Mix-order Attention Networks for Image Restoration

Tao Dai[1], Yalei Lv[2], Bin Chen[2], Zhi Wang[2], Zexuan Zhu[1], Shu-Tao Xia[2]
[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
daitao.edu@gmail.com,lyl20mails.tsinghua.edu.cn,cb17@tsinghua.org.cn,zhuzx@szu.edu.cn,xiast@sz.tsinghua.edu.cn

ABSTRACT

Convolutional neural networks (CNNs) have obtained great success in image restoration tasks, like single image denoising, demosaicing, and super-resolution. However, most existing CNN-based methods neglect the diversity of image contents and degradations in the corrupted images and treat channel-wise features equally, thus hindering the representation ability of CNNs. To address this issue, we propose a deep mix-order attention networks (MAN) to extract features that capture rich feature statistics within networks. Our MAN is mainly built on simple residual blocks and our mix-order channel attention (MOCA) module, which further consists of feature gating and feature pooling blocks to capture different types of semantic information. With our MOCA, our MAN can be flexible to handle various types of image contents and degradations. Besides, our MAN can be generalized to different image restoration tasks, like image denoising, super-resolution, and demosaicing. Extensive experiments demonstrate that our method obtains favorably against state-of-the-art methods in terms of quantitative and qualitative metrics.

CCS CONCEPTS

• **Computing methodologies → Reconstruction**.

KEYWORDS

image restoration, convolutional neural networks, attention

ACM Reference Format:
Tao Dai[1], Yalei Lv[2], Bin Chen[2], Zhi Wang[2], Zexuan Zhu[1], Shu-Tao Xia[2]. 2021. Mix-order Attention Networks for Image Restoration. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475205

1 INTRODUCTION

Image restoration is a fundamental problem in image processing community, which aims to recover high-quality images from their corrupted low-quality images. However, image restoration is a highly ill-posed problem since image degradation processes are irreversible. Based on the type of degradations, image restoration can be further divided into different subtasks, including image denoising [36, 39, 42], image super-resolution (SR) [4, 32, 43], compression artifacts reduction (CAR) [2, 5, 39], and other applications [1, 8]. To date, due to the powerful feature representational ability of convolutional neural networks (CNNs), a plenty of CNN-based methods have been developed to reconstruct missing information from the corrupted low-quality images.

The early CNN-based image restoration methods like SRCNN [6] and ARCNN [5] adopt shallow-layer networks and obtain remarkable performance against previous works. To ease the training difficulty of deeper networks, Zhang et al. proposed DnCNN [39] with residual learning for image denoising and compression artifacts reduction. Later, IRCNN [41] was developed by introducing the denoiser prior for fast image restoration. To further improve the performance, more sophisticated models [25, 30, 45] were proposed. Among them, Mao et al. designed a very deep encoder-decoder network with skip connections, while Tai et al. designed a very deep persistent memory network. Recently, attention-based deep methods [4, 36, 43, 45] have achieved impressive performance by exploring the feature correlations of intermediate layers. Typical methods [4, 24, 43] introduce channel attention (CA) mechanism to adaptively rescale channel-wise features by explicitly modeling interdependencies between channels, and thus allow the network to focus on more useful channels. By contrast, Zhang et al. [45] exploit spatial-wise feature correlations by considering long-range dependencies in the whole feature map. These CNN-based methods have achieved remarkable performance for image restoration tasks.

However, there exist several issues in the existing CNN-based methods. **First**, the diversity of image contents and degradations have not been fully considered. Most of them extract features from the degraded images with the same convolutional filters, which results in inflexibility to handling a wide variety of image contents and degradations. **Second**, rich feature statistics of intermediate layers have not fully captured. For example, RCAN [43] only utilizes the first-order feature statistics by global average pooling, while SAN [4] only exploits the second-order feature statistics by covariance pooling as a channel descriptor. However, using only first-order or second-order feature statistics is limited in representing the global distribution of channel-wise feature responses, thus hindering the representational ability of CNNs. Meanwhile, recent works [19, 38] have also shown that higher-order statistics are also helpful to improve discriminative ability of CNNs.

To address the above issues, we propose a deep mix-order attention networks (MAN) by exploring rich feature statistics of intermediate layers to enhance the feature correlation learning ability of CNNs. As shown in Fig. 1, the proposed MAN is built on simple residual blocks (RB) and the proposed mix-order channel

The first two authors contribute equally to this work.
Corresponding author: Bin Chen and Zexuan Zhu.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475205

attention (MOCA) module for exhaustively capturing hierarchical features. Unlike previous works [4, 43] that only use only first- or second-order features statistics, we exploit a mixture of k-order (e.g., k=1,2,3,4) feature statistics as the channel descriptor, thus fully capturing the global distribution of channel-wise responses. Moreover, to adapt our method to varying inputs and degradations, we design a feature gating block to adaptively select specific k-order channel attention for better feature correlation learning. In this way, our MOCA allow the network to not only capture rich feature statistical information, but also emphasize more informative features, thus improving the representational ability of CNNs.

In summary, the main contributions are listed as follows:

- We propose a deep mix-order attention networks (MAN) for image restoration by exploring rich feature statistical information. Our MAN is built on residual blocks and mix-order channel attention (MOCA) module to extract features.
- We propose mix-order channel attention (MOCA) for better feature correlation learning. Our MOCA contains feature pooling block to capture rich feature statistics of intermediate layers, and feature gating block to adapt to varying inputs and degradations, thus being more flexible for image restoration. Based on MOCA, we can obtain better feature representation ability and thus achieve accurate image restoration.
- We demonstrate with extensive experiments that the proposed MAN is effective for different image restoration tasks. Our MAN obtains impressive performance against state-of-the-art methods for image denoising, super-resolution, and demosaicing in terms of both quantitative and qualitative metrics.

