Neurocomputing 517 (2023) 71–80

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

MSPNet: Multi-stage progressive network for image denoising

Yu Bai [a,b], Meiqin Liu [a,b,*], Chao Yao [c], Chunyu Lin [a,b], Yao Zhao [a,b]

[a] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[b] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China
[c] School of Computer & Communication Engineering, University of Science and Technology Beijing, 100083, Beijing, China

ARTICLE INFO

Article history:
Received 21 October 2021
Revised 30 August 2022
Accepted 15 September 2022
Available online 22 September 2022
Communicated by Zidong Wang

Keywords:
Multi-stage
Image denoising
Criss-cross attention
Encoder-decoder

ABSTRACT

Image denoising which aims to restore a high-quality image from the noisy version is one of the most challenging tasks in the low-level computer vision tasks. In this paper, we propose a multi-stage progressive denoising network (MSPNet) and decompose the denoising task into some sub-tasks to progressively remove noise. Specifically, MSPNet is composed of three denoising stages. Each stage combines a feature extraction module (FEM) and a mutual-learning fusion module (MFM). In the feature extraction module, an encoder-decoder architecture is employed to learn non-local contextualized features, and the channel attention blocks (CAB) are utilized to retain the local information of the image. In the mutual-learning fusion module, the criss-cross attention is introduced to balance the image spatial details and the contextualized information. Compared with the state-of-the-art works, experimental results show that MSPNet achieves notable improvements on both objective and subjective evaluations.

1. Introduction

Image denoising plays an important role in the low-level computer vision tasks, and it has attracted much attention from both academia and industry. Theoretically, image denoising is a special case of an inverse problem to restore clean images from noisy observation. It can be used as a preprocessing step for subsequent high-level computer vision tasks.

Many traditional denoising methods [1–4], are mainly attributed to the famous block-matching 3D (BM3D) [1] framework, which combines the non-local similarity characteristic of natural images and the sparse representation in the transform domain. And these methods assume that noise is independent and identically distributed. However, the strong assumption inevitably leads to inferior performance in real-world noise.

Recent state-of-the-art methods [5–10] employ convolutional neural networks (CNNs) to implicitly learn more general priors by capturing natural image statistics from large-scale data. CNN-based methods over others primarily attributes to many network modules and functional units including residual learning (DnCNN [11], MemNet [12]), attention (RIDNet [13], MIRNet [8]) and dense connections (RDN [14]). These methods are designed as single-stage, which lack the flexibility of image denoising. The multi-stage networks are widely used in pose-estimation [15,16], action segmentation [17,18] and image restoration [9,19] and so on. Nevertheless, it is claimed that some architectural bottlenecks limit the performance of the existing multi-stage frameworks [9]. Either encoder-decoder structure or single-scale network is only effective to obtain broad large-scale information, or to maintain the local information. There are rare architectures to employ both of them. *Zamir et al.* [9] considered this problem, but the non-local contextualized information and local details were not well fused. Therefore, it is very important to fuse the non-local and local information.

In this paper, image denoising is considered as a process to gradually learn the degradation function. Thus, we propose a multi-stage progressive denoising network, named MSPNet. The denoising processing is decomposed into some sub-tasks to progressively restore the clean image. In each stage, we design a parallel structure including a single-scale branch and an encoder-decoder branch. Considered information fusion demands the long-term dependency of the image, the criss-cross global attention [20] based non-local design is introduced to achieve feature fusion of contextualized information and spatial image details. Extensive experiments based on several benchmark datasets show our MSPNet can significantly improve the denoising performance on both synthetic and real noisy images.

The contributions of this paper are summarized in the following aspects:

* Corresponding author at:Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China.
E-mail address: mqliu@bjtu.edu.cn (M. Liu).

https://doi.org/10.1016/j.neucom.2022.09.098
0925-2312/

- We design a multi-stage progressive denoising network and decompose the whole denoising process into several sub-tasks to progressively remove noise.
- In each denoising stage, we employ a parallel network structure to simultaneously obtain non-local contextualized information and image spatial details.
- At the end of each stage, the criss-cross attention based on non-local design is introduced to fuse the image local details and non-local contextualized information.

2. Related Work

2.1. Single-stage methods

As widely adopted in denoising networks, single-stage architectures utilize different kinds of functional components to obtain performance gains. For example, DnCNN was proposed by *Zhang et al.* [11], where the residual learning and batch normalization were utilized to enhance the deep neural network learning and denoising. Considering the long-term dependence on the images, an end-to-end memory network based on residual connections was proposed [12] by *Tai et al.*, where both long-term and short-term memories were adopted to capture different levels of information in noisy images. RIDNet [13] firstly utilized the attention mechanism for the image denoising task, where the dependence among channels was employed to remove the real noise. Besides, RNAN [21] utilized residual non-local attention for high-quality image restoration. CycleISP [22] employed a channel attention-based framework that modeled the camera imaging pipeline in forward and reverse directions. For image denoising in a multi-scale feature space, SADNet [23] built an encoder-decoder architecture based on a deformable convolution unit to capture multi-scale features of the noisy image. MIRNet [8] was a multi-scale residual architecture and introduced both channel attention and spatial attention to further improve the performance of real image denoising. *Li et al.* [24] employed an enhanced encoder-decoder network to capture the image contextualized information for image deraining.

2.2. Multi-stage methods

A number of previous works have verified that the multi-stage network can achieve better performance than the single-stage counterparts in high-level vision tasks, such as pose estimation [15,16] and action segmentation [17,18]. For example, *Li et al.* [16] proposed the single-stage module design, cross-stage feature aggregation and coarse-to-fine supervision to improve the denoising performance. *farha et al.* [17] introduced a multi-stage architecture for the temporal action segmentation task. Recently, the multi-stage design is also utilized in low-level tasks. For example, some restoration works based on multi-stage employed a light-weight sub-network to progressively recover clean images. *Ren et al.* [25] presented a progressive ResNet (PRN) to take advantage of recursive computation. For handling large blur variations across different spatial locations, *Suin et al.* [19] proposed an efficient pixel adaptive and feature attentive design to adaptively remove motion blur. To balance the spatial details and high-level contextualized information, *Zamir et al.* [9] proposed a multi-stage architecture to progressively learn restoration functions for the degraded inputs.

3. Proposed Method

In this section, we first give the whole framework of the MSPNet. And then we detail the structure of the feature extraction module (FEM) and mutual-learning fusion module (MFM) as follows.

3.1. Network Architecture

The framework of MSPNet is shown in Fig. 1. It contains three stages to gradually remove the noise. Given a noisy image $y \in \mathbb{R}^{C_{in}\times W\times H}$ [1], the basic feature $F_1$of y is extracted by the shallow layer with a 3×3 convolutional layer and a channel attention block (CAB) in the first denoising stage,

$$F_1 = CAB(W * y), \tag{1}$$

where *CAB* represents a channel attention block, W is the convolution kernel to expand the number of the feature maps.

Every denoising stage includes a feature extraction module (FEM) and a mutual-learning fusion module (MFM). FEM is a parallel combination of encoder-decoder branch and single-scale branch. The single-scale branch based on channel attention is employed to capture the local information F_1^d, The encoder-decoder branch is utilized to extract rich contextualized features F_1^n,

$$\begin{aligned} F_1^d &= f_S^1(F_1), \\ F_1^n &= f_E^1(F_1), \end{aligned} \tag{2}$$

where f_S^1 represents the extraction function of single-scale branch, composed of series of channel attention blocks (CABs). f_E^1 represents the extraction function of encoder-decoder branch.

