International Journal of Information Technologies and Systems Approach
Volume 16 • Issue 3

An Efficient Self-Refinement and Reconstruction Network for Image Denoising

Jinqiang Xue, Jiangnan University, China
Qin Wu, Jiangnan University, China*

ABSTRACT

Recent works tend to design effective but deep and complex denoising networks, which usually ignored the industrial requirement of efficiency. In this paper, an effective and efficient self-refinement and reconstruction network (SRRNet) is proposed for image denoising. It is based on the encoder-decoder architecture and three improvements are introduced to solve the problem. Specifically, four novel residual connections of different types are proposed as building blocks to maintain original contextual details. A high-resolution reconstruction module is introduced to connect cross-level encoders and corresponding decoders, so as to boost information flow and result in realistic clear image. And multi-scale dual attention is used for suppressing noise and enhancing beneficial dependency. SRRNet achieves PSNR of 39.83 dB and 39.96 dB on SIDD and DND respectively. Compared with other works, the accuracy is higher and the complexity is lower. Extensive experiments in real-world image denoising and Gaussian noise removal prove that SRRNet balances performance and temporal cost better.

KEYWORDS

Attention Mechanism, Deep Learning, Image Denoising, Image Processing, Reconstruction Module, Residual Connection

INTRODUCTION

Image noise reduction task aims to remove useless noisy information from a given degraded noisy image and restore a clear image close to the real world. As a dense prediction task with pixel-by-pixel output with infinite possible outcomes for complex noise scenes in reality, image denoising is challenging to some extent. With the successful development of convolutional neural networks (CNNs) and deep learning, recent outstanding approaches employ CNNs to adaptively capture the essential correlation between noisy images and clear images from large-scale data sets and apply the trained prior parameters to reconstruct noisy images into clear images close to the real world.

In order to expand the receptive field and better extract contextual details of the feature, many works (Chang et al., 2020; Guo et al., 2019) designed U-shaped encoder–decoder-based (Ronneberger et al., 2015; Isola et al., 2017) architectures to hierarchically extract deep feature maps and reconstruct

DOI: 10.4018/IJITSA.321456 *Corresponding Author

a clear image from coarse to fine. Other works (Zamir et al., 2020a; Zamir et al., 2020b; Anwar & Barnes, 2019) paid attention to maintaining details of high resolution rather than using downsampling to expand the receptive field and process the feature map at the original resolution. Recently, a novel design that stacks several subnetworks and constructs a multistage network (Zamir et al., 2021) was proposed to progressively restore clear image stage by stage.

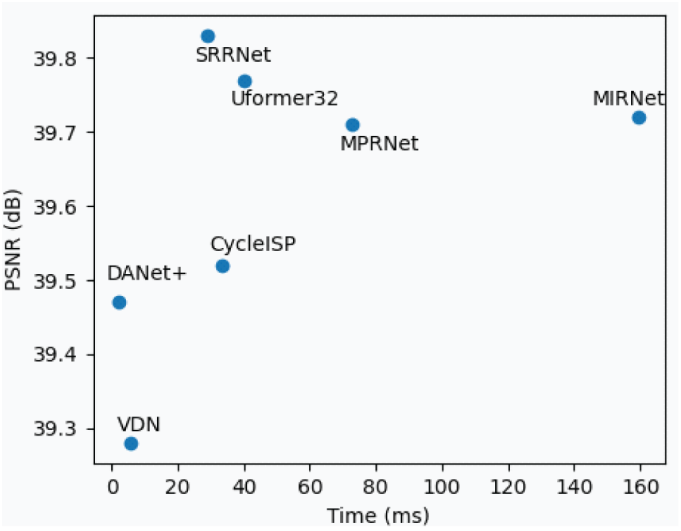
On the one hand, the essential properties of image-denoising tasks are explored, and specific denoisers are designed (Cheng et al., 2021). On the other hand, thanks to the successful development of self-attention in Natural Language Processing (NLP), the convolution block is replaced with the shifted windows (Swin) Transformer block (Wang et al., 2021; Liu et al., 2021) to capture long-range dependency and construct generalized denoiser.

However, the encoder–decoder-based methods are efficient, but the result is relatively poor, and other methods proved effective but very time-consuming. Due to the development of industrial cameras and mobile phones, the requirement for recovering clear images at little temporal cost is rapidly growing. Balancing performance and temporal cost needs to be addressed urgently. Therefore, the motivation and objective of this study were improving the traditional encoder–decoder-based architecture and exploring effective and efficient modules to make up for the deficiencies of accuracy so as to achieve a balance between performance and temporal requirements. This study was expected to encourage further research to explore effective and efficient denoising algorithms, considering the specific implementation of the algorithm in applied products.

To solve this problem, this study reinforced the interaction of information flow on the basis of the traditional encoder-decoder structure. Specifically, cross-level encoders are used to progressively extract self-refined features from coarse to fine. And the corresponding decoders with high-resolution reconstruction modules are passed to restore clear images hierarchically without losing the original characteristics. Then noise and signal of deep levels are discriminated without destroying the structure by multiscale dual attention blocks.

As vividly illustrated in Figure 1, the proposed self-refinement and reconstruction network (SRRNet) achieved excellent denoising accuracy with little temporal cost. The primary contributions of this paper are as below:

Figure 1. Comparison of the peak signal-to-noise ratio (PSNR) and temporal cost between the proposed SRRNet and other methods on the SIDD (Abdelhamed et al., 2018) data set with a Nvidia 3090 GPU. Note: The proposed SRRNet balances the performance and the temporal cost better than other works.



- A fast encoder–decoder-based self-refinement and reconstruction network (SRRNet) is proposed for image-denoising, which balances the performance and the temporal cost.
- A contextual self-refinement block (CSRB) is designed as the building block, which boosts information exchange and self-refining contextual details.
- A high-resolution reconstruction module (HRRM) is explored to reconstruct clear and high-resolution features under the guidance of a shallow information flow.
- A multiscale dual attention block (MDAB) is introduced to capture cross-scale information and concentrate on useful local details at different dimensions. A large number of comparative and ablation experiments are conducted to confirm the efficiency and effectiveness of SRRNets both in real-world image denoising and synthetic Gaussian denoising (Zhou et al., 2020).

This article is organized as follows. The Related Works section introduces the latest related image-denoising algorithms and analyzes the improvements of this study compared to other works. The Self-Refinement and Reconstruction Network section presents the details of the proposed SRRNet, and it includes four parts: overall pipeline, CSRB, HRRM, and MDAB. The Experiments section conducts many quantitative experiments to evaluate performance among SRRNet and other methods in real-world denoising tasks and Gaussian noise removal tasks. And a series of qualitative experiments are performed on each module to demonstrate the effectiveness of each design. Finally, the Conclusion section summarizes the conclusions, deficiencies analysis, and future outlook.