2 RELATED WORKS

2.1 CNN-based Image Restoration

Recently, CNN-based image restoration algorithms show superior performance over the traditional ones. In the early works, Vincent et al. [34] simply stacked auto-encoder for image denoising. Dong et al. [7] proposed shallow SRCNN for image super-resolution. Later, Zhang et al. [39] proposed deeper CNNs to obtain better denoising performance with residual learning. VDSR [15] increases the network depth to a very high level, and proves that the network depth is essential for image super-resolution task. EDSR [22] removes unnecessary batch normalization module in residual networks and increases the number of channels to further improve the model. Other recent works [4, 23, 45] attempt to improve the performance by exploiting feature correlation. Among them, Zhang et al. have recently proposed a powerful image restoration method, named RNAN [45], by exploiting spatial non-local attention. Other works like RDN [46] and MemNet [30] form deep networks based on dense blocks and concentrates on exploiting all hierarchical features from all convolutional layers. Although these works have achieved significant progress in image restoration, most CNN-based methods focus on designing deeper network architectures, while neglecting the diversity of inputs and degradations, thus hindering the discriminative ability of CNNs. By contrast, we exploit rich feature statistical information of intermediate layers to enhance the ability of feature expressions.

2.2 Attention Mechanism

Attention mechanism is a common phenomenon in the visual field, that is, the human visual system will adaptively process visual information and focus on salient areas while ignoring irrelevant information. Similarly, a plenty of of recent CNN-based models have introduced attention mechanism to improve the performance and achieved great success in different computer vision tasks. For example, Wang et al. [35] proposed the residual attention network for image classification. The trunk-and-mask attention mechanism is composed of multiple attention modules, and can be easily scaled up to hundreds of layers with great performance. Wang et al. [36] proposed a non-local neural network by exploring spatial non-local attention. Hu et al. [12] proposed the SENet to exploit channel-wise relationships, which can effectively improve the representational power of CNNs and achieve significant performance improvements on image classification tasks. Recently, several works, such as NLRN [23] and RNAN [45], RCAN [46] and SAN [4] have proposed to investigate the effects of spatial attention or channel attention for image restoration tasks. However, these attention-based methods explore only first- or second-order feature statistics, while ignoring rich higher-order features statistics, which are also helpful for enhancing the representational ability. Here, we propose a deep mix-order attention networks with distinguished power for image contents and degradations.

3 MIX-ORDER ATTENTION NETWORKS

3.1 The Overall Framework

The overall framework of our mix-order attention networks (MAN) is shown in Fig. 1, which mainly consists of residual block (RB) and mix-order channel attention (MOCA) module. Specifically, the first and last convolutional layers in our MAN serve as shallow feature extractor and reconstruction layer, respectively. The stacked residual block (RB) and mix-order channel attention (MOCA) aims to extract features that capture rich feature statistics. In our MOCA, feature pooling block aims to capture rich feature statistics, while feature gating block allows our MAN adapt to varying image contents and degradations.

The high-quality image and its corresponding low-quality image (*e.g.*, noisy) are denoted as X_H and X_L. Thus, the recovered images X_R by our MAN can be expressed as

$$X_R = H_{MAN}(X_L), \tag{1}$$

where $H_{MAN}(\cdot)$ is the function of our MAN.

The MAN is optimized with a loss function. As previous works show, different loss functions have been adopted, like L_1 loss [22], L_2 loss [30, 39, 41] and adversarial loss [18]. To verify the effectiveness of our MAN, we adopt the same loss function as previous works (*e.g.*, L_2 loss). Specifically, given a series of training images with N low-quality images and its corresponding high-quality counterparts, which is denoted by $\{X_L^i, X_H^i\}_{i=i}^N$, the goal of training our MAN is to optimized the L_2 loss function as

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||H_{MAN}(X_L^i) - X_H^i||_2, \tag{2}$$

where Θ is the parameter set of MAN. The loss function is optimized by stochastic gradient descent algorithm.

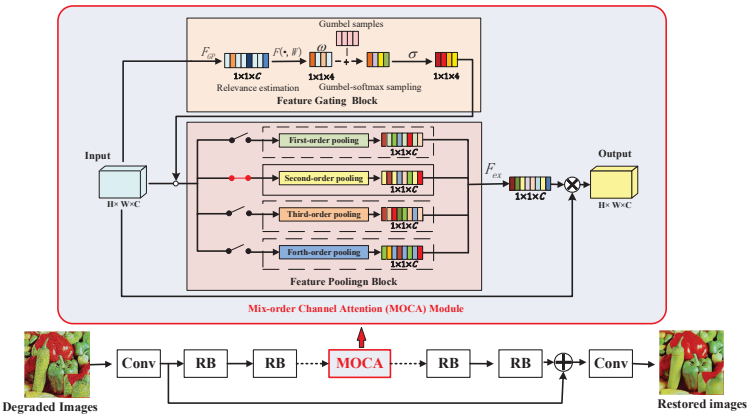


Figure 1: The framework of our mix-order attention networks for image restoration. 'CONV', 'RB', 'MOCA' is convolutional layer, residual block, and mix-order channel attention (MOCA) module, respectively. Our MOCA consists of feature gating block and feature pooling block to capture semantic information and rich feature statistics of intermediate layers.

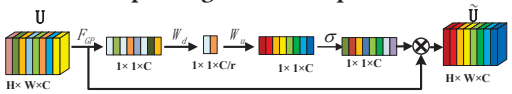


Figure 2: Channel attention which uses global average pooling to generate channel-wise statistics.