After obtaining image detail features F_1^d and contextualized features F_1^n , a mutual-learning fusion module (MFM) is designed to fuse the two features. The process is represented as,

$$M_1 = \mathrm{MFM}(F_1^d, F_1^n) \tag{3}$$

where M_1 is the output features of MFM.

To achieve the joint training of multiple stages, M_1 is added to the second stage by skip connection with the shallow features $F\prime_2$ extracted by the first CAB at the second stage. The process is formulated as,

$$F_2 = F\prime_2 + M_1 \tag{4}$$

where F_2 is input to the single-scale branch and encoder-decoder branch to get M_2. With the similar operation, the feature M_3 is output from MFM in the third stage.

In three stages, the clean image X_1, X_2 and X_3 are reconstructed with features M_1, M_2 and M_3 by a 3×3 convolutional layer. Thus, MSPNet can get three clean images with different quality to meet the different applications.

3.2. Feature Extraction Module (FEM)

The idea behind FEM is that image spatial details and non-local contextualized information are beneficial for removing the noise in the denoising process [8]. To extract these rich features, we propose a parallel architecture including a single-scale branch based on channel attention and an encoder-decoder branch. Our single-scale branch is operated on the full-resolution to obtain spatially precise and local information. Moreover, our encoder-decoder branch is progressively operated on different resolutions to extract semantically reliable and non-local information.

[1] C_{in}, W and H are respectively the channel number, width, and the height of the input image y.

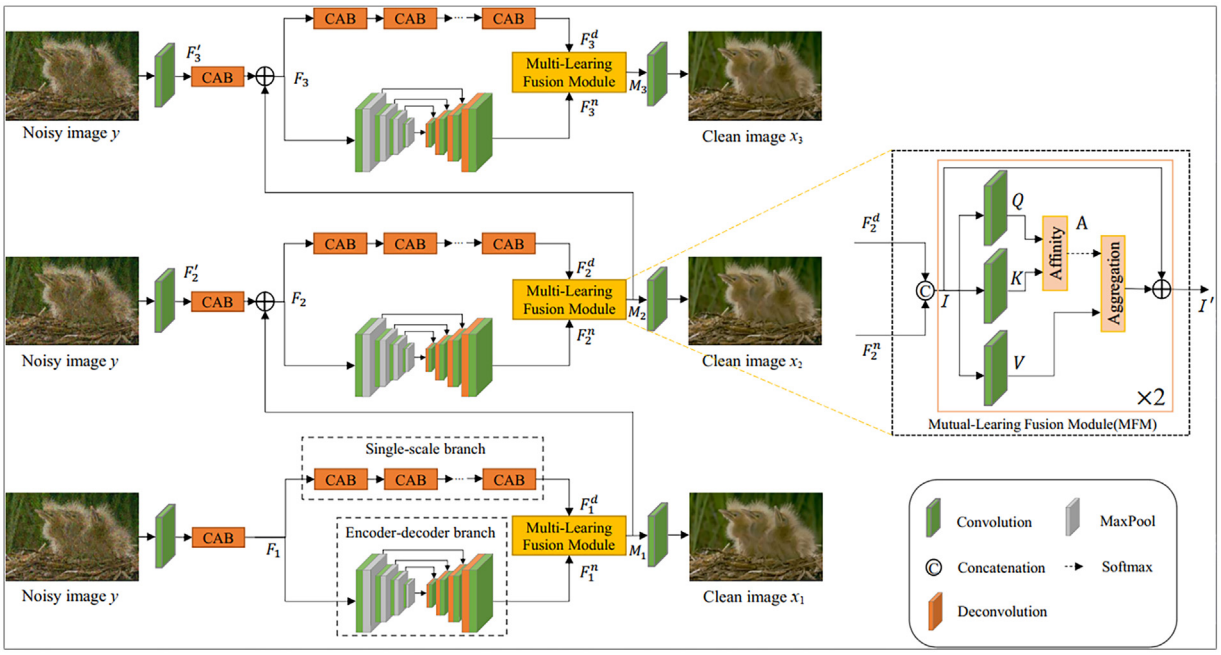


Fig. 1. The architecture of MSPNet.

3.2.1. Single-scale branch

In order to learn the spatial details of the image, our single-scale branch is designed with several stacked channel attention blocks (CABs). Every channel attention block (CAB) is a single-scale structure and the size of the feature maps does not change during the image processing. Besides, some works [8,9] have proven that the depth of the CNN-based model is highly correlated with its performance, so we employ several CAB modules in our single-scale branch. Taking the first stage as an example, the spatial image details F_1^d are captured by several stacked CAB modules,

$$F_1^d = CAB_m(CAB_{m-1}(\cdots CAB_1(F_1))) \tag{5}$$

where $CAB_1, CAB_2, \cdots, CAB_m$ denote m stacked CAB modules. F_1 is the basic feature of input image y achieved by Eq. (1).

The structure of CAB is shown in Fig. 2. The feature $f \in \mathbb{R}^{C\times W\times H}$ is input to two convolutional layers with ReLU function to obtain the feature h. Then, channel attention calculates and analyzes the weights of local information h. Specifically, the global average pooling (GAP) is applied to h to get the statistical quantity z_c of the c-th channel of feature h,

$$z_c = GAP(h_c) = \frac{1}{H\times W}\sum_{i=1}^{W}\sum_{j=1}^{H} h_c(i,j), c \in \{0,1,\cdots,C\} \tag{6}$$

where $h_c(i,j)$ represents the value of the c-th channel of feature h with the coordinates (i,j). H and W represent the spatial dimensions of h.

Channel statistical quantity z of all channels can be represented as,

$$z = [z_1, z_2, \cdots, z_C] \tag{7}$$

where [,] denotes concatenation operation. Next, the convolutional operation and the sigmoid activation function are adopted to obtain channel attention s,

$$s = \sigma(W_2 * (\delta(W_1 * z))) \tag{8}$$

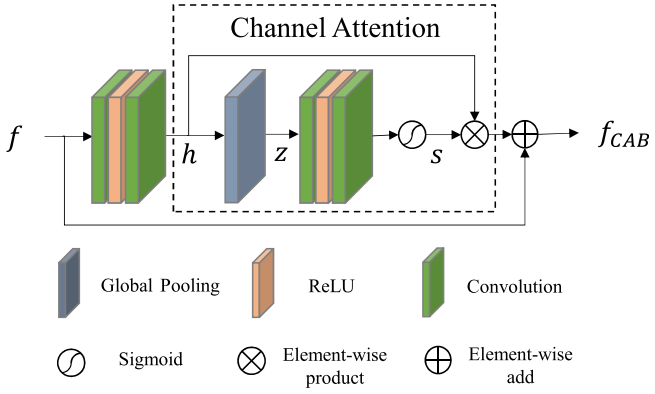


Fig. 2. The structure of channel attention block(CAB).

where W_1 and W_2 represent the convolution operations, σ represents the sigmoid function. δ represents the ReLU activation function.

The output of the CAB, f_{CAB} is obtained by element multiplying operation,

$$f_{CAB} = s \cdot h + f \tag{9}$$

where $[\cdot]$ denotes the element multiplying operation.