RELATED WORKS

In the U-shaped encoder–decoder-based architecture, the resolution of features shrinks in half by each downsampling layer, resulting in the unavoidable loss of some spatial details. This is the reason why those single encoder–decoder-based methods tend to achieve limited performance.

One solution is multistage progressive restoration (Zamir et al., 2021; Tu et al., 2022), which stacks several U-Nets to progressively restore a clearer image. Then the cross-stage feature fusions are used to aggregate the corresponding encoders and decoders between low-stage and high-stage, aiming at guiding the high-stage to restore more degraded features. These kinds of methods have been proven to be effective but sacrifice large computational and temporal costs.

Other works (Wang et al., 2021; Chang et al., 2020; Fan et al., 2022) introduce novel convolution modules or other technologies to the single encoder-decoder architecture, such as dilated convolution, deformable convolution, and transformer block. Spatial Adaptive Network (SADNet) takes advantage of deformable convolution on each decoder and uses the dilated convolution to capture multiscale features at deep level. However, its performance is not as good as other recent works. Differently, U-Shaped Transformer-32 (Uformer32) applies a transformer (Vaswani et al., 2017) to the encoder–decoder-based denoising network and achieves better performance and robustness than other convolution-based methods at different benchmarks, yet the temporal cost is much higher.

Earlier methods simply add (Kim et al., 2020) or concatenate (Yue et al., 2019, 2020) feature maps of encoders and decoders at each level, thus with limited results. Recently, Noise Basis Network (NBNet) has improved skip-connection modules of encoder–decoder-based architecture. By using several convolution blocks and a subspace attention module to better reconstruct the signal vectors, the computational cost is saved, and details of low-level features are maintained. However, NBNet is still limited in that fixing a basis signal vector is hard to deal with in different noisy scenarios.

Different from the multistage architecture and high-resolution iterative architecture, this paper adopts a more lightweight encoder-decoder architecture and replaces traditional skip connections between each encoder and decoder of the corresponding level by an efficient HRRM. The deep-level feature maps are enhanced by a novel multiscale attention unit. Based on the design of residual skip-connections (He et al., 2016), a simple but effective basic building block is introduced to further self-refine the contextual details throughout the iterative process.

SELF-REFINEMENT AND RECONSTRUCTION NETWORK

Overall Pipeline

As illustrated in Figure 2, the proposed self-refinement and reconstruction network is an encoder–decoder-based architecture. Specifically, when an original degraded noisy image N is input, SRRNet first uses a 3×3 convolutional layer to extract a shallow feature map F from the noisy image. Then the feature map is fed into an encoder-decoder architecture of four levels.

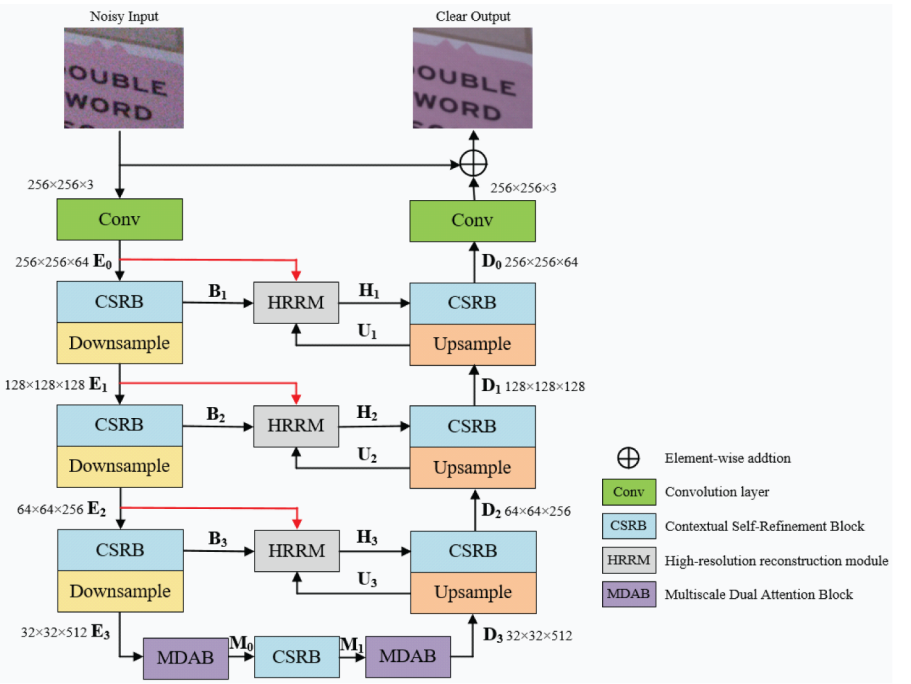
Each encoder consists of a Contextual Self-Refinement Block and a downsampling layer, which is a 3×3 convolution layer followed by a pixel unshuffle function (Shi et al., 2016). Each decoder includes an upsample layer, a high-resolution reconstruction module, and a Contextual Self-Refinement Block. A 3×3 convolution followed by a pixel shuffle function is used to upsample. Two Multiscale Dual-Attention Blocks are set in front and back of the bottleneck, respectively. It is worth emphasizing that a high-resolution reconstruction module not only fuses features from the corresponding encoder and decoder but also skip-connects shallow information from the upper-level encoder.

Finally, a 3×3 convolution is passed to restore features from the last decoder layer to a pixel-level residual feature R, and then it is added to the input image N to generate a clear output image C, that is, $C = N + R$.

Contextual Self-Refinement Block

Most existing methods simply stack several residual blocks as the basic building block. The residual skip-connection branch is conducive to maintaining early information and improving gradient descent. However, it inevitably causes the disappearance of contextual details during the iteration

Figure 2. The architecture of the proposed SRRNet. Note: Three key components are included: contextual self-refinement block (CSRB), high-resolution reconstruction module (HRRM), and multiscale dual attention (MDAB). Encoding features, decoding features, bridging features, features of HRRM and features of upsample are named E_k, D_k, B_k, H_k, and U_k, respectively, where k∈{1,2,3}. M_0 and M_1 are feature flow of MDAB).



process and outputs a smooth and blurry image. Therefore, this study used four self-refinement residual skip-connections at different levels. As illustrated in Figure 3, the first two used the improved residual blocks to refine shallow features. The third one was the long-range residual addition after an instance normalization to fuse original and stable features. The last one utilized a longer-range residual multiplication, which was treated as an attention weight to self-refine contextual relations.

