw19, cyh20, liuzh18}@mails.tsinghua.edu.cn, wanghaoqian@tsinghua.edu.cn, yulun100@gmail.com

Abstract

The feedback mechanism in the human visual system extracts high-level semantics from noisy scenes. It then guides low-level noise removal, which has not been fully explored in image denoising networks based on deep learning. The commonly used fully-supervised network optimizes parameters through paired training data. However, unpaired images without noise-free labels are ubiquitous in the real world. Therefore, we proposed a multi-scale selective feedback network (MSFN) with dual loss. We allow shallow layers to access valuable contextual information from the following deep layers selectively between two adjacent time steps. Iterative refinement mechanism can remove complex noise from coarse to fine. The dual regression is designed to reconstruct noisy images to establish closed-loop supervision that is training-friendly for unpaired data. We use the dual loss to optimize the primary clean-to-noisy task and the dual noisy-to-clean task simultaneously. Extensive experiments prove that our method achieves state-of-the-art results and shows better adaptability on real-world images than the existing methods.

1 Introduction

There are complex unknown noises during processing, storage, and transmission in the real-world image acquisition system. The complicated electrical noise overwhelms the image details and causes the image quality to deteriorate. Most of the existing denoising methods are based on known synthesized noise and often have poor performance in real-world images [Kim et al.2020, Anwar and Barnes2019].

Known noisy images have infinite noisy-to-clean mappings in the solution space, making image denoising an ill-posed task. Deep learning networks can learn complex end-to-end mappings and have been widely used in image denoising tasks [Zhang et al.2018, Zhang et al.2020]. The recursive structure deepens the network and increases the receptive field [Tai et al.2017, Liu et al.2018]. Although residual learning alleviates the disappearance of gradients and accelerates the optimization of loss [Zhang et al.2019, Zhang et al.2020], too many skip connections reuse destructive noise information. The bottom layer cannot access valuable contextual information. A feed-forward network that only performs one-step prediction has poor fitting ability to complex noise. The feedback mechanism is widespread in the human visual system and has been widely used in high-level computer vision tasks. The top-down structure forces the shallow units to carry high-level semantic information and, in turn, guides noise removal. Single-to-single and multiple-to-multiple feedback structures have been explored for image super-resolution [Li et al.2019b, Li et al.2019a] tasks but are rarely used for image denoising. Severely degraded images put forward higher requirements on pixel prediction. Image denoising needs to balance valuable high-level semantic information with accurate low-level image details.

It is difficult for a single-stage supervised image denoising network to find the best mapping from noise to clean in the infinite solution space. The two-stage methods for image denoising include noise estimation and removal. But the two-step structure increases network complexity and accumulates errors inevitably. In network optimization, most end-to-end networks strive to minimize the difference between clean labels and the predicted images. However, it is challenging to predict complex degradation with only one-step supervision in the clean domain. Due to the lack of noise-free labels, many unpaired real-world noise images that are easily available cannot participate in general fully-supervised training, which limits network performance and adaptability to real
noise. Networks that rely excessively on high-quality images are likely to overfit the synthetic noise. Dual learning [He et al.2016] uses closed-loop dual tasks to optimize the network. Adding additional supervision in the noise domain can make the network more robust to various noise. Some methods used for high-level tasks such as CycleGAN [Zhu et al.2017] and DualGAN [Yi et al.2017] discard a large number of task-independent image details, so they cannot be directly transferred to the image-to-image denoising task.

We design a novel multi-scale selective feedback network (MSFN) with the dual loss for real image denoising based on the above discussion. The commonly used one-step prediction is replaced by a multi-level iterative prediction, which guides image restoration from coarse to fine adaptively. As shown in Fig. 1(a), the multi-scale selective feedback block (MSFB) is unfolded in time. We count each iteration into the total loss to ensure that the hidden unit to be fed back contains reliable high-level functions. Using too many high-level semantic features will overwhelm the original low-level information. The selective feedback mechanism is proposed to analyze principal feature components. We reduce feature redundancy through principal component analysis (PCA), which captures the most representative information. The multi-scale selective block (MSB) integrates multi-scale features for changing receptive fields dynamically. The step-by-step learning strategy is proven to be more suitable for fitting various real-world degradations accurately.