3.2.2. Encoder-decoder branch

In order to learn the non-local information of the image, we introduce a U-shaped encoder-decoder network shown in Fig. 3. Supposed the input feature with a size of 64×64, max pooling operation with stride 2 is used for down-sampling and the channel numbers are doubled to reduce the decay of information caused by the down-sampling operation. The process is represented as,

$$\begin{aligned} x_{k+1}^{\downarrow} &= MaxPool(x_k), \\ x_{i+1} &= H_{k+1}(x_{k+1}^{\downarrow}), k \in \{0,1,2,3\} \end{aligned} \tag{10}$$

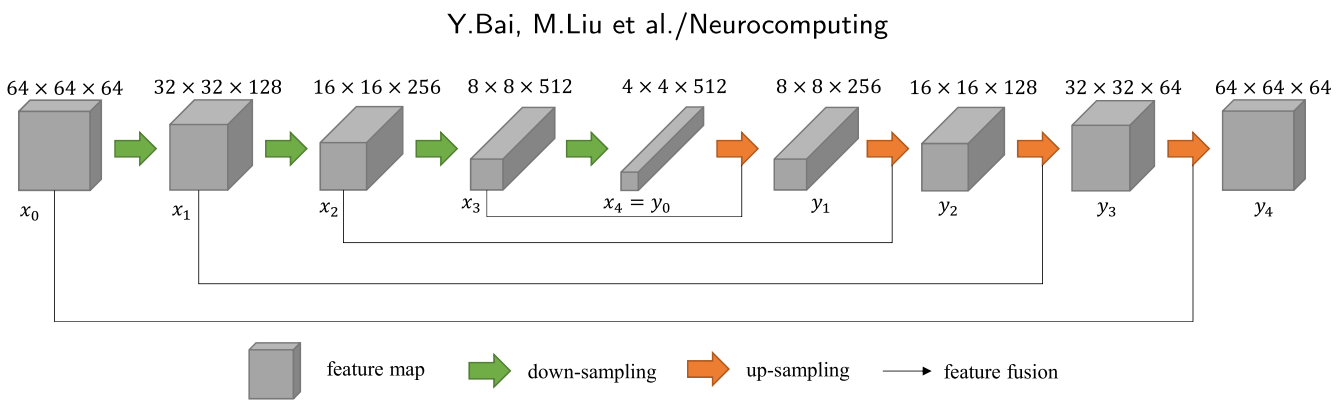


Fig. 3. The structure of encoder-decoder branch based on U-Net.

where x_k denotes the feature map after the k-th down-sampling operation, $x_{k+1}^{\downarrow}$ denotes the intermediate feature in the $(k+1)$-th down-sampling operation, H_{k+1} represents the convolutional operation. It is noted that x_4 is the low-resolution feature of x_0 which is the shallow feature F_1.

After obtaining the multi-scale features x_k $(k \in \{0,1,2,3\})$ from down-sampling operation, the up-sampling process adopts the deconvolution operation to enlarge the resolution of the features. It is formulated as,

$$y_{k'} = W_{k'}^{T} * [x_{4-k'}, P(y_{k'-1})], k' \in \{1,2,3,4\} \tag{11}$$

where [,] represents concatenation. $W_{k'}^{T}$ represents the deconvolution operation in the k'-th up-sampling operation. When $k' = 1, y_0$ equals to x_4. In fact, the decoded feature y_4 is the contextualized features F_1^n. $P(\,)$ represents the padding operation to enlarge the resolution of features $y_{k'-1}$ and then concatenate two features with different resolutions. Here, four down-sampling and up-sampling operations are adopted to capture the rich non-local contextualized information.

3.3. Mutual-Learning Fusion Module (MFM)

In order to fuse the image spatial details F_1^d and non-local contextualized information F_1^n, we design mutual-learning fusion module (MFM). MFM utilizes two successive criss-cross attention (CC-attention) [20] to obtain non-local dependency and avoid the lack of memory. Fig. 4 is the structure of CC-attention.

Given a feature map $I \in \mathbb{R}^{C\times W\times H}$, two convolutional layers with a kernel size of 1×1 are used to obtain feature map Q and K respectively. $\{Q,K\} \in \mathbb{R}^{C'\times W\times H}$. C' is smaller than C for dimension reduction. The attention map $A \in \mathbb{R}^{(H+W-1)\times(H\times W)}$ is generated by affinity and softmax operations. For each position u of Q, the set $\omega_u \in \mathbb{R}^{(H+W-1)\times C'}$ is obtained from K which is in the same row or column as u. $\omega_{u,i}$ represents the i-th element of ω_u, the affinity operation is represented as,

$$d_{u,i} = Q_u \omega_{u,i}^{T} \tag{12}$$

where $d_{u,i}$ is an element of $D \in \mathbb{R}^{(H+W-1)\times(H\times W)}$, and denotes the correlation between feature Q_u and $\omega_{u,i}$.

Another convolutional layer with a kernel size of 1×1 is used to obtain $V \in \mathbb{R}^{C\times W\times H}$. For each position u of V, we can obtain the set $\phi \in \mathbb{R}^{(H+W-1)\times(H\times W)}$ from V with the same row or column as position u. The output of the CC-attention module is obtained by an aggregation operation,

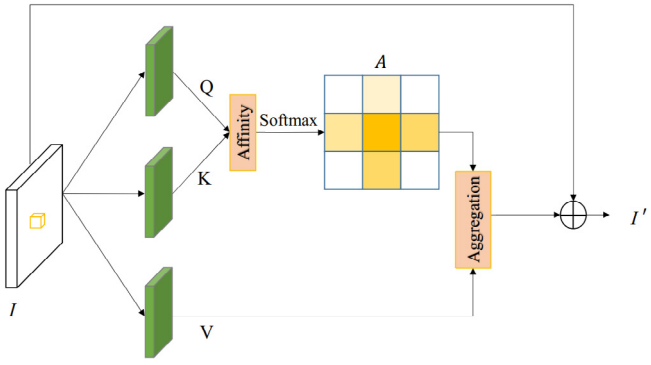


Fig. 4. The structure of criss-cross attention.

$$I'_u = \sum_{i=0}^{W+H-1} A_{u,i}\phi_{u,i} + I_u \tag{13}$$

where I'_u is a feature vector of $I' \in \mathbb{R}^{C\times W\times H}$. $A_{u,i}$ is the scalar value of A with coordination (u,i). Obviously, the long-term dependency of all pixels can be captured by stacking some CC-attention modules [20].

3.4. Loss function

For end to end training, we utilize the loss function L to measure the difference between the denoised image x_j $(j \in \{1,2,3\})$ and ground-truth x_{gt}, which is formulated as follows,

$$L = \sum_{j=1}^{3} L_{Char}(x_j, x_{gt}) \tag{14}$$

where L_{Char} denotes the Charbonnier loss of each denoising stage shown as follows,

$$L_{Char} = \sqrt{||x_j, x_{gt}||^2 + e^2}, j \in \{1,2,3\} \tag{15}$$

where e is a constant.

4. Experiments

In this section, we demonstrate the effectiveness of our method on both synthetic datasets and real noisy datasets.

