The Visual Computer
https://doi.org/10.1007/s00371-024-03599-6

ORIGINAL ARTICLE

A novel single-stage network for accurate image restoration

Hu Gao[1] · Jing Yang[1] · Ying Zhang[1] · Ning Wang[1] · Jingfan Yang[1] · Depeng Dang[1]

Accepted: 30 July 2024

Abstract
Image restoration is the task of aiming to obtain a high-quality image from a corrupt input image, such as deblurring and deraining. In image restoration, it is typically necessary to maintain a complex balance between spatial details and contextual information. Although a multi-stage network can optimally balance these competing goals and achieve significant performance, this also increases the system's complexity. In this paper, we propose a mountain-shaped single-stage design, which achieves the performance of multi-stage networks through a plug-and-play feature fusion middleware. Specifically, we propose a plug-and-play feature fusion middleware mechanism as an information exchange component between the encoder-decoder architectural levels. It seamlessly integrates upper-layer information into the adjacent lower layer, sequentially down to the lowest layer. Finally, all information is fused into the original image resolution manipulation level. This preserves spatial details and integrates contextual information, ensuring high-quality image restoration. Simultaneously, we propose a multi-head attention middle block as a bridge between the encoder and decoder to capture more global information and surpass the limitations of the receptive field of CNNs. In order to achieve low system complexity, we removes or replaces unnecessary nonlinear activation functions. Extensive experiments demonstrate that our approach, named as M3SNet, outperforms previous state-of-the-art models while using less than half the computational costs, for several image restoration tasks, such as image deraining and deblurring. The code and the pre-trained models will be released at https://github.com/Tombs98/M3SNet.

Keywords Image restoration · Single-stage · Feature fusion middleware · Multi-head attention middle block

1 Introduction

Image degradation is a common issue that occurs during image acquisition due to a variety of factors such as camera limitations, environmental conditions, and human factors. For instance, smartphone cameras with narrow apertures, small sensors, and limited dynamic range can produce blurred and noisy images due to device shaking caused by body movements. Similarly, images captured in adverse weather conditions can be affected by haze and rain. Most classical image restoration tasks can be formulated as:

$$\mathbf{L} = \mathbf{D}(\mathbf{H}) + \gamma \tag{1}$$

where $\mathbf{L}$ denotes an observed low-quality image, $\mathbf{H}$ refers to its corresponding high-quality image, and $\mathbf{D}(\cdot)$, γ indicate the degradation function and the noise during the imaging and transmission processes, respectively. This formulation can signify different image restoration tasks when $\mathbf{D}(\cdot)$ varies (Fig. 1).

Image restoration aims to recover the high-quality clean image $\mathbf{H}$ from its degraded image $\mathbf{L}$. It is a highly ill-posed problem as there are many candidates for any original input. In order to restrict the infinite feasible candidates space to natural images, traditional methods [1–7] explicitly design appropriately priors for the given kind of restoration problem, such as domain-relevant priors and task-relevant priors. Then, the potential high-quality image can be obtained by solving a maximum a posteriori (MAP) problem:

$$\hat{\mathbf{H}} = \arg\max_{\mathbf{H}} \log P(\mathbf{L}|\mathbf{H}) + \log P(\mathbf{H}) \tag{2}$$

where $P(\mathbf{L}|\mathbf{H})$ represents the probability of observing the degraded image $\mathbf{L}$ given the clean image $\mathbf{H}$, and $P(\mathbf{H})$ represents the prior distribution of the clean image $\mathbf{H}$. This can also be expressed as a constrained maximum likelihood estimation:

✉ Depeng Dang
ddepeng@bnu.edu.cn

[1] School of Artificial Intelligence, Beijing Normal University, Beijing 100000, China

Published online: 26 August 2024

Fig. 1 Visualized results of M3SNet on various image restoration tasks. Left: degraded image. Right: the predicted result of M3SNet. From top to bottom: image deblurring, and image deraining task respectively

$$\hat{\mathbf{H}} = \arg\min_{\mathbf{H}} \|\mathbf{L} - \mathbf{D}(\mathbf{H})\|^2 + \lambda\Psi(\mathbf{H}) \tag{3}$$

where fidelity term $|\mathbf{L} - \mathbf{D}(\mathbf{H})|^2$ serves as an approximation for the likelihood $P(\mathbf{L}|\mathbf{H})$, while the regularization term $\lambda\Psi(\mathbf{H})$ represents either the priors of the latent image $\mathbf{H}$ or the constraints on the solution. The aim is to express the fidelity of the reconstructed image to the original input while simultaneously considering the prior knowledge or constraints imposed on the solution

While designing effective priors for image restoration can be challenging and may not be generalizable. With large-scale data, deep models such as Convolutional Neural Networks(CNNs) [8–17] and Transformer [18–22] have emerged as the preferred choice due to their ability to implicitly learn more general priors by capturing natural image statistics and achieving state-of-the-art (SOTA) performance in image restoration. The performance gain of these deep learning models over conventional restoration approaches is primarily attributed to their model design, which includes numerous network modules and functional units for image restoration, such as recursive residual learning [23], transformer [18, 19, 21], encoder-decoders [12, 13, 24], multi-scale models [25–27], and generative models [28–30].

Nevertheless, most of these models for low-level vision problems are based on a single-stage design, which ignores the interactions that exist between spatial details and contextualized information. To address this limitation, [8–11] proposes a multi-stage architecture in which contextualized features are first learned through an encoder-decoder architecture and subsequently integrated with a high-resolution branch to preserve local information. Despite its good performance, this method requires refining the results from the previous stage in the later stage, leading to a high level of system complexity.

Based on the information presented, a natural question that comes to mind is whether it is feasible to use a single-stage architecture to reduce system complexity and achieve the same balance between spatial details and contextualized information as a multi-stage architecture while maintaining the SOTA performance. To achieve this objective, we propose a mountain-shaped single-stage image restoration architecture, called M3SNet, with several key components. (1). We utilize NAFNet [12] as the baseline architecture and concentrate on modifying the network model to attain multi-stage functionality. By emitting the information transfer between the multi-stage and eliminating the nonlinear activation function from the network structure, we are able to reduce the system's complexity. (2). A plug and play feature fusion middleware (FFM) mechanism has been added to facilitate multi-scale information fusion between encoder and decoder blocks from different layers, resulting in the acquisition of more contextual information. Additionally, this approach enables manipulation of the original image resolution, thereby aiding in the preservation of spatial details. As a basic feature fusion block, it can be plug-and-play in various other image restoration networks to improve the model representation. (3). A multi-head attention middle block (MHAMB) is the bridge between the encoder and decoder that surpass the limitations of the receptive field of CNNs and capture more global information.