In the residual block, the Rectified Linear Units (ReLU) activation function was replaced by Parametric Rectified Linear Units (PReLU). Compared with stacking several residual blocks, the combination of these four different self-refinement branches can accelerate convergence, stabilize gradient, and maintain more contextual details. More details of ablation studies are discussed in the Ablation Studies and Discussion section.

High-Resolution Reconstruction Module

Among the encoder–decoder-based methods, a common and simple practice is employing a skip connection to aggregate features of each encoder and the corresponding decoder. Some use concatenation (Yue et al., 2019, 2020), and others use addition (Kim et al., 2020). But both demonstrate limited effectiveness because only fusing feature maps of the same scale makes information flow inflexible and leads to losing spatial details of high resolution.

Inspired by the dense connection among cross-level features (Cho et al., 2021), the proposed HRRM not only combined the features from the current encoder (named B_k) and the features after the upsampling in the corresponding decoder (named U_k) but also included features from the previous encoder layer (named E_{k-1}). Extensive ablation studies in the Ablation Studies and Discussion section proved that combining high-resolution features from encoders into each encoder-decoder pair was conducive to gradually recovering the lost information.

As shown in Figure 4, firstly, B_k was processed by a 3×3 convolution and concatenated with U_k. Secondly, channel attention (CA) (Hu et al., 2018) was used to distinguish the weight of different channels in the combined feature map. CA consisted of a spatial-wise average pooling (SAP) to pool a feature map from a spatial resolution of Height×Width to a single pixel, a convolution layer with a PReLU activation function to squeeze channels, and another convolution layer followed by sigmoid to excite channels. Finally, the last 3×3 convolution was used to reconstruct and enhance E_{k-1}, and a long-range residual addition was used to aggregate with the aforementioned feature map as the output of HRRM. The overall pipeline of HRRM was formulated as:

$$Curr = CA\left(Cat\left(Conv\left(B_k\right), U_k\right)\right) \tag{1}$$

$$H_k = Curr + Conv\left(E_{k-1}\right) \tag{2}$$

Figure 3. The proposed CSRB. Note: It contains four self-refinement residual skip-connections. The input and output are E_{k-1} and B_k, respectively, at the encoding stage, while they become M_0 and M_1 at the bottleneck stage, and H_k and D_{k-1} at the decoding stage. All of these symbols correspond to the symbols in Figure 2, respectively.

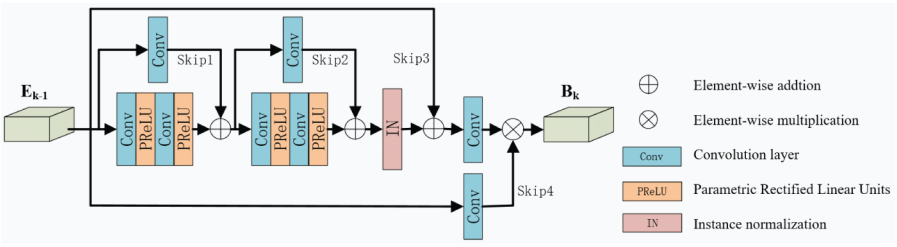
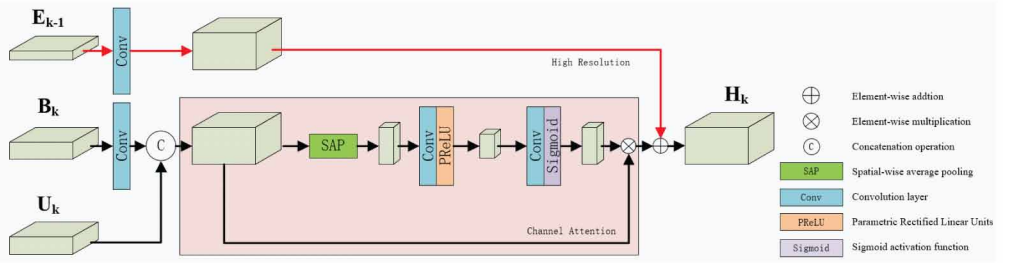


Figure 4. The proposed HRRM. Note: The output of the first feature extraction layer is treated as E_0. And high-resolution information flows are highlighted in red. E_{k-1}, B_k, U_k, and H_k are correspond to the symbols in Figure 2, respectively.



where $k \in \{1,2,3\}$ is the level of the encoder-decoder structure, $Conv(\cdot)$ is the 3×3 convolution operation (LeCun et al., 1998), $Curr$ means the combination of current level features, $CA(\cdot)$ means CA mechanism, and $Cat(\cdot,\cdot)$ denotes concatenate operation. The output of the first convolution was E_0.

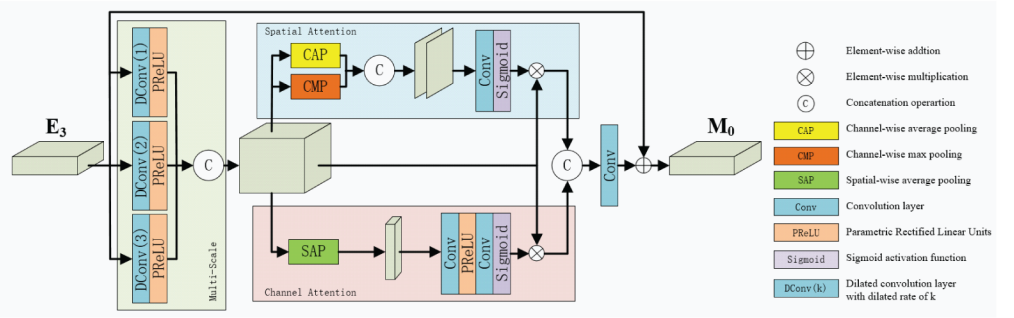
The long-range and cross-level connection could boost the interaction of information flow, learn low-level spatial details, and reconstruct high-resolution features from coarse to fine. The reason why this study did not aggregate features from the low-resolution (deep-level) encoders is that the deeper feature maps lose half resolution. Integrating the feature map of low resolution will lose more signal details and make trouble to the subsequent training schedule to some extent.

Multiscale Dual Attention Block

Inspired by the success of dilated convolution (Yu & Koltun, 2015) and attention mechanism (Liang et al., 2021; Zhang et al., 2018b, 2019) in image restoration, this paper proposes MDAB to distinguish signal features and noisy features at the deep level. It consists of two components, one for expanding receptive fields without downsampling, and the other for sharing information along the channel-wise and spatial-wise dimensions.

Dilated convolution is able to enlarge the receptive field without destroying the structures of the image or increasing the number of parameters. Therefore, as depicted in Figure 5, the module first applied three different dilated convolution layers and concatenated these outputs to facilitate the aggregation of cross-scale dependencies.