4.1. Dataset and Evaluation Metrics

For the denoising of synthetic noisy images, we adopt *DIV2K* [26] which contains 800 images with 2*K* resolution as our training

dataset. Different levels of AWGN are added to the clean images. For the training of real noisy images, we use the *SIDD* [27] Medium dataset. For test datasets, we adopt *BSD68* and *Kodak24* in the synthetic noise situation, and *SIDD* [27] validation dataset and *DnD* [28] dataset in the real noise situation. PSNR, RMSE and SSIM are employed to evaluate the performance. The best and second-best results are **highlighted** and underlined respectively in the following experiments.

4.2. Experiment Setup

We randomly rotate and flip the image horizontally and vertically for data augmentation. In each training batch, 16 patches with the size of 64×64 are input to the model in the synthetic image denoising, and 16 patches with the size of 128×128 are used for real image denoising. We train our model by the ADAM optimizer [29] with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. For synthetic image denoising, The initial learning rate is 1×10^{-4} and then halved after 1×10^5 iterations. And for real image denoising, the initial learning rate is 2×10^{-4} and is decreased to 1×10^{-6} in the cosine annealing strategy [30]. All experiments are implemented in the PyTorch framework and trained by one Nvidia GeForce RTX 3090.

4.3. Ablation Study

4.3.1. Stage analysis

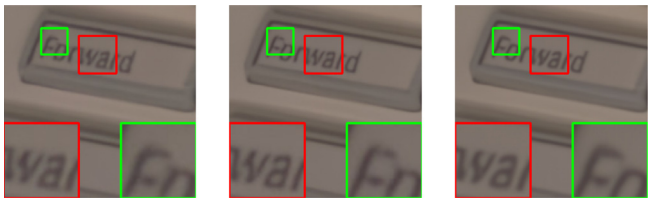
To compare the denoising performance of different stages, we implement comparative experiments on *SIDD* [27] dataset shown in Table 1. Here, MSPNet-1, MSPNet-2, and MSPNet-3 represent the first, second, and third stages of MSPNet respectively. The PSNR of MSPNet-1 is 39.55 dB. The PSNR of MSPNet-2 and MSPNet-3 increase by 0.17 dB and 0.06 dB. It indicates that most of the noise has been removed in the first stage. With the increase of the stage numbers, the denoising performance is further improved to some extent. Fig. 5 is their subjective results. It indicates that the text is well reconstructed from MSPNet-1 and other sharper information is reconstructed from MSPNet-2 and MSPNet-3. Hence, the performance of MSPNet does not linearly increase with the stage number. And it can be applied to different situations.

4.3.2. Model Analysis

Comparative experiments are implemented to compare some combinations of CABs, U-Net, and CC-attention. The experimental configurations and experimental results are shown in Table 2. Model D_1 consists of CABs and CC-attention, model D_2 is composed of U-Net and CC-attention, and D_3 consists of the CABs and U-Net. Compared D_3 with MSPNet, the effectiveness of CC-attention is demonstrated. Compared with D_1, MSPNet achieves 0.16 dB gains and demonstrates the importance of non-local contextualized information. When compared with D_2, MSPNet achieves 0.06 dB gains and demonstrates the effectiveness of local information. Moreover, we perform ablation experiments on different combination models with CABs and U-Net for different stages on *BSD68* dataset. The experimental configurations and results are shown in Table 3. Specifically, CABs $\cup$ U-Net represents the parallel structure that contains channel attention blocks and U-Net. It is noted that each model contains CC-attention module at the end of each stage and the column of "original image" represents whether the original image is obtained at this stage. It indicates that MSPNet can get better denoising performance with more stages. Moreover, from the comparison results of model C_2 and model C_3, C_3 achieves 0.06 dB and 0.03 dB gains over C_2 at stage 2 and stage 3. The effectiveness of CABs demonstrates the importance of local information during the denoising process. The original image is only available for the first stage in model C_4. Table 3 shows that model C_4 achieves 28.36 dB on *BSD68* dataset with σ=50 and is 0.11 dB lower than MSPNet. This indicates that the original image is very useful to improve the performance of models in each denoising stage.

Table 1
Evaluations on the number of stages on *SIDD* dataset.

Models	MSPNet-1	MSPNet-2	MSPNet-3
Parameters(M)	18.2	36.4	54.6
PSNR(dB)	39.55	39.72	**39.78**



(a) MSPNet-1 (b) MSPNet-2 (c) MSPNet-3

Fig. 5. Subjective results of different stages.

4.4. Synthetic Noisy Images

For experiments of synthetic noisy images, *Kodak24*, *BSD68* and *Urban100* datasets are used as the test datasets. They all contain gray-scale and color-scale images. We generate noisy color images by adding AWGN with different noise levels $\sigma = 30, 50$ and 70 to the groundtruth.

4.4.1. Gray-scale Image Denoising

The PSNR and RMSE results are listed in Table 4. When compared with the traditional method BM3D [1], MSPNet achieves 1.02 dB performance gains on *Kodak24* dataset with $\sigma = 50$. When compared with classic CNN-based denoising methods, MSPNet performs the best on all datasets with all nosie levels. Taking *Kodak24* with $\sigma = 50$ as a example, MSPNet achieves 0.50 dB and 0.33 dB performance gains over DnCNN [11] and MemNet [12]. When compared with RIDNet, MSPNet still surpasses 0.22 dB. In addition, our MSPNet still performs well for high-resolution images in *Urban100* dataset, and achieves 0.07 dB ($\sigma = 30$), 0.24 dB ($\sigma = 50$) and 0.34 dB ($\sigma = 70$) gains over RDN [14]. This is mainly because of the effective fusion of spatial details and contextualized information. Besides, RMSE values of our MSPNet with all noise levels are also the lowest, and also demonstrate the effectiveness of MSPNet.

Visual gray-scale denoised results of different methods are shown in Fig. 6 and Fig. 7. BM3D preserves the image structure to some degree but fails to remove noise deeply as shown in Fig. 6 (c) and Fig. 7 (c). BM3D could not well handle image textures and causes lots of artifacts and blurs. DnCNN and FFDNet oversmooth the edges and confuse the foreground and background. RIDNet restores more clean images, but the textures and details are destroyed during the denoising process and could not handle the background shown in Fig. 6 (f). Our MSPNet can recover sharper edges and cleaner smooth areas. The zebra stripes and text are very clean shown in Fig. 6 (g) and Fig. 7 (g).

4.4.2. Color-scale Image Denoising

The denoising results of color-scale images evaluated by PSNR and RMSE are listed in Table 5. MSPNet achieves the highest PSNR and the lowest RMSE on all datasets. Taking $\sigma = 50$as an example,

Table 2
Comparative experiments on *BSD68* dataset.

Models	U-Net	CABs	CC-attention	PSNR(dB)
D_1		✓	✓	28.31
D_2	✓		✓	28.41
D_3	✓	✓		28.46
MSPNet	✓	✓	✓	**28.47**

Table 3
Models evaluations of different combination on *BSD68* dataset.

Models	Combination	Stage	Orignal image	PSNR(dB)
C_1	Stage1: CABs	1	✓	28.19
	Stage2: CABs	2	✓	28.29
	Stage3: CABs	3	✓	28.33
C_2	Stage1: U-Net	1	✓	28.30
	Stage2: U-Net	2	✓	28.36
	Stage3: U-Net	3	✓	28.41
C_3	Stage1: CABs ∪ U-Net	1	✓	28.35
	Stage2: U-Net	2	✓	28.42
	Stage3: U-Net	3	✓	28.44
C_4	Stage1: CABs ∪ U-Net	1	✓	28.19
	Stage2: CABs ∪ U-Net	2	×	28.34
	Stage3: CABs ∪ U-Net	3	×	28.36

Table 4
Denoising results (PSNR/RMSE) of synthetic gray-scale images.