The main contributions of this work are:

(1) A novel single-stage approach capable of generating outputs that are contextually enriched and spatially accurate similar to a multi-stage architecture. Our architecture reduces system complexity due to its single-stage design

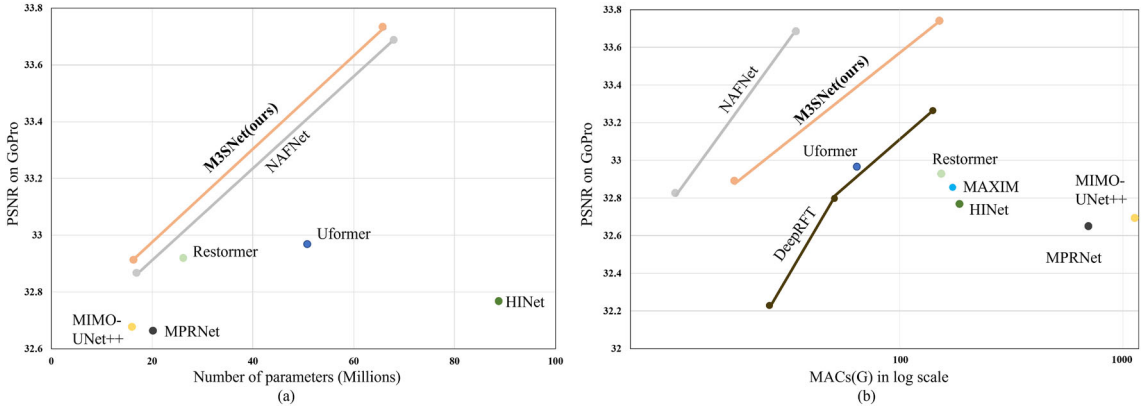


Fig. 2 PSNR vs. computational cost on Image Deblurring. Under different parameter capacities, our model achieves state-of-the-art. In addition, our model involves only a relatively small number of multiply-accumulate operations (MACs)

eliminates the need for information to be passed between stages.

(2) A feature fusion middleware mechanism (FFM) that enables the exchange of information across multiple scales while preserving the fine details from the input image to the output image. It can be used as a general and efficient plug-in module with few lightweight parameters for various image restoration networks.

(3) A multi-head attention middle block (MHAMB) that is capable of aggregating local and non-local pixel interactions.

(4) We demonstrate the versatility of M3SNet by setting new state-of-the-art on 6 synthetic and real-world datasets for various restoration tasks (image deraining and deblurring) while maintaining low complexity (see Fig. 2). Further, we provide detailed analysis, qualitative results, and generalization tests.

2 Related work

Image degradation is a common occurrence caused by camera equipment and a variety of environmental factors. Depending on the specific degradation phenomenon, different image restoration tasks are proposed, e.g., deblurring and deraining. Early image restoration work was mainly based on manually crafting some prior knowledge, such as total variation and self-similarity [1–7]. With the rise of deep learning, data-driven methods like CNN [8, 31–41] and Transformer [19–21, 42, 43] have become the dominant approach for image restoration due to their impressive performance. These methods can be categorized as either single-stage or multi-stage based on their architectural design.

2.1 Single-stage architecture

In recent years, the majority of image restoration research has focused on single-stage architecture. Among these architectures, the encoder-decoder based U-Net [8, 12, 13, 21, 30, 44–46] and dual network structure [47–51] are mainly included.

Encoder-Decoder Approaches. In recent years, encoder-decoder have gained great attention from researchers in the field of image restoration thanks to its ability to capture multi-scale information. To construct an effective and efficient Transformer-based architecture for image restoration, [21] introduce a novel locally-enhanced window and multi-scale restoration modulator to create a hierarchical encoder-decoder network. [37] utilize selective kernel feature fusion to realize the information exchange of different scales and information aggregation based on attention. [27] develops a simple yet effective boosted decoder to progressively restore the haze-free image by incorporating the strengthen-operate-subtract boosting strategy in the decoder. By eliminating or substituting the nonlinear activation function, [12] establishes a simple baseline that yields measurable outcomes while requiring fewer computing resources.

Dual Network Approaches. The Dual Networks architecture is designed with two parallel branches that separately estimate the structure and detail components of the target signals from the input. These components are then combined to reconstruct the final results according to the specific task formulation module. This architecture was first proposed by [52] and has since inspired a lot of subsequent work, including in the areas of image dehazing [47–49], image deraining [53], image denoising [50], and image super-resolution/deblurring [51]. The Dual Networks approach has proven to be effective in addressing various low-level vision problems, by enabling a better separation of the structure and detail information, leading to improved performance in terms of both accuracy

and computational efficiency. Furthermore, the flexibility of this architecture makes it adaptable to different types of data, making it a popular choice in the field of image restoration.

Despite the significant achievements made by these networks, it remains a challenge to effectively balance these competing goals of preserving spatial details and contextualized information while recovering images.

2.2 Multi-stage architecture

The multi-stage networks are shown to be more effective than their single-stage counterparts in high level vision problems [54–57]. In recent years, there have been some attempts [8, 10, 11, 58–61] to apply multi-stage networks to image restoration. They aims to break down the image restoration process into several manageable stages, enabling the use of lightweight subnetworks to progressively restore clear images. This approach facilitates the capture of both spatial details and contextualized information by individual subnetworks at each stage. To prevent the production of suboptimal results that may arise from using the same subnetwork at each stage, a supervisory attention mechanism was proposed along with the adoption of distinct subnetwork structures [8]. Additionally, [60] present a novel self-supervised event-guided deep hierarchical Multi-patch Network to handle blurry images and videos through fine-to-coarse hierarchical localized representations. Nevertheless, this approach elevates the complexity of the system as refining the previous stage's results is required in subsequent stages.

3 Method

Our primary goal is to create a single-stage network architecture that can efficiently handle the challenging task of image restoration by balancing the need for spatial details and context information, all while using fewer computational resources. The M3SNet is built upon a U-Net architecture, as shown in Fig. 3. As is apparent from the figure, in contrast to the traditional U-Net network, we have inverted the architecture and introduced two key components: **(a)** the feature fusion middleware (FFM) and **(b)** the multi-head attention middle block (MHAMB). The model's architecture takes on a mountain-like shape, and we liken the image restoration process to climbing a mountain.

Overall Pipeline. Given a degraded image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, M3SNet first applies a 3×3 convolutional layer to extract shallow feature maps $\mathbf{F_0} \in \mathbb{R}^{H\times W\times C}$ (H, W, C are the feature map height, width, and channel number, respectively). Next these shallow features $\mathbf{F_0}$ pass through $4-level$ encoder-decoder and one multi-head attention middle block, yielding deep features $\mathbf{F_{DF}} \in \mathbb{R}^{H\times W\times C}$. Each layer contains multiple feature fusion middleware between the encoder and decoder to capture multi-scale information and retain spatial details. Finally, we apply convolution to deep features $\mathbf{F_{DF}}$ and generate a residual image $\mathbf{R} \in R^{H\times W\times 3}$ to which degraded image is added to obtain the restored image:$\hat{\mathbf{I}} = \mathbf{R} + \mathbf{I}$. We optimize the proposed network using PSNR loss:

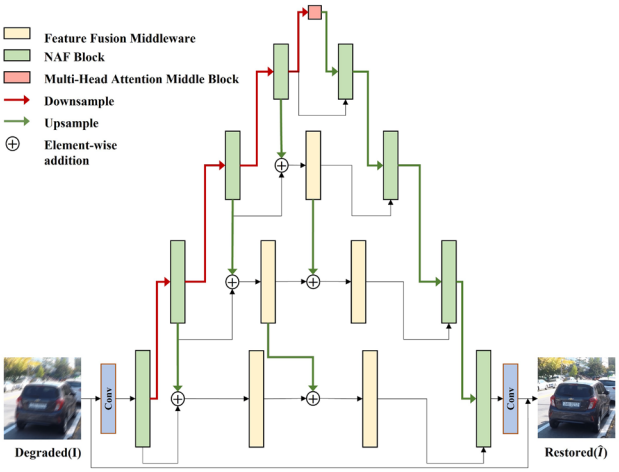


Fig. 3 Architecture of M3SNet for image restoration

$$PSNR = 10 \cdot log_{10} \cdot \frac{(2^n - 1)^2}{||\hat{\mathbf{I}} - \dot{\mathbf{I}}||^2} \tag{4}$$

where $\dot{\mathbf{I}}$ denotes the ground-truth image.

