Applied Intelligence (2021) 51:935–951
https://doi.org/10.1007/s10489-020-01869-z

Attention augmented multi-scale network for single image super-resolution

Chengyi Xiong[1,2] · Xiaodi Shi[1] · Zhirong Gao[3] · Ge Wang[4]

Published online: 10 September 2020

Abstract

Multi-scale convolution can be used in a deep neural network (DNN) to obtain a set of features in parallel with different perceptive fields, which is beneficial to reduce network depth and lower training difficulty. Also, the attention mechanism has great advantages to strengthen representation power of a DNN. In this paper, we propose an attention augmented multi-scale network (AAMN) for single image super-resolution (SISR), in which deep features from different scales are discriminatively aggregated to improve performance. Specifically, the statistics of features at different scales are first computed by global average pooling operation, and then used as a guidance to learn the optimal weight allocation for the subsequent feature recalibration and aggregation. Meanwhile, we adopt feature fusion at two levels to further boost reconstruction power, one of which is intra-group local hierarchical feature fusion (LHFF), and the other is inter-group global hierarchical feature fusion (GHFF). Extensive experiments on public standard datasets indicate the superiority of our AAMN over the state-of-the-art models, in terms of not only quantitative and qualitative evaluation but also model complexity and efficiency.

Keywords Single image super-resolution · Attention mechanism · Multi-scale convolution · Feature recalibration and aggregation · Local hierarchical feature fusion · Global hierarchical feature fusion.

1 Introduction

Single Image Super-Resolution (SISR), which works to construct a high-quality high-resolution (HR) image based on a corresponding low-resolution (LR) one, has attracted increasing attention in both academia and industry. With an efficient and cost-saving SISR technology, the obtained HR image is expected to equip itself with pleasing visual comfort and detailed information, which is needed in many fields, such as security and surveillance imaging [56], medical or satellite imaging [27, 44] etc. There could be many super-resolution results for a given LR one due to the lack of information, thus SISR is a challenging and ill-posed problem. So far, a large number of SISR schemes were proposed, which are generally classified as interpolation-based [3, 6, 29, 34], reconstruction-based [7, 53] and learning-based methods [2, 9–11, 20, 23, 37–39, 49, 51].

Recently, deep convolutional neural networks (DCNNs) are widely favored to solve the ill-posed SISR problem, due to the excellent representation ability of DCNNs. The DCNNs-based approaches are characterized by data-driven modeling a non-linear relationship between LR image and its HR counterpart for the optimal solution, which have shown great superiority over the traditional schemes in terms of both image recovery quality and speed. Dong et al. [9] first introduced a shollow-layer CNN based method (SRCNN) for SISR reconstruction leading to deep learning-based SISR methods. Indeed, an increasing number of deep

✉ Chengyi Xiong
xiongcy@mail.scuec.edu.cn

Xiaodi Shi
sxd0071@gmail.com

Zhirong Gao
gaozhirong@mail.scuec.edu.cn

Ge Wang
wangg6@rpi.edu

1 School of Electronic and Information Engineering, South-Central University for Nationalities, Wuhan, 430074 China

2 Hubei Key Laboratory of Intelligent Wireless Communication, South-Central University for Nationalities, Wuhan, 430074 China

3 School of Computer Science, South-Central University for Nationalities, Wuhan, 430074 China

4 Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

architectures are proposed successively. On one hand, some of these learning-based methods focus on wider or deeper network designs to improve the SISR performance. With a residual connection, VDSR [20] holds larger receptive fields to learn more information from an LR image than SRCNN [9]. Both DRCN [21] and DRRN [41] adopt a recursive-supervision structure to expand network depth, which can extract more complicated features and efficiently control the number of parameters. Lim et al. [30] proposed a very deep model (EDSR) by repeating multiple residual blocks monotonously for better characterization power. On the other hand, features of different phases along network depth can be layered, and combined to offer benefits for many high-level vision tasks [16, 17, 50]. Along this direction, some researchers introduced multi-path jump connections to combine features of different layers for SISR, e.g., DCSCN [47], SRDenseNet [43], RDN [55] and SRDCNN [23]. Although increasing network depth and introducing dense skip connections of layered features can significantly improve performance, it can also cause problems. First, deepening network could suffer from training difficulty with the increased number of parameters. Second, over-dense skip connections would make networks too complex to be practical due to the resultant slow speed and heavy memory burden.

In a deep network design, an important consideration for increasing the depth of network is to obtain a wider range of reference information. To reduce network depth, multi-scale methods such as GoogleNet [40], have attracted attention. With a similar multi-branch structure, Li et al. [28] designed a residual block (MSRB) to extract features on different scales, which can yield more effective results with a shallower network than most deeper ones. Similarly, MSFF [13] is a compressed multi-scale fusion network with structural sparsity, taking five branches to extract rich scale information by different kernel sizes. Also, Qin et al. [36] constructed a multi-scale feature fusion block, in which multi-scale features from four streams are progressively integrated to work cooperatively. The multi-stream structure not only captures affluent scale information to deal well with objects of varying sizes in an image, but also helps reduce the network complexity and speed up the model training while improving reconstruction quality.

Another point that needs to be noted in learning-based methods is that in a deep network there is not only redundancy of information, but also different degrees of importance for features from different channels and scales. How to effectively harness these features is critical to achieve high performance. Introducing an attention mechanism is an efficient solution to address this issue. Zhang et al. [54] constructed a very deep residual-in-residual network (e.g., over 400 layers), by interpolating an attention structure to selectively retain or discard feature channels under the same scale. Instead of first-order statistics in RCAN [54], Dai et.al [8] exploited the second-order properties to filter features, and employ region-level non-local operations to capture long-range context dependencies.

Based on the effectiveness of the multi-scale network in reducing network complexity and that of attentional mechanism in features harness, here we propose an attention augmented multi-scale network (AAMN) for SISR, in which the attention strategy is adopted to discriminatively aggregate the features from different scales. The proposed AAMN exploits global average pooling to compute the statistics of features on different scales, all of which are further combined as a guidance to learn the optimal weight allocation for the subsequent feature recalibration and aggregation. Meanwhile, a two-level feature fusion technique is employed to further boost reconstruction power of the network, including intra-group local hierarchical feature fusion (LHFF) and inter-group global hierarchical feature fusion (GHFF) respectively. As presented in Fig. 1, the proposed AAMN can achieve 0.14 dB performance gain relative to MSRN [28] with the roughly same setting, or 0.20 dB gain with the same size of MSRN. Extensive experiments on public standard datasets show that the proposed AAMN can obtain comparable results as popular EDSR [30] and RDN [55], while parameters of about 85% and 72% are reduced, respectively. This proves that our AAMN can take advantage of multi-scale and attention strategies for accurate reconstruction with a relatively lightweight network.

Our main contributions are listed as follows:

– An attention augmented multi-scale network (AAMN) for SISR is proposed. Benefiting from the multi-scale and attention strategies, the AAMN shows superiority in balancing performance and complexity.

