Single Image Super-Resolution via a Holistic Attention Network

Ben Niu[1,⋆], Weilei Wen[2,3,⋆], Wenqi Ren[3], Xiangde Zhang[1], Lianping Yang[1,†], Shuzhen Wang[2], Kaihao Zhang[5], Xiaochun Cao[3,4], and Haifeng Shen[6]

[1]Northeastern University [2]Xidian University [3]SKLOIS, IIE, CAS
[5] Peng Cheng Laboratory, Cyberspace Security Research Center, China
[5]ANU [6]AI Labs, Didi Chuxing, China

Abstract. Informative features play a crucial role in the single image super-resolution task. Channel attention has been demonstrated to be effective for preserving information-rich features in each layer. However, channel attention treats each convolution layer as a separate process that misses the correlation among different layers. To address this problem, we propose a new holistic attention network (HAN), which consists of a layer attention module (LAM) and a channel-spatial attention module (CSAM), to model the holistic interdependencies among layers, channels, and positions. Specifically, the proposed LAM adaptively emphasizes hierarchical features by considering correlations among layers. Meanwhile, CSAM learns the confidence at all the positions of each channel to selectively capture more informative features. Extensive experiments demonstrate that the proposed HAN performs favorably against the state-of-the-art single image super-resolution approaches.

Keywords: Super-Resolution, Holistic Attention, Layer Attention, Channel-Spatial Attention

1 Introduction

Single image super-resolution (SISR) is an important task in computer vision and image processing. Given a low-resolution image, the goal of super-resolution (SR) is to generate a high-resolution (HR) image with necessary edge structures and texture details. The advance of SISR will immediately benefit many application fields, such as video surveillance and pedestrian detection.

SRCNN [3] is an unprecedented work to tackle the SR problem by learning the mapping function from LR input to HR output using convolutional neural networks (CNNs). Afterwards, numerous deep CNN-based methods [26,27] have been proposed in recent years and generate a significant progress. The superior reconstruction performance of CNNs based methods are mainly from deep architecture and residual learning [7]. Networks with very deep layers have

[⋆] Equal contribution

[†] Corresponding author

larger receptive fields and are able to provide a powerful capability to learn a complicated mapping between the LR input and the HR counterpart. Due to the residual learning, the depth of the SR networks are going to deeper since residual learning could efficiently alleviate the gradient vanishing and exploding problems.

Though significant progress have been made, we note that the texture details of the LR image often tend to be smoothed in the super-resolved result since most existing CNN-based SR methods neglect the feature correlation of intermediate layers. Therefore, generating detailed textures is still a non-trivial problem in the SR task. Although the results obtained by using channel attention [40,2] retain some detailed information, these channel attention-based approaches struggle in preserving informative textures and restoring natural details since they treat the feature maps at different layers equally and result in lossing some detail parts in the reconstructed image.

To address these problems, we present a novel approach termed as holistic attention network (HAN) that is capable of exploring the correlations among hierarchical layers, channels of each layer, and all positions of each channel. Therefore, HAN is able to stimulate the representational power of CNNs. Specifically, we propose a layer attention module (LAM) and a channel-spatial attention module (CSAM) in the HAN for more powerful feature expression and correlation learning. These two sub-attention modules are inspired by channel attention [40] which weighs the internal features of each layer to make the network pay more attention to information-rich feature channels. However, we notice that channel attention cannot weight the features from multi-scale layers. Especially the long-term information from the shallow layers are easily weakened. Although the shallow features can be recycled via skip connections, they are treated equally with deep features across layers after long skip connection, hence hindering the representational ability of CNNs. To solve this problem, we consider exploring the interrelationship among features at hierarchical levels, and propose a layer attention module (LAM). On the other hand, channel attention neglects that the importance of different positions in each feature map varies significantly. Therefore, we also propose a channel-spatial attention module (CSAM) to collaboratively improve the discrimination ability of the proposed SR network.

Our contributions in this paper are summarized as follows:

- We propose a novel super-resolution algorithm named Holistic Attention Network (HAN), which enhances the representational ability of feature representations for super-resolution.
- We introduce a layer attention module (LAM) to learn the weights for hierarchical features by considering correlations of multi-scale layers. Meanwhile, a channel-spatial attention module (CSAM) is presented to learn the channel and spatial interdependencies of features in each layer.
- The proposed two attention modules collaboratively improve the SR results by modeling informative features among hierarchical layers, channels, and positions. Extensive experiments demonstrate that our algorithm performs favorably against the state-of-the-art SISR approaches.

2 Related Work

Numerous algorithms and models have been proposed to solve the problem of image SR, which can be roughly divided into two categories. One is the traditional algorithm [35,12,11], the other one is the deep learning model based on neural network [15,4,19,22,41,16,30,31]. Due to the limitation of space, we only introduce the SR algorithms based on deep CNN.

Deep CNN for super-resolution. Dong et al. [3] proposed a CNN architecture named SRCNN, which was the pioneering work to apply deep learning to single image super-resolution. Since SRCNN successfully applied deep learning network to SR task, various efficient and deeper architectures have been proposed for SR. Wang et al. [33]combined the domain knowledge of sparse coding with a deep CNN and trained a cascade network to recover images progressively. To alleviate the phenomenon of gradient explosion and reduce the complexity of the model, DRCN [16] and DRRN [30] were proposed by using a recursive convolutional network. Lai et al. [19] proposed a LapSR network which employs a pyramidal framework to progressively generate ×8 images by three sub-networks. Lim et al. [22] modified the ResNet [7] by removing batch normalization (BN) layers, which greatly improves the SR effect.