3.1 Feature fusion middleware (FFM)

By incorporating an encoder and decoder network in the initial stage, followed by a network operating at the original image input resolution in the final stage, the multi-stage network can produce high-quality images with accurate spatial details and reliable contextual information. However, the latter stage of this process requires revising the results of the previous stage, which adds a slight level of complexity to the system. While a single-stage network has relatively less complexity, it may struggle to balance spatial details and context information effectively. Therefore, we are exploring a middleware mechanism for feature fusion that enables a single-stage architecture to achieve the same functionality as a multi-stage architecture. As a basic feature fusion block, it can be plug-and-play in various other image restoration networks to improve the model representation.

As shown in Fig. 4a, the feature fusion middleware(FFM) is a nonlinear activation-free block (NAFBlock) with upsample and feature fusion. We have introduced the FFM between the encoder-decoder architectural levels to integrate upper-layer information into adjacent lower layers. This integration takes place sequentially, from the highest layer down to the lowest, until all information is fused into the original image

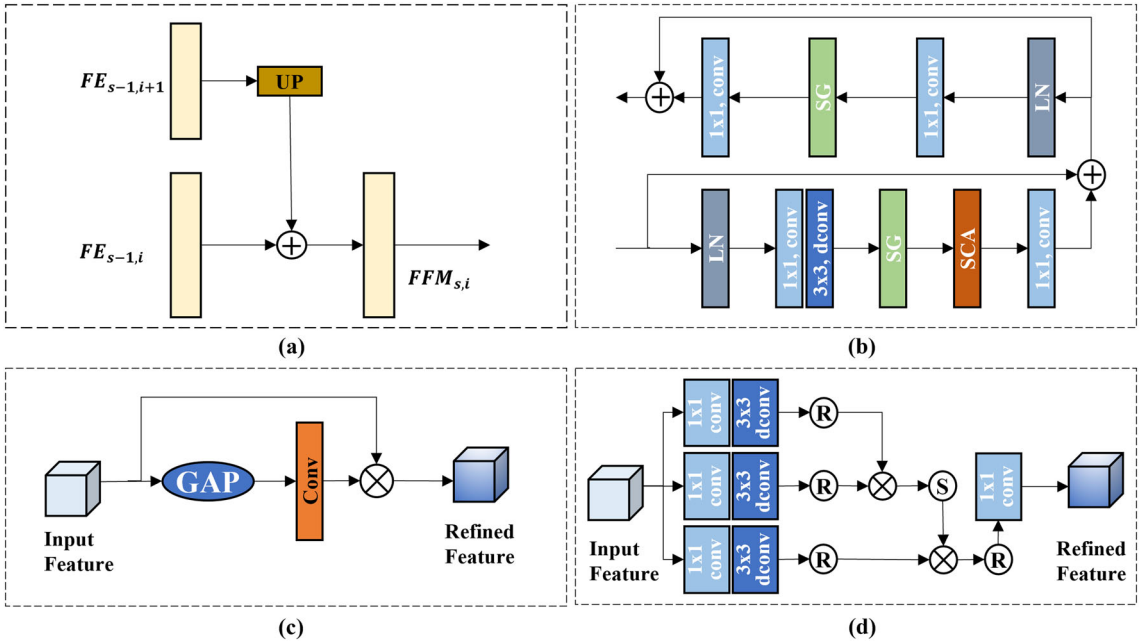


Fig. 4 (**a**) Feature fusion middleware (FFM) that enables the exchange of information across multiple scales while preserving the fine details. (**b**) The architecture of nonlinear activation free block (NAFBlock). (**c**) Simplified Channel Attention (SCA). (**d**) Multi-head attention middle block (MHAMB) that captures more global information

resolution manipulation level. This approach enhances the network's capacity to capture and fuse multi-scale features, ranging from simple patterns at low levels, such as corner or edge/color connections, to more complex higher-level features, such as significant variations and specific objects. As a result, this structure preserves spatial details while integrating contextual information, ultimately ensuring high-quality image restoration. Formally, let $\mathbf{FE_i} \in \mathbb{R}^{\frac{H}{i^2} \times \frac{W}{i^2} \times i^2 C}$ be the output in the i-th level encoder ($i = 1, 2, 3, 4$). At each level, the feature fusion information $FFM_{s,i}$ is given as:

$$\begin{aligned} FFM_{1,i} &= H_{Naf_{1,i}}(FE_i \oplus UP(FE_{i+1})) \\ FFM_{s,i} &= H_{Naf_{s,i}}(UP(FFM_{s-1,i+1}) \oplus FFM_{s-1,i}) \end{aligned} \tag{5}$$

where $\oplus$ denote the element-wise addition, $UP(\cdot)$ represents the up-sampling operation and $H_{Naf_{s,i}}(\cdot)$ is the s-th FFM in the i-th level.

This design offers two benefits. Firstly, the feature fusion middleware integrates multi-scale information. For example, the FFM of the third level fuses the encoder information of the third and fourth level, and the FFM of the second layer fuses the information of the second, third and fourth levels, so that the network model can capture abundant context information. Secondly, the feature fusion middleware in the 1_{th} layer operates on the original image resolution, without employing any subsampling operation, thereby enabling the network model to acquire detailed spatial information of high-resolution features.

NAFBlock. NAFBlock [12] is a variant of the U-Net network that simplifies the system by replacing or removing the nonlinear activation function. Figure 4b illustrates the process of obtaining an output$Y$ from an input X using Layer Normalization, Convolution, Simple Gate, and Simplified Channel Attention. Express as follows:

$$\begin{aligned} X_1 &= X + C_1(SCA(SG(C_3(C_1(LN(X)))))) \\ Y &= X_1 + C_1(SG(C_1(LN(X_1)))) \\ SG &= X_{f1} \times X_{f2} \end{aligned} \tag{6}$$

where C_1 is the 1×1 convolution, C_3 is the 3×3 depth-wise convolution, GAP is the adaptive average pool, $X_{f1}, X_{f2} \in R^{H \times W \times \frac{C}{2}}$ are obtained by dividing X_{f3} into channel dimensions, and $SCA(\cdot)$ is shown in Fig. 4c.

Finally, the depth features $\mathbf{F_{DF}}$ are obtained through this single-stage architecture, as demonstrated below:

$$\begin{aligned} FD_i &= H_{Naf_{\text{-}1,i}}(FD_{i+1} + FFM_{\text{-}1,i}) \\ F_{DF} &= FD_1 \end{aligned} \tag{7}$$

where FD_i is the output in the i-th level decoder, and -1 indicates that this is the last feature fusion middleware at this level.