As shown in Fig. 1(b), we decompose the noise removal task into closed-loop dual tasks. The primary task predicts noise-free images \( \hat{Y} \) from the noise image \( X \), and the dual regression task restores the original noise image \( \hat{X} \) from \( \hat{Y} \) reversely. We introduce a weighting mechanism to design dual loss to guide network optimization. The closed-loop dual-domain supervision can reduce the solution space and alleviate the model’s excessive dependence on clean images. For paired data, there is a complete dual loss. For unpaired data, the loss function only includes supervision in the noise domain. Unpaired real-world images can directly participate in network training. We can expand diversified noise types in the training set freely and effectively, which reduces the risk of network overfitting to a specific noise level. In summary, our contribution has the following three points:

- The proposed MSFN allows shallow layers to access valuable information from the following deep layers selectively. Combining bottom-up multi-level intermediate supervision and the top-down feedback mechanism can remove complex noise from coarse to fine.
- We design a dual weighted loss combining clean-to-noisy and noisy-to-clean tasks. The closed-loop multi-domain supervision is easier to learn the best mapping in an infinite solution space. Dual loss can perform end-to-end training on unpaired data, which has better generalization and adaptability to real-world noise images.
- Dual loss and feedback learning strategies can be adapted to different denoising tasks, including complex degraded and unsupervised images. Extensive experiments prove that our proposed network achieved the best denoising performance in multiple synthetic noise images and paired or unpaired real-world noise images.

### 2 Proposed Method

#### 2.1 Overall Network Architecture

The feedback network contains two essential elements: it-erability and the reverse connection from deep to shallow. As shown in Fig. 3(a), we designed the multi-scale selection block (MSB) to extract features with multiple receptive fields and the selective feedback module (SFM) to capture valuable deep features to assist shallow denoising adaptively. Our proposed multi-scale selective feedback network (MSFN) can be expanded into an independent convolutional neural network with \( T \) iterations. We train two regression tasks in the noise domain and the clean domain jointly at each time step.

We define the input image of each iteration in MSFN as \( I_{noisy} \). Firstly, we use two convolutional layers with convolution kernel sizes of 3 \( \times \) 3 and 1 \( \times \) 1 to extract the initial shallow feature \( F_{N,0} \), which can be expressed as:

\[
F_{N,0} = MSFE(I_{noisy}), \tag{1}
\]

where \( MSFE(\cdot) \) represents the shallow feature extraction (SFE) function. After that, the initial feature is transferred to the recurrent structure that contains the stacked MSBs. The MSB fuses multi-scale information by changing receptive fields and selects useful features for image denoising adaptively and dynamically. The MSB is detailed in Section 2.3.

As shown in Fig. 3(a), when \( t = 0 \), there is no high-level information fed back from the previous step. The feed-forward shallow layers access useful information from the following layers when \( 0 < t \leq T \). Assuming that the number of stacked MSBs in the feed-forward network is \( B \), then the feature \( F_{N,B} \) output by the last MSB in each time step is expressed as:

\[
F_{N,B} = MSFB(F_{N,0}), \tag{2}
\]

where \( MSFB(\cdot) \) is the function that combines \( B \) MSBs and \( M \) SFMs. The principal components of high-level features from the previous time step are fused with the shallow features adaptively. Dynamically aggregated context allows top-down and bottom-up real-time knowledge exchange.

At the end of each time step, the reconstruction function is defined as \( M_R(\cdot) \), which includes two convolutional layers and a residual skip connection. The final output of our denoised image in the \( t \) -th iteration can be expressed as:

\[
l_{clean} = M_R(F_{N,B}) + I_{noisy}. \tag{3}
\]

We visualize the iterative restoration process in Fig. 2. The structural edges and textures are refined step by step, proving that the feedback hierarchical learning strategy can reconstruct high-quality details from coarse to fine.

In each time step, we design a dual regression task to constrain the pixel prediction further. As shown by the red arrows...
where the DM function $M_{DM}()$ remaps the clean image to the image $I_{noisy}$ in the noise domain. The degradation process of the $t$-th iteration is defined as:

$$
\hat{I}_{noisy} = M_{DM}(I_{clean}'),
$$

$$
M_{DM}(I_{clean}') = M_{DB}^{t}(M_{DB}^{t-1}(\ldots M_{DB}^{1}(I_{clean}')),
$$

where $M_{DB}()$ represents the DB, and the convolution kernel sizes of the convolutional layer are $3 \times 3$, $3 \times 3$, and $1 \times 1$, respectively, and the activation layer LeakyReLU follows the first convolutional layer. We set $k = 8$ and the DM output degraded noisy image predicted by the dual regression task.