Figure 5. The details of the proposed MDAB. Note: It includes multiscale feature extraction and dual-attention branches. The input and output of the first MDAB is E_3 and M_0, and they become M_1 and D_3 in the second one. All of these symbols correspond to the symbols in Figure 2, respectively.



Then, channel attention and spatial attention (Woo et al., 2018) were applied to the concatenation of the abovementioned multiscale feature maps to calculate appropriate weights for different channels and pixels. Specifically, the channel attention branch was the same as the one mentioned in the High-Resolution Reconstruction Module section. And the spatial attention branch used a channel-wise global average pooling and a channel-wise global max pooling followed by a concatenation operation to encode global contextual information. Then the excitation operation passed a convolution layer and a sigmoid activation function to calculate the spatial-wise attention weight. The dual attentions can suppress degraded information and enhance clear information at the channel dimension and spatial dimension.

Finally, the module concatenated the output of the two abovementioned branches, and convolution was used to refine the details of the feature map. A residual connection was added to maintain original spatial and contextual details. Given an input X_0 with a size of Height×Width×Channel, the overall pipeline of MDAB was formulated as:

$$MS = Cat\left(DConv\left(1, X_0\right), DConv\left(2, X_0\right), DConv\left(3, X_0\right)\right) \tag{3}$$

$$MDAB\left(X_0\right) = Conv\left(Cat\left(CA\left(MS\right), SA\left(MS\right)\right)\right) + X_0 \tag{4}$$

where $Cat(\cdot,\cdot)$ denotes concatenate operation and $DConv(\cdot, i)$ means that dilated convolution (Yu & Koltun, 2015) with dilation rate equals *i*, and *MS* is the multiscale feature map.

Different from Multi-scale Image Restoration Network (MIRNet) (Zamir et al., 2020b), the proposed MDAB reinforced two attention modules by multiscale refined features, and it was conducted on the deep level of encoder-decoder architecture rather than the single-scale iteration process. Compared with other works (Chen et al., 2021), another difference lies in the overall layers, which are reduced from five to four layers, considering the disadvantage of destroying the structure caused by downsampling. Therefore, the proposed MDAB efficiently aggregated multiscale details and captured contextual locality in both channel-wise and spatial-wise dimensions, and thereby, the model achieved better results with a simpler structure. The ablation studies in the Ablation Studies and Discussion section verified this conclusion.

Experiments

In this section, details regarding a large number of comparison experiments and ablation experiments conducted to evaluate the effectiveness of the proposed SRRNet on both real-world image-denoising tasks and synthetic Gaussian noise removal tasks are reported. This section also describes the full implementation details and demonstrates the results on different denoising benchmarks. Finally, many ablation studies are performed to confirm the advantage of each proposed CSRB, MDAB, and HRRM, and discuss the superiority of the proposed method in balancing the performance and temporal cost. In each table, the best and the second-best peak signal-to-noise ratios (PSNRs) and structural similarities (SSIMs) of the evaluated methods are highlighted and underlined.

Experimental Setting

The proposed SRRNet was implemented in the Ubuntu 20.04 operating system with a Nvidia 3090 GPU. The model was trained on noisy-clear image pairs with a resolution of 256×256. The first features extraction changed the channel number of the original image to 64, each encoder halved the size of the feature map and doubled the number of channels, and each decoder performed the opposite operation. Therefore, the size of each feature map at a different level was 256×256×64, 128×128×128, 64×64×256, and 32×32×512, respectively. To better utilize the parallel computing

efficiency of a GPU, each level of the network only adopted a single CSRB, rather than setting more blocks and less channels.

For real-world image denoising, this study used SIDD (Abdelhamed et al., 2018), DND (Plotz & Roth, 2017), and Renoir (Anaya & Barbu, 2018) data sets. As to synthetic Gaussian denoising, this study used the combination of DIV2K (Agustsson & Timofte, 2017), WED (Ma et al., 2016), Flickr2K (Chen et al., 2017), and BSD500 (Martin et al., 2001) as the training set to learn more general priors, and evaluated performance on Set12 (Zhang et al., 2017a), BSD68, Urban100 (Huang et al., 2015), Kodak24 (Franzen, 1999), and McMaster (Zhang et al., 2011), respectively.

For training configuration, the batch size was fixed to 30. The minibatch iteration was set to 200,000. Flip and rotation were randomly applied for data augmentation. An AdamW optimizer was used with an initial learning rate of 0.0002, weight decay of 0, beta1 of 0.9, beta2 of 0.999, and epsilon of 1×10^{-8}. The learning rate was gradually reduced to 1×10^{-7}, using the One-cycle learning rate scheduler with cosine annealing strategy; cycle momentum was 0.85. The model was trained with Charbonnier loss function (Charbonnier et al., 1994) as follows:

$$L_{Char} = \sqrt{\left\| \widehat{X} - Y \right\|^2 + \varepsilon^2} \tag{5}$$

where $\widehat{X}$ and Y represent the prediction and the ground truth, respectively, and ε is set to 0.001.

Results on the SIDD Benchmark

Smartphone Image Denoising Data set (SIDD) consists of 30,000 noisy images from different scenarios generated by five smartphone cameras, which is treated as the benchmark for real-world noise reduction tasks. Table 1 illustrates the comparison results between the proposed method and other related methods, MIRNet, MPRNet (Zamir et al., 2021), NBNet, Uformer32, etc. Compared with the other methods, the accuracy of the proposed SRRNet exceeded NBNet by 0.14 dB, and exceeeds SADNet by 0.43 dB. Furthermore, the denoising accuracy of SRRNet was higher than that of the transformer-based method like Uformer32. SRRNet+ estimates the results of SRRNet with a test-time data augmentation named geometric self-ensemble (GSE) (Lim et al., 2017).

The visualization of noise reduction resulting from different methods is provided in Figure 6. The proposed method's ability to remove noise was significantly better than that of Color Block Matching 3D (CBM3D) (Dabov et al., 2007) and Real Image Denoising Network (RIDNet), and the proposed method preserves the original stripe structure better than MPRNet and MIRNet. It proved that the proposed SRRNet was able to remove noisy interference better and preserves the original color and stripes.

Results on the DND Benchmark

The Darmstadt Noise Dataset (DND) includes 50 noisy-clear image pairs in different real-world scenarios, where the clear data are taken at a low film speed level, while the other noisy data are generated at a higher film speed level. All image pairs are postprocessed later, and the DND benchmark website is available for online quantitative comparisons. It is worth mentioning that the DND benchmark consists only of testing data and provides online judgment for evaluating denoising performance. Therefore, SIDD and Renoir data sets are combined as a training set (Chen et al., 2021).

Table 1 demonstrates the comparison of various methods. The accuracy of the proposed SRRNet exceeds NBNet by 0.11 dB and exceeds SADNet by 0.41 dB. Especially the proposed method also performs better than the well-accepted transformer-based method Uformer32.