Table 1 Dataset description

Tasks	Deraining							Deblurring		Denoising
Datasets	Rain14000 [62]	Rain1800 [63]	Rain12 [64]	Rain800 [65]	Rain1200 [66]	Rain100H [63]	Rain100L [63]	GoPro [67]	HIDE [68]	SIDD [69]
Train Samples	11200	1800	12	700	0	0	0	2103	0	320
Test Samples	0	0	0	100	1200	100	100	1111	2025	40
Testset Rename	–	–	–	Test100	Test1200	Rain100H	Rain100L	GoPro	HIDE	SIDD

3.2 Multi-head attention middle block (MHAMB)

The transformer model [19–21] has gained popularity in image restoration tasks due to its capability to capture global information, as evidenced by its increasing usage in recent years. The computation on a global scale results in a quadratic complexity in relation to the number of tokens as shown in Eq. 8, rendering it inadequate for the representation of high-resolution images.

$$\mathcal{O}_{MSA} = 4hwC^2 + 2(hw)^2C \tag{8}$$

To alleviate this issue, [19, 20, 43] etc., proposed various methods to reduce complexity. In this paper, we propose a multi-head attention middle block (MHAMB) as the bridge of the encoder-decoder, shown in Fig. 4d. MHAMB utilizes global self-attention to process and integrate the feature map information that is generated by the last layer of the encoder. This approach is particularly efficient in handling large images because convolution downsamples space, while attention can effectively process smaller resolutions for better performance. From the last layer of the encoder output FE_4, our MHAMB first encode channel-wise context by applying a 1×1 convolution and a 3×3 depth-wise convolution to generate the $query, key$ and $value$ matrices $\mathbf{Q} \in \mathbb{R}^{H\times W\times C}$, $\mathbf{K} \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{V} \in \mathbb{R}^{H\times W\times C}$ as follows:

$$\begin{aligned} Q &= C_3(C_1(FE_4)) \\ K &= C_3(C_1(FE_4)) \\ V &= C_3(C_1(FE_4)) \end{aligned} \tag{9}$$

Inspired by [19], we reshape $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ to $\hat{\mathbf{Q}} \in \mathbb{R}^{(HW)\times \frac{C}{h}\times h}$, $\hat{\mathbf{K}} \in \mathbb{R}^{(HW)\times \frac{C}{h}\times h}$, $\hat{\mathbf{V}} \in \mathbb{R}^{(HW)\times \frac{C}{h}\times h}$ to apply SA across channels rather than spatial dimensions to reduce the computation complexity, where h is the number of head. Next,we calculate similarities of pixel pairs between all the reshaped queries and keys as:

$$Attention(\hat{Q}, \hat{K}, \hat{V}) = SoftMax\left(\frac{\hat{Q}\hat{K}}{\beta}\right)\hat{V} \tag{10}$$

where β is a learning scaling parameter used to adjust the magnitude of the dot product of $\hat{Q}$ and $\hat{K}$ prior to the application of the softmax function. As we use the multi-head strategy, we finally concatenate all the outputs of multi-head attention and reshape the attention matrix back to its original dimensions of $\mathbb{R}^{H\times W\times C}$, and then get the final result by applying a 1×1 convolution. The resulting output is then added to FE_4 and passed to the decoder. This allows the model to incorporate self-attention in the lowest-resolution feature maps, generating richer feature representations that

improve the overall performance of the model, as shown in the experiments below.

4 Experiments

We evaluate the proposed M3SNet on benchmark datasets for three image restoration tasks, including **(a)** image deraining, **(b)** image deblurring, and **(c)** image denoising.

4.1 Datasets and evaluation protocol

We use PSNR and SSIM as quality assessment metrics. To report the reduction in error for each method relative to the best-performing method, we convert PSNR to RMSE ($RMSE \propto \sqrt{10^{-PSNR/10}}$) and SSIM to DSSIM (DSSIM = (1 - SSIM)/2). The datasets are summarized in Table 1.

Image deraining. Our derain model is trained on a collection of 13,712 clean-rain image pairs obtained from multiple datasets [62, 63, 65, 66]. We assess the model's performance on various test sets, including Test100 [65], Test1200 [66], Rain100H [63], and Rain100L [63].

Image deblurring. To perform image deblurring, we utilize the GoPro [67] dataset, which consists of 2,103 image pairs for training and 1,111 pairs for evaluation. Additionally, we assess the generalizability of our approach by applying the GoPro-trained model directly to the test images of the HIDE dataset. The HIDE dataset is designed specifically for human-aware motion deblurring, and its test set comprises 2,025 images.

Image denoising. For training our image denoising model, we utilize the SIDD dataset [69], which consists of 320 high-resolution images for training and 1,280 patches from 40 high-resolution images for evaluation. It's important to note that SIDD datasets comprise real-world images.

4.2 Implementation details

We train the proposed M3SNet without any pre-training and separate models for different image restoration tasks. We utilize the following block configurations in our network for each level: [1, 1, 1, 28] blocks for the encoder, [1, 1, 1, 1] blocks for the decoder, [2, 2, 1, 0] blocks for the FFM. And one MHAMB for the bridge of the encoder and decoder. We train models with Adam [74] optimizer($\beta_1 = 0.9, \beta_2 = 0.999$) and PSNR loss for 5×10^5 iterations with the initial learning rate 1×10^{-3} gradually reduced to 1×10^{-7} with the cosine annealing [75]. We extract patches of size 256×256 from training images, and the batch size is set to 32. We adopt TLC [76] to solve the issue of performance degradation caused by training on patched images and testing on the full image. For data augmentation, we perform horizontal and vertical flips.

4.3 Image deraining results

In our image deraining task, we compute the PSNR/SSIM scores using the Y channel in the YCbCr color space, which is consistent with previous works such as [8, 71, 73]. Our method has been demonstrated to outperform existing approaches significantly and consistently, as presented in Table 2. Specifically, our method achieves a remarkable improvement of 0.93 dB and a 10.2% error reduction on average across all datasets when compared to the best CNN-based method, SPAIR [73]. And achieves a improvement of 0.2 dB and a 3% error reduction compared to the best Transformer-based method, Restormer [19]. Moreover, we can achieve up to 2.76 dB improvement over HINet [9] on individual datasets, such as Rain100L. Compared to our baseline network NAFNet [12], we see significant performance gains across all datasets, with an average improvement of 1.11 dB. This proves that our proposed FFM and MHAMB can learn an enriched set of hybrid features, which combines local and non-local information which is vital for high-quality image deraining.

In addition to quantitative evaluations, Fig. 5 presents qualitative results that demonstrate the effectiveness of our M3SNet in removing rain streaks of various orientations and magnitudes while preserving the structural content of the images.

4.4 Image deblurring result

The performance evaluation of image deblurring approaches on the GoPro [67] and HIDE [68] datasets is presented in Table 3. Our M3SNet outperformed other methods, with a performance gain of 0.09dB when averaging across all datasets [67, 68]. Specifically, compared to our baseline network NAFNet [12], we improve 0.08 dB and 0.12 dB at 32 widths and 64 widths, respectively. Compared with previous CNN-based models [8], this progress can be much more obvious. Compared with previous best Transformer-based method, Restormer-local [19], we improve 0.17 dB on the GoPro dataset. It is worth noting that even though our network is trained solely on the GoPro Dataset, it still achieves state-of-the-art results (31.49 dB in PSNR) on the HIDE dataset. This demonstrates its impressive generalization capability.