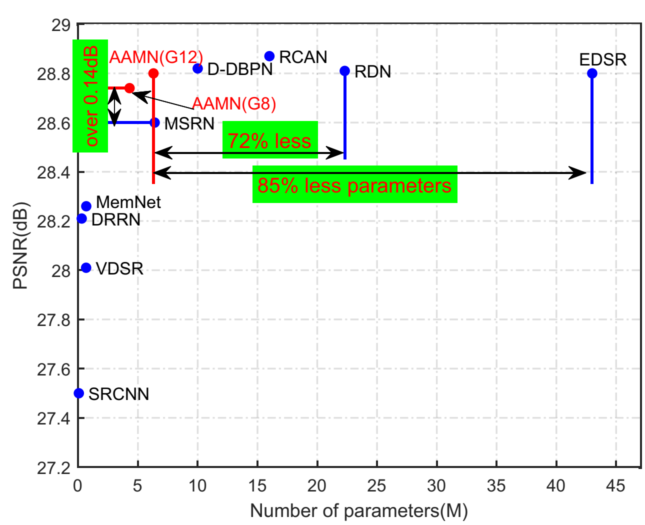


Fig. 1 Comparison in terms of model size and quantitative results with the state-of-the-art methods (x4 on Set14)

- The attention strategy is incoperated into the multi-scale network to discriminatively and effectively aggregate features from different scales and improve the super-resolution performance.
- Extensive experimental results on public standard datasets demonstrate that the proposed AAMN is competitive when compared to the state of the art algorithms, especially obtaining results comparable to that with EDSR and RDN, while the number of parameters is reduced by 85% and 72%, respectively.

2 Related work

Recently, SISR technology has drawn increasing attention from different domains for its active applications and impressive performance. SISR can mainly be grouped into two parts: traditional methods and data-driven learning methods. Here we briefly focus on the multi-scale strategy based SISR methods and attention mechanism, to which our proposed network relates.

2.1 Multi-scale SISR models

Multi-scale features are defined generally as ones obtained by convolution kernels of varying sizes, which are different from MDSR [30], where scales refer to the input with multiple sizes. Scale features can be roundly grouped into two categories: features of layers with different depths, and those of multi-branches with various kernel sizes.

On one hand, recent DCNNs-based SISR methods mostly focus on gradually widening or deepening network arcgitechtures to increase SR reconstruction ability, e.g., VDSR [20], DRCN [21], DRRN [41] and EDSR [30]. Since features along network depth have different receptive fields, they can be regarded as scale features, and are usually named as hierarchical features [55]. Hence, integrating them by multi-path jump connections can either promote information transfer or improve multi-level joint characterization, such as DenseNet [18]. For SISR, Tai et al. [42] proposed a memory block, and assemble outputs from different recursive units followed by a gate unit to maintain a long-term memory. RDN [55] densely reused features locally and globaly to enable full feature utilization. Similarily, SRDCNN [23] contains many densely-connected modules to enhance usage of layered features. Further, sharing the similar connected pattern as DLA [50] in a high-level task, Ma et al. [31] presented a dense discriminative network, resorting to the attention mechanism to gradually fuse features in a tree style.

On the other hand, multi-scale features are usually analyzed in the independent multi-stream network, such as Inception [40]. Fan et al. [13] designed a network (MSFF) for SISR, in which each MSFF module serves as a multi-scale feature extractor to help recover high-frequency details. MSRN [28] introduces two filter sizes (such as 3×3, 5×5) in each block to adaptively extract features on different scales via two branches. MSRN exceeds quite a few methods in quantitative performance, and with lower complexity. Wang et al. [46] presented a new way for up-sampling, in which multiple paths with different scale kernels are applied to predict many HR images, and each scale branch would be removed or preserved by a learned weight according to its contribution. Inspired by MSRN, Qin et al. [36] suggested a basic block, employing four intertwined fused paths to explore texture structures. With a multi-scale concept, Wan et al. [45] designed a new module, utilizing up-sampling layers to progressively fuse feature maps of hierarchy. However, these ways treat scale information from paralleled branches equally, neglecting to their dissimilarities and redundancy.

2.2 Attention mechanism for fusion

In the visual perception of human, due to limited capability of processing the entire field of view, humans would pay orientated attention to specific areas to favor information which is needed first. Then, this information is used to guide next focusing points. Inspired by such an idea, many attempts were made to selectively focus on the most useful information for different vision tasks, including image classification, image generation, lip reading and semantic segmentation [12]. Hu et al. [15] constructed a “Squeeze-and- Excitation” (SE) block to readjust features of different channels for enhanced discriminative power of CNN. For saliency detection, Kuen et al. [24] incorporated attention idea into a recurrent structure, employing an iterative way to find sub-regions for gradual saliency refinement. The attention has been demonstrated effectiveness in guiding feature learning, and becomes important in the SISR field. Zhang et al. [54] exploited feature correlations by the attention mechanism like in the SE block [15], to emphasize the informative features and filter useless ones. Inspired by this, Dai et al. [8] utilized higher-order feature statistics for feature selection. Bai et al. [48] proposed a attention-based way to adjust the original convolution. However, most of these attention-based methods just deal with single-scale features, lacking rich information, and suffer from high model complexity and unaffordable computing cost, which hinder their real applications.

3 Proposed AAMN method

In this paper, we propose an attention augmented multi-scale network (AAMN) for SISR, employing an attention-driven strategy to guide feature selection and aggregation among multiple branches. Specifically, we synthesize information of more than one scale to reasonably refine features at each scale, instead of that of a single scale. By this way, we strenthen the reconstruction ability of our network to an extreme while with low complexity. Meanwhile, a two-level fusion technique is employed to further boost reconstruction power of AAMN, including intra-group local hierarchical feature fusion (LHFF) and inter-group global hierarchical feature fusion (GHFF). We first give an overall description for the proposed AAMN accompanying with an optimization target, and further analyze its modules in detail.

3.1 Overall network framework

As shown in Fig. 2, the proposed AAMN pipeline mainly consists of four parts, namely shallow features extraction stage (SFES), local information collection group (LICG) based deep feature extraction stage (DFES), multi-level feature fusion stage (MFFS) and reconstruction stage (RS). All parts cooperate well to take full advanyage of features with various perceptive fields to achieve an accurate SISR reconstruction. Given I_{LR} and O_{SR} as input and output of AAMN, which are an LR image and the predicted SR image respectively. Similar to [19], we take a convolutional layer to obtain shallow features H_0, which can be used for deep feature extraction and auxiliary global residual connection (GRC). Formally, we have

$$H_0 = F_{SFES}(I_{LR}) \tag{1}$$

where F_{SFES} denotes a convolution operation. Then, H_0 is fed into the next deep feature extraction stage (DFES) to explore complicated features, which involves G local information collection groups (LICGs). The output of the g-th LICG H_g can be formulated as

$$H_g = F_{LICG}(H_{g-1}) \tag{2}$$

where F_{LICG} denotes the operation of LICG, and $g = 1, 2, ..., G$. Hierarchical information at different stages contributes to the final reconstructed result. Hence, we globally merge low-level and high-level information from all LICGs in MFFS (denoted as F_{MFFS}). The final deep features H_{DF} generated by the MFFS can be expressed as

$$H_{DF} = F_{MFFS}(H_0, H_1, ..., H_g, ..., H_G) \tag{3}$$

After the deep feature generation, it is the turn of the reconstruction stage to convert the features into the super-resolution (SR) image. Following [55], we choose a sub-pixel layer [39] followed by a convolution (Conv) layer to upscale features H_{DF} into a larger and better SR image, whose size matches that of the target (HR image). Consequently, the restored SR image O_{SR} from our network can be discribed as

$$O_{SR} = F_{RS}(H_{DF}) = F_{AAMN}(I_{LR}) \tag{4}$$

where F_{RS} and F_{AAMN} are functions of reconstruction stage (RS) and our AAMN respectively.