In addition to above MSE minimizing based methods, perceptual constraints are proposed to achieve better visual quality [28]. SRGAN [20] uses a generative adversarial networks (GAN) to predict high-resolution outputs by introducing a multi-task loss including a MSE loss, a perceptual loss [14], and an adversarial loss [5]. Zhang et al. [42] further transferred textures from reference images according to the textural similarity to enhance textures. However, the aforementioned models either result in the loss of detailed textures in intermediate features due to the very deep depth, or produce some unpleasing artifacts or inauthentic textures. In contrast, we propose a holistic attention network consists of a layer attention and a channel-spatial attention to investigate the interaction of different layers, channels, and positions.

Attention mechanism. Attention mechanisms direct the operational focus of deep neural networks to areas where there is more information. In short, they help the network ignore irrelevant information and focus on important information [8,9]. Recently, attention mechanism has been successfully applied into deep CNN based image enhancement methods. Zhang et al. [40] proposed a residual channel attention network (RCAN) in which residual channel attention blocks (RCAB) allow the network to focus on the more informative channels. Woo et al. [34] proposed channel attention (CA) and spatial attention (SA) modules to exploit both inter-channel and inter-spatial relationship of feature maps. Kim et al. [17] introduced a residual attention module for SR which is composed of residual blocks and spatial channel attention for learning the inter-channel and intra-channel correlations. More recently, Dai et al. [2] presented a second-order channel attention (SOCA) module to adaptively refine features using second-order feature statistics.

However, these attention based methods only consider the channel and spatial correlations while ignore the interdependencies between multi-scale layers. To

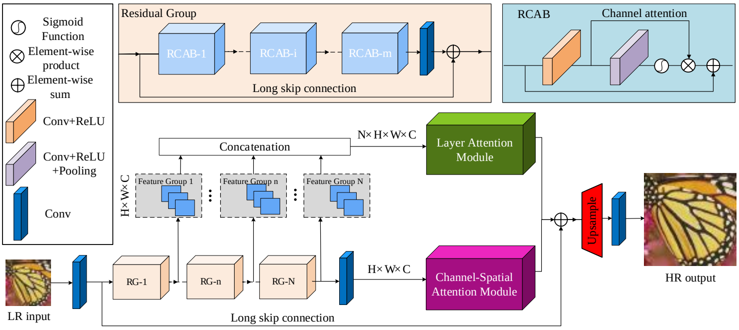

Fig. 1. Network architecture of the proposed holistic attention network(HAN). Given a low-resolution image, the first convolutional layer of the HAN extracts a set of shallow feature maps. Then a series of residual groups further extract deeper feature representations of the low-resolution input. We propose a layer attention module (LAM) to learn the correlations of each output from RGs and a channel-spatial attention module (CSAM) to investigate the interdependencies between channels and pixels. Finally, an upsampling block produces the high-resolution image

solve this problem, we propose a layer attention module (LAM) to exploit the nonlinear feature interactions among hierarchical layers.

3 Holistic Attention Network (HAN) for SR

In this section, we first present the overview of HAN network for SISR. Then we give the detailed configurations of the proposed layer attention module (LAM) and channel-spatial attention module (CSAM).

3.1 Network Architecture

As shown in Figure 1, our proposed HAN consists of four parts: feature extraction, layer attention module, channel-spatial attention module, and the final reconstruction block.

Features extraction. Given a LR input I_{LR}, a convolutional layer is used to extract the shallow feature F_0 of the LR input

$$F_0 = \text{Conv}(I_{LR}). \tag{1}$$

Then we use the backbone of the RCAN [40] to extract the intermediate features F_i of the LR input

$$F_i = H_{RB_i}(F_{i-1}), \quad i = 1, 2, ..., N, \tag{2}$$

where H_{RB_i} represents the i-th residual group (RG) in the RCAN, N is the number of the residual groups. Therefore, except F_N is the final output of RCAN network backbone, all other feature maps are intermediate outputs.

Holistic attention. After extracting hierarchical features F_i by a set of residual groups, we further conduct a holistic feature weighting, which includes: *i*) layer attention of hierarchical features, and *ii*) channel-spatial attention of the last layer of RCAN.

The proposed layer attention makes full use of features from all the preceding layers and can be represented as

$$F_L = H_{LA}(\text{concatenate}(F_1, F_2, ..., F_N)), \tag{3}$$

where H_{LA} represents the LAM which learns the feature correlation matrix of all the features from RGs' output and then weights the fused intermediate features F_i capitalized on the correlation matrix (see Section 3.2). As a results, LAM enables the high contribution feature layers to be enhanced and the redundant ones to be suppressed.

In addition, channel-spatial attention aims to modulate features for adaptively capturing more important information of inter-channel and intra-channel for the final reconstruction, which can be written as

$$F_{CS} = H_{CSA}(F_N), \tag{4}$$

where H_{CSA} represents the CSAM to produce channel-spatial attention for discriminately abtaining feature information, F_{CS} denotes the filtered features after channel-spatial attention (details can be found in Section 3.3). Although we can filter all the intermediate features of F_i using CSAM, we only modulate the last feature layer of F_N as a trade-off between accuracy and speed.

Image reconstruction. After obtaining features from both LAM and CSAM, we integrate the layer attention and channel-spatial attention units by element-wise summation. Then, we employ the sub-pixel convolution [29] as the last up-sampling module, which converts the scale sampling with a given magnification factor by pixel translation. We perform the sub-pixel convolution operation to aggregate low-resolution feature maps and simultaneously impose projection to high dimensional space to reconstruct the HR image. We formulate the process as follows

$$I_{SR} = U_{\uparrow}(F_0 + F_L + F_{CS}), \tag{5}$$

where $U_{\uparrow}$ represents the operation of sub-pixel convolution, and I_{SR} is the reconstructed SR result. The long skip connection is introduced in HAN to stabilize the training of the proposed deep network, *i.e.,* the sub-pixel upsampling block takes $F_0 + F_L + F_{CS}$ as input.