3.2 Mix-order Channel Attention Module

In this section, we introduce the novel mix-order channel attention module, which models feature interdependencies based on rich feature statistics.

3.2.1 Revisiting Channel Attention. By treating channel-wise features unequally, channel attention mechanism focus on more informative channels. Based on SENet [12], channel attention is divided into squeeze and excitation operation. In squeeze operation, channel-wise global spatial information is *squeezed* into a channel descriptor by global average pooling. As shown in Fig. 2, let $U = [u_1, u_2, ..., u_C]$ denote the input, *i.e.*, a feature map has C channels with spatial size of $H \times W$. The channel descriptor $z \in \mathbb{R}^C$ can be generated by shrinking U in spatial dimension. The c-th element of z is calculated as

$$z_c = F_{GP}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \tag{3}$$

where $F_{GP}(\cdot)$ is the global average pooling function, $u_c(i, j)$ is the pixel at position (i, j) of c-th channel. Such channel descriptor can be thought as a collection of the spatial information, and is used to represent the global distribution of the whole c-th channel.

In order to make full use of the information aggregated by global average pooling, the excitation operation is introduced to capture channel-wise interdependencies. A simple gating mechanism with a sigmoid activation is introduced, and we obtain the final channel statistics s as

$$s = F_{ex}(z, W) = \sigma(W_u \delta(W_d z)), \tag{4}$$

where W_u and W_d are weight sets of the two stacked convolutional layers in the channel attention; σ refers to sigmoid function and δ refers to ReLU [28] function. To decrease model complexity and aid generalisation, a bottleneck is used with two 1×1 convolutional layers around the non-linearity, *i.e.*, a convolutional layer with parameter W_d acts as channel-downscaling with reduction ratio r, while a a convolutional layer with parameter W_u acts as channel-upscaling layer to increase dimension with ratio r.

Finally, the c-th channel feature map u_c is rescaled by the final channel statistics s

$$\tilde{u}_c = s_c \cdot u_c, \tag{5}$$

where s_c refers to the scaling factor. With such channel attention, we can explicitly modelling channel interdependencies to recalibrate features.

3.2.2 Feature Pooling Block. Recent works [4, 46] have introduced channel attention for image super-resolution, and obtained remarkable performance. However, these works exploit only first- or second-order feature statistics while ignoring the rich higher-order feature statistics, which are also shown to be helpful to improve discriminative ability of CNNs [19, 38].

Base on such observations, here, we design feature pooling block to exploit a mixture of k-order (*e.g.*, k=1,2,3,4) features statistics, which is more suitable to represent the global distribution of each channel. Specifically, we take first-order statistics: Average, second-order statistics: Standard Deviation, third-order statistics: Skewness and forth-order statistics: Kurtosis into consideration. Given an input feature map $U \in \mathbb{R}^{H \times W \times C}$, by shrinking U in spatial dimension, the c-th element of k-order channel descriptor z_2 is calculated by

$$z_{1,c} = \mu_c = F_1(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \tag{6}$$

$$z_{2,c} = F_2(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (u_c(i, j) - \mu_c)^2, \tag{7}$$

$$z_{3,c} = F_3(u_c) = \mathbf{E}\left[\left(\frac{U - \mu}{\sigma}\right)^3\right], \tag{8}$$

$$z_{4,c} = F_4(u_c) = \mathbf{E}\left[\left(\frac{U - \mu}{\sigma}\right)^4\right], \tag{9}$$

where $F_k(\cdot)$ denotes m-order feature pooling function ($k = 1, 2, 3, 4$). $u_c(i, j)$ is the pixel at position (i, j) of c-th channel; where μ is the mean and σ is the standard deviation of U, $\mathbf{E}$ is the expectation operator and can be calculated by $\mathbf{E}(X) = \sum_{i=1}^{k} p_i x_i$, X is a random variable occurring with probabilities $p_1, p_2, \cdots, p_k$.

Natural images often contain various objects and different texture, resulting in very complex distributions of convolutional activations. As a result, a single k-order statistic cannot fully capture statistical information of features. A natural idea to overcome the above limitation is to ensemble multiple k-order statistics. Thus, we attempt to mix different k-order statistics with a feature gating mechanism to capture more complex statistical information. How to design the gate becomes a very important problem. First, in order to prevent the mode from collapsing into trivial solutions that are independent of input features, such as always choosing certain statistics, the randomness of the gate is very important. Second, the gate needs to make a discrete decision while still providing gradient to estimate relevance. Third, the computing cost of the gate should be relatively low. To this end, we design a feature gating block to make our MAN adaptive to varying image contents and degradations.

3.2.3 Feature Gating Block. As shown in Fig. 1, the feature gating block contains two parts: the first part estimates the relevance of these different k-order statistics, while the second part makes a discrete decision by sampling using Gumbel-Softmax.

Relavance Estimation. The aim of the gate's first part is to estimate the relevance of different higher-order statistics given the input features map $U \in \mathbb{R}^{H \times W \times C}$. Based on recent study [33], much of the information in features is caught by the statistics of channel interdependencies. Therefore, we just utilize channel-wise statistics aggregated by global average pooling, which is same as Equ. (6) and compresses the input feature map into a $1 \times 1 \times C$ channel descriptor. To fully capture the interdependencies between channels, a non-linear function of two fully-connected layers connected with a ReLU [28] activation function is added. The output ω of this part is the relevance score of different statistics, a vector containing unnormalized weights for choosing from statistics of different orders.

$$\omega = F(z, W) = W_2 \delta(W_1 z), \tag{10}$$

where z is the channel descriptor, δ is the ReLU function, $W_1 \in \mathbb{R}^{d \times C}$, $W_2 \in \mathbb{R}^{4 \times d}$ are parameters of two fully-connected layers and d is the dimension of the hidden layer.