In the primary regression, we get $T$ predicted noise-free images. Simultaneously, other $T$ noise images are restored by dual regression tasks. We design a novel asymmetric dual loss to optimize the dual model jointly. As shown in Fig. 3(a), the dual loss is weighted by $Loss_{ primal}$ in the clean domain and $Loss_{dual}$ in the noisy domain, which can be defined as:

$$
L(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left[ \lambda_1 \left| \left| I_{GT} - I_{clean}' \right| \right|_{1} + \lambda_2 \left| \left| I_{noisy} - \hat{I}_{noisy} \right| \right|_{1} \right],
$$

where $\lambda_1$ and $\lambda_2$ can guide the task optimization with attention.

### 2.2 Selective Feedback Module

Integration of all unprocessed high-level feedback will bring information redundancy, which will overwhelm the original shallow features. We improved the multi-level feedback and showed the internal implementation details in Fig 3(b).

Firstly, we use the selective mechanism (SM) to analyze the principal feature components. The correlated high-dimensional features are transformed into mutually independent low-dimensional information. We perform principal component analysis (PCA) on the fused features. We decompose the singular value of the feature $X \in \mathbb{R}^{h \times w \times c}$ as:

$$
X = U_{(h \times w) \times (h \times w)} A_{(h \times w) \times (c/r)} V_{(c/r) \times (c/r)}^{T},
$$

where $\hat{X} \in \mathbb{R}^{(h \times w) \times (c/r)}$. For the $t$-th iteration, the selected potential clean feature $F_{C,b}^{t}$ to be fed back is expressed as:

$$
F_{C,b}^{t} = M_{PCA}\left([F_{N,b-1}^{t-1}, \ldots, F_{N,B-1}^{t-1}, F_{N,B}^{t-1}]\right),
$$

where $F_{N,b-1}^{t-1}, \ldots, F_{N,B-1}^{t-1}, F_{N,B}^{t-1}$ is the high-level feedback features from different depths. The number of feedback branches is defined as $m$, i.e., $m = B - b$. These high-level components are concatenated into a whole in the channel direction.

For the $b$-th MSB, we fuse the feedback feature $F_{C,b}^{t}$ and feed-forward feature $F_{N,b}^{t}$ by the fusion block (FB) as:

$$
F_{S,b}^{t} = M_{FB}\left(F_{C,b}^{t}, F_{N,b}^{t}\right),
$$

where the $F_{S,b}^{t}$ is the refined low-level features.
2.3 Multi-scale Selective Block
The changing receptive field brings more contextual information. We design a selective mechanism to adjust the self-attention of the multi-scale feature. As shown in Fig 3(c), the MSB extracts features through three convolutional layers with different convolution kernels and aggregates multiple branches according to the significance dynamically.

Specifically, the MSB contains three branches with convolution kernels of $3 \times 3$, $5 \times 5$, and $7 \times 7$. For the $b$-th MSB, the multi-scale information of feature $F_{S,b}^t$ is described as:

$$f_3 = M_{3 \times 3}^t(F_{S,b}^t),$$ (10)

$$f_5 = M_{5 \times 5}^t(F_{S,b}^t),$$ (11)

$$f_7 = M_{7 \times 7}^t(F_{S,b}^t),$$ (12)

where $f_3$, $f_5$, and $f_7$ are the outputs of each path.

We use the global average pooling and LeakyReLU activation function to squeeze and excite the fused feature. The output global feature descriptor $F_{f,b}$ is:

$$F_{f,b} = M_{\text{GAP}}^t(M_{1 \times 1}^t(\text{LeakyReLU}(f_3 + f_5 + f_7))).$$ (13)

The self-attention mapping (SAM) recalibrates the correlation of each branch. The function $M_{\text{SAM}}$ includes convolution and the softmax regression to map the descriptor as:

$$[a_3, a_5, a_7] = M_{\text{SAM}}(F_{f,b}^t),$$ (14)

where $a_3$, $a_5$, and $a_7$ are the attention vectors. The valuable and representative information at each scale is remapped as:

$$F_{S,b}^t = f_3 a_3 + f_5 a_5 + f_7 a_7.$$ (15)

The multi-scale features are selected and gathered for critical cross-layer and cross-scale information interaction.