Methods	*Kodak24*			*BSD68*			*Urban100*		
	30	50	70	30	50	70	30	50	70
BM3D [1]	29.13/0.31	26.99/0.51	25.73/0.68	27.76/0.43	25.62/0.70	24.44/0.92	28.75/0.34	25.94/0.65	24.27/0.95
RED [31]	29.77/0.27	27.66/0.44	26.39/0.59	28.50/0.36	26.37/0.59	25.10/0.79	29.18/0.31	26.51/0.57	24.82/0.84
DnCNN [11]	29.62/0.28	27.51/0.45	26.08/0.63	28.36/0.37	26.23/0.61	24.90/0.83	28.88/0.33	26.28/0.60	24.36/0.93
MemNet [12]	29.72/0.27	27.68/0.45	26.42/0.58	28.43/0.37	26.35/0.59	25.09/0.79	29.10/0.31	26.65/0.55	25.01/0.80
IRCNN [32]	29.53/0.28	27.45/0.46	N/A	28.26/0.38	26.15/0.62	N/A	28.85/0.33	26.24/0.61	N/A
FFDNet [33]	29.70/0.27	27.63/0.44	26.34/0.59	28.39/0.37	26.30/0.60	25.04/0.80	29.03/0.32	26.52/0.57	24.86/0.83
RIDNet [13]	29.90/0.26	27.79/0.42	26.51/0.57	28.54/0.36	26.40/0.58	25.12/0.78	N/A	N/A	N/A
RDN [14]	30.00/0.26	27.85/0.42	26.54/0.57	28.56/0.36	26.41/0.58	25.10/0.79	30.01/0.25	27.40/0.46	25.64/0.70
MSPNet(ours)	**30.06/0.25**	**28.01/0.40**	**26.59/0.56**	**28.64/0.35**	**26.55/0.56**	**25.31/0.75**	**30.09/0.25**	**27.64/0.44**	**25.98/0.64**

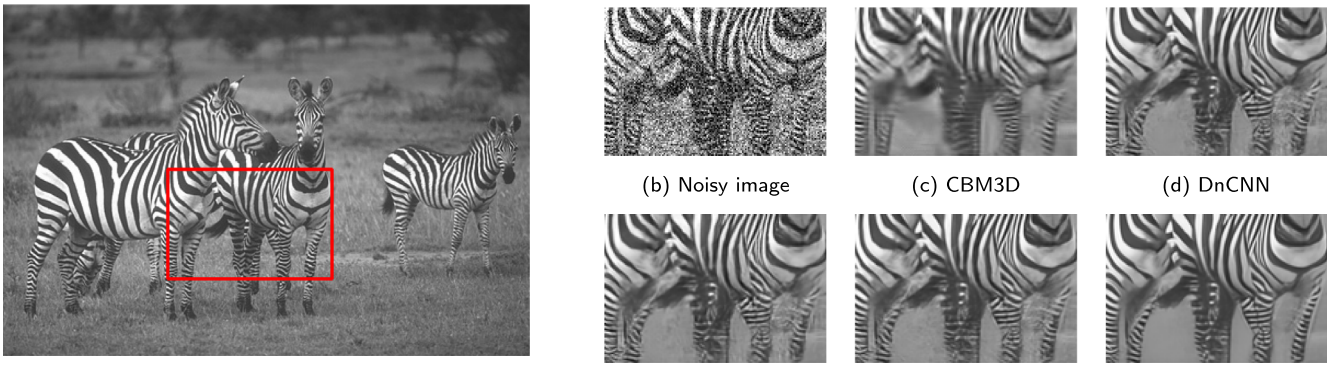


Fig. 6. Synthetic gray-scale image denoising results on *BSD68* with noise level σ = 50.

when compared with CBM3D, MSPNet surpasses 1.11 dB on *Kodak24* dataset. When compared with CNN-based methods, MSPNet also achieves super performance. Making a comparison with RDN, our MSPNet achieves 0.08 dB and 0.16 dB gains on *Kodak24* and *BSD68* datasets respectively. The rich contextualized information and spatial details in images are effective in the denoising task and suggest the improvement of the fusion of the two kinds of information. Moreover, MSPNet surpasses 0.05 dB, 0.08 dB and 0.14 dB over RDN with $\sigma = 30, 50$ and 70 on *Kodak24* dataset. We can find that MSPNet can obtain more performance gains with the increase of noise level.

The subjective results of each methods on images are visualized in Fig. 8 and Fig. 9. We analyze the edge value and IQI from the subjective figures. In fact, The clothing textures and the birds' feathers are difficult to be separated in the heavy noise situation. CBM3D produces artifacts in the smooth area and is difficult to recover clear edges as shown in Fig. 8 (c) and Fig. 9 (c). The classic CNN-based methods tend to remove the details along with the noise

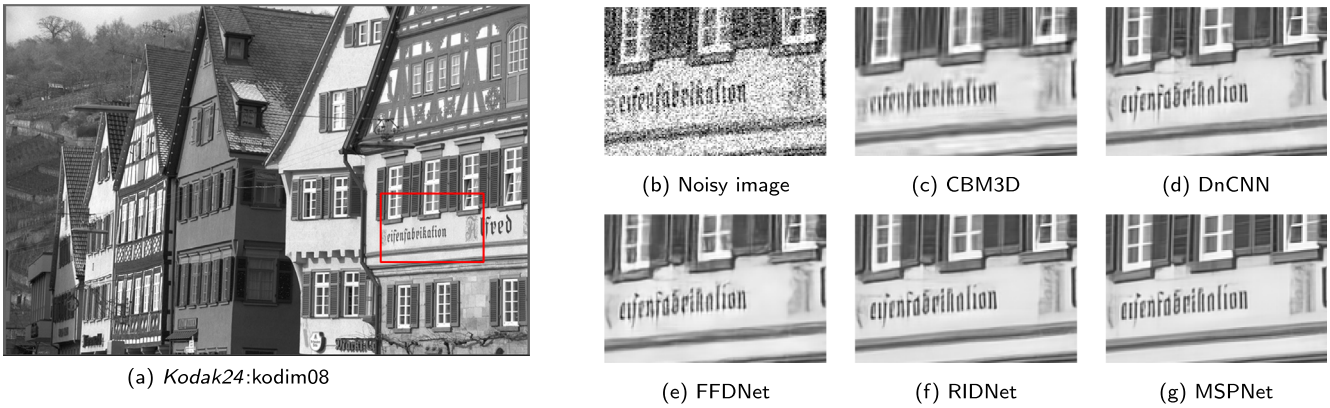


Fig. 7. Synthetic gray-scale image denoising results on *Kodak24* with noise level σ = 50.

Table 5
Denoising results (PSNR/RMSE) of synthetic color-scale images.