Loss function. Since we employ the RCAN network as the backbone of the proposed method, only L_1 distance is selected as our loss function as in [40] for a fair comparison

$$L(\Theta) = \frac{1}{m}\sum_{i=1}^{m}\left\|H_{HAN}(I_{LR}^i) - I_{HR}^i\right\|_1 = \frac{1}{m}\sum_{i=1}^{m}\left\|I_{SR}^i - I_{HR}^i\right\|_1, \tag{6}$$

where H_{HAN}, Θ, and m denote the function of the proposed HAN, the learned parameter of the HAN, and the number of training pairs, respectively. Note that

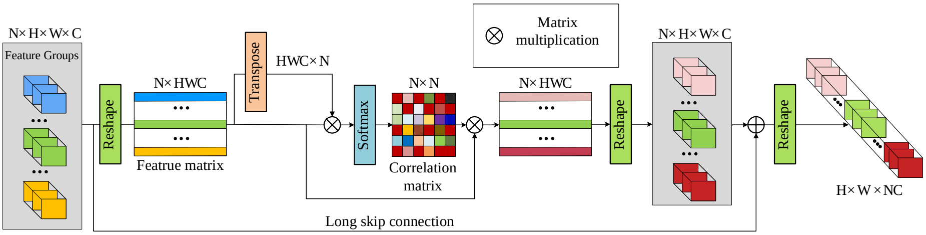

Fig. 2. Architecture of the proposed layer attention module

we do not use other sophisticated loss functions such as adversarial loss [5] and perceptual loss [14]. We show that simply using the naive image intensity loss $L(\Theta)$ can already achieve competitive results as demonstrated in Section 4.

3.2 Layer Attention Module

Although dense connections [10] and skip connections [7] allow shallow information to be bypassed to deep layers, these operations do not exploit interdependencies between the different layers. In contrast, we treat the feature maps from each layer as a response to a specific class, and the responses from different layers are related to each other. By obtaining the dependencies between features of different depths, the network can allocate different attention weights to features of different depths and automatically improve the representation ability of extracted features. Therefore, we propose an innovative LAM that learns the relationship between features of different depths, which automatically improve the feature representation ability.

The structure of the proposed layer attention is shown in Figure 2. The input of the module is the extracted intermediate feature groups FGs, with the dimension of $N \times H \times W \times C$, from N residual groups. Then, we reshape the feature groups FGs into a 2D matrix with the dimension of $N \times HWC$, and apply matrix multiplication with the corresponding transpose to calculate the correlation $W_{la} = w_{i,j=1}^{N}$ between different layers

$$w_{j,i} = \delta(\varphi(FG)_i \cdot (\varphi(FG))_j^{\mathrm{T}}), \quad i,j = 1,2,...,N, \tag{7}$$

where $\delta(\cdot)$ and $\varphi(\cdot)$ denote the softmax and reshape operations, $x_{i,j}$ represents the correlation index between i-th and j-th feature groups. Finally, we multiply the reshaped feature groups FGs by the predicted correlation matrix with a scale factor α, and add the input features FGs

$$F_{L_j} = \alpha \sum_{i=1}^{N} w_{i,j} FG_i + FG_j, \tag{8}$$

where α is initialized to 0 and is automatically assigned by the network in the following epochs. As a result, the weighted sum of features allow the main parts of network to focus on more informative layers of the intermediate LR features.

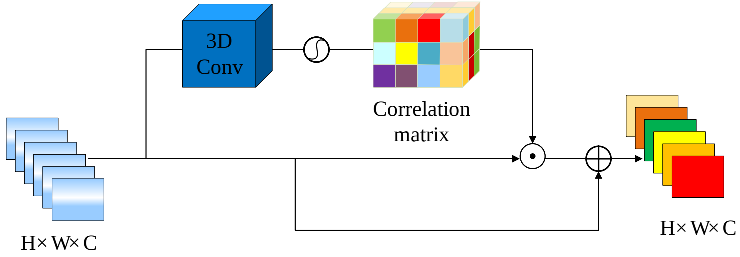

Fig. 3. Architecture of the proposed channel-spatial attention module

3.3 Channel-Spatial Attention

The existing spatial attention mechanisms [34,17] mainly focuse on the scale dimension of the feature, with little uptake of channel dimension information, while the recent channel attention mechanisms [40,41,2] ignore the scale information. To solve this problem, we propose a novel channel-spatial attention mechanism (CSAM) that contains responses from all dimensions of the feature maps. Note that although we can perform the CSAM for all the feature groups FGs extracted from RCAN, we only modulate the last feature group of F_N for a trade-off between accuracy and speed as shown in Figure 1.

The architecture of the proposed CSAM is shown in Figure 3. Given the last layer feature maps $F_N \in R^{H \times W \times C}$, we feed F_N to a 3D convolution layer [13] to generate attention map by capturing joint channel and spatial features. We operate the 3D convolution via convolving 3D kernels with the cube constructed from multiple neighboring channels of F_N. Specifically, we perform 3D convolutions with kernel size of $3 \times 3 \times 3$ with step size of 1 (*i.e.*, three groups of consecutive channels are convolved with a set of 3D kernels respectively), resulting in three groups of channel-spatial attention maps W_{csa}. By doing so, our CSAM can extract powerful representations to describe inter-channel and intra-channel information in continuous channels.