Figure 6 displays some of the deblurred images produced by our method. Our model recovered clearer images that were closer to the ground truth than those by others.

4.5 Image denosing result

We compare the RGB image denoising results with other SOTA methods on SIDD [69], show in Table 4. Our method obtains considerable gains over the state-of-the-art

Table 2 Image deraining results

Methods	Test100 [70]		Test1200 [71]		Rain100H [72]		Rain100L [72]		Average			
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑		SSIM ↑	
PreNet [10]	24.81	0.851	31.36	0.911	26.77	0.858	32.44	0.950	29.42	(39.9%)	0.897	(28.2%)
MSPFN [71]	27.50	0.876	32.39	0.916	28.66	0.860	32.40	0.933	30.75	(29.9%)	0.903	(23.7%)
MPRNet [8]	30.27	0.907	32.91	0.916	30.51	0.890	37.20	0.965	32.73	(12.0%)	0.921	(6.3%)
SPAIR [73]	30.35	0.909	33.04	0.922	30.95	0.893	37.30	0.978	32.91	(10.2%)	**0.926**	(0.0%)
NAFNet [12]	30.25	0.908	32.92	0.917	30.40	0.891	37.40	0.964	32.73	(12.0%)	0.921	(6.3%)
HINet [9]	30.29	0.906	33.05	0.919	30.65	0.894	37.28	0.970	32.81	(11.2%)	0.922	(5.1%)
U2Former [22]	–	–	33.48	**0.926**	30.87	0.893	39.31	0.982	–	–	–	–
MDARNet [17]	28.98	0.892	33.08	0.919	29.71	0.884	35.68	0.961	31.86	(20.1%)	0.914	(15.7%)
SDLNet [16]	–	–	–	–	30.83	0.891	39.52	0.981	–	–	–	–
Restormer [19]	**31.32**	**0.910**	33.19	**0.926**	**31.06**	**0.895**	38.99	0.975	33.64	(3.0%)	**0.926**	(0%)
M3SNet-32	31.29	0.903	33.46	0.924	30.64	0.892	39.62	0.984	33.75	(1.0%)	**0.926**	(0.0%)
M3SNet-64	31.25	0.901	**33.52**	0.925	30.54	0.889	**40.04**	**0.985**	**33.84**	(0.0%)	0.925	(1.3%)

The best and second best scores are **highlighted** and underlined. Our M3SNet is better than the state-of-the-art by 0.2 dB

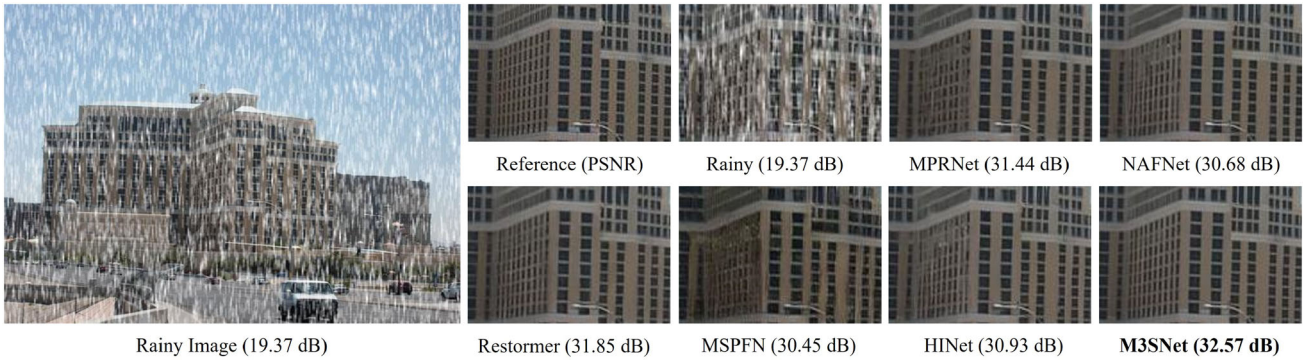


Fig. 5 **Image deraining** example. The outputs of M3SNet exhibit no traces of rain streaks on both image samples. M3SNet also recovers the most detailed images

Table 3 Image deblurring results

Methods	GoPro [67]		HIDE [68]		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
MT-RNN [77]	31.15	0.945	29.15	0.918	30.15	0.932
DMPHN [61]	31.20	0.940	29.09	0.924	30.15	0.932
Suin *etal*. [78]	31.85	0.948	29.98	0.930	30.92	0.939
SPAIR [73]	32.06	0.953	30.29	0.931	31.18	0.942
MIMO-UNet++ [44]	32.45	0.957	29.99	0.930	31.22	0.944
MPRNet [8]	32.66	0.959	30.96	0.939	31.81	0.949
MPRNet-local [8]	33.31	0.964	31.19	0.945	32.25	0.955
Restormer [19]	32.92	0.961	31.22	0.942	32.07	0.952
Restormer-local [19]	33.57	0.966	**31.49**	0.945	32.53	0.956
Uformer [21]	32.97	**0.967**	30.83	**0.952**	31.90	**0.960**
HINet [9]	32.71	–	–	–	–	–
HINet-local [9]	33.08	0.962	–	–	–	–
MSFS-Net [79]	32.73	0.959	31.05	0.941	31.99	0.950
MSFS-Net-local [79]	33.46	0.964	31.30	0.943	32.38	0.954
NAFNet-32 [12]	32.83	0.960	–	–	–	–
NAFNet-64 [12]	33.62	**0.967**	–	–	–	–
M3SNet-32 (ours)	32.91	0.965	30.92	0.948	31.92	0.957
M3SNet-64 (ours)	**33.74**	**0.967**	**31.49**	0.951	**32.62**	0.959

The best and second best scores are highlighted and underlined
The proposed M3SNet is trained only on the GoPro dataset but achieves a 0.09 dB improvement over the state of the art on the average of the effects on both datasets

Fig. 6 **Image deblurring** example on the GoPro dataset [67]. Compared to the state-of-the-art methods, our M3SNet restores sharper and perceptually-faithful images

Table 4 Image denoisng results

Methods	MPRNet [8]	Restormer [19]	NAFNet [12]	M3SNet
PSNR	39.71	40.02	40.30	**40.55**
SSIM	0.958	0.960	0.962	**0.962**

The best and second best scores are highlighted and underlined

Table 5 Ablation analysis for FFM and MHAMB on the benchmarks

Methods	Test100 [70]		Test1200 [71]		Rain100H [72]		Rain100L [72]		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
w/o FFM	30.37	0.899	33.02	0.918	30.45	0.892	37.47	0.968	32.83	0.920
w FFM	31.29	0.903	33.46	0.924	30.64	0.892	39.62	0.984	33.75	0.926
w/o MHAMB	31.17	0.901	33.41	0.924	30.59	0.891	39.48	0.980	33.66	0.924
w MHAMB	31.29	0.903	33.46	0.924	30.64	0.892	39.62	0.984	33.75	0.926

Table 6 Plug-and-play study of FFM. We pluged FFM into various image restoration networks to verify its improvement on the model effect

Model	PSNR	SSIM
Restormer [19]	32.92	0.961
Restormer + FFM	33.03	0.962
Uformer [21]	32.97	0.967
Uformer + FFM	33.04	0.967
NAFNet [12]	32.83	0.961
NAFNet + FFM	32.90	0.963

approaches, i.e., 0.25 dB over NAFNet [12], and 0.53 dB over the Transformer-based method Restormer [19].