Methods	*Kodak24*			*BSD68*			*Urban100*		
	30	50	70	30	50	70	30	50	70
CBM3D [2]	30.89/0.21	28.63/0.35	27.27/0.48	29.73/0.27	27.38/0.47	26.00/0.64	30.36/0.23	27.94/0.41	26.31/0.60
RED [31]	29.71/0.27	27.62/0.44	26.36/0.59	28.46/0.36	26.35/0.59	25.09/0.79	29.02/0.32	26.40/0.58	24.74/0.86
DnCNN [11]	31.39/0.19	29.16/0.31	27.64/0.44	30.40/0.23	28.01/0.40	26.56/0.56	30.28/0.24	28.16/0.39	26.17/0.62
MemNet [12]	29.67/0.28	27.65/0.44	26.40/0.58	28.39/0.37	26.33/0.59	25.08/0.79	28.93/0.33	26.53/0.57	24.93/0.82
IRCNN [32]	31.24/0.19	28.93/0.33	N/A	30.22/0.24	27.86/0.42	N/A	30.28/0.24	27.69/0.43	N/A
FFDNet [33]	31.39/0.19	29.10/0.31	27.68/0.44	30.31/0.24	27.96/0.41	26.53/0.57	30.53/0.23	28.05/0.40	26.39/0.59
RIDNet [13]	31.64/0.17	29.25/0.30	27.94/0.41	30.47/0.23	28.12/0.39	26.69/0.55	N/A	N/A	N/A
RDN [14]	31.94/0.16	29.66/0.28	28.20/0.39	30.67/0.22	28.31/0.38	26.85/0.53	**31.69/0.17**	29.29/0.30	27.63/0.44
MSPNet(ours)	**31.99/0.16**	**29.74/0.27**	**28.34/0.37**	**30.76/0.21**	**28.47/0.36**	**27.03/0.50**	31.64/0.17	**29.40/0.29**	**27.66/0.44**

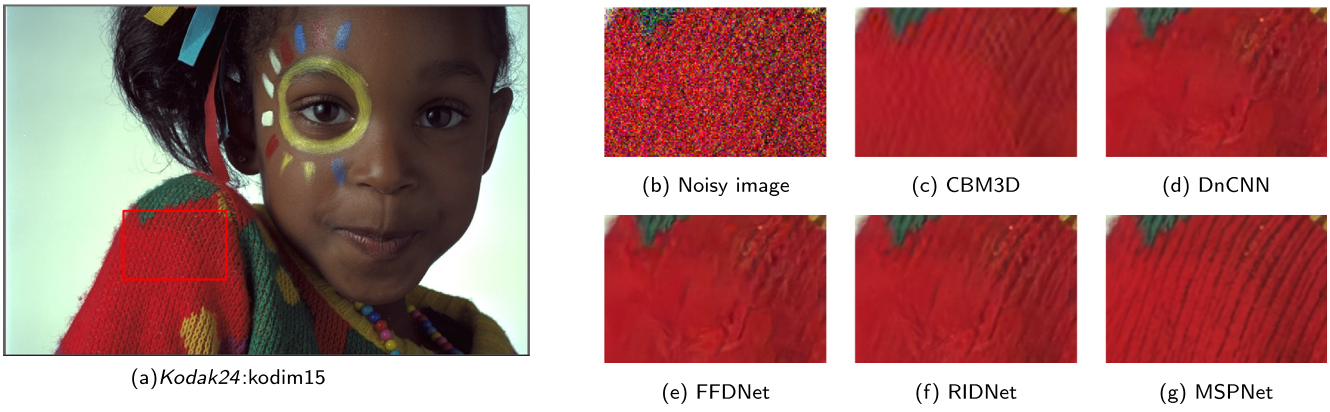


Fig. 8. Synthetic color-scale image denoising results on *Kodak24* with noise level σ = 50.

resulting in over smoothing artifacts. The textures of the girl's sweater are not be recovered well as shown in Fig. 8 (d), (e) and (f). The details of chicken feathers are destroyed to some extent as shown in Fig. 9 (d), (e) and (f). From the Fig. 8 (g) and Fig. 9 (g), we find that MSPNet can restore vivid textures without blurring the details. Therefore, the edge value and IQI of MSPNet is better than other methods, while they can't restore visual-pleasing edge details.

4.4.3. Real Noisy Images

We also conduct a series of experiments to evaluate the denoising performance of MSPNet on real noisy images. *DnD* [28] and *SIDD* [27] datasets are adopted as our test datasets. *DnD* dataset contains 50 real noisy images and needs to be submitted the denoised images to the *DnD* official website for the test. And *SIDD* validation dataset contains 1280 noisy-clean image pairs with the resolutions of 256 × 256.

Comparison methods contain several outstanding works, i.e., CBM3D [1], DnCNN [11], CBDNet [34], RIDNet [13], MIRNet [8] and MPRNet [9]. The objective evaluation results on two datasets are shown in Table 6. The traditional method CBM3D cannot get good performance and achieves 25.56 dB on *SIDD* dataset. When compared with early CNN-based models such as CBDNet and DnCNN, MSPNet achieves huge improvements. When compared with RIDNet with feature attention, MSPNet surpasses 1.07 dB on *SIDD* dataset and 0.49 dB on *DnD* dataset. For *SIDD* dataset, MSPNet achieves the best performance.

Meanwhile, we further validate the parameters and running speed, it can be seen that MSPNet has only 19% GFLOPs of MIRNet. Compared with DANet+, MSPNet has 86% parameters and achieves 0.31 dB performance gains.

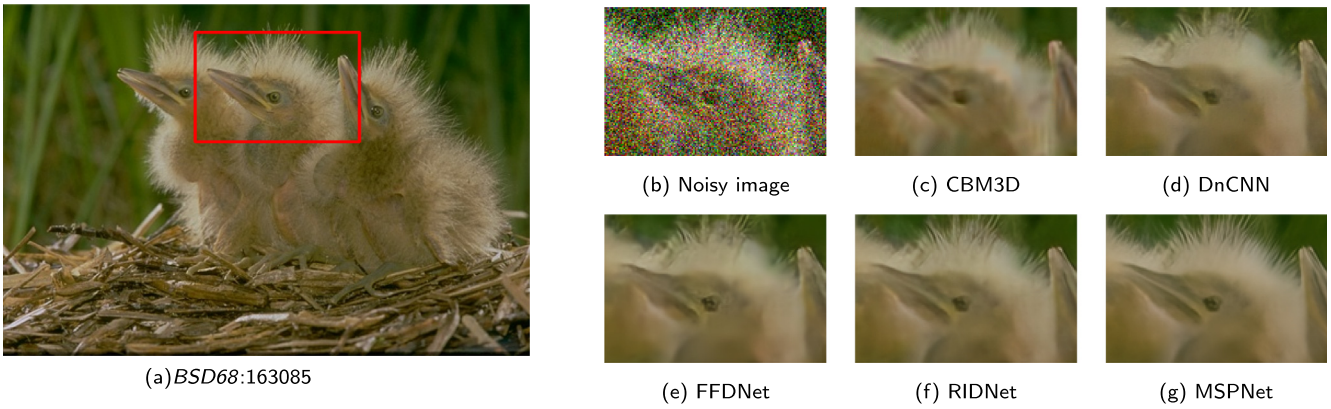


Fig. 9. Synthetic color-scale image denoising results on *BSD68* with noise level σ = 50.