3 Experiments

3.1 Datasets

Synthetic Noisy Images. The DIV2K [Timofte et al.2017] and Flickr2K [Lim et al.2017] datasets are used to generate synthetic noise images. There are 3100 images used for training and 350 images for verification. We add white Gaussian noise with levels of $\sigma = 10, 30, 50, 70$ for paired data. We will evaluate the trained network on common benchmark datasets, including BSD68, Kodak24, and Urban100.

Real Noisy Images. The real-world noise images come from the SIDD [Abdelhamed and Lin2018]. Specifically, SIDD includes 10 static scenes of 5 smartphones with different shooting pipeline settings, which are under different lighting conditions. The SIDD-Medium dataset that we used consists of 320 image pairs. We additionally use three other public real-world noise data sets for model testing, including PolyU [Xu et al.2018a], Nam [Nam et al.2016], and Darmstadt Noise Dataset (DND) [Plotz and Roth2017]. The unpaired data used for adaptive training come from the SIDD-Full dataset.

To verify the proposed noise domain supervision effectiveness, we select 200 images that do not overlap with the original training set. It should be noted that unpaired data do not use clean reference images in the training process.

Table 1: Quantitative results about color image denoising. Best results are highlighted.

<table>
<thead>
<tr>
<th>Method</th>
<th>Kodak24</th>
<th>BSD68</th>
<th>Urban100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>CBM3D [Dabov et al.2007]</td>
<td>35.71</td>
<td>28.98</td>
<td>25.33</td>
</tr>
<tr>
<td>TNRD [Chen and Pock2017]</td>
<td>34.33</td>
<td>28.83</td>
<td>27.17</td>
</tr>
<tr>
<td>DrCNN [Zhang et al.2017a]</td>
<td>36.98</td>
<td>31.39</td>
<td>29.16</td>
</tr>
<tr>
<td>IRCNN [Zhang et al.2017b]</td>
<td>36.70</td>
<td>31.24</td>
<td>28.93</td>
</tr>
<tr>
<td>FFDNet [Zhang et al.2017c]</td>
<td>36.81</td>
<td>31.39</td>
<td>29.10</td>
</tr>
<tr>
<td>MSFN (Ours)</td>
<td>37.49</td>
<td>32.08</td>
<td>29.81</td>
</tr>
</tbody>
</table>

Figure 4: Visual comparisons between MSFN and other state-of-the-art denoising methods on the SIDD benchmark.
We evaluate different methods including DnCNN-B [Zhang et al. 2017a], Noise Clinic (NC) [Lebrun and Colom2015], MCWNNM [Xu et al. 2017], RDN [Zhang et al. 2020], FFDNet+ [Zhang et al. 2017c], TWSC [Xu et al. 2018b], CBDNet [Guo et al. 2019], RIDNet [Anwar and Barnes2019], VDN [Yue et al. 2019], MISPNet [Ours], MSFN-U (Ours) and MSFN (Ours).

<table>
<thead>
<tr>
<th>Method</th>
<th>PolyU</th>
<th>Nam</th>
</tr>
</thead>
<tbody>
<tr>
<td>DnCNN−B [Zhang et al. 2017a]</td>
<td>36.84 / 0.895</td>
<td>34.95 / 0.885</td>
</tr>
<tr>
<td>NC [Lebrun and Colom2015]</td>
<td>37.89 / 0.946</td>
<td>37.89 / 0.952</td>
</tr>
<tr>
<td>MCWNNM [Xu et al. 2017]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>RDN [Zhang et al. 2020]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>FFDNet+ [Zhang et al. 2017c]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>TWSC [Xu et al. 2018b]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>CBDNet [Guo et al. 2019]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>RIDNet [Anwar and Barnes2019]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>VDN [Yue et al. 2019]</td>
<td>39.17 / 0.951</td>
<td>39.17 / 0.957</td>
</tr>
<tr>
<td>MISPNet [Ours]</td>
<td>40.12 / 0.974</td>
<td>40.12 / 0.974</td>
</tr>
<tr>
<td>MSFN-U (Ours)</td>
<td>40.12 / 0.974</td>
<td>40.12 / 0.974</td>
</tr>
</tbody>
</table>

Table 3: Quantitative comparison on PolyU and Nam.

3.2 Implementation Details

The MSFN has 30 MSBs and 4 SFMs in the sub-network at each time step. We set up a multiple-to-multiple feedback connection (m = 4). The reduction factor r in PCA is 16, and the DM has 8 DBs. The weighting coefficients in dual loss are \( \lambda_1 = 0.9, \lambda_2 = 0.1 \). In the synthetic datasets, 16 patches cropped to 96 \times 96 form a training batch, while in the real-world datasets, the setting is 32 patches that cropped to 128 \times 128. The number of feature channels is 64, except for the input and output layer. We set the initial learning rate as 1 \times 10^{-4} and use the ADAM optimization method with parameter \( \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8} \). All models are implemented in PyTorch and trained on NVIDIA GeForce RTX 2080 Ti GPU.