In addition, we perform element-wise multiplication with the attention map W_{csa} and the input feature F_N. Finally, multiply the weighted result by a scale factor β, and then add the input feature F_N to obtain the weighted features

$$F_{CS} = \beta\sigma(W_{csa}) \odot F_N + F_N, \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function, $\odot$ is the element-wise product, the scale factor β is initialized as 0 and progressively improved in the follow iterations. As a results, F_{CS} is the weighted sum of all channel-spatial position features as well as the original features. Compared with conventional spatial attention and channel attention, our CSAM adaptively learns the inter-channel and intra-channel feature responses by explicitly modelling channel-wise and spatial feature interdependencies.

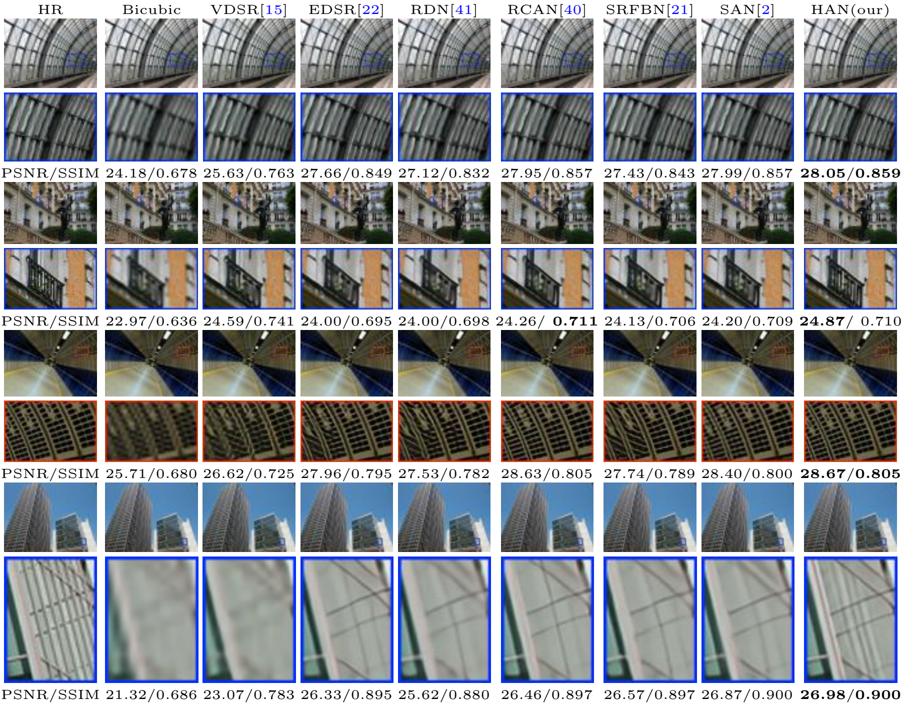
HR
Bicubic
VDSR[15]
EDSR[22]
RDN[41]
RCAN[40]
SRFBN[21]
SAN[2]
HAN(our)
PSNR/SSIM 24.18/0.678 25.63/0.763 27.66/0.849 27.12/0.832 27.95/0.857 27.43/0.843 27.99/0.857 28.05/0.859
PSNR/SSIM 22.97/0.636 24.59/0.741 24.00/0.695 24.00/0.698 24.26/ 0.711 24.13/0.706 24.20/0.709 24.87/ 0.710
PSNR/SSIM 25.71/0.680 26.62/0.725 27.96/0.795 27.53/0.782 28.63/0.805 27.74/0.789 28.40/0.800 28.67/0.805
PSNR/SSIM 21.32/0.686 23.07/0.783 26.33/0.895 25.62/0.880 26.46/0.897 26.57/0.897 26.87/0.900 26.98/0.900

Fig. 4. Visual comparison for 4× SR with BI degradation model on the Urban100 datasets. The best results are highlighted. Our method obtains better visual quality and recovers more image details compared with other state-of-the-art SR methods

4 Experiments

In this section, we first analyze the contributions of the proposed two attention modules. We then compare our HAN with state-of-the-art algorithms on five benchmark datasets. The implementation code will be made available to the public. Results on more images can be found in the supplementary material.

4.1 Settings

Datasets. We selecte DIV2K [32] as the training set as like in [40,2,41,22]. For the testing set, we choose five standard datasets: Set5 [1], Set14 [36], B100 [23], Urban100 [11], and Manga109 [24]. Degraded data was obtained by bilinear interpolation and Blur-downscale Degradation model. Following [40], the reconstruct RGB results by the proposed HAN are first converted to YCbCr space, and then we only consider the luminance channel to calculate PSNR and SSIM in our experiments.

Implementation Details. We implement the proposed network using PyTorch platform and use the pre-trained RCAN (×2), (×3), (×4), (×8) model

Table 1. Effectiveness of the proposed LAM and CSAM for image super-resolution

	baseline	w/o CSAM	w/o LAM	Ours
PSNR/SSIM	31.22/0.9173	31.38/0.9175	31.28/0.9174	**31.42/0.9177**

Table 2. Ablation study about using different numbers of RGs

	Set5	Set14	B100	Urban100	Manga100
RCAN	32.63	28.87	27.77	26.82	31.22
HAN 3RGs	32.63	28.89	27.79	26.82	31.40
HAN 6RGs	**32.64**	**28.90**	27.79	26.84	**31.42**
HAN 10RGs	**32.64**	**28.90**	**27.80**	**26.85**	**31.42**

to initialize the corresponding holistic attention networks, respectively. In our network, patch size is set as 64 × 64. We use ADAM [18] optimizer with a batch size 16 for training. The learning rate is set as 10^{-5}. Default values of β_1 and β_2 are used, which are 0.9 and 0.999, respectively, and we set $\epsilon = 10^{-8}$. We do not use any regularization operations such as batch normalization and group normalization in our network. In addition to random rotation and translation, we do not apply other data augmentation methods in the training. The input of the LAM is selected as the outputs of all residual groups of RCAN, we use $N = 10$ residual groups in out network. For all the results reported in the paper, we train the network for 250 epochs, which takes about two days on an Nvidia GTX 1080Ti GPU.