This study presents the visual comparison results on the DND benchmark in Figure 7. The proposed SRRNet can reconstruct a clear image while keeping the original sharpness and textures.

Table 1. Comparison experiments of real-world image noise removal tasks on SIDD and DND benchmarks

Methods	Publication	SIDD		DND	
		PSNR	SSIM	PSNR	SSIM
BM3D (Dabov et al., 2007)	TIP 2007	25.65	0.685	34.51	0.851
CBDNet (Guo et al., 2019)	CVPR 2019	30.78	0.801	38.06	0.942
RIDNet (Anwar & Barnes, 2019)	ICCV 2019	38.71	0.951	39.26	0.953
AINDNet (Kim et al., 2020)	CVPR 2020	38.95	0.952	39.37	0.951
VDN (Yue et al., 2019)	NeurIPS 2019	39.28	0.956	39.38	0.952
SADNet (Chang et al., 2020)	ECCV 2020	39.46	0.956	39.59	0.952
DANet (Yue et al, 2020)	ECCV 2020	39.47	0.957	39.58	0.955
CycleISP (Zamir et al., 2020a)	CVPR 2020	39.52	0.957	39.56	0.956
MPRNet (Zamir et al., 2021)	CVPR 2021	39.71	0.958	39.80	0.954
MIRNet (Zamir et al., 2020b)	ECCV 2020	39.72	0.959	39.88	0.956
SRMNet (Fan et al., 2022)	EUSIPCU 2022	39.72	0.959	39.44	0.951
NBNet (Cheng et al., 2021)	CVPR 2021	39.75	0.959	39.89	0.955
Uformer32 (Wang et al., 2021)	Arxiv 2021	39.77	**0.970**	39.96	0.956
SRRNet	–	39.83	**0.970**	39.96	0.956
SRRNet+	–	**39.89**	**0.970**	**40.00**	**0.957**

Figure 6. Real image denoising comparisons on SIDD. Note: The proposed SRRNet can preserve the original colors and stripes better than other methods.

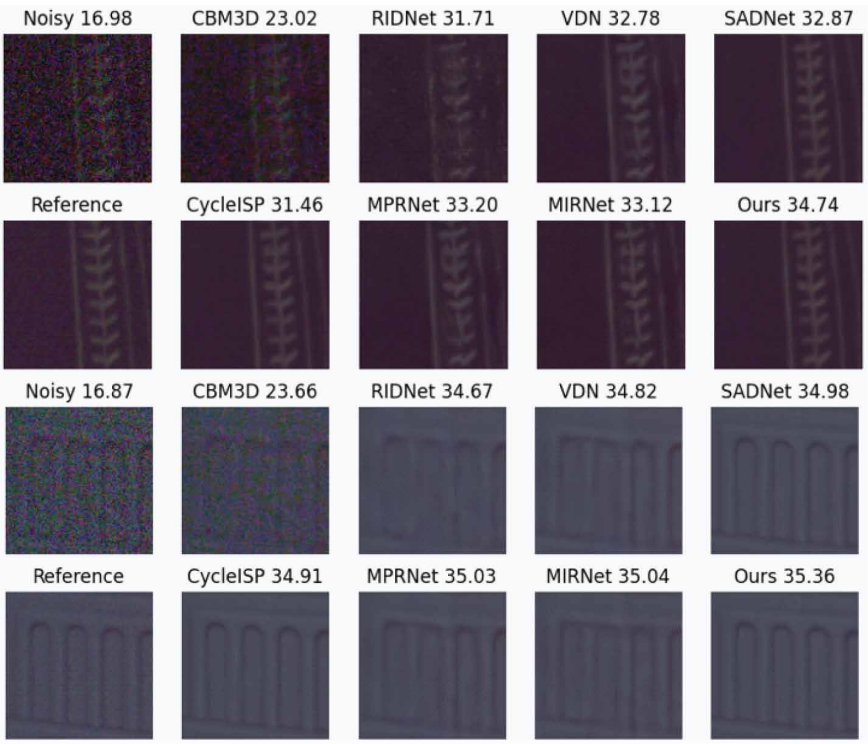
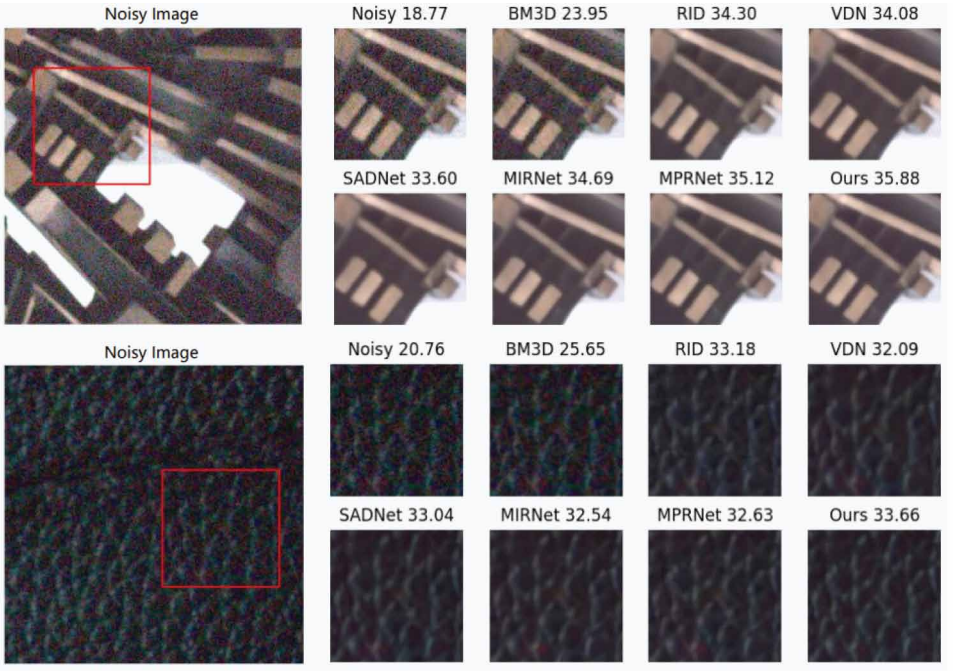


Figure 7. Visual comparison on DND online judgment for real-world image denoising



Results of Synthetic Gaussian Noise Removal

This paper also evaluated the proposed SRRNet on several synthetic Gaussian denoising data sets. The training data included 800 clear images from DIV2K, 2,650 images from Flickr2K, 4,744 images from WED, and 400 images from BSD500. The training details are the same as those given in the Experimental Setting section, except for the training image pairs.