We denote the target HR image as O_{HR} and choose the L1 loss [30] to optimize our AAMN for fair contrast, instead of L2 loss [41] and perceptual loss [26]. Given N pairs of images as our training dataset, which can be denoted as $\{I^i_{LR}, O^i_{HR}\}^N_{i=1}$, where a pair of images consists of a LR input and its HR counterpart. The ultimate optimizing target

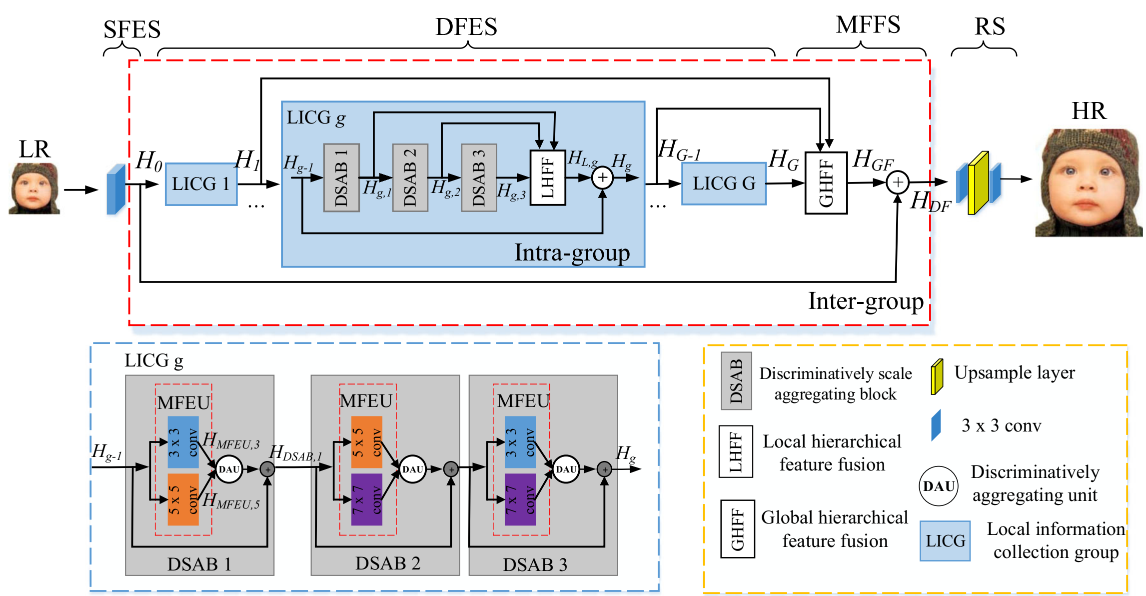


Fig. 2 Architecture of the attention augmented multi-scale network (AAMN), the local information collection group (LICG) and the discriminatively scale aggregating block (DSAB)

is defined as

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|O_{SR}^{i} - O_{HR}^{i}\right\|_1 = \frac{1}{N}\sum_{i=1}^{N}\left\|F_{AAMN}(I_{LR}^{i}) - O_{HR}^{i}\right\|_1 \tag{5}$$

where O_{SR}^{i} refers to the predicted SR image for the corresponding I_{LR}^{i}, and Θ represents the learnable parameters of AAMN.

3.2 Local information collection group (LICG)

Hierarchical features play an irreplaceable role in low-level tasks (e.g., SISR) [55], so we construct a local information collection group (LICG) as a basic module for deep feature extraction stage (DFES). The LICG fully assembles features under different receptive fields from different stages using an intra-group local hierarchical feature fusion (LHFF) technology. As is shown in Fig. 2, the LICG contains three discriminatively scale aggregating blocks (DSABs) followed by a local residual connection (LRC). We use the subscript g to represent the g-th LICG. Let H_{g-1} and H_g be the input and output of LICG. Let $H_{g,m}$ be the output of the m-th DSAB in LICG, and the outputs of three blocks can be respectively formulated as

$$\begin{aligned} H_{g,1} &= F_{DSAB,1}(H_{g-1}) \\ H_{g,2} &= F_{DSAB,2}(H_{g,1}) \\ H_{g,3} &= F_{DSAB,3}(H_{g,2}) \end{aligned} \tag{6}$$

where $F_{DSAB,m}$ ($m = 1, 2, 3$) denotes the m-th DSAB block. To cover various scale features of hierarchy, we perform a local hierarchical feature fusion (LHFF) operation for all the outputs mentioned above by skip connections within the LICG. Thus, the output of LHFF is formulated as:

$$H_{L,g} = F_{LHFF}(H_{g,1}, H_{g,2}, H_{g,3}) = W_{1\times1}([H_{g,1}, H_{g,2}, H_{g,3}]) \tag{7}$$

where $H_{L,g}$ stands for the output of LFHH, and $W_{1\times1}$ denotes a 1×1 convolution. ($[H_{g,1}, H_{g,2}, H_{g,3}]$) refers to the concatenated output of feature maps generated by three DSABs. Then, we resort to a local residual connection (LRC) to save the output of the g-th LICG as

$$H_g = F_{LICG}(H_{g-1}) = H_{g-1} + H_{L,g} \tag{8}$$

The LRC not only lowers training difficulty of the deep network but also allows a direct flow of low-frequency information from LR images.

The LICG can exploit and unite scale features of different levels, which is conducive to further enhance the multi-scale representation power of AAMN. Compared to multi-scale residual blocks in MSRN [28], our LICG captures wider-range dependencies of images by stacking DSABs, and fully harnesses scale information of hierarchy by skip connections. Different from RDN [55], which connects multiple results from simple Conv layers, the LICG aims to collect outputs of multi-scale residual blocks (DSABs) for information aggregation. In addition, although there is no dense and direct access from the preceding layer to all the subsequent layers as in RDN, the LICG also enables sufficient information flow and supplement by combination of residual connections and skip connections at a lower computational complexity. Also, instead of repeating general residual blocks simply as in RCAN [54], our LICG takes the multi-branch residual block as a basic module to capture more scale information of an image, and collects outputs of blocks for local fusion to fully reuse rich and hierarchical features. Before making a further step towards the discriminatively scale aggregating blocks (DSABs) to prove advantages of LICG, let us introduce the attention mechanism simply.

3.3 Attention mechanism (AM)

To select the most effective features in a full image, the global information is needed as a guidance to attach corresponding weights to features in accordance of their contributions. Due to the local operation in CNN, each output value can not summarize the holistic dependencies of the whole image. Let the input be $X = [x_1, ..., x_c, ..., x_C]$, which contains C feature maps with dimension of $H \times W$. The attention mechanism is described in Fig. 3. In the first step, a global average pooling [15] is applied to obtain the global statistics, which are denoted as $Z = [z_1, ..., z_c, ..., z_C]$. Then, the c-th element of Z is defined as

$$z_c = F_{GAP}(x_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i, j) \tag{9}$$

where $x_c(i, j)$ is the value at the position of (i, j) in the c-th feature map x_c , with F_{GAP} denoting a global average pooling (GAP) operation. Next, we model the nonlinear interaction and non-mutually-exclusive relationship among different feature channels, i.e., an attention learning (AL) process. The operation can be formulated as

$$W = F_{AL}(Z) = s(W_2\delta(W_1 Z)) \tag{10}$$

where F_{AL} means an AL process to learn the appropriate weight for each channel, and s and δ are the sigmoid function and ReLU function [35] respectively. The W_1 and W_2 are parameters of two full connected (FC) layers. The sigmoid gating allows a selection of multiple important channels, not a one-hot activation [15]. Assume that Z has C channels. The output of the first FC layer has $\frac{C}{r}$ channels (r is a reduction ratio), and that of the second is back to C channels, to reduce model complexity. Finally, we apply the learned weight set W to reweight the input, so that the c-th feature map y_c of output Y is computed by

$$y_c = w_c x_c \tag{11}$$

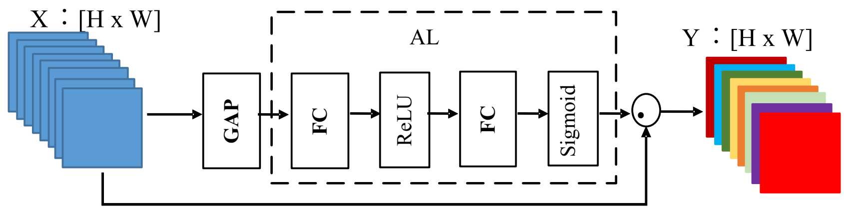


Fig. 3 Attention module (AM) where ⊙ denotes element-wise product

where x_c and w_c mean the c-th map of input X and the corresponding weight factor, respectively. By this way, the input is rescaled adaptively to focus on the important features and neglect the others.