4.2 Ablation Study about the Proposed LAM and CSAM

The proposed LAM and CSAM ensure that the proposed SR method generate the feature correlations between hierarchical layers, channels, and locations. One may wonder whether the LAM and CSAM help SISR. To verify the performance of these two attention mechanisms, we compare the method without using LAM and CSAM in Table 1, where we conduct experiments on the Manga109 dataset with the magnification factor of ×4.

Table 1 shows the quantitative evaluations. Compared with the baseline method which is identical to the proposed network except for the absence of these two modules LAM and CSAM. CSAM achieves better results by up to 0.06 dB in terms of PSNR, while LAM promotes 0.16 dB on the test dataset. In addition, the improvement of using both LAM and CSAM is significant as the proposed algorithm improves 0.2 dB, which demonstrates the effectiveness of the proposed layer attention and channel-spatial attention blocks. Figure 4 further shows that using the LAM and CSAM is able to generate the results with clearer structures and details.

Table 3. Quantitative results with BI degradation model. The best and second best results are highlighted in **bold** and underlined

Methods	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9339
SRCNN [3]	×2	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946	35.60	0.9663
FSRCNN [4]	×2	37.05	0.9560	32.66	0.9090	31.53	0.8920	29.88	0.9020	36.67	0.9710
VDSR [15]	×2	37.53	0.9590	33.05	0.9130	31.90	0.8960	30.77	0.9140	37.22	0.9750
LapSRN [19]	×2	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9101	37.27	0.9740
MemNet [31]	×2	37.78	0.9597	33.28	0.9142	32.08	0.8978	31.31	0.9195	37.72	0.9740
EDSR [22]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
SRMDNF [38]	×2	37.79	0.9601	33.32	0.9159	32.05	0.8985	31.33	0.9204	38.07	0.9761
D-DBPN [6]	×2	38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324	38.89	0.9775
RDN [41]	×2	38.24	0.9614	34.01	0.9212	32.34	0.9017	32.89	0.9353	39.18	0.9780
RCAN [40]	×2	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
SRFBN [21]	×2	38.11	0.9609	33.82	0.9196	32.29	0.9010	32.62	0.9328	39.08	0.9779
SAN [2]	×2	38.31	**0.9620**	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
HAN(ours)	×2	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
HAN+(ours)	×2	**38.33**	0.9617	**34.24**	**0.9224**	**32.45**	**0.9030**	**33.53**	**0.9398**	**39.62**	**0.9787**
Bicubic	×3	30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349	26.95	0.8556
SRCNN [3]	×3	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989	30.48	0.9117
FSRCNN [4]	×3	33.18	0.9140	29.37	0.8240	28.53	0.7910	26.43	0.8080	31.10	0.9210
VDSR [15]	×3	33.67	0.9210	29.78	0.8320	28.83	0.7990	27.14	0.8290	32.01	0.9340
LapSRN [19]	×3	33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280	32.21	0.9350
MemNet [31]	×3	34.09	0.9248	30.00	0.8350	28.96	0.8001	27.56	0.8376	32.51	0.9369
EDSR [22]	×3	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
SRMDNF [38]	×3	34.12	0.9254	30.04	0.8382	28.97	0.8025	27.57	0.8398	33.00	0.9403
RDN [41]	×3	34.71	0.9296	30.57	0.8468	29.26	0.8093	28.80	0.8653	34.13	0.9484
RCAN [40]	×3	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
SRFBN [21]	×3	34.70	0.9292	30.51	0.8461	29.24	0.8084	28.73	0.8641	34.18	0.9481
SAN [2]	×3	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
HAN(ours)	×3	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
HAN+(ours)	×3	**34.85**	**0.9305**	**30.77**	**0.8495**	**29.39**	**0.8120**	**29.30**	**0.8735**	**34.80**	**0.9514**
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRCNN [3]	×4	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221	27.58	0.8555
FSRCNN [4]	×4	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280	27.90	0.8610
VDSR [15]	×4	31.35	0.8830	28.02	0.7680	27.29	0.0726	25.18	0.7540	28.83	0.8870
LapSRN [19]	×4	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet [31]	×4	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
EDSR [22]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
SRMDNF [38]	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
D-DBPN [6]	×4	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	30.91	0.9137
RDN [41]	×4	32.47	0.8990	28.81	0.7871	27.72	0.7419	26.61	0.8028	31.00	0.9151
RCAN [40]	×4	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
SRFBN [21]	×4	32.47	0.8983	28.81	0.7868	27.72	0.7409	26.60	0.8015	31.15	0.9160
SAN [2]	×4	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
HAN(ours)	×4	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
HAN+(ours)	×4	**32.75**	**0.9016**	**28.99**	**0.7907**	**27.85**	**0.7454**	**27.02**	**0.8131**	**31.73**	**0.9207**
Bicubic	×8	24.40	0.6580	23.10	0.5660	23.67	0.5480	20.74	0.5160	21.47	0.6500
SRCNN [3]	×8	25.33	0.6900	23.76	0.5910	24.13	0.5660	21.29	0.5440	22.46	0.6950
FSRCNN [4]	×8	20.13	0.5520	19.75	0.4820	24.21	0.5680	21.32	0.5380	22.39	0.6730
SCN [33]	×8	25.59	0.7071	24.02	0.6028	24.30	0.5698	21.52	0.5571	22.68	0.6963
VDSR [15]	×8	25.93	0.7240	24.26	0.6140	24.49	0.5830	21.70	0.5710	23.16	0.7250
LapSRN [19]	×8	26.15	0.7380	24.35	0.6200	24.54	0.5860	21.81	0.5810	23.39	0.7350
MemNet [31]	×8	26.16	0.7414	24.38	0.6199	24.58	0.5842	21.89	0.5825	23.56	0.7387
MSLapSRN[19]	×8	26.34	0.7558	24.57	0.6273	24.65	0.5895	22.06	0.5963	23.90	0.7564
EDSR [22]	×8	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221	24.69	0.7841
D-DBPN [6]	×8	27.21	0.7840	25.13	0.6480	24.88	0.6010	22.73	0.6312	25.14	0.7987
RCAN [40]	×8	27.31	0.7878	25.23	0.6511	24.98	0.6058	23.00	0.6452	25.24	0.8029
SAN [2]	×8	27.22	0.7829	25.14	0.6476	24.88	0.6011	22.70	0.6314	24.85	0.7906
HAN(ours)	×8	27.33	0.7884	25.24	0.6510	24.98	0.6059	22.98	0.6437	25.20	0.8011
HAN+(ours)	×8	**27.47**	**0.7920**	**25.39**	**0.6552**	**25.04**	**0.6075**	**23.20**	**0.6518**	**25.54**	**0.8080**