Gumbel Sampling. The aim of the second part is to make a discrete decision based onvance scores.

We choose the Gum the relevance scores. A simple attempt is to choose the maximum of the relevance scores. However, this approach loses the uncertainty of probability and is not differentiable. Therefore, adding noise is a common way to introduce such randomness. Then we can choose from these four options proportional to the relebel distribution for the noise, because it has a great property named Gumbel-Max trick [9]. Let v be a K-dimensional discrete random variable with probabilities $[\alpha_1, \alpha_2, ..., \alpha_K]$. The Gumbel-Max trick offers a simple and efficient way to draw samples v from a discrete distribution with probabilities α

$$v = one\_hot\left(\arg\max_k \left[G_k + \log \alpha_k\right]\right), \tag{11}$$

where $G_1, G_2, ..., G_K$ are a sequence of i.i.d. Gumbel random variables, which can be sampled using inverse transform sampling by $Z \sim \text{Uniform}(0, 1)$ and computed by $G = -\log(-\log(Z))$.

A drawback of this approach is that the argmax operation is not continuous. Replacing the argmax operation with a softmax function, a continuous relaxation of the Gumbel-Max trick has been proposed [14]. Then, samples from the Gumbel-Softmax relaxation can be expressed as

$$v = one\_hot\left(\text{softmax}\left[(\mathrm{G_k} + \log \alpha_\mathrm{k})/\tau\right]\right), \tag{12}$$

where τ is the temperature of the softmax.

4 EXPERIMENTS

4.1 Setup

To verify the effectiveness of our MAN, we apply our method to different image datasets on different image restoration tasks, including image denoising, demosaicing, and super-resolution. Following the previous works [27, 45], we set the same settings for image denoising, demosaicing, compression artifacts reduction, and super-resolution for fair comparison. Specifically, we use 800 high-resolution traing images from DIV2K dataset [31] as training set. All results are evaluated by PSNR and SSIM [37] metrics on Y channel of transformed YCbCr space.

In our MAN, the kernel size and channel number of all convolutional layers is set to 3×3 and $C = 64$ except for those in higher-order channel attention module, which use 1×1 convolutional layer and set $C = 32$. Our MAN contains 32 residual blocks (RBs) in total, which contain one mix-order channel attention (MOCA) module in each RB. For our MOCA, we set order factor as $k = 4$ empirically.

When training, we augment the 800 training images by randomly rotating 90°, 180°, 270° and horizontally flipping. In each batch, we crop the input as patches with size 48×48. Our model is trained by ADAM optimizer [16] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. The learning rate is initialized with 10^{-4} and then decreased to half every 200 epochs. We implement our models on the Pytorch framework [29] on Nvidia 2080Ti GPUs.

4.2 Ablation Study

Table 1: PSNR results (×4) of MAN w/o FGB with varying k.

Method	Set5	Set14	BSD100	Urban100	Manga109
MAN w/o FGB ($k = 1$)	31.898	28.415	27.460	25.682	29.885
MAN w/o FGB ($k = 2$)	32.060	28.430	27.469	25.722	29.990
MAN w/o FGB ($k = 3$)	32.104	28.474	27.508	25.833	30.162
MAN w/o FGB ($k = 4$)	32.094	28.446	27.501	25.845	30.142

4.2.1 Effects of high-order feature statistics. To explore the roles of high-order feature statistics, we test our MAN without feature gating block (FGB) (denoted as MAN w/o FGB) on image super-resolution tasks (4×), and report the PSNR results in Table 1, from which we see that $k = 2, 3, 4$ obtains consistently better performance than $k = 1$. This indicates that high-order feature statistics are also helpful in restoring image structures. Besides, $k = 3$ obtains the best

results on Set5 and Set14, while $k = 4$ performs best on Urban100, which implies that the optimal k is related to image contents.

Table 2: PSNR results(×4) of MAN with MOCA with varying k.

Method	Set5	Set14	BSD100	Urban100	Manga109
MAN w/ MOCA ($k = 1$)	31.898	28.415	27.460	25.682	29.885
MAN w/ MOCA ($k = 2$)	32.130	28.480	27.513	25.863	30.086
MAN w/ MOCA ($k = 3$)	32.138	28.518	27.529	25.879	30.225
MAN w/ MOCA ($k = 4$)	32.144	28.519	27.534	25.928	30.267

Table 3: PSNR results (×4) of normalization in MOCA.

	Set5	Set14	BSD100	Urban100	Manga109
MOCA w/ normalization	32.097	28.465	27.491	25.807	30.024
MOCA w/o normalization	32.144	28.519	27.534	25.928	30.627

4.2.2 Effects of MOCA. Similarly, we test MAN with MOCA (denoted as MAN w/ MOCA) on image super-resolution tasks (4×), and report PSNR results on Table 2, from which we see that $k = 4$ obtains the best results on different datasets. Thus, we set $k = 4$ in MAN empirically. Moreover, we can find that $k = 2, 3, 4$ performs better than $k = 1$. This is mainly because natural images usually contain rich texture structures with complex statistical characteristics. Thus, our MOCA with a mixture of k-order feature statistics allows MAN to focus on more informative features, thus improving the feature expression ability.

4.2.3 Effects of normalization in MOCA. In our MOCA, different k-order statistics may cover a large range of amplitude, which may influence the final performance. To answer this question, we test our MOCA with/without normalization, and report the PSNR results in Table 3, from which we can see that MOCA with normalization obtains consistently worse performance than MOCA without normalization. The possible reason is that the amplitude of different k-order statistics also contain rich latent semantic information, while normalization can degrade such information. Therefore, our MOCA does not perform normalization operations in our MAN.