3.4 Discriminatively scale aggregating blocks (DSABs)

Now, we go deep into the discriminative scale aggregation for three different blocks. First, we combine three convolutional kernels (e.g., 3×3, 5×5, and 7×7) in pairs into triplicates, which are used for three discriminatively scale aggregating blocks (DSABs) respectively. Then, in each block two paralleled branches are utilized to generate features with different perceptive fields, as shown in Fig. 2. After that, we use an attention-driven strategy to distinctively aggregate scale features. To reduce the number of parameters, we replace the kernel of 5×5 with two kernels of 3×3, and the kernel of 7×7 with three 3×3.

Taking the first discriminatively scale aggregating block (DSAB 1) in the g-th LICG as an example (see Fig. 2), in which the input is given as H_{g-1}. The block consists of two parts: multi-scale feature extraction unit (MFEU) and discriminatively aggregating unit (DAU) based on the attention mechanism (Fig. 3). DSAB 1 mainly serves as a merger of information produced by convolutional kernels of 3×3 and 5×5. The outputs of MFEU can be computed by

$$\begin{aligned} H_{MFEU,3} &= F_{MFEU,3}(H_{g-1}) \\ H_{MFEU,5} &= F_{MFEU,5}(H_{g-1}) \end{aligned} \tag{12}$$

where $H_{MFEU,3}$ and $H_{MFEU,5}$ denote scale features from two paths respectively, with $F_{MFEU,3}$ and $F_{MFEU,5}$ being the corresponding functions. The multi-scale structure is capable of capturing more comprehensive structural and contextual information.

Dealing with all features equally limits the discriminative learning ability of CNNs, and leads to unbalanced source allocation. Therefore, a DAU (Fig. 4) is used to fuse scale features according to their contribution indexes. We first adaptively assign proper weights to different scales for feature selection, and then fuse these recalibrated features. This method contributes to strengthen aggregation for information with different properties. First, we apply the global average pooling (GAP) operation [15] for all scale features to obtain their corresponding and global compressed statistics, which can be formulated as

$$\begin{aligned} H_{GAP,3} &= F_{GAP}(H_{MFEU,3}) \\ H_{GAP,5} &= F_{GAP}(H_{MFEU,5}) \end{aligned} \tag{13}$$

where $H_{GAP,3}$ and $H_{GAP,5}$ represent outputs of the GAP operation for dual-path scale features. Then, these statistics are concatenated as a guidance to allow a reasonable weight allocation for each scale. The process is represented as

$$H_{GAP} = Concate(H_{GAP,3}, H_{GAP,5}) \tag{14}$$

$$H_{AL} = F_{AL}(H_{GAP}) \tag{15}$$

where $Concate(\cdot)$ is a concatenation operation for two-scale feature maps along channel dimension. Next, the overall GAP outputs are feed into a series of convolution operations called the attention learning (AL) process in Fig. 3, which models interaction among the statistics for different scales to decide weight values H_{AL}. Further, we split H_{AL} back into two parts to recalibrate the corresponding scale features respectively (see the split operation and element-wise product operation in Fig. 4). The re-adjusted features for two branches are named as $H_{R,3}$ and $H_{R,5}$, and input into a 1×1 convolutional layer $(W_{1\times1})$ for feature aggregation and dimension reduction

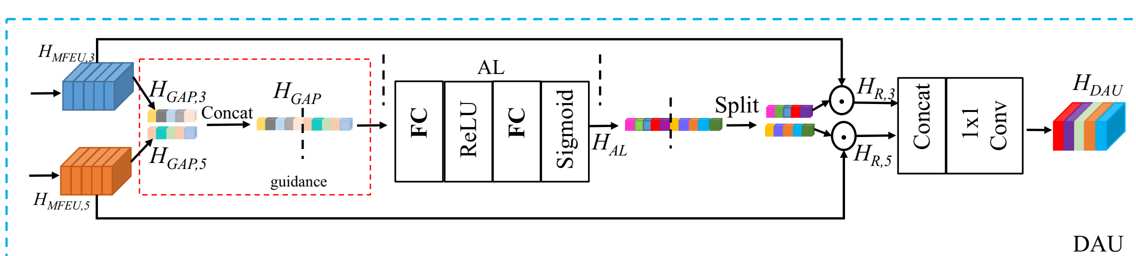


Fig. 4 Discriminatively aggregating unit (DAU) through attention learning where ⊙ denotes element-wise product

simultaneously:

$$H_{DAU} = F_{1\times1}(H_{R,3}, H_{R,5}) = W_{1\times1}([H_{R,3}, H_{R,5}]) \quad (16)$$

where the $[H_{R,3}, H_{R,5}]$ refers to a concatenation output, with H_{DAU} being the output of feature aggregation for 3×3 and 5×5 scales. Finally, to enhance information flow and stabilize network training, we obtain the final result with help of a skip residual connection:

$$H_{DSAB,1} = F_{DSAB,1}(H_{g-1}) = H_{DAU} + H_{g-1} \quad (17)$$

where $F_{DSAB,1}$ and $H_{DSAB,1}$ denote the operation and output of the first DSAB, respectively.

Here, we perform the global pooling operation (GAP) on scale features to keep their own properties independent and not spoiled, instead of working on the fused features as in [31]. Then, these statistics $H_{GAP,3}$ and $H_{GAP,5}$ are further concatenated as a guidance in the form of H_{GAP}, which synthesizes global and comprehensive information from different scale features, and properly decides which scale information should be removed or preserved. It is worth noting that, different from previous attention-based SISR models [8, 54], which mainly select important information on single scale, we combine information from multiple scales to guide feature selection. Furthermore, when compared to MSRN, which fuses scale features by a naïve 1×1 convolution, the proposed method has two advantages. One is that the AAMN motivates an excellent feature selection by taking multi-branch information as a guide to learn weight allocation for each scale. The other one is that the multi-path guidance provides an implicit supervision for subsequent feature aggregation by a mechanism of selection and fusion. In this way, we enhance discriminative learning and multi-scale joint representation power of the network.

As shown in Fig. 2, DSAB 2 and DSAB 3 are in charge of merging information produced with the convolutional kernels of 5×5 and 7×7, 3×3 and 7×7. Due to the fact that three blocks share the same aggregating mechanism, we will not describe the DSAB 2 and DSAB 3 further.