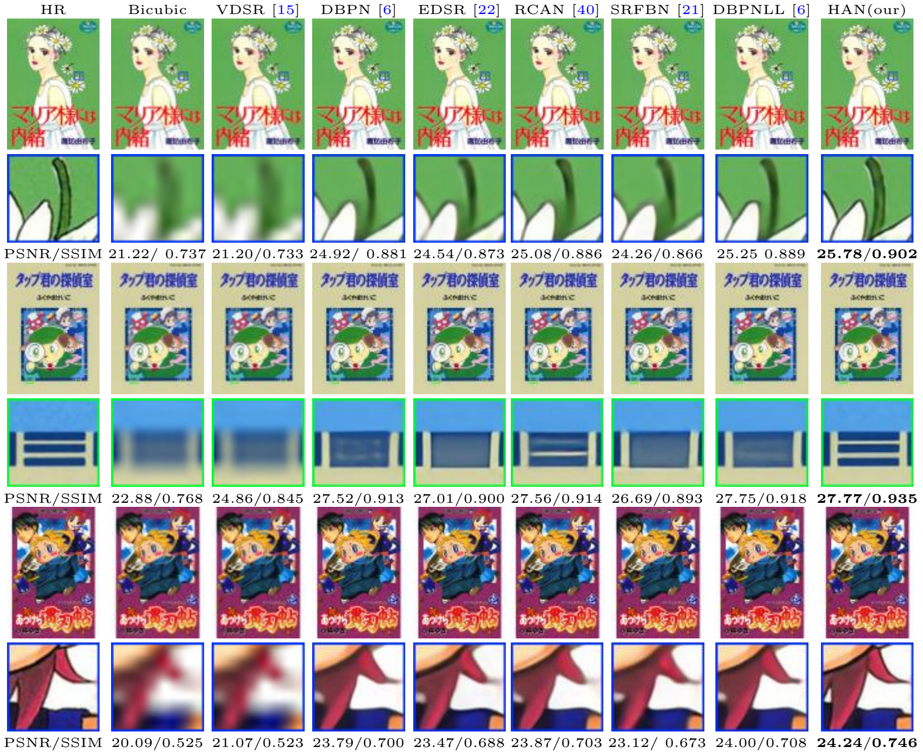

Fig. 5. Visual comparison for 8× SR with BI model on the Manga109 dataset.The best results are highlighted

4.3 Ablation Study about the Number of Residual Group

We conduct an ablation study about feeding different numbers of RGs to the proposed LAM. Specifically, we apply severally three, six, and ten RGs to the LAM, and we evaluate our model on five standard datasets. As shown in Table 2, we compare our three models with RCAN, although using fewer RGs, our algorithm still generates higher PSNR values than the baseline of RCAN. This ablation study demonstrates the effectiveness of the proposed LAM.

4.4 Ablation Study about the Number of CSAM

In the paper, the channel-spatial attention module (CSAM) can extract powerful representations to describe inter-channel and intra-channel information in continuous channels. We conduct an ablation study about using different numbers of CSAM. We use one, three, five, and ten CSAMs in RGs. As shown in Table5, with the increase of CSAM, the values of PSNR are increasing on the testing datasets. This ablation study demonstrates the effectiveness of the proposed CSAM.