4.6 Ablation studies

Effectiveness of FFM. To examine the effect of the FFM, we present the deraining results of w/o FFM in Table. 5. We can see that the derained images obtained from the method utilizing the FFM exhibit higher PSNR and SSIM values compared to the derained images produced by the method that does not employ the FFM.

To demonstrate the compatibility of our FFM method as a plug-and-play component for other image restoration models, we showcase the deblurring results achieved by integrating FFM into top-performing models (e.g. Resformer [19], Uformer [21], and NAFNet [12]) in Table. 6. We can clearly see the performance of both three models are improved by our approach. For a visually intuitive understanding the effect of such FFM, we further use high-pass filtering (HPF) to visualize learned features in Fig. 7. Compared to original image restoration model without adding FFM, the addition of FFM can better help to reconstruct finer detail features and improve the potential restoration quality. Since FFM seamlessly integrates information from the upper layers of the encoder-decoder structure into the adjacent lower layers and fuses all the information to the resolution manipulation level of the original image, this allows the model to better capture local and non-local information, thus facilitating more accurate representation for achieving high-quality output.

Effectiveness of MHAMB. To evaluate the effectiveness of MHAMB, we perform experiments based on w/o MHAMB in Table. 5. Compared to not applying MHAMB, MHAMB provides additional performance benefits thanks to the capability to capture global information.

In Fig. 8, we have provided visual comparisons of M3SNet w/ and wo/ the MHAMB. This study validates the recovered results of the model with the MHAMB tend to be clearer since it enables more global features to be fully used during the restoration process.

4.7 Resource efficient

Deep learning models have become increasingly complex in order to achieve higher accuracy. However, larger models require more resources and may not be practical in certain contexts. Therefore, there is a need to design lightweight image restoration models that can achieve high accuracy. In our work, we design a mountain-shaped single-stage network. This architecture optimizes the balance between spatial details and contextual information while minimizing the computational resources required to restore images.

Our M3SNet has been shown to outperform other models, as demonstrated in Table 7. Despite having 0.6M higher parameters than MIMO-unet++ [44], our proposed M3SNet-32 still achieves better performance, while using significantly

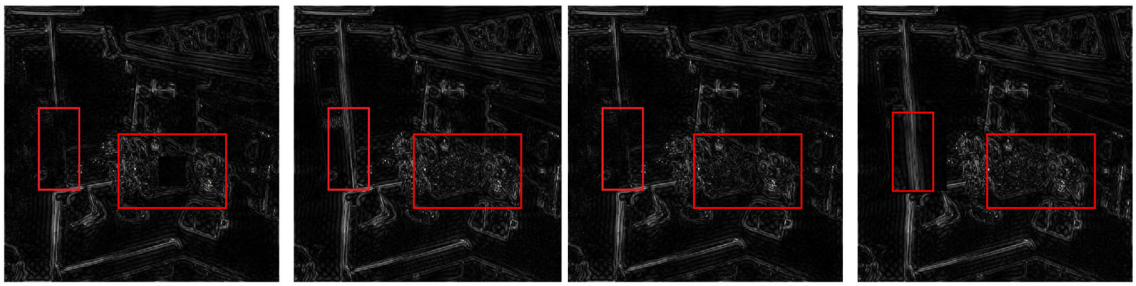


Fig. 7 Visualization of feature maps. Our proposed FFM can be pluged into existing image restoration models to generate more precise high-frequency details

Fig. 8 Effect of the MHAMB on image deraining. **a** Input corrupt image, Compared with (**b**), the image restored by M3SNet w/ MHAMB (**c**) can remove much more rain and recover the numbers accurately, **d** target highquality image

Table 7 The evaluation of model computational complexity. This is conducted with an input size of 256 × 256, on an NVIDIA 1060 GPU

Method		PSNR	Params(M)	MACs(G)
Multi-Stage	MIMO-UNet++ [44]	32.68	16.1	1235
	MPRNet [8]	32.66	20.1	778
	E-StackMPN [60]	33.56	118.2	472
Single-Stage	HINet [9]	32.77	88.7	171
	Restormer [19]	32.92	26.13	140
	Uformer [21]	32.97	50.88	89.5
	M3SNet-32 (ours)	32.91	16.7	37
	M3SNet-64 (ours)	33.74	66.3	146

fewer computational resources, with MACs approximately 40 times smaller than that of MIMO-unet++. Considering all factors, including model parameters, MACs, and performance, our model is the optimal choice.

5 Conclusion

In this paper, we present a mountain-shaped single-stage network that effectively captures multi-scale feature information and minimizes the computational resources required for image restoration. Our design is guided by the principle of balancing the competing goals of contextual information and spatial details while recovering images. To this end, we propose a feature fusion middleware mechanism that enables seamless information exchange between the encoder-decoder architecture's different levels. This approach smoothly combines upper-layer information with adjacent lower-layer information and eventually integrates all information to the original image resolution manipulation level. As a basic feature fusion block, it can be plug-and-play in various other image restoration networks to improve the model representation. To overcome the limitations of CNNs' receptive fields and capture more global information, we utilize a multi-head attention middle block as the bridge of our encoder-decoder architecture. Furthermore, to maintain computational efficiency and lightweight model size, we replace or remove nonlinear activation functions and instead use multiplication. Our extensive experiments on multiple benchmark datasets demonstrate that our M3SNet model significantly outperforms existing methods while utilizing low computational resources.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC1523303; in part by the National Natural Science Foundation of China under Grant 61672102.

Author Contributions Hu Gao: Writing, Methodology, Reviewing. Jing Yang: Reviewing, Supervision. Ying Zhang: Reviewing, Supervision. Ning Wang: Reviewing. Jingfan Yang: Reviewing. Depeng Dang: Conceptualization, Supervision.

Data availibility The data used in this paper are all from public datasets.

Declarations

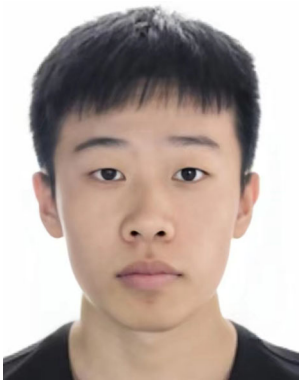
Conflict of interest The authors declare that there is no confict of interest.

Ethical approval and informed consent As this study involved a secondary analysis of publicly available data, ethical approval was not sought. All participants provided written informed consent. They were informed about the study's purpose, potential risks and benefts, confdentiality measures, and their right to withdraw at any time without any consequences.