3.5 Multi-level feature fusion stage (MFFS)

The residual connections in LICGs help overcome gradient disappearance as network deepening, and ease training. While the features from different LICGs are also hierarchical, most of which would vanish gradually when being delivered along network depth [28]. To take full advantage of low- and high-level multi-scale information for more accurate reconstruction, we perform fusion for features from all the LICGs followed by a global residual connection. We first integrate these features by a 1×1 and 3×3 convolution (F_{GHFF}) as H_{GF} :

$$\begin{aligned} H_{GF} &= F_{GHFF}(H_1, ..., H_g, H_G) \\ &= W_{3\times3}(W_{1\times1}([H_1, ..., H_g, H_G])) \end{aligned} \quad (18)$$

where $[H_1, ..., H_g, H_G]$ stands for a concatenation of shallow and deep features from LICG $1, ..., g, ..., G$, $W_{1\times1}$ and $W_{3\times3}$ denote 1×1 and 3×3 convolutions respectively. To enhance information flow and alleviate gradient vanishment, we add a global residual connection (GRN) to obtain the final output H_{DF} as

$$H_{DF} = H_{GF} + H_0 \quad (19)$$

GHFF could make the best of layered features from different stages to guide information restoration, which is neglected in RCAN [54]. Furthermore, different from global fusion in MSRN [28], AAMN uses a global residual connection (GRC) to hold the final fused result. GRC not only speeds up the training process but also boosts expression power of the network. Finally, sharing similar ways of fusion, AAMN fully utilizes features from multi-scale residual modules, while RDN [55] devotes to reusing those from the dense residual block (RDB). Different from the intra-group LHFF technology within LICG, GHFF targets an inter-group pattern. With the two-level fusion technique, AAMN covers longer-range structural and contextual information to ensure width and depth of information, which further boosts reconstruction power of the network.

4 Implementation setup

In our implementation, we set different convolution kernel sizes for specific modules. Concretely, kernels of 3×3 and 5×5 are used in DSAB1, 5×5 and 7×7 in DSAB 2, and 3×3 and 7×7 in DSAB 3, respectively. We use the size of 3×3 for other kernels except for fusion and attention with 1×1.

For each LICG, we stack three DSABs followed by a 1×1 filter to obtain its outputs. Similar to [54], the reduction ratio r in the attention module is set to 16. Furthermore, our network contains an adjustable number of LICGs, which is denoted as G. Except for extra notes, all layers have 64-channel filters. Finally, we choose ESPCN [39] for reconstruction, followed by a three-filter convolutional layer, producing the final RGB HR image.

Table 1 Effectiveness analysis of DAU (x4 on Set5)

Schemes	DAU_A	DAU_B	DAU_C (ours)
Set5	32.40	32.35	**32.43**

5 Experiments

5.1 Experimental settings

Our training dataset derives from 800 high-resolution (HR) images in DIV2K [1] dataset, in which 800 images for training, 100 images for validation, and 100 images for test. We perform a Bicubic down-sample on 800 images at three scale factors ($\times 2$, $\times 3$, and $\times 4$) to gain LR input patches, which cooperate with HR images in pairs. All training images are augmented by randomly rotating 90°, 180°, 270° and flipping horizontally to increase diversity of data. For each mini-batch, we randomly crop 16 LR images with size of 48×48 as inputs. Our model is optimize by Adam [22] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The learning rate is initialized 10^{-4} for the AAMN network, and reduced half every 200 epochs.

For evaluation, we utilize five standard benchmark datasets: Set5 [4], Set14 [52], BSD100 [32], Urban 100 [32] and Manga 109 [33], which involve various characteristics and resolutions. And with PSNR and structural similarity as a metric, the proposed method is evaluated quantitatively on luminance channel, namely Y part of transformed YCbCr space, and higher value of which means more accurate restoration. All experiments are conducted on PyTorch platform with a NVIDIA GTX 1080Ti GPU.

5.2 Analysis of DAU and G

To verify advantages of the discriminatively aggregating unit (DAU) on SISR reconstruction task, we design another two variants to replace DAU, called DAU_A and DAU_B for comparison. Similar to [31], DAU_A first fuses the concatenated features with 1×1 convolution, then apply an attention mechanism to select the informative features. While, based on the motivation of choosing the best for combination, DAU_B firstly select the best parts for different scales respectively, then followed by a 1×1 fusion. But the DAU_B doesn't consider if these best parts directly can be fused well to work best. Consequently, the DAU_C synthesizes multi-scale information as a guidance to learn the weight allocation for feature selection, which enables an implicit supervision for later feature aggregation. Different to DAU_A, we compute global statistics for all the scale features to keep their own characteristics undestroyed. With the number of LICG as 12, Table 1 shows performance comparison between the DAU-based and the other two methods on Set5 in 6×10^5 iterations. It can be found that the DAU_C can obtain better results, increasing by 0.03 dB than DAU_A and 0.08dB than DAU_B, which shows effectiveness of the proposed DAU.

Network depth plays a significant role on improving reconstruction performance. Consequently, we study the influence of number of LICGs (i.e., G), with G increasing from 8 to 12. Figure 5 shows convergence analysis of proposed AAMN with different G for scaling factor $\times 4$ on Set14 and DIV2K. As depicted in graph Fig. 5, higher value of G will contribute to performance improvement due to the excellent non-linear abstract ability of deep network. So does in Fig. 5, as G increasing from 8 to 12, the loss will tend to converge more quickly. Meanwhile, it can be observed that the learning curves will tend to make a slight difference after G=10, which can be called saturated phenomenon of deeper network. In addition,

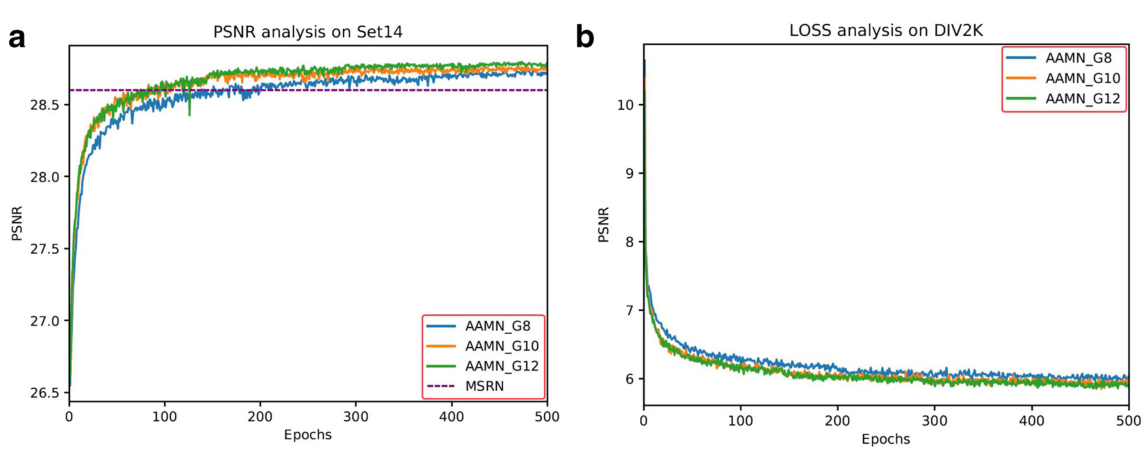


Fig. 5 Convergence analysis on Set14 and DIV2K with scaling factor $\times 4$

Table 2 Ablation investigations of multi-scale structure (Multi-scale), discriminatively aggregating unit (DAU), local hierarchical feature fusion (LHFF), global hierarchical feature fusion (GHFF) and global residual connection (GRC)

Schemes	Base	A	B	C	D	E	F	G	H
Multi-scale	×	✓	✓	✓	✓	✓	✓	✓	✓
DAU	×	×	✓	×	×	×	✓	✓	✓
LHFF	×	×	×	✓	×	×	✓	✓	✓
GHFF	×	×	×	×	✓	×	×	✓	✓
GRC	×	×	×	×	×	✓	×	×	✓
PSNR	32.09	32.19	32.31	32.24	32.30	32.28	32.32	32.33	32.37

The best results on Set5 ($\times 4$) in 200 epochs are shown

although with only G=8 LICGs, the relatively shallow AAMN is still superior to the MSRN [28], from which our design motivation derives. In later sections, we set G=12 for all experiments to balance the performance and model parameters.