4.3 Image Denoising

For image denoising, we follow the settings as in RNAN [45], and evaluate our MAN on standard benchmarks, including Kodak24 (http://r0k.us/graphics/kodak/), BSD68 [26], and Urban100 [13]. Specifically, noisy images are produced by adding additive white Gaussian noise (AWGN) with standard deviation $\sigma = 10, 30, 50, 70$. We compare our MAN with state-of-the-art image denoising methods, including CBM3D [3], TNRD [2], RED [25], DnCNN [39], MemNet [30], IRCNN [40], FFDNet [41], RNAN [45].

All the results are reported in Table 4, from which we can see that our MAN obtains the best performance at different datasets and noise levels in most cases. Our mix-order channel attention allows networks to focus on informative parts, thus being effective in image denoising. Compared with RNAN, which is considered as one of the most powerful denoising methods, our MAN achieves consistently better results. For example, our MAN can achieve over 0.31 and 0.25 dB gains over RNAN at $\sigma = 50$ and $\sigma = 70$. These observations demonstrate the effectiveness of our proposed mix-order channel attention.

For visual quality, we evaluate different denoising methods on BSD68 and Urban100, and show visual results in Fig. 3, from which we can observe that our MAN with mix-order channel attention produces better visual quality with recovering more image details. Take images '223061' and 'img044' as an example, our MAN generates the most faithful restoration results (*e.g.*, lines) than others.

4.4 Image Demosaicing

Following the same settings in RNAN [45], we conduct experiments on McMaster [40], Kodak24, BSD68, and Urban100 for image demosaicing. We compare our MAN with recent state-of-the-art demosaicing methods: IRCNN [40] and RNAN [46], which work well in demosaicing. All the results are listed in Table 5, from which we see that mosaic corruptions significantly degrade the image quality. IRCNN and RNAN remove mosaic corruptions to some degree, and thus obtain relatively high-quality restoration. By contrast, our MAN obtains significantly better performance than IRCNN and RNAN on different datasets. Compared with RNAN, the PSNR gains of our MAN is at least 0.30 dB on average, and even up to nearly 1 dB on Urban100. These observations demonstrate the effectiveness of our mix-order channel attention in handling the mosaic corruptions.

The visual results are shown in Fig. 4, from which we observe that our MAN and RNAN can effectively reduce mosaic degradations and produce similar visual results, both of which have significantly better visual quality than IRCNN. Compared with RNAN, ours produces less artifacts, thus leading to better visual quality. With mix-order channel attention, our MAN eliminates most of artifacts and reconstructs more accurate color.

4.5 Image Super Resolution

We further apply our MAN on image super-resolution (SR), and compare MAN with other state-of-the-art SR methods: LapSRN [17], MemNet [30], SRMDNF [39], DBPN [10] , RDN [46], EDSR [21], NLRN [23] SRFBN [20], OISR [11], RCAN [44], and RNAN [45]. Besides, we introduce self-ensemble strategy to further improve the performance of our MAN (denoted as MAN+).

All the results of compared SR methods are reported in Table 6, from which we see that our MAN+ achieves the best performance on different benchmarks in most cases. Without self-ensemble strategy, our MAN , RNAN and RCAN are the three best SR methods, and outperforms other SR methods. Note that the parameter number of our MAN is 3.87 M, far smaller than 7.5 M in RNAN, and 16 M in RCAN. Specifically, the network depth of our MAN (about 66 convolutional layers), is far shallower than that of RNAN (about 120 convolutional layers) and RCAN (about 400 convolutional layers). It implies that our mix-order channel attention can make full use of informative features. These observations verify the effectiveness of our MAN with mix-order channel attention.

We also compare visual results of different SR methods. As shown in Fig. 5, we can observe that our MAN and RCAN produce visually pleasing results with finer image structures, and outperform other methods. For example, RNAN recovers lines in windows areas of 'img_025' with wrong direction, while our MAN produce more

Table 4: Quantitative evaluation of state-of-the-art approaches on color image denoising. Best results are highlighted

Method	Kodak24				BSD68				Urban100			
	10	30	50	70	10	30	50	70	10	30	50	70
CBM3D	36.57	30.89	28.63	27.27	35.91	29.73	27.38	26.00	36.00	30.36	27.94	26.31
TNRD	34.33	28.83	27.17	24.94	33.36	27.64	25.96	23.83	33.60	27.40	25.52	22.63
RED	34.91	29.71	27.62	26.36	33.89	28.46	26.35	25.09	34.59	29.02	26.40	24.74
DnCNN	36.98	31.39	29.16	27.64	36.31	30.40	28.01	26.56	36.21	30.28	28.16	26.17
MemNet	N/A	29.67	27.65	26.40	N/A	28.39	26.33	25.08	N/A	28.93	26.53	24.93
IRCNN	36.70	31.24	28.93	N/A	36.06	30.22	27.86	N/A	35.81	30.28	27.69	N/A
FFDNet	36.81	31.39	29.10	27.68	36.14	30.31	27.96	26.53	35.77	30.53	28.05	26.39
RNAN	37.24	31.86	29.58	**28.15**	36.43	30.63	28.27	26.83	36.59	31.50	29.08	27.45
MAN	**37.32**	**31.91**	**29.63**	28.14	**36.50**	**30.70**	**28.34**	**26.87**	**36.75**	**31.76**	**29.41**	**27.71**