Table 6
Quantitative results on *SIDD* and *DnD* datasets

Methods	Parameters(M)	GFLOPs	*SIDD* dataset		*DnD* dataset	
			PSNR	SSIM	PSNR	SSIM
BM3D [1]	-	-	25.65	0.685	34.51	0.851
CBDNet [34]	4.3	80.7	30.78	0.801	38.06	0.942
DnCNN [11]	0.7		26.33	0.583	32.43	0.790
RIDNet [13]	1.5	196	38.71	0.951	39.26	0.953
DGNSCNet [6]	2.1		39.31	0.955	39.43	0.953
SADNet [23]	4.3	-	39.46	0.957	39.59	0.952
VDN [35]	7.8	99	39.28	0.956	39.38	0.952
Path-Restore [36]	-	-	-	-	39.00	0.954
MIRNet [8]	31.8	1572	39.72	0.959	**39.88**	**0.959**
MPRNet [9]	20.1	1176	39.71	0.958	39.80	0.954
DANet+ [37]	63.1	66	39.47	0.957	39.58	0.955
MSPNet(Ours)	54.6	298	**39.78**	**0.959**	39.75	0.954

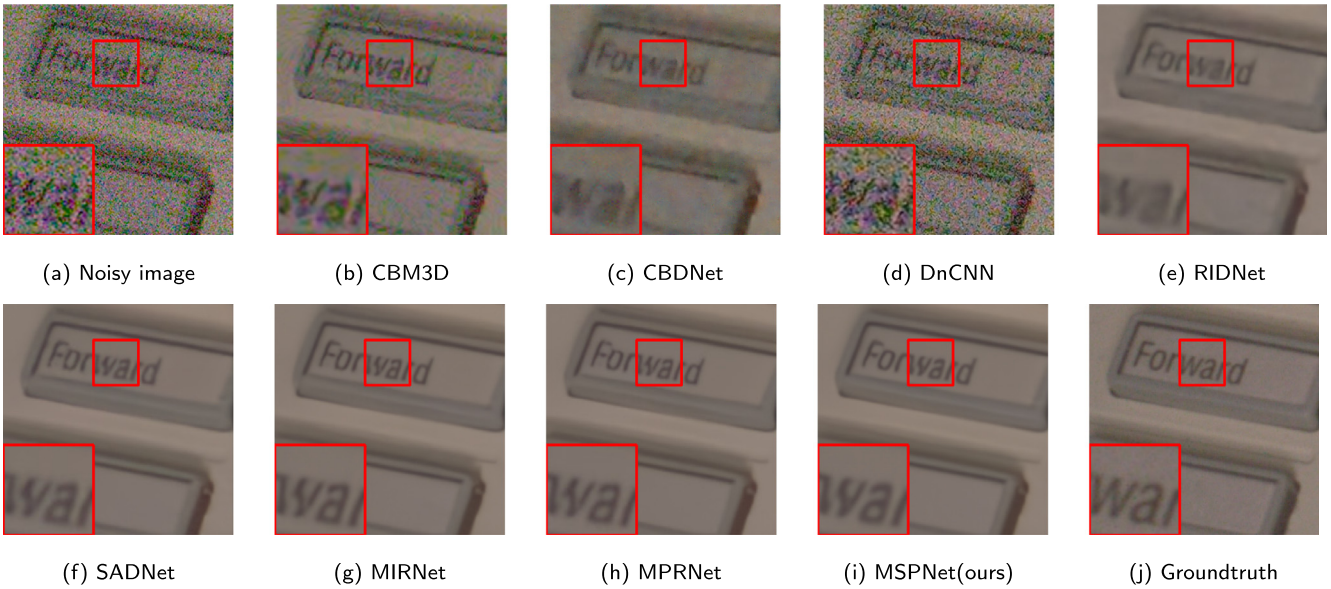


Fig. 10. Visual comparison results of real noisy images on *SIDD* dataset.

Fig. 10 and Fig. 11 show the visual comparison results on *SIDD* and *DnD* datasets. As shown in Fig. 10 (b) and Fig. 11 (b), there are lots of noises in the background reconstructed by CBM3D and DnCNN. CBDNet and RIDNet blur the edges and over-smooth the texture. The text is very blurry as shown in Fig. 10 (c) and (e). Although SADNet restores many pleasing images, it still corrodes the edges with residual noise. Our MSPNet can effectively remove the noise and maintain clear edges as shown in Fig. 10 (i) and Fig. 11 (i).

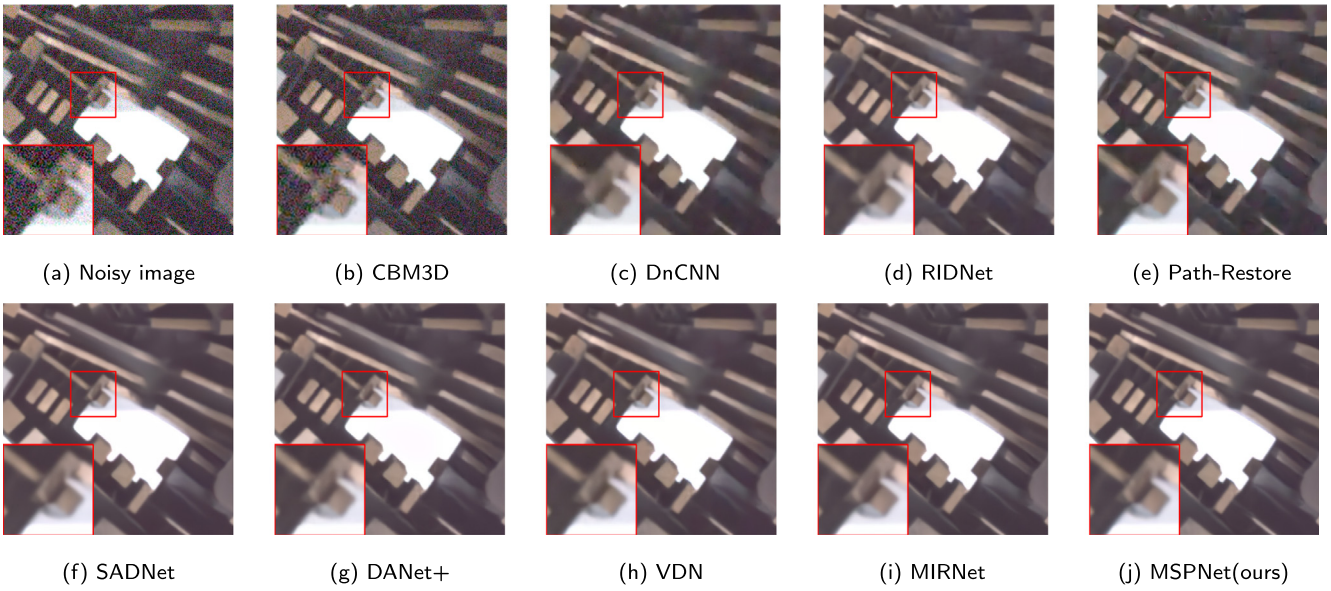


Fig. 11. Visual comparison results of real noisy images on *DnD* dataset.

5. Conclusions

In this paper, MSPNet is proposed to progressively remove the noise. It contains three denoising stages. Every stage includes a parallel structure with an encoder-decoder branch and a single-scale branch. The criss-cross attention is designed to fuse features of contextualized information and spatial details. We conduct ablation study to evaluate the effectiveness of stage numbers, CABs, U-Net and CC-attention. Compared with the state-of-the-art works, MSPNet achieves objective and subjective improvements on both synthetic noisy image and real noisy images.