5.3 Ablation experiments

In this section, we decompose our AAMN to explore effectiveness of each module, namely the multi-scale structure (Multi-scale), discriminatively aggregating unit (DAU), intra-group local hierarchical features fusion (LHFF), inter-group global hierarchical features fusion (GHFF) and global residual connection (GRC). Here, the LHFF operation is always tied with the local residual connection (LRC) together to promote information flow. Our base model (denoted as Base) remove these five parts, with a single path to replace two-branch structure. Specifically, we just reserve a largest kernel (5×5 path in DSAB 1, 7×7 path in DSAB 2, and 7×7 path in DSAB 3). So in the Base scheme, each DSAB will be a simple residual block. Setting the number of LICG as 12, the best results on Set5 for scale factor $\times 4$ in 200 epochs are summarized as Table 2

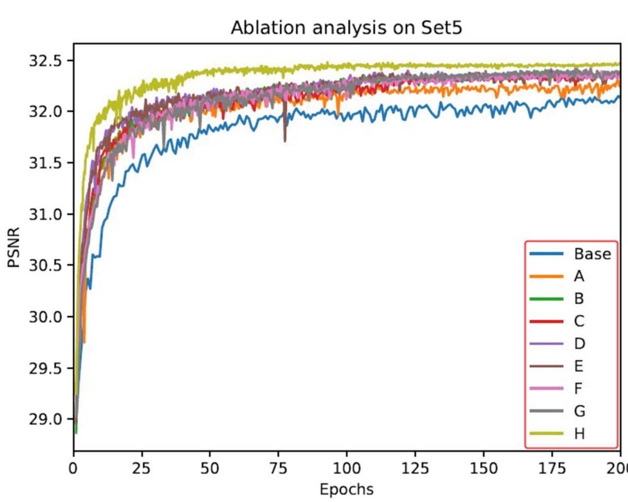


Fig. 6 Convergence analysis on Set5 with scaling factor $\times 4$

From the Table 2, the Base network performs poorly, just gaining PSNR=32.09 dB. Then, we add each module to the Base respectively to investigate the effectiveness of five modules, resulting in A, B, C D and E schemes in Table 2. Compare to the baseline, the A increases by 0.10 dB, owing to that the multi-scale structure can capture more abundant scale structure and provide adequate clues to guide infommation recovery. Furthermore, the B model with DAU makes about 0.22 dB improvements on PSNR over the Base because of advantages of scale aggreagting technology. Also, the other three models (C, D, E) achieve obvious quantitative performance gains (shown from the 5-th to 7-th column in Table 2) This is mainly because the hierarchical feature fusion and residual learning contribute a lot to information reuse and supplement.

Next, we further add three components (LHFF, GHFF and GRC) to the B model successively to validate effectiveness of combined modules, yielding F, G and H schemes in Table 2. Comparing the results, we can find that the model with multiple components performs better than the rest. Among them, the AAMN obtain the best performance when equipping with five modules, which is also presented in Fig. 6 (visualization of convergence process for these nine schemes on Set5). Furthermore, we can see that the residual learning and integration of layered features help to stabilize training process, which makes the AAMN avoid too much PSNR fluctuation.

5.4 Model complexity

Table 3 shows comparisons with some mainstream approaches with scaling factor of 4 on Set14, in term of model parameters and quantitative performance. First, compare to MSRN [28], by which our method is inspired. Keeping the same number of blocks (G=12) as MSRN, our AAMN can gain higher PSNR by a large margin (0.20 dB) with slightly less parameters. This Here, "M" denotes million. Then, ours can achieve comparable performance

Table 3 Comparison of model size and quantitative results

Methods	SRCNN	DRRN	DCAN	EDSR	MSRN	D-DBPN	RDN	Ours
Parameters	0.05M	0.3M	1.7 M	43M	6.4M	10M	22.3M	6.3M
PSNR/dB	27.49	28.21	28.30	28.80	28.60	**28.82**	<u>28.81</u>	28.80

The best quantitative performance is shown in **bold**, and the second best is <u>underlined</u>

as EDSR [30], D-DPBN [14] and RDN [55], while just possess much fewer parameters and storage. So the proposed method AAMN can performs better on keeping a trade-off between the performance and model complexity, when compared with these state-of-the-arts. Consequently, the combination of attention and multi-scale branch allows the AAMN to be more practical in real world.

5.5 Running time evaluations

In this section, we make comparisons with 9 typical approaches in terms of running time to demonstrate efficiency of AAMN. They include SRCNN [9], FSRCNN [10], VDSR [20], LapSRN [25], DCAN [48], DRRN [41], EDSR [30], MSRN [28] and RDN [55]. Here, we do not consider SAN [8], which takes much time and storage to conducted, so less suitable to real applications. The evaluation results are exhibited in Fig. 7, which describes the trade-off between the mean PSNR and running time on Set5 dataset at scaling factor x4. With the comparable performance, the AAMN is slightly slower than EDSR with an acceptable rate, and holds a faster speed when compared to RDN. Although not the fastest, the AAMN outperforms the rest of methods by a large margin (at least over 0.20 dB) in reconstruction performance at relative fast inference speed. The reason is that the AAMN need some time to

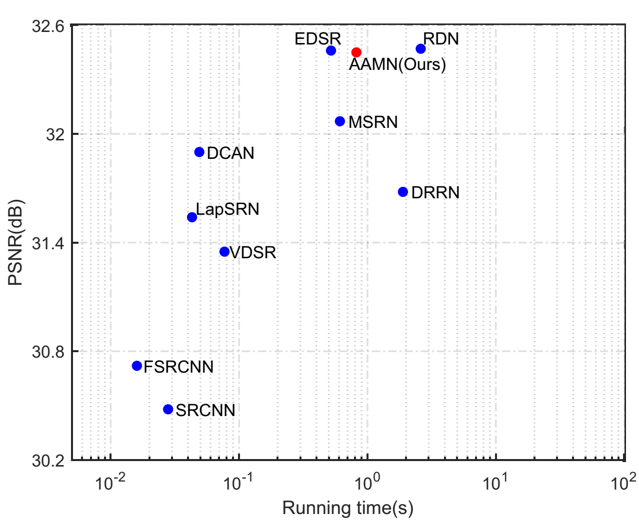


Fig. 7 The trade-off of between the reconstructed accuracy and running time on Set5 for scaling factor ×4. Ours is marked in red

generate global statistics and reweight the features in DAU process. In summary, the proposed model can restore an accurate SR image with high efficiency.