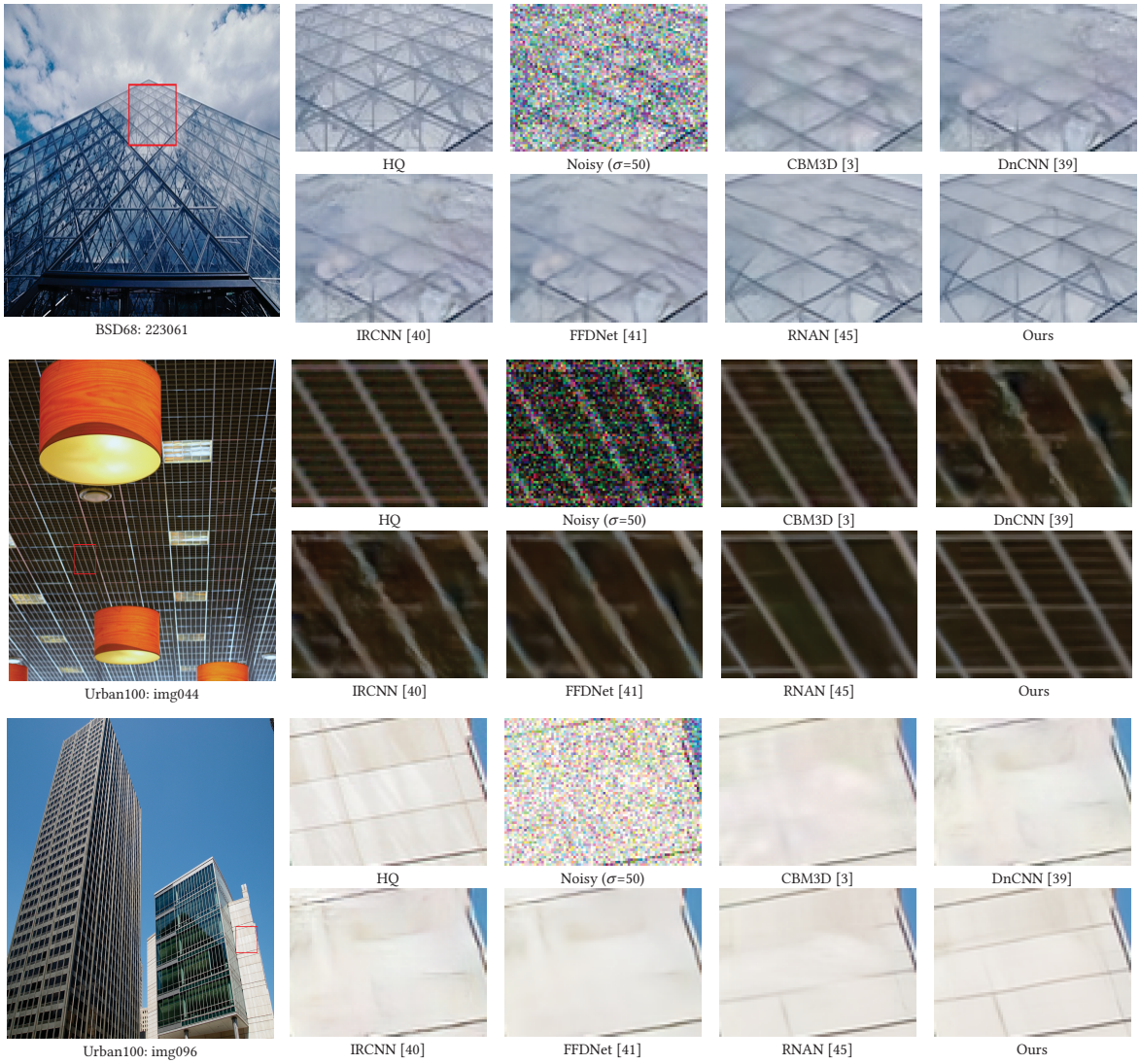


Figure 3: Visual comparison for color image denoising with noise level σ = 50

faithful results. These results further demonstrate the effectiveness of our MAN with mix-order channel attention.

4.6 Model Size Analyses

Take image denoising as an example, we compare the model size of our MAN with other advanced image denoising approaches in Table 7, from which we can observe that our MAN with 32

Table 5: Quantitative evaluation of state-of-the-art approaches on color image demosaicing. Best results are highlighted

Method	McMaster18		Kodak24		BSD68		Urban100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Mosaiced	9.17	0.1674	8.56	0.0682	8.43	0.0850	7.48	0.1195
IRCNN	37.47	0.9615	40.41	0.9807	39.96	0.9850	36.64	0.9743
RNAN	39.71	0.9725	43.09	0.9902	42.50	0.9929	39.75	0.9848
MAN	**40.05**	**0.9739**	**43.37**	**0.9905**	**42.90**	**0.9934**	**40.69**	**0.9851**