References

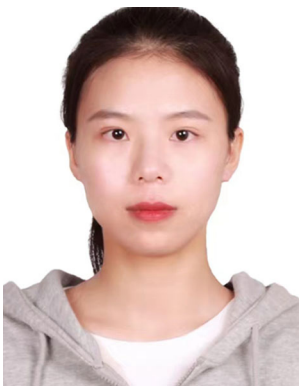
1. Rudin,L. I., Osher,S., Fatemi,E.: Nonlinear total variation based noise removal algorithms, Physica D Nonlinear Phenomena, 1992
2. Song, C.Z., Mumford, D.: Prior learning and gibbs reaction-diffusion. TPAMI **19**(11), 1236–1250 (1997)
3. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. TPAMI **12**(7), 629–639 (2002)
4. Roth,S., Black,M. J.: Fields of experts: A framework for learning image priors, in CVPR, 2005
5. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. TPAMI **32**(6), 1127 (2010)
6. Dong, W., Zhang, L., Shi, G., Wu, X.: Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. TIP **20**(7), 1838–1857 (2011)
7. He,K., Sun,J., Tang,X.: Single image haze removal using dark channel prior. TPAMI, 2011
8. Zamir,S. W., Arora,A., Khan,S., Hayat,M., Khan,F. S., Yang,M.-H., Shao,L.: Multi-stage progressive image restoration, in CVPR, 2021
9. Chen,L., Lu,X., Zhang,J., Chu,X., Chen,C.: Hinet: Half instance normalization network for image restoration, in CVPR, 2021
10. Ren,D., Zuo,W., Hu,Q., Zhu,P. F., Meng,D.: Progressive image deraining networks: A better and simpler baseline, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3932–3941, 2019
11. Li,X., Wu,J., Lin,Z., Liu,H., Zha,H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining, in European Conference on Computer Vision, 2018
12. Chen,L., Chu,X., Zhang,X., Sun,J.: Simple baselines for image restoration, arXiv preprint arXiv:2204.04676, 2022
13. Chu,X., Chen,L., Yu,W.: Nafssr: Stereo image super-resolution using nafnet, in CVPR, 2022
14. Pan,J., Sun,D., Zhang,J., Tang,J., Yang,J., Tai,Y. W., Yang,M. H.: Dual convolutional neural networks for low-level vision, IJCV, 2022
15. Lahiri, A., Bairagya, S., Bera, S., Haldar, S., Biswas, P.K.: Lightweight modules for efficient deep learning based image restoration. IEEE Trans. Circuits Syst. Video Technol. **31**(4), 1395–1410 (2020)
16. Wu, W., Liu, Y., Li, Z.: Subband differentiated learning network for rain streak removal. IEEE Trans. Circuits Syst. Video Technol. **33**(9), 4675–4688 (2023)
17. Hao, Z., Gai, S., Li, P.: Multi-scale self-calibrated dual-attention lightweight residual dense deraining network based on monogenic wavelets. IEEE Trans. Circuits Syst. Video Technol. **33**(6), 2642–2655 (2023)
18. Zhang,J., Zhang,Y., Gu,J., Zhang,Y., Kong,L., Yuan,X.: Accurate image restoration with attention retractable transformer, in ICLR, 2023
19. Zamir,S. W., Arora,A., Khan,S., Hayat,M., Khan,F. S., Yang,M.-H.: Restormer: Efficient transformer for high-resolution image restoration, in CVPR, 2022
20. Tsai,F.-J., Peng,Y.-T., Lin,Y.-Y., Tsai,C.-C., Lin,C.-W. : Stripformer: Strip transformer for fast image deblurring, in ECCV, 2022
21. Wang,Z., Cun,X., Bao,J., Zhou,W., Liu,J., Li,H.: Uformer: A general u-shaped transformer for image restoration, in CVPR, 2022
22. Feng,X., Ji,H., Pei,W., Li,J., Lu,G., Zhang,D.: U2-former: Nested u-shaped transformer for image restoration via multi-view contrastive learning, IEEE Trans. Circ. Syst. Video Technol. 1–1, 2023
23. Wang,X., Yu,K., Wu,S., Gu,J., Liu,Y., Dong,C., Loy, C. C., Qiao,Y., Tang,X.: Esrgan: Enhanced super-resolution generative adversarial networks, in ECCV Workshops, 2018
24. Ronneberger,O., Fischer,P., Brox,T.: U-net: Convolutional networks for biomedical image segmentation, in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015
25. Deng, H., Qian, G., Luo, D., Lv, X., Liu, H., Li, H.: Mrs-net: an image inpainting algorithm with multi-scale residual attention fusion. Appl. Intell. **53**, 07 (2022)
26. Kim,K., Lee,S., Cho,S.: Mssnet: Multi-scale-stage network for single image deblurring, arXiv, arXiv:2202.09652
27. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.-H.: Multi-scale boosted dehazing network with dense feature fusion, in. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) **2020**, 2154–2164 (2020)
28. Kupyn,O., Budzan,V., Mykhailych,M., Mishkin,D., Matas,J.: Deblurgan: Blind motion deblurring using conditional adversarial networks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8183–8192, 2017
29. Zhang,K., Luo,W., Zhong,Y., Ma,L., Stenger,B., Liu,W., Li,H.: Deblurring by realistic blurring, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2734–2743, 2020
30. Kupyn,O., Martyniuk,T., Wu,J., Wang,Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 8877–8886, 2019
31. Xue, T., Ma, P.: Tc-net: transformer combined with cnn for image denoising. Appl. Intell. **53**(6), 6753–6762 (2023)
32. Anwar,S., Barnes,N.: Densely residual laplacian super-resolution, TPAMI, 2020
33. Asif, M., Chen, L., Song, H., Yang, J., Frangi, A.F.: An automatic framework for endoscopic image restoration and enhancement. Appl. Intell. **51**, 1959–1971 (2021)
34. Zhang,Y., Li,K., Li,K., Wang,L., Zhong,B., Fu,Y.: Image super-resolution using very deep residual channel attention networks, in ECCV, 2018
35. Zhang,Y., Tian,Y., Kong,Y., Zhong,B., Fu,Y.: Residual dense network for image restoration, TPAMI, 2020
36. Dudhane,A., Zamir,S. W., Khan,S., Khan,F. S., Yang,M.-H.: Burst image restoration and enhancement, in CVPR, 2022
37. Zamir,S. W., Arora,A., Khan,S., Hayat,M., Khan,F. S., Yang,M.-H., Shao,L.: Learning enriched features for fast image restoration and enhancement, TPAMI, 2022
38. Zhang,D., Xu,Y., Ma,L., Li,X., Zhang,X., Peng,Y., Chen,Y.: Srenet: Structure recovery ensemble network for single image