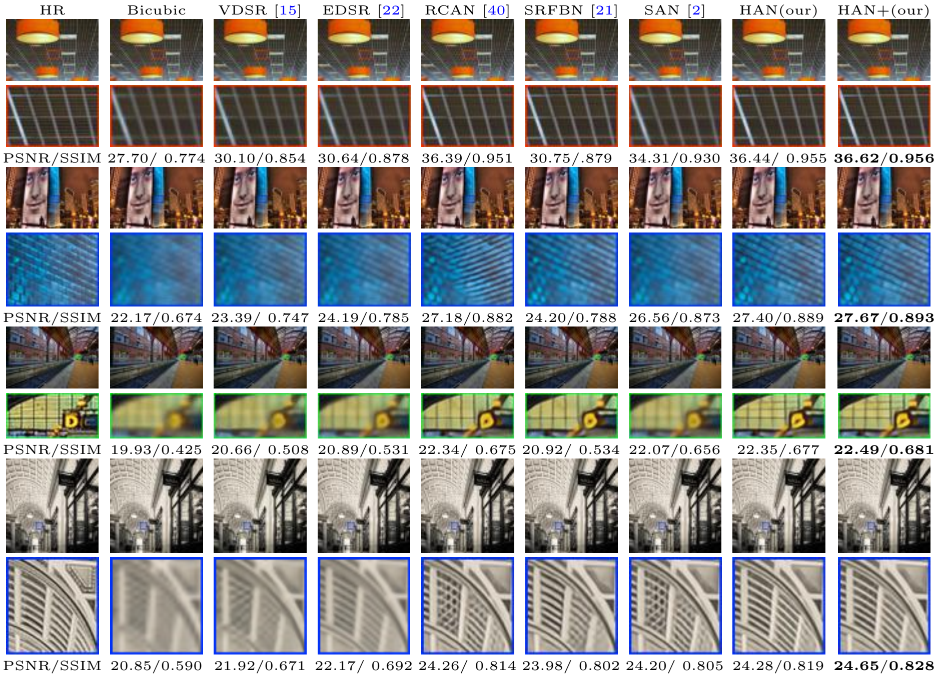

Fig. 6. Visual comparison for 3× SR with BD model on the Urban100 dataset. The best results are highlighted

4.5 Results with Bicubic (BI) Degradation Model

We compare the proposed algorithm with 11 state-of-the-art methods: SRCNN [3], FSRCNN [4], VDSR [15], LapSRN [19], MemNet [31], SRMDNF [38], D-DBPN [6], RDN [41], EDSR [22], SRFBN [21] and SAN [2]. We provide more comparisons in supplementary material. Following [22,2,40], we also propose self-ensemble model and donate it as HAN+.

Quantitative results. Table 3 shows the comparison of 2×, 3×, 4×, and 8× SR quantitative results. Compared to existing methods, our HAN+ performs best on all the scales of reconstructed test datasets. Without using self-ensemble, our network HAN still obtains great gain compared with the recent SR methods. In particular, our model is much better than SAN which also uses the same backbone network of RCAN and has more computationally intensive attention module. Specifically, when we compare the reconstruction results at ×8 scale on the Set5 dataset, the proposed HAN advances 0.11 dB in terms of PSNR than the competitive SAN.

To further evaluate the proposed HAN, we conduct experiments on the large test sets of B100, Urban100, and Manga109. Our algorithm still performs favorably against the state-of-the-art methods. For example, the super-resolved

Table 4. Quantitative results with BD degradation model. The best and second best results are highlighted in **bold** and underlined

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×3	28.78	0.8308	26.38	0.7271	26.33	0.6918	23.52	0.6862	25.46	0.8149
SPMSR [25]	×3	32.21	0.9001	28.89	0.8105	28.13	0.7740	25.84	0.7856	29.64	0.9003
SRCNN [3]	×3	32.05	0.8944	28.80	0.8074	28.13	0.7736	25.70	0.7770	29.47	0.8924
FSRCNN [4]	×3	26.23	0.8124	24.44	0.7106	24.86	0.6832	22.04	0.6745	23.04	0.7927
VDSR [15]	×3	33.25	0.9150	29.46	0.8244	28.57	0.7893	26.61	0.8136	31.06	0.9234
IRCNN [37]	×3	33.38	0.9182	29.63	0.8281	28.65	0.7922	26.77	0.8154	31.15	0.9245
SRMDNF [38]	×3	34.01	0.9242	30.11	0.8364	28.98	0.8009	27.50	0.8370	32.97	0.9391
RDN [41]	×3	34.58	0.9280	30.53	0.8447	29.23	0.8079	28.46	0.8582	33.97	0.9465
RCAN [40]	×3	34.70	0.9288	30.63	0.8462	29.32	0.8093	28.81	0.8647	34.38	0.9483
SRFBN [21]	×3	34.66	0.9283	30.48	0.8439	29.21	0.8069	28.48	0.8581	34.07	0.9466
SAN [2]	×3	34.75	0.9290	30.68	0.8466	29.33	0.8101	28.83	0.8646	34.46	0.9487
HAN(ours)	×3	34.76	0.9294	30.70	0.8475	29.34	0.8106	28.99	0.8676	34.56	0.9494
HAN+(ours)	×3	**34.85**	**0.9300**	**30.79**	**0.8487**	**29.41**	**0.8116**	**29.21**	**0.8710**	**34.87**	**0.9509**

Table 5. Ablation study about using different numbers of CSAMs

	Set5	Set14	B100	Urban100	Manga100
HAN(1 CSAM)	32.64	28.90	27.80	26.85	31.42
HAN(3 CSAM)	32.67	28.91	27.80	26.89	**31.46**
HAN(5 CSAM)	**32.69**	28.91	27.80	26.89	31.43
HAN(10 CSAM)	32.67	**28.91**	**27.80**	**26.89**	31.43

results by the proposed HAN is 0.06 dB and 0.35 dB higher than the very recent work of SAN for the 4× and 8× scales, respectively.

Visual results. We also show visual comparisons of various methods on the Urban100 dataset for 4× SR in Figure 4. As shown, most compared SR networks cannot recover the grids of buildings accurately and suffer from unpleasant blurring artifacts. In contrast, the proposed HAN obtains clearer details and reconstructs sharper high-frequency textures.