This paper used the same additive white Gaussian noise generation (AWGN) method as the related work Zhou et al. (2020). The Gaussian noise and the noisy image were generated by:

$$N^1 = O + n^1, n_{ij}^1 \sim N\left(0, \left(\frac{\sigma}{255}\right)^2\right) \tag{6}$$

$$N = C + N^1 \tag{7}$$

where O is the zero mask sharing the equivalent image shape with the clear image, C is the given clear image, and σ is the noise level. The noise levels of 15, 25, and 50 were employed in the experiment.

Table 2 lists the results of DnCNN (Bae et al., 2017), FFDNet (Zhang et al., 2018a), IRCNN (Zhang et al., 2017b), FOCNet (Jia et al., 2019), MWCNN (Liu et al., 2019), DeamNet (Ren et al., 2021), and SRRNet on grayscale Gaussian denoising data sets, where the proposed SRRNet outperforms other methods on Set12, BSD68, and Urban100.

Furthermore, the results on color Gaussian noise with different noise levels also verified that the proposed approach surpasses other methods on color images, including CBSD68, Urban100, Kodak24, and McMaster, as shown in Table 3. This proves the superiority of the proposed SRRNet that it can

Table 2. The average peak signal-to-noise ratios (dB) results of gray-scale synthetic Gaussian image noise removal tasks compared to other competitive methods

Data Sets	σ	DnCNN	FFDNet	IRCNN	FOCNet	MWCNN	DeamNet	Ours
Set12	15	32.67	32.75	32.76	33.07	33.15	33.19	**32.26**
	25	30.25	30.43	30.37	30.73	30.79	30.81	**30.92**
	50	27.18	27.32	27.12	27.68	27.74	27.74	**27.82**
BSD68	15	31.62	31.63	31.63	31.83	31.86	**31.91**	**31.91**
	25	29.16	29.19	29.15	29.38	29.41	29.44	**29.45**
	50	26.23	26.29	26.19	26.50	26.53	26.54	**26.55**
Urban100	15	32.28	32.40	32.46	33.15	33.17	33.37	**33.45**
	25	29.80	29.90	29.80	30.64	30.66	30.85	**31.00**
	50	26.35	26.50	26.22	27.40	27.42	27.53	**27.74**

Table 3. The average peak signal-to-noise ratio (dB) results of color-scale synthetic Gaussian image noise removal tasks compared to other competitive methods

Data Sets	σ	IRCNN	FFDNet	DnCNN	DSNet	RPCNN	BRDNet	Ours
CBSD68	15	33.86	33.87	33.90	33.91	–	34.10	**34.29**
	25	31.16	31.21	31.24	31.28	31.24	31.43	**31.68**
	50	27.86	27.96	27.95	28.05	28.06	28.16	**28.47**
Kodak24	15	34.69	34.63	34.60	34.63	–	34.88	**35.18**
	25	32.18	32.13	32.14	32.16	32.34	32.41	**32.80**
	50	28.93	28.98	28.95	29.05	29.25	29.22	**29.71**
McMaster	15	34.58	34.66	33.45	34.67	–	35.08	**35.31**
	25	32.18	32.35	31.52	32.40	32.33	32.75	**33.08**
	50	28.91	29.18	28.62	29.28	29.33	29.52	**30.02**
Urban100	15	33.78	33.83	32.98	–	–	34.42	**34.75**
	25	31.20	31.40	30.81	–	31.81	31.99	**32.56**
	50	27.70	28.05	27.59	–	28.62	28.56	**29.54**

remove synthetic Gaussian noise in many scenarios better than IRCNN, FFDNet, DnCNN, DSNet (Peng et al., 2019), RPCNN (Xia & Chakrabarti, 2020), BRDNet (Tian et al., 2020).

Ablation Studies and Discussion

A small version of SRRNet was built for ablation study, and the experiments are performed on SIDD benchmarks to validate the contribution of each module. The simplified design halved the channel numbers and training iterations. The other experimental settings are the same as what was described in the Experimental Setting section.

Ablation on Contextual Self-Refinement Block

Four different residual skip-connections were ablated, and the comparison is presented in Table 4. The PSNR increased 0.10 dB after the application of more skip-connections of different types. As the

Table 4. Ablation results of the proposed CSRB

Skip1	✓	✓	✓	✓
Skip2	×	✓	✓	✓
Skip3	×	×	✓	✓
Skip4	×	×	×	✓
PSNR	39.51	39.55	39.56	39.61

crucial building block, CSRB resulted in a 0.22 dB improvement in the complete ablation experiment, which confirmed the contribution of CSRB.

Ablation on Multiscale Dual Attention

The core design of MDAB includes multiscale feature extraction and dual attention mechanism. Table 5 shows that both of them worked well for noise removal at different levels, and the combination of them was the best choice.

Ablation on High-Resolution Reconstruction Modules

Compared with concatenation-based skip-connection between encoders and decoders, HRRM was able to maintain more high-resolution details. Table 6 verifies the effectiveness of the additional channel attention module and the feature fusion of the high-resolution.

The Overall Ablation Results

Table 7 illustrates the ablation results of different components with complete experimental configuration, including parameters, FLOPs, and PSNRs. Each module contributed to the increase of PSNRs, which demonstrated the effectiveness.

Figure 8 shows local visualized images (in red rectangle) generated by different ablation models. Traditional U-shaped Network (UNet) may lose stripe details, but the original stripe details are better preserved, and a clearer image is reconstructed after the proposed modules are introduced.

Table 5. Ablation results of the proposed MDAB

Multiscale	×	✓	×	✓
Dual Attention	×	×	✓	✓
PSNR	39.49	39.55	39.51	39.61

Table 6. Ablation results of the proposed HRRM

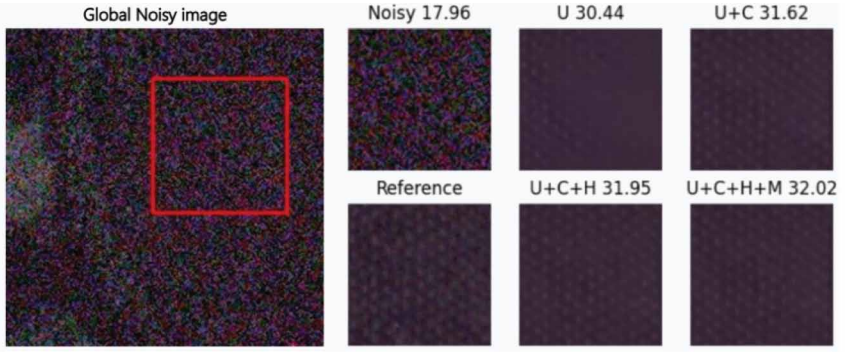
Concatenation (Concat)	✓	✓	✓
Channel Attention (CA)	×	✓	✓
High Resolution (HR)	×	×	✓
PSNR	39.56	39.59	39.61

Table 7. Ablation study of different components with complete experimental configurations

CSRB	×	✓	✓	✓	✓
MDAB	×	×	✓	✓	✓
HRRM	×	×	×	✓	✓
GSE	×	×	×	×	✓
Params (M)	14.55	27.16	55.17	56.16	56.16
FLOPs (G)	63.60	187.02	214.21	223.90	223.90
PSNR	39.53	39.75	39.78	39.83	39.89

Note: The PSNR (dB) values are based on a SIDD data set.