5.6 Comparisons with state-of-the-art approaches

For model evaluation, We apply Bicubic down-sampling operation to generate LR images from the original HR images. We compare the proposed model with 11 state-of-the-art methods on five standard test datasets to evaluate effectiveness of AAMN from a quantitative perspective. These methods includes Bicubic, SRCNN [9], FSRCNN [10], VDSR [20], DRRN [41], DCAN [48], SRDCNN [23], LapSRN [25], EDSR [30], ,D-DPBN [14] and RDN [55]. Here, we leave out three models: MDSR [30] which takes multi-scale patches as input, RCAN [54] whose number of layers is quadruple of ours (i.e., over 400 layers in RCAN), and SAN [8] which take too much time and memory. Following [30], we apply a self-ensemble strategy to our network, resulting in AAMN+. Table 4 reports quantitative comparisons for scaling factors ×2, ×3 and ×4. Results of these are gained by executing their publicly available source codes.

As is shown in Table 4, our AAMN+ can almost perform best on all the datasets for three scaling factors. Although without the self-ensemble strategy, ours can still reach comparable level as EDSR [30], D-DPBN [14], and RDN [55], while with much lower model parameters (see Table 3) and storage consumption. It indicates our AAMN can hold a great balance for quantitative performance and model complexity. Furthermore, the AAMN can obtain more enhanced results by a remarkable margin (at least 0.08dB) on five datasets for all the scaling factors, when compared to the rest approaches. There are mainly some reasons accountable for this. First, the multi-scale structure allows AAMN to explore more diverse structure and texture dependencies, thus can provide adequate clues to guide information restoration. Second, the scale aggregating unit (DAU) can comprehensively synthesize features at different scales to assign the reasonable weight value for each scale, which can motivate these recalibrated features to be effectively fused. Third, two-level fusion technique (LHFF and GHFF) allows to aggregate information from different phases maximallly for more accurate reconstruction.

Table 4 Quantitative results of BI degradation model for scaling factors ×2, ×3 and ×4

Methods	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	BSD100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	x2	33.66/0.9299	30.24/0.8688	29.56/0.8431	26.88/0.8403	30.80/0.9339
SRCNN [9]	x2	36.66/0.9542	32.45/0.9067	31.36/0.8879	29.50/0.8946	35.60/0.9663
FSRCNN [10]	x2	37.05/0.9560	32.66/0.9090	31.53/0.8920	29.88/0.9020	36.67/0.9710
VDSR [20]	x2	37.53/0.9590	33.05/0.9130	31.90/0.8960	30.77/0.9140	37.22/0.9750
SRDCNN [23]	x2	37.26/0.9573	32.69/0.8986	31.55/0.8908	-/-	-/-
LapSRN [25]	x2	37.52/0.9591	33.08/0.9130	31.08/0.8950	30.41/0.9101	37.27/0.9740
DRRN [41]	x2	37.74/0.9591	33.23/0.9136	32.05/0.8973	31.23/0.9188	37.60/0.9736
DCAN [48]	x2	37.96/0.9614	33.29/0.9160	32.15/0.9001	31.51/0.9230	-/-
EDSR [30]	x2	38.11/0.9602	33.92/0.9195	32.32/0.9013	32.93/0.9351	-/-
D-DPBN [14]	x2	38.09/0.9600	33.85/0.9190	32.27/0.9000	32.55/0.9324	38.89/0.9775
RDN [55]	x2	**38.24**/0.9614	**34.01**/0.9212	32.34/**0.9017**	32.89/0.9353	39.18/**0.9780**
AAMN (Ours)	x2	38.19/0.9611	33.88/0.9200	32.31/0.9013	32.86/0.9347	39.14/0.9774
AAMN+ (Ours)	x2	**38.24/0.9616**	33.97/**0.9213**	**32.35/0.9017**	**33.05/0.9360**	**39.32/0.9780**
Bicubic	x3	30.39/0.8682	27.55/0.7742	27.21/0.7385	24.46/0.7349	26.95/0.8556
SRCNN [9]	x3	32.75/0.9090	29.30/0.8215	28.41/0.7863	26.24/0.7989	30.48/0.9117
FSRCNN [10]	x3	33.18/0.9140	29.37/0.8240	28.53/0.7910	26.43/0.8080	31.10/0.9210
VDSR [20]	x3	33.67/0.9210	29.78/0.8320	28.83/0.7990	27.14/0.8290	32.01/0.9340
SRDCNN [23]	x3	33.59/0.9234	29.54/0.8244	28.80/0.7973	-/-	-/-
LapSRN [25]	x3	33.82/0.9227	29.87/0.8320	28.82/0.7980	27.07/0.8280	32.21/0.9350
DRRN [41]	x3	34.03/0.9244	29.96/0.8349	28.95/0.8004	27.53/0.8378	32.42/0.9359
DCAN [48]	x3	34.16/0.9263	29.94/0.8364	28.94/0.8020	27.41/0.8371	-/-
EDSR [30]	x3	34.65/0.9280	30.52/0.8462	29.25/0.8093	28.80/0.8653	-/-
D-DPBN [14]	x3	-/-	-/-	-/-	-/-	-/-
RDN [55]	x3	34.71/0.9296	30.57/0.8468	29.26/0.8093	28.80/0.8652	34.13/0.9484
AAMN (Ours)	x3	34.67/0.9292	30.53/0.8460	29.24/0.8087	28.77/0.8638	34.07/0.9477
AAMN+ (Ours)	x3	**34.74/0.9297**	**30.62/0.8472**	**29.29/0.8096**	**28.94/0.8664**	**34.33/0.9491**
Bicubic	x4	28.42/0.8104	26.00/0.7027	25.96/0.6675	23.14/0.6577	24.89/0.7866
SRCNN [9]	x4	30.48/0.8628	27.50/0.7513	26.90/0.7101	24.52/0.7221	27.58/0.8555
FSRCNN [10]	x4	30.72/0.8660	27.61/0.7550	26.98/0.7150	24.62/0.7280	27.90/0.8610
VDSR [20]	x4	31.35/0.8830	28.02/0.7680	27.29/0.0726	25.18/0.7540	28.83/0.8870
SRDCNN [23]	x4	31.16/0.8788	27.85/0.7644	27.08/0.7090	-/-	-/-
LapSRN [25]	x4	31.54/0.8850	28.19/0.7720	27.32/0.7270	25.21/0.7560	29.09/0.8900
DRRN [41]	x4	31.68/0.8888	28.21/0.7721	27.38/0.7284	25.44/0.7638	29.18/0.8914
DCAN [48]	x4	31.90/0.8921	28.30/0.7772	27.44/0.7340	25.57/0.7720	-/-
EDSR [30]	x4	32.46/0.8968	28.80/0.7876	27.71/0.7420	26.64/0.8033	-/-
D-DPBN [14]	x4	32.47/0.8980	28.82/0.7860	27.72/0.7400	26.38/0.7946	30.91/0.9137
RDN [55]	x4	32.47/0.8990	28.81/0.7871	27.72/0.7419	26.61/0.8028	31.00/0.9151
AAMN (Ours)	x4	32.45/0.8982	28.80/0.7865	27.71/0.7409	26.58/0.8013	30.92/0.9141
AAMN+ (Ours)	x4	**32.57/0.8995**	**28.86/0.7878**	**27.76/0.7420**	**26.77/0.8049**	**31.24/0.9171**

The best performance is shown in **bold**, and the second best is underlined

Furthermore, we make comparisons between the proposed AAMN with some similar networks with multi-scale concept for SISR on Set14 for all scaling factors, and the evaluations are listed in Table 5. These comparisons contain PSRN, network depth and parameters. This networks include MSFF [13], MSRN [28], MSRCAN [5] and PRNet [45]. Among these, MSFF [13] presents a compressed multi-scale feature fusion module, utilizing five