Table 6: Quantitative results on SR benchmark datasets. The best results and second results are highlighted and underline.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
LapSRN [17]	×2	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9101	37.27	0.9740
MemNet [30]	×2	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195	37.72	0.9740
SRMDNF [42]	×2	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	38.07	0.9761
DBPN [10]	×2	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
RDN [46]	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
EDSR [21]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
NLRN [23]	×2	38.00	0.9603	33.46	0.9159	32.19	0.8992	31.81	0.9249	–	–
SRFBN [20]	×2	38.11	0.9609	33.82	0.9196	32.29	0.9010	32.62	0.9328	39.08	0.9779
OISR [11]	×2	38.21	0.9612	33.94	0.9206	32.36	0.9019	33.03	0.9365	–	–
RCAN [44]	×2	38.27	0.9614	34.12	0.9216	32.41	**0.9027**	33.34	0.9384	39.44	0.9786
RNAN [46]	×2	38.17	0.9611	33.87	0.9207	32.32	0.9014	32.73	0.9340	39.23	0.9785
MAN	×2	38.26	0.9614	34.14	0.9225	32.37	0.9021	33.12	0.9369	39.23	0.9785
MAN+	×2	**38.31**	**0.9616**	**34.17**	**0.9230**	**32.42**	0.9026	**33.34**	**0.9385**	**39.51**	0.9789
LapSRN [17]	×3	33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280	32.21	0.9350
MemNet [30]	×3	34.09	0.9248	30.00	0.8350	28.96	0.8001	27.56	0.8376	32.51	0.9369
SRMDNF [42]	×3	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
RDN [46]	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
EDSR [21]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
NLRN [23]	×3	34.27	0.9266	30.16	0.8374	29.06	0.8026	27.93	0.8453	-	-
SRFBN [20]	×3	34.70	0.9292	30.51	0.8461	29.24	0.8084	28.73	0.8641	34.18	0.9481
OISR [11]	×3	34.72	0.9297	30.57	0.8470	29.29	0.8103	28.95	0.8680	-	-
RCAN [44]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
MAN	×3	34.79	0.9300	30.59	0.8472	29.28	0.8106	28.91	0.8671	34.22	0.9489
MAN+	×3	**34.86**	**0.9306**	**30.72**	**0.8491**	**29.36**	**0.8118**	**29.18**	**0.8711**	**34.58**	**0.9506**
LapSRN [17]	×4	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet [30]	×4	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
SRMDNF [42]	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
DBPN [10]	×4	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
RDN [46]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
EDSR [21]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
NLRN [23]	×4	31.92	0.8916	28.36	0.7745	27.48	0.7306	25.79	0.7729	-	-
SRFBN [20]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
OISR [11]	×4	32.53	0.8992	28.86	0.7878	27.75	0.7428	26.79	0.8068	-	-
RCAN [44]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
RNAN [46]	×4	32.49	0.8982	28.83	0.7878	27.72	0.7421	26.61	0.8023	31.09	0.9149
MAN	×4	32.58	0.8992	28.79	0.7871	27.73	0.7424	26.70	0.8046	31.01	0.9154
MAN+	×4	**32.69**	0.9007	**28.93**	**0.7895**	**27.81**	**0.7440**	**26.95**	**0.8097**	**31.42**	**0.9189**

Table 7: Model size comparison

Methods	RED	DnCNN	MemNet	RNAN	MAN
Parameters	4131K	672K	677K	7409K	**3877K**
PSNR (dB)	26.40	28.16	26.53	29.15	**29.40**

residual blocks and mix-order channel attention modules obtains the best denoising performance with a much lighter architecture. Compared with the state-of-the-art RNAN, our MAN has achieved significantly better results with much smaller model size. Such observations demonstrate the great superiority of our MAN with mix-order channel attention module.

5 CONCLUSIONS

In this paper, we propose deep mix-order attention networks (MAN) for accurate image restoration. The networks is built on stacking residual blocks and mix-order channel attention modules (MOCA), which extract attention-aware features that capture rich feature

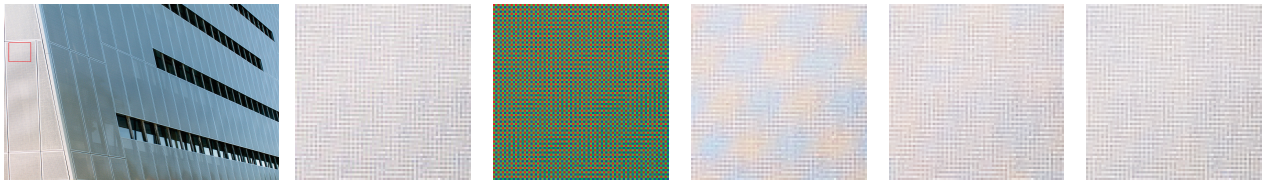
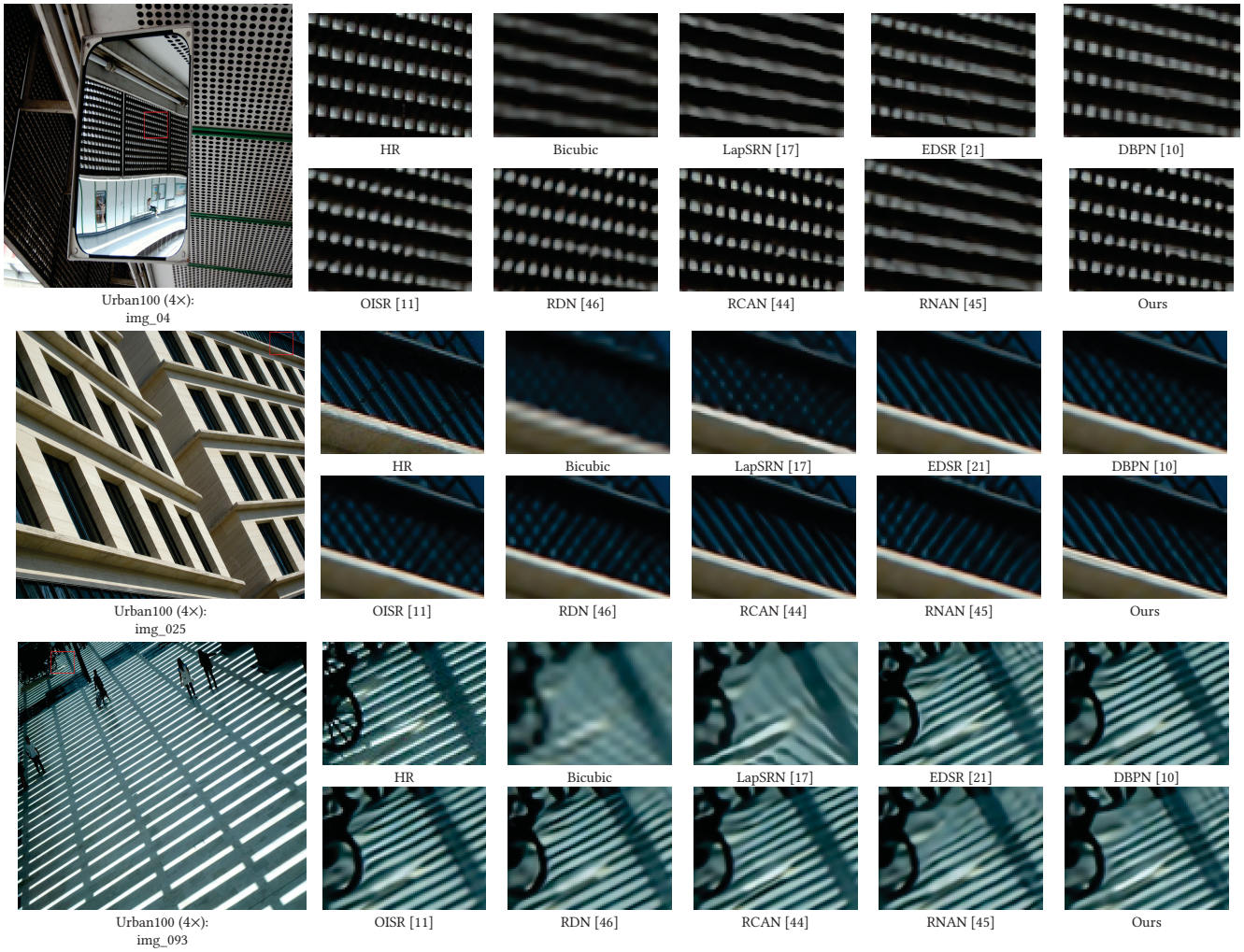