Figure 8. Results of local visualized details and PSNR. Note: Denoising results with and without the proposed CSRB (C), HRRM (H), and MDAB (M) modules. UNet (U) easily loses stripe details in image denoising, and all of the proposed modules are conducive to alleviating this problem.



Discussion of Feasibility and Complexity

This section further analyzes the feasibility of the proposed method and compares it with other works to analyze its advantages and practical application value. The comparison results of PSNRs, FLOPs, and the temporal cost with an image size of 256×256×3 are provided in Table 8. It is worth emphasizing that compared with the high-resolution (single-scale) -based method MIRNet, the multistage method MPRNet, and the transformer-based model Uformer32, the proposed single encoder–decoder-based

Table 8. Comparing the PSNRs of SIDD, FLOPs, and time among MIRNet, MPRNet, Uformer32, SRMNet, MAXIM, and ours

Method	PSNR (dB)	FLOPs (G)	Proportion	Time (ms)	Speedup
MIRNet	39.72	788.04	100%	159.64	1×
MPRNet	39.71	573.88	72.8%	72.91	2.2×
Uformer32	39.77	40.86	5.2%	40.08	4.0×
SRMNet	39.72	285.36	36.2%	39.31	4.1×
MAXIM	39.96	339.20	43.0%	107.67	1.5×
SRRNet	39.83	223.90	28.4%	28.99	5.5×

Note: FLOPs and Time are tested with the input size of 256 × 256 × 3 on Nvidia 3090 GPU. Proportion and Speedup are the ratio of FLOPs and multiplier of speed compared to MIRNet.

SRRNet achieves the highest accuracy with minimal temporal cost. Compared with MIRNet, the proposed method only used 28.4% of the FLOPs and had a speed increase of 5.5 times. Although the FLOPs of transformer methods like Uformer32 were smaller, SRRNet worked about 1.4 times faster and had better performance than Uformer32.

Compared with the most recent works, Selective Residual M-shaped Network (SRMNet) (Fan et al., 2022) and Multi-Axis Multi-layer perceptron (MAXIM) (Tu et al., 2022), the proposed SRRNet was superior to SRMNet in denoising accuracy, computational cost, and temporal cost. Although its accuracy was a little lower than MAXIM, the FLOPs and time were only 66.0% and 26.9% of MAXIM, respectively. This study also observed that when the experiments are performed on GPUs with poorer computing power (such as Nvidia 1080Ti), the advantage of the temporal cost of the proposed SRRNet was greater. It proved that the traditional encoder-decoder architecture also had the potential to achieve excellent denoising performance while meeting the needs of little complexity and fast inference after introducing effective and efficient design. The proposed SRRNet balanced the effectiveness and temporal cost better in image-denoising tasks.

CONCLUSION

This paper summarizes the problem that recent image denoisers have in balancing effectiveness and efficiency. The inference speed of traditional U-shaped encoder–decoder-based networks is fast but shows limited performance and up-to-date deep networks achieve high accuracy but sacrifice computation and inference time. To this end, an SRRNet is proposed for efficient image denoising. It is based on a simple encoder-decoder architecture and introduces three improvements named CSRB, HRRM, and MDAB so as to make up for the deficiencies of accuracy. Extensive experiments prove that the proposed method achieves excellent performance in real-world denoising tasks and Gaussian noise removal tasks with minimal computational cost and temporal cost. The study is expected to encourage further research to explore effective and efficient denoising algorithms, considering the specific implementation of the algorithm in applied products.

The limitation of this study is that the proposed method is not applied to more image restoration tasks, such as image deblurring, image deraining, super-resolution, etc. Compared to the recent deep and complex networks, its restoration performance and robustness are not necessarily guaranteed. This study also observes that replacing a convolution layer with a transformer layer is another choice, and its robustness is higher in many scenarios, but it is time-consuming. Therefore, further exploration of faster and more lightweight transformer blocks and reconstruction modules for image restoration is necessary for the next work.

AUTHOR NOTE

The data used to support the findings of this study are included within the article. The authors declare that there is no conflict of interest regarding the publication of this paper. This research is supported by the National Natural Science Foundation of China (No. 61972180).

REFERENCES

Abdelhamed, A., Lin, S., & Brown, M. S. (2018, June 18–23). A high-quality denoising dataset for smartphone cameras. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2018.00182

Agustsson, E., & Timofte, R. (2017, July 21–26). Ntire 2017 challenge on single image super-resolution: Dataset and study. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. doi:10.1109/CVPRW.2017.150

Anaya, J., & Barbu, A. (2018). Renoir–A dataset for real low-light image noise reduction. *Journal of Visual Communication and Image Representation*, *51*, 144–154. doi:10.1016/j.jvcir.2018.01.012

Anwar, S., & Barnes, N. (2019, October 27–November 2). Real image denoising with feature attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. doi:10.1109/ICCV.2019.00325

Bae, W., Yoo, J., & Chul, Ye., J. (2017, July 21–26). Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. doi:10.1109/CVPRW.2017.152

Chang, M., Li, Q., Feng, H., & Xu, Z. (2020, August 23–28). Spatial-adaptive network for single image denoising. In *16th European Conference Computer Vision–ECCV 2020, Proceedings Part XXX*. Springer International Publishing. doi:10.1007/978-3-030-58577-8_11

Charbonnier, P., Blanc-Feraud, L., Aubert, G., & Barlaud, M. (1994, November 13–16). Two deterministic half-quadratic regularization algorithms for computed imaging. *Proceedings of the 1st International Conference on Image Processing*. doi:10.1109/ICIP.1994.413553

Chen, Y. L., Huang, T. W., Chang, K. H., Tsai, Y. C., Chen, H. T., & Chen, B. Y. (2017, March 24–31). *Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study* [Conference session]. In 2017 IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, CA. doi:10.1109/WACV.2017.32

Cheng, S., Wang, Y., Huang, H., Liu, D., Fan, H., & Liu, S. (2021, June 20–25). NBNet: Noise basis learning for image denoising with subspace projection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR46437.2021.00486