Table 5 Comparisons of Quantitative result, model depth and parameters (for x4) with some similar multi-scale methods on Set14

Method	x2	x3	x4	Depth	Para.
MSFF [13]	33.04/0.913	29.80/0.832	28.07/0.768	10	0.047M
MSRN [28]	33.74/0.917	30.34/0.839	28.60/0.775	25	6.4M
MSRCAN [5]	33.37/0.915	30.09/0.837	28.33/0.775	24	4.4M
PRNet [45]	**34.02/0.921**	**30.57/0.847**	**28.86/0.788**	232	17.5M
AAMN (Ours)	33.88/0.920	30.53/0.846	28.80/0.787	113	6.3M

The best performance is shown in **bold**, and the second best is underlined (PSNR /SSIM)

branches with different kernel sizes to capture rich structure information. Then, MSRN [28] introduces two paths to adaptively detect features at different scales, and integrate them into a multi-scale grid block by a skip connection. MSRCAN [5] places a attention module after the fused scale features to select key information in both LR and HR space, which is equivalent to the mentioned DAU_A scheme (Table 1). And the PRNet [45] presents a new way to gradually aggregate feature maps of hierarchy and different sizes for image reconstruction by a series of up-sampling operations. From the results, we can find that the AAMN performs well with reasonable model complexity. here, "M" denotes million. Especially, although keeping a lighter architecture, the AAMN still reach a comparable level as PRNet. Furthermore, involving dense-connected up-sampling operations, the PRNet would occupy too much storage and time. All of these prove effectiveness and applicability of AAMN.

5.7 Qualitative analysis

To further validate the advantages of our AAMN, we continue to make qualitative result analyses on some representative images from the standard datasests. For the scaling factor of ×3 and ×4, we choose two images for comparison respectively (shown in Fig. 8 and Fig. 9). It can be shown that the AAMN can gain a sharper and visually comfortable HR image than all the comparative methods. More important, benefiting from the multi-scale structure and scale aggregating unit, the AAMN is capable to distinguish minute details from a complex structure. Take the image "barbara" as an example. Most of the compared models can just predict a single direction of the grid structure. However, our AAMN can restore a complete grid shape. Also, for the image "img_024" and "YumeiroCooking", the AAMN can gain clear strip shapes, while the rest models tend to generate artifact or twining

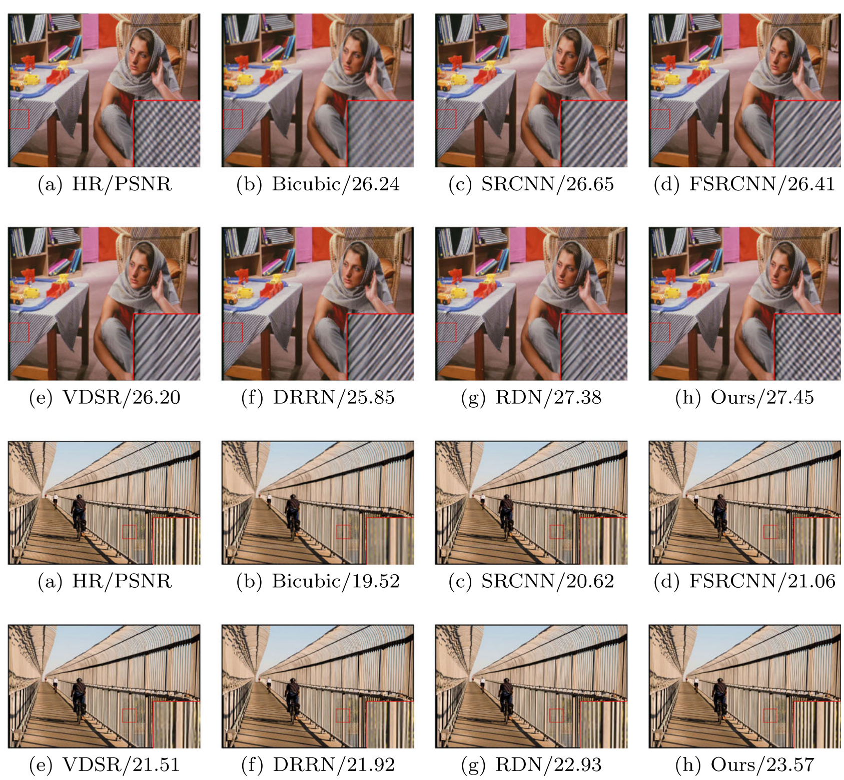


Fig. 8 Visual result comparisons for "barbara" (top) from Set14 and "img_024"(bottom) from Urban100 at scaling factor ×3. (Zoom in for best view)

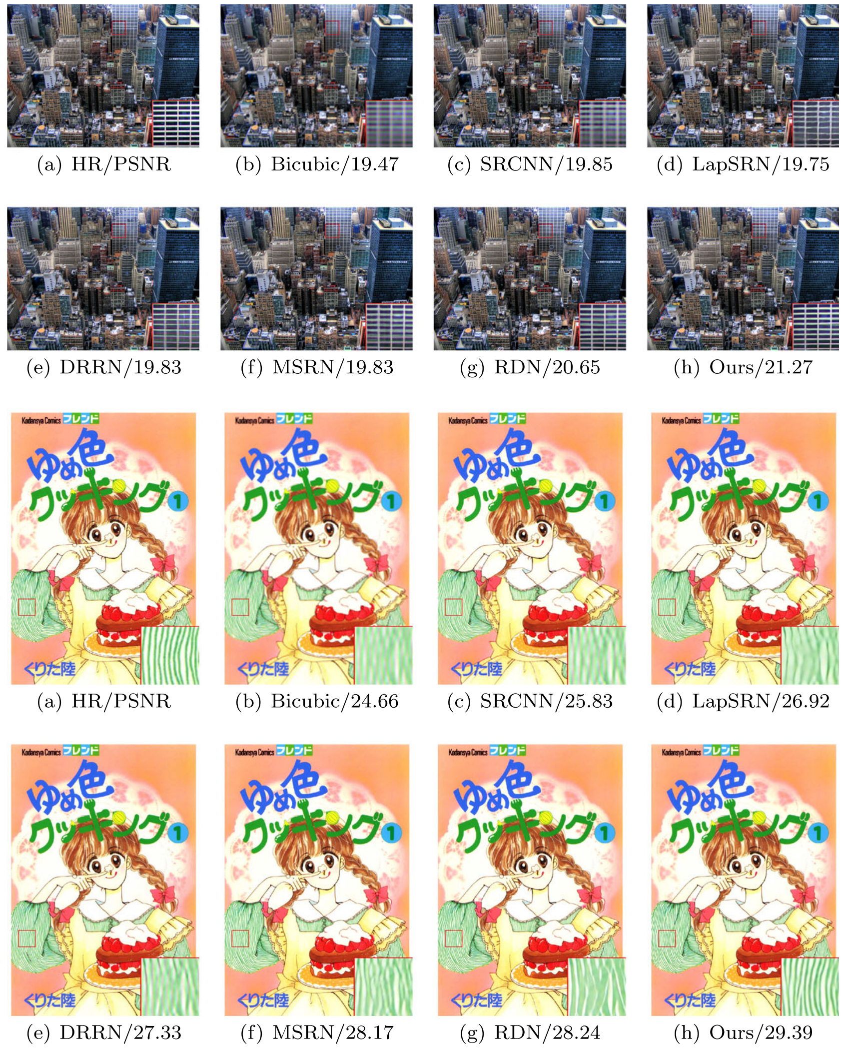