deraining, Applied Intelligence, 2024. Available: https://api.semanticscholar.org/CorpusID:268743378
39. Su, C., Wu, X., Guo, Y.: Restoration of turbulence-degraded images using the modified convolutional neural network. Appl. Intell. **53**, 07 (2022)
40. Sheng, B., Li, P., Fang, X., Tan, P., Wu, E.: Depth-aware motion deblurring using loopy belief propagation. IEEE Trans. Circuits Syst. Video Technol. **30**(4), 955–969 (2020)
41. Lin, X., Sun, S., Huang, W., Sheng, B., Li, P., Feng, D.D.: Eapt: Efficient attention pyramid transformer for image processing. IEEE Trans. Multimedia **25**, 50–61 (2023)
42. Conde,M. V., Choi,U.-J., Burchi,M., Timofte,R.: Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration, in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2022
43. Liang,J., Cao,J., Sun,G., Zhang,K., Van Gool,L., Timofte,R.: Swinir: Image restoration using swin transformer, arXiv preprint arXiv:2108.10257, 2021
44. Cho,S. J., Ji,S. W., Hong,J. P., Jung,S. W., Ko,S. J.: Rethinking coarse-to-fine approach in single image deblurring, in ICCV, 2021
45. Yue,Z., Zhao,Q., Zhang,L., Meng,D.: Dual adversarial network: Toward real-world noise removal and noise generation, in ECCV, August 2020
46. Zhang, K., Li, Y., Zuo, W., Zhang, L., Gool, L.V., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. TPAMI **44**, 6360–6376 (2020)
47. Zhu,H., Xi,P., Chandrasekhar,V., Li,L., Lim,J. H.: Dehazegan: When image dehazing meets differential programming, in Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18, 2018
48. Guo,T., Li,X., Cherukuri,V., Monga,V.: Dense scene information estimation network for dehazing, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019
49. Yang,A., Wang,H., Ji,Z., Pang,Y., Shao,L.: Dual-path in dual-path network for single image dehazing, in Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19, 2019
50. Tian, C., Xu, Y., Zuo, W., Du, B., Lin, C.-W., Zhang, D.: Designing and training of a dual cnn for image denoising. Knowl.-Based Syst. **226**, 106949 (2021)
51. Singh,V., Ramnath,K., Mittal,A.: Refining high-frequencies for sharper super-resolution and deblurring, Comput. Vis. Image Underst. 2020
52. Pan,J., Liu,S., Sun,D., Zhang,J., Liu,Y., Ren,J., Li, Z., Tang,J., Lu,H., Tai, Y. W. a.:Learning dual convolutional neural networks for low-level vision, in CVPR, 2018
53. Siyuan,L. I., Ren,W., Zhang,J., Yu,J., Guo,X.: Fast single image rain removal via a deep decomposition-composition network, 2018
54. Li,W., Wang,Z., Yin,B., Peng,Q., Du,Y., Xiao,T., Yu,G., Lu,H., Wei,Y., Sun,J.: Rethinking on multi-stage networks for human pose estimation, arXiv, 2019 arXiv:1901.00148
55. Cheng,B., Chen,L.-C., Wei,Y., Zhu,Y., Huang,Z., Xiong,J., Huang,T., Hwu,W. mei W., Shi,H.: Spgnet: Semantic prediction guidance for scene parsing, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 5217–5227, 2019
56. Ghosh,P., Yao,Y., Davis,L. S., Divakaran,A.: Stacked spatio-temporal graph convolutional networks for action segmentation, 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 565–574, 2018
57. Li,S.-J., AbuFarha,Y., Liu,Y., Cheng,M.-M., Gall,J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1, 2020
58. Tao,X., Gao,H., Wang,Y., Shen,X., Wang,J., Jia,J.: Scale-recurrent network for deep image deblurring, CVPR, 2018
59. Fu,X., Liang,B., Huang,Y., Ding,X., Paisley,J.: Lightweight pyramid networks for image deraining, IEEE Trans. Neural Netw. Learn. Syst. 2018
60. Zhang,H., Zhang,L., Dai,Y., Li,H., Koniusz,P.: Event-guided multi-patch network with self-supervision for non-uniform motion deblurring, Int. J. Comput. Vis. 1–18, 2022
61. Zhang,H., Dai,Y., Li,H., Koniusz,P.: Deep stacked hierarchical multi-patch network for image deblurring, in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019
62. Fu,X., Huang,J., Zeng,D., Huang,Y., Ding,X., Paisley,J.: Removing rain from single images via a deep detail network, in CVPR, 2017
63. Yang,W., Tan,R. T., Feng,J., Liu,J., Guo,Z., Yan,S.: Deep joint rain detection and removal from a single image, CVPR, 2017
64. Li,Y., Tan,R. T., Guo,X., Lu,J., Brown,M. S.: Rain streak removal using layer priors, in CVPR, 2016
65. Zhang, H., Sindagi, V.A., Patel, V.M.: Image de-raining using a conditional generative adversarial network. IEEE Trans. Circuits Syst. Video Technol. **30**, 3943–3956 (2017)
66. Zhang,H., Patel,V. M.: Density-aware single image de-raining using a multi-stream dense network, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 695–704, 2018
67. Nah,S., Kim,T. H., Lee,K. M.: Deep multi-scale convolutional neural network for dynamic scene deblurring, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 257–265, 2016
68. Shen,Z., Wang,W., Lu,X., Shen,J., Ling,H., Xu,T., Shao,L.: Human-aware motion deblurring, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 5571–5580, 2019
69. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras, in. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. **2018**, 1692–1700 (2018)
70. Zhang, H., Sindagi, V.A., Patel, V.M.: Image de-raining using a conditional generative adversarial network. IEEE Trans. Circuits Syst. Video Technol. **30**, 3943–3956 (2017)
71. Jiang,K., Wang,Z., Yi,P., Chen,C., Huang,B., Luo,Y., Ma,J., Jiang,J.: Multi-scale progressive fusion network for single image deraining, CVPR, 2020
72. Yang,W., Tan,R. T., Feng,J., Liu,J., Guo,Z., Yan,S.: Deep joint rain detection and removal from a single image, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1685–1694, 2016
73. Purohit,K., Suin,M., Rajagopalan,A. N., Boddeti,V. N.: Spatially-adaptive image restoration using distortion-guided networks, CoRR, vol. abs/2108.08617, 2021
74. Kingma,D., Ba,J.: Adam: A method for stochastic optimization, Comput. Sci., 2014
75. Loshchilov,I., Hutter,F.: Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, 2016
76. Chu,X., Chen,L., Chen,C., Lu,X.: Improving image restoration by revisiting global information aggregation, in ECCV, 2021
77. Park,D., Kang,D. U., Kim,J., Chun,S. Y.: Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training, in ECCV, 2019
78. Suin, M., Purohit, K., Rajagopalan,A. N.: Spatially-attentive patch-hierarchical network for adaptive motion deblurring, CVPR . 3603–3612. (2020)
79. Zhang, Y., Li, Q., Qi, M., Liu, D., Kong, J., Wang, J.: Multi-scale frequency separation network for image deblurring. IEEE Trans. Circuits Syst. Video Technol. **33**(10), 5525–5537 (2023)



Hu Gao is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. His research interests include image restoration and image enhancement.

Jing Yang received the master's degree from Capital Normal University School of Information Engineering, China, in 2019. She is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. Her main research interests include natural language processing and deep learning.



Ying Zhang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. Her research interests include relation extraction and multi modal fusion.

Ning Wang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. His research interests include machine translation and multi modal fusion.

Jingfan Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. His research interests include image restoration and image enhancement.

Depeng Dang receive the Ph.D. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2003. From July 2003 to June 2005, he did his postdoctoral research with the Department of Computer Science and Technology, Tsinghua University, China. Currently, he is a full professor and supervisor of Ph.D. students in computer science and technology from Beijing Normal University, China. His research interests include natural language processing and deep learning.