CRediT authorship contribution statement

Yu Bai: Conceptualization, Methodology, Software. **Meiqin Liu:** Data curation, Writing - original draft. **Chao Yao:** Visualization, Investigation. **Chunyu Lin:** Supervision. **Yao Zhao:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (61972028, 61902022, 62120106009) and the Fundamental Research Funds for the Central Universities (FRF-IDRY-20-038).

References

[1] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, Karen Egiazarian, Image denoising by sparse 3D transform-domain collaborative filtering, IEEE Transactions on Image Processing 16 (8) (2007) 2080–2095.
[2] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space. In Proceedings of the IEEE International Conference on Image Processing, volume 1, pages I–313, 2007.
[3] Yan Liu, Yi Zhang, Low-dose CT restoration via stacked sparse denoising autoencoders, Neurocomputing 284 (2018) 80–89.
[4] Mahmud Hasan, Mahmoud R El-Sakka, Improved BM3D image denoising using SSIM-optimized Wiener filter, Journal on Image and Video Processing 2018 (1) (2018) 1–12.
[5] Shengjie Chen, Shuo Chen, Zhenhua Guo, Yushen Zuo, Low-resolution palmprint image denoising by generative adversarial networks, Neurocomputing 358 (2019) 275–284.
[6] Chuncheng Wang, Chao Ren, Xiaohai He, Linbo Qing, Deep recursive network for image denoising with global non-linear smoothness constraint prior, Neurocomputing 426 (2021) 147–161.
[7] Wu Fangfang, Tao Huang, Weisheng Dong, Guangming Shi, Zhonglong Zheng, Xin Li, Toward blind joint demosaicing and denoi-sing of raw color filter array data, Neurocomputing 453 (8) (2021).
[8] 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. ArXiv Preprint ArXiv:2003.06792, 2020.
[9] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 14821–14831, 2021.
[10] Hong Song, Lei Chen, Yutao Cui, Qiang Li, Qi Wang, Jingfan Fan, Jian Yang, Le Zhang, Denoising of MR and CT images using cascaded multi-supervision convolutional neural networks with progressive training, Neurocomputing 469 (2022) 354–365.
[11] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, Lei Zhang, Beyond a gaussian denoiser: residual learning of deep CNN for image denoising, IEEE Transactions on Image Processing 26 (7) (2017) 3142–3155.
[12] Ying Tai, Jian Yang, Xiaoming Liu, Chunyan Xu, MemNet: a persistent memory network for image restoration, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4539–4547.
[13] Saeed Anwar, Nick Barnes, Real image denoising with feature attention, in: In Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3155–3164.
[14] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP(99):1–1, 2020.
[15] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7103–7112, 2018.
[16] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148, 2019.
[17] Yazan Abu Farha and Jurgen Gall. MS-TCN: multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3575–3584, 2019.
[18] Pallabi Ghosh, Yi Yao, Larry Davis, Ajay Divakaran, Stacked spatio-temporal graph convolutional networks for action segmentation, in: Proceedings of the Winter Conference on Applications of Computer Vision, 2020, pp. 576–585.
[19] Maitreya Suin, Kuldeep Purohit, A.N. Rajagopalan, Spatially-attentive patch-hierarchical network for adaptive motion deblurring, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3606–3615.
[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, pages 603–612, 2019.

[21] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In Proceedings of the International Conference on Learning Representations, 2018.
[22] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: real image restoration via improved data synthesis. In Proceedings of the European Conference on Computer Vision, pages 2696–2705, 2020.
[23] Meng Chang, Qi Li, Huajun Feng, Zhihai Xu, Spatial-adaptive network for single image denoising, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 171–187.
[24] Guanbin Li, Xiang He, Wei Zhang, Huiyou Chang, Le Dong, and Liang Lin. Non-locally enhanced encoder-decoder network for single image deraining. In Proceedings of the ACM International Conference on Multimedia, pages 1056–1064, 2018.
[25] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, Deyu Meng, Progressive image deraining networks: a better and simpler baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3937–3946.
[26] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 126–135, 2017.
[27] Abdelrahman Abdelhamed, Stephen Lin, Michael S Brown, A high-quality denoising dataset for smartphone cameras, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1692–1700.
[28] Tobias Plotz, Stefan Roth, Benchmarking denoising algorithms with real photographs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1586–1595.
[29] Kingma Da. A method for stochastic optimization. ArXiv Preprint ArXiv:1412.6980, 2014.
[30] Ilya Loshchilov, S.G.D.R. Frank Hutter, stochastic gradient descent with warm restarts. ArXiv Preprint ArXiv:1608.03983, 2016.
[31] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. ArXiv Preprint ArXiv:1603.09056, 2016.
[32] Kai Zhang, Wangmeng Zuo, Gu. Shuhang, Lei Zhang, Learning deep CNN denoiser prior for image restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3929–3938.
[33] Kai Zhang, Wangmeng Zuo, Lei Zhang, FFDNet: toward a fast and flexible solution for CNN-based image denoising, IEEE Transactions on Image Processing 27 (9) (2018) 4608–4622.
[34] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, Lei Zhang, Toward convolutional blind denoising of real photographs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1712–1722.
[35] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: toward blind noise modeling and removal. ArXiv Preprint ArXiv:1908.11314, 2019.
[36] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: learning network path selection for image restoration. ArXiv Preprint ArXiv:1904.10343, 2019.
[37] Zongsheng Yue, Qian Zhao, Lei Zhang, Deyu Meng, Dual adversarial network: toward real-world noise removal and noise generation, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 41–58.

Yu Bai received the B.S. degree from Shanxi University, China, in 2020. He is currently pursuing the M.E. degree at the Institute of Information Science, Beijing Jiaotong University, Beijing, China. His research interests include image denoising and image super-resolution.

Meiqin Liu received the M.E. degree and doctor's degree from Beijing Jiaotong University (BJTU), China, in 2007 and 2018. She is currently an assistant professor at institute of information and science in BJTU. From 2014 to 2015, she was a visiting scholar in Simon Fraser University, Canada. Her research interests include image/video compression and image processing.

Chao Yao received the doctor's degree from Beijing Jiaotong University (BJTU), Beijing, China, in 2016. He is currently an assistant professor with University of Science and Technology Beijing. His research interests include image/video compression and image processing.

Chunyu Lin received the doctor's degree from Beijing Jiaotong University (BJTU), Beijing, China, in 2011. From 2009 to 2010, he was a Visiting Researcher with the ICT Group, Delft University of Technology, The Netherlands. From 2011 to 2012, he was a PostDoctoral Researcher with the Multimedia Laboratory, Gent University, Belgium. He is currently a Full Professor with BJTU. His research interests include image/video compression and robust transmission, 3-D video coding, virtual reality video processing.

Yao Zhao received the B.S. degree from Radio Engineering Department, Fuzhou University, China, in 1989, the M.E. degree from Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), China, in 1996. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He serves on the Editorial Board of several international journals, including as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE SIGNAL P ROCESSING LETTERS and an Area Editor for Signal Processing: Image Communication (Elsevier). He was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013. He was named as a Distinguished Young Scholar by the National Science Foundation of China in 2010.