Cho, S. J., Ji, S. W., Hong, J. P., Jung, S. W., & Ko, S. J. (2021, October 10-17). Rethinking coarse-to-fine approach in single image deblurring. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. doi:10.1109/ICCV48922.2021.00460

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, *16*(8), 2080–2095. doi:10.1109/TIP.2007.901238 PMID:17688213

Fan, C. M., Liu, T. J., Liu, K. H., & Chiu, C. H. (2022, August 29–September 2). *Selective residual M-Net for real image denoising* [Conference session]. In 2022 30th European Signal Processing Conference, Belgrade, Serbia. doi:10.23919/EUSIPCO55093.2022.9909521

Franzen, R. (1999). *Kodak lossless true color image suite*. Kodak Lossless True Color Image Suite. https://r0k.us/graphics/kodak

Guo, S., Yan, Z., Zhang, K., Zuo, W., & Zhang, L. (2019, June 15–20). Toward convolutional blind denoising of real photographs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2019.00181

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June 27–30). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2016.90

Hu, J., Shen, L., & Sun, G. (2018, June 18–23). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2018.00745

Huang, J. B., Singh, A., & Ahuja, N. (2015, June 7–12). Single image super-resolution from transformed self-exemplars. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2015.7299156

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017, July 21–26). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2017.632

Jia, X., Liu, S., Feng, X., & Zhang, L. (2019, June 15–20). Focnet: A fractional optimal control network for image denoising. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2019.00621

Kim, Y., Soh, J. W., Park, G. Y., & Cho, N. I. (2020, June 13–19). Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR42600.2020.00354

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. doi:10.1109/5.726791

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021, October 11–17). Swinir: Image restoration using Swin transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. doi:10.1109/ICCVW54120.2021.00210

Lim, B., Son, S., Kim, H., Nah, S., & Lee, M., K. (2017, July 21–26). Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. doi:10.1109/CVPRW.2017.151

Liu, P., Zhang, H., Lian, W., & Zuo, W. (2019). Multi-level wavelet convolutional neural networks. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 74973–74985. doi:10.1109/ACCESS.2019.2921451

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021, October 10–17). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. doi:10.1109/ICCV48922.2021.00986

Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., & Zhang, L. (2016). Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, *26*(2), 1004–1016. doi:10.1109/TIP.2016.2631888 PMID:27893392

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001, July 7–14). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings of the Eighth IEEE International Conference on Computer Vision*. doi:10.1109/ICCV.2001.937655

Peng, Y., Zhang, L., Liu, S., Wu, X., Zhang, Y., & Wang, X. (2019). Dilated residual networks with symmetric skip connection for image denoising. *Neurocomputing*, *345*, 67–76. doi:10.1016/j.neucom.2018.12.075

Plotz, T., & Roth, S. (2017, July 21–26). Benchmarking denoising algorithms with real photographs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2017.294

Ren, C., He, X., Wang, C., & Zhao, Z. (2021, June 20–25). Adaptive consistency prior based deep network for image denoising. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR46437.2021.00849

Ronneberger, O., Fischer, P., & Brox, T. (2015, October 5–9). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention: 18th International Conference, Proceedings, Part III 18.* Springer International Publishing. doi:10.1007/978-3-319-24574-4_28

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016, June 27–30). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2016.207

Tian, C., Xu, Y., & Zuo, W. (2020). Image denoising using deep CNN with batch renormalization. *Neural Networks*, *121*, 461–473. doi:10.1016/j.neunet.2019.08.022 PMID:31629201

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., & Li, Y. (2022, June 18–24). Maxim: Multi-axis MLP for image processing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR52688.2022.00568

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. ArXiv. doi:10.48550/arXiv.1706.03762

Wang, Z., Cun, X., Bao, J., & Liu, J. (2021). *Uformer: A general U-shaped transformer for image restoration*. ArXiv. doi:10.48550/arXiv.2106.03106

Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018, September 8–14). Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision*. doi:10.1007/978-3-030-01234-2_1

Xia, Z., & Chakrabarti, A. (2020). Identifying recurring patterns with deep neural networks for natural image denoising. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. doi:10.1007/978-3-030-01234-2_1

Yu, F., & Koltun, V. (2015). *Multi-scale context aggregation by dilated convolutions*. ArXiv. doi:10.48850/arXiv.1511.07122

Yue, Z., Yong, H., Zhao, Q., Meng, D., & Zhang, L. (2019). *Variational denoising network: Toward blind noise modeling and removal*. ArXiv. doi:10.48850/arXiv.1908.11314

Yue, Z., Zhao, Q., Zhang, L., & Meng, D. (2020, August 23–28). Dual adversarial network: Toward real-world noise removal and noise generation. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part X 16.* Springer International Publishing. doi:10.1007/978-3-030-58607-2_3

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2020, June 13–19). Cycleisp: Real image restoration via improved data synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR42600.2020.00277

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2020, August 23–28). Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XXV 16.* Springer International Publishing. doi:10.1007/978-3-030-58595-2_30

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., Yang, M. H., & Shao, L. (2021, June 20–25). Multi-stage progressive image restoration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR46437.2021.01458

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, *26*(7), 3142–3155. doi:10.1109/TIP.2017.2662206 PMID:28166495

Zhang, K., Zuo, W., Gu, S., & Zhang, L. (2017, July 21–26). Learning deep CNN denoiser prior for image restoration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2017.300

Zhang, K., Zuo, W., & Zhang, L. (2018). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, *27*(9), 4608–4622. doi:10.1109/TIP.2018.2839891 PMID:29993717

Zhang, L., Wu, X., Buades, A., & Li, X. (2011). Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic Imaging*, *20*(2), 023016–023016. doi:10.1117/1.3600632

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018, September 8-14). Image super-resolution using very deep residual channel attention networks. *Proceedings of the European Conference on Computer Vision*. doi:10.1007/978-3-030-01234-2_18

Zhang, Y., Li, K., Li, K., Zhong, B., & Fu, Y. (2019). *Residual non-local attention networks for image restoration*. ArXiv. doi:10.48550/arXiv.1903.10082

Zhou, Y., Jiao, J., Huang, H., Wang, Y., Wang, J., Shi, H., & Huang, T. (2020, April). When awgn-based denoiser meets real noises. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(7), 13074–13081. doi:10.1609/aaai.v34i07.7009

Jinqiang Xue, born in 1998. He is an M.S. candidate at Jiangnan University, and the member of CCF. His research interests include deep learning and computer vision.

Qin Wu received her MS degrees in Computer Science and Ph.D. degree in mathematics from West Virginia University, Morgantown, WV, USA, in 2011. She is an associate professor with Jiangnan University, Wuxi, China. Her current research interests include computer vision and machine learning.