Take the first and fourth images in Figure 4 as example, VDSR and EDSR fail to generate the clear structures. The results generated by the recent work of RCAN, SRFBN, and SAN still contain noticeable artifacts caused by spatial aliasing. In contrast, our approach effectively suppresses such artifacts through the proposed two attention modules. As shown, our method accurately reconstructs the grid patterns on windows in the first row and the parallel straight lines on the building in the fourth image.

For 8× SR, we also show the super-resolved results by different SR methods in Figure 5. As show, it is challenging to predict HR images from bicubic-upsampled input by VDSR and EDSR. Even the state-of-the-art methods of RCAN and SRFBN cannot super-resolve the fine structures well. In contrast, our HAN reconstructs high-quality HR images for 8× results by using cross-scale layer attention and channel-spatial attention modules on the limited information.

4.6 Results with Blur-downscale Degradation (BD) Model

Quantitative results. Following the protocols of [38,37,41], we further compare the SR results on images with blur-downscale degradation model. We compare the proposed method with nine state-of-the-art super-resolution methods: SPMSR [25], SRCNN [3], FSRCNN [4], VDSR [15], IRCNN [37], SRMD [39], RDN [41], RCAN [40],SRFBN [21] and SAN [2]. Quantitative results on the 3× SR are reported in Table 4. As shown, both the proposed HAN and HAN+ perform favorably against existing methods. In particular, our HAN+ yields the best quantitative results and HAN obtains the second best scores for all the datasets, 0.06-0.2 dB PSNR better than the attention-based methods of RCAN and SAN and 0.2-0.8 dB better than the recently proposed SRFBN.

Visual quality. In Figure 6, we show visual results on images from the Urban 100 dataset with blur-downscale degradation model by a scale factor of 3. Both the full images and the cropped regions are shown for comparison. We find that our proposed HAN is able to recover structured details that were missing in the LR image by properly exploiting the layer, channel, and spatial attention in the feature space.

As shown, VDSR and EDSR suffer from unpleasant blurring artifacts and some results even are out of shape. RCAN alleviate it to a certain extent, but still misses some details and structures. SRFBN and SAN also fail to recover these structured details. In contrast, our proposed HAN effectively suppresses artifacts and exploits the scene details and the internal natural image statistics to super-resolve the high-frequency contents.

5 Conclusions

In this paper, we propose a holistic attention network for single image super-resolution, which adaptively learns the global dependencies among different depths, channels, and positions using the self-attention mechanism. Specifically, the layer attention module captures the long-distance dependencies among hierarchical layers. Meanwhile, the channel-spatial attention module incorporates the channel and contextual information in each layer. These two attention modules are collaboratively applied to multi-level features and then more informative features can be captured. Extensive experimental results on benchmark datasets demonstrate that the proposed model performs favorably against the state-of-the-art SR algorithms in terms of accuracy and visual quality.

Acknowledgements: This work is supported by the National Key R&D Program of China under Grant 2019YFB1406500, National Natural Science Foundation of China (No. 61971016, U1605252, 61771369), Fundamental Research Funds of Central Universities (Grant No. N160504007), Beijing Natural Science Foundation (No. L182057), Peng Cheng Laboratory Project of Guangdong Province PCL2018KP004, and the Shaanxi Provincial Natural Science Basic Research Plan (2019JM-557).

References

1. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC (2012)
2. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: CVPR (2019)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV (2014)
4. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV (2016)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
6. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: CVPR (2018)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
9. Hu, Y., Li, J., Huang, Y., Gao, X.: Channel-wise and spatial feature modulation network for single image super-resolution. IEEE Transactions on Circuits and Systems for Video Technology (2019)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
11. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015)
12. Huang, S., Sun, J., Yang, Y., Fang, Y., Lin, P., Que, Y.: Robust single-image super-resolution based on adaptive edge-preserving smoothing regularization. TIP **27**(6), 2650–2663 (2018)
13. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. TPAMI **35**(1), 221–231 (2012)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
15. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
16. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: CVPR (2016)
17. Kim, J.H., Choi, J.H., Cheon, M., Lee, J.S.: Ram: Residual attention module for single image super-resolution. arXiv preprint arXiv:1811.12043 (2018)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: CVPR (2017)
20. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
21. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: CVPR (2019)
22. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPR (2017)

23. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
24. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76**(20), 21811–21838 (2017)
25. Peleg, T., Elad, M.: A statistical prediction model based on sparse representations for single image super-resolution. TIP **23**(6), 2569–2582 (2014)
26. Ren, W., Yang, J., Deng, S., Wipf, D., Cao, X., Tong, X.: Face video deblurring using 3d facial priors. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9388–9397 (2019)
27. Ren, W., Zhang, J., Ma, L., Pan, J., Cao, X., Zuo, W., Liu, W., Yang, M.H.: Deep non-blind deconvolution via generalized low-rank approximation. In: Advances in Neural Information Processing Systems. pp. 297–307 (2018)
28. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: ICCV (2017)
29. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR (2016)
30. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR (2017)
31. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: ICCV (2017)
32. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPRW (2017)
33. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: ICCV (2015)
34. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: ECCV (2018)
35. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution as sparse representation of raw image patches. In: CVPR (2008)
36. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces (2010)
37. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep cnn denoiser prior for image restoration. In: CVPR (2017)
38. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: CVPR (2018)
39. Zhang, L., Wu, X.: An edge-guided image interpolation algorithm via directional filtering and data fusion. TIP **15**(8), 2226–2238 (2006)
40. 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)
41. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
42. Zhang, Z., Wang, Z., Lin, Z., Qi, H.: Image super-resolution by neural texture transfer. In: CVPR (2019)