3.3 Comparisons with Other Methods

We perform qualitative and quantitative comparisons on the standard benchmarks. The metrics are peak signal-to-noise ratio (PSNR) and structural similarity index metric (SSIM).

### Quantitative Comparison

#### Synthetic Noise

We evaluate different methods including CBM3D [Dabov et al. 2007], TNRD [Chen and Pock2017], RED [Mao et al. 2016], DnCNN [Zhang et al. 2017a], MemNet [Tai et al. 2017], IRCNN [Zhang et al. 2017b], FFDNet [Zhang et al. 2017c], and RNAN [Zhang et al. 2019]. The quantitative and qualitative comparison results of all methods on the three benchmark data sets of Kodak24, BSD68, and Urban100 are shown in Tab. 1. The proposed MSFN achieves the best performance on all noise levels and all benchmark datasets, which increases the highest PSNR by more than 0.2 dB. Our method shows superiority on severely degraded images with a noise level of \( \sigma = 70 \). On the Urban100 dataset, MSFN increases the highest PSNR from 27.45 dB to 27.98 dB. The iterative feedback mechanism reduces the difficulty of training and accelerates convergence step by step.

#### Real-World Noise

We compare the performance of nine real-world denoising methods on PolyU and Nam, and seven methods on SIDD and DND, including DnCNN-B [Zhang et al. 2017a], Noise Clinic (NC) [Lebrun and Colom2015], MCWNNM [Xu et al. 2017], RDN [Zhang et al. 2020], FFDNet+ [Zhang et al. 2017c], TWSC [Xu et al. 2018b], CBDNet [Guo et al. 2019], RIDNet [Anwar and Barnes2019], VDN [Yue et al. 2019], MISPNet [Ours], MSFN-U (Ours), and MSFN (Ours).

The dual loss can supervise images in the noise domain independently so that we can add unpaired data for training. As shown in Tab. 2 and Tab. 3, the proposed MSFN shows significant superiority and good generalization on the real-world images denoising. Compared with the latest MIRNet, our network achieved a performance gain of 0.26 dB on SIDD and 0.14 dB on DND. The performance on PolyU and Nam also exceeds other methods by a large margin, improving at least 1.64 dB and 1.13 dB, respectively. The network fine-tuned by 200 unpaired noise images is defined as MSFN-U, which improves the model performance on the 4 datasets by at least 0.11 dB. The MSFN based on dual loss can make full use of unpaired training data and does not rely on noise-free labels excessively. Experiments prove that training with paired and unpaired data is optimal.
In this paper, we propose a novel network for real-world image denoising, called multi-scale selective feedback network (MSFN) with dual loss. By propagating the principal components of the high-level hierarchical features to the shallow layers, the selective feedback module (SFM) enriches the early representation learning effectively. Iterative multi-scale feature reusing in multiple time steps is conducive to fitting complex degradation from coarse to fine. The proposed dual loss method provides a more strict closed-loop supervision on both the noise domain and the clean domain for paired and unpaired data. The MSFN outputs refined noise-free images from coarse to fine and refits the original noise as much as possible. The unpaired real-world noise images can be used for model training and fine-tuning. The effective expansion of real-world images further increases the model adaptability. It can alleviate the model’s excessive dependence on labels and noise overfitting.

4 Conclusion

In this paper, we propose a novel network for real-world image denoising, called multi-scale selective feedback network (MSFN) with dual loss. By propagating the principal components of the high-level hierarchical features to the shallow layers, the selective feedback module (SFM) enriches the early representation learning effectively. Iterative multi-scale feature reusing in multiple time steps is conducive to fitting complex degradation from coarse to fine. The proposed dual loss method provides a more strict closed-loop supervision on both the noise domain and the clean domain for paired and unpaired data. The comprehensively experimental results prove that our method is superior to the state-of-the-art methods and has better applicability to real-world data.
Acknowledgments
This work is partially supported by the NSFC fund (61831014), in part by the Shenzhen Science and Technology Project under Grant (ZDYBH201900000002, JCJ20180508152042002, GGF201704071461462).

References