Figure 4: Visual image demosaicing results

Figure 5: Visual comparison for 4× SR on Urban100 dataset



statistical information. Furthermore, we design feature gating mechanism to adapt our MAN to varying image contents and degradations. Experiments demonstrate the effectiveness of our MAN on image restoration tasks with more visually pleasing results.

6 ACKNOWLEDGEMENT

This work is supported in part by the National Key Research and Development Program of China under Grant 2018YFB1800204, the National Natureal Science Foundation of China under Grant (61771273, 61871272), the Natural Science Foundation of Guangdong Province (2021A1515011807), the Shenzhen Fundamental Research Program under Grant JCYJ20190808173617147, the Shenzhen Science and Technology Program (Grant No. RCYX20200714114523079).

REFERENCES

[1] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. 2018. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3291–3300.

[2] Yunjin Chen and Thomas Pock. 2017. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *TPAMI* (2017).

[3] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. 2007. Color image denoising via sparse 3D collaborative filtering with grouping constraint in luminance-chrominance space. In *ICIP*.

[4] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11065–11074.

[5] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2015. Compression artifacts reduction by a deep convolutional network. In *ICCV*.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *ECCV*.

[7] C. Dong, C. C. Loy, K. He, and X. Tang. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (2016), 295–307.

[8] Raanan Fattal. 2008. Single image dehazing. *ACM transactions on graphics (TOG)* 27, 3 (2008), 1–9.

[9] Emil Julius Gumbel. 1948. *Statistical theory of extreme values and some practical applications: a series of lectures*. Vol. 33. US Government Printing Office.

[10] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. 2018. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1664–1673.

[11] Xiangyu He, Zitao Mo, Peisong Wang, Yang Liu, Mingyuan Yang, and Jian Cheng. 2019. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1732–1741.

[12] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[13] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5197–5206.

[14] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1646–1654.

[16] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *CVPR*.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*.

[19] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. 2017. Is second-order information helpful for large-scale visual recognition?. In *Proceedings of the IEEE International Conference on Computer Vision*. 2070–2078.

[20] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. 2019. Feedback network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3867–3876.

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPRW*.

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*. 136–144.

[23] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. 2018. Non-local recurrent network for image restoration. In *NeurIPS*.

[24] Yalei Lv, Tao Dai, Bin Chen, Jian Lu, Shu-Tao Xia, and Jingchao Cao. 2021. HOCA: Higher-Order Channel Attention for Single Image Super-Resolution. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1605–1609.

[25] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. 2016. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *NeurIPS*.

[26] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2. IEEE, 416–423.

[27] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S Huang, and Honghui Shi. 2020. Pyramid attention networks for image restoration. *arXiv preprint arXiv:2004.13824* (2020).

[28] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*. 807–814.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[30] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. 2017. MemNet: A Persistent Memory Network for Image Restoration. In *ICCV*.

[31] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*. 114–125.

[32] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. 2017. Image Super-Resolution Using Dense Skip Connections. In *ICCV*.

[33] Andreas Veit and Serge Belongie. 2018. Convolutional networks with adaptive inference graphs. In *Proceedings of the European Conference on Computer Vision*. 3–18.

[34] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *ICML*.

[35] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local Neural Networks. In *CVPR*.

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[38] Kaicheng Yu and Mathieu Salzmann. 2018. Statistically-motivated second-order pooling. In *Proceedings of the European Conference on Computer Vision*. 600–616.

[39] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP* (2017).

[40] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. 2017. Learning Deep CNN Denoiser Prior for Image Restoration. In *CVPR*.

[41] Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2017. FFDNet: Toward a fast and flexible solution for CNN based image denoising. *arXiv preprint arXiv:1710.04026* (2017).

[42] Kai Zhang, Wangmeng Zuo, and Lei Zhang. 2018. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3262–3271.

[43] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*. 286–301.

[44] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*.

[45] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. 2019. Residual non-local attention networks for image restoration. In *ICLR*.

[46] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2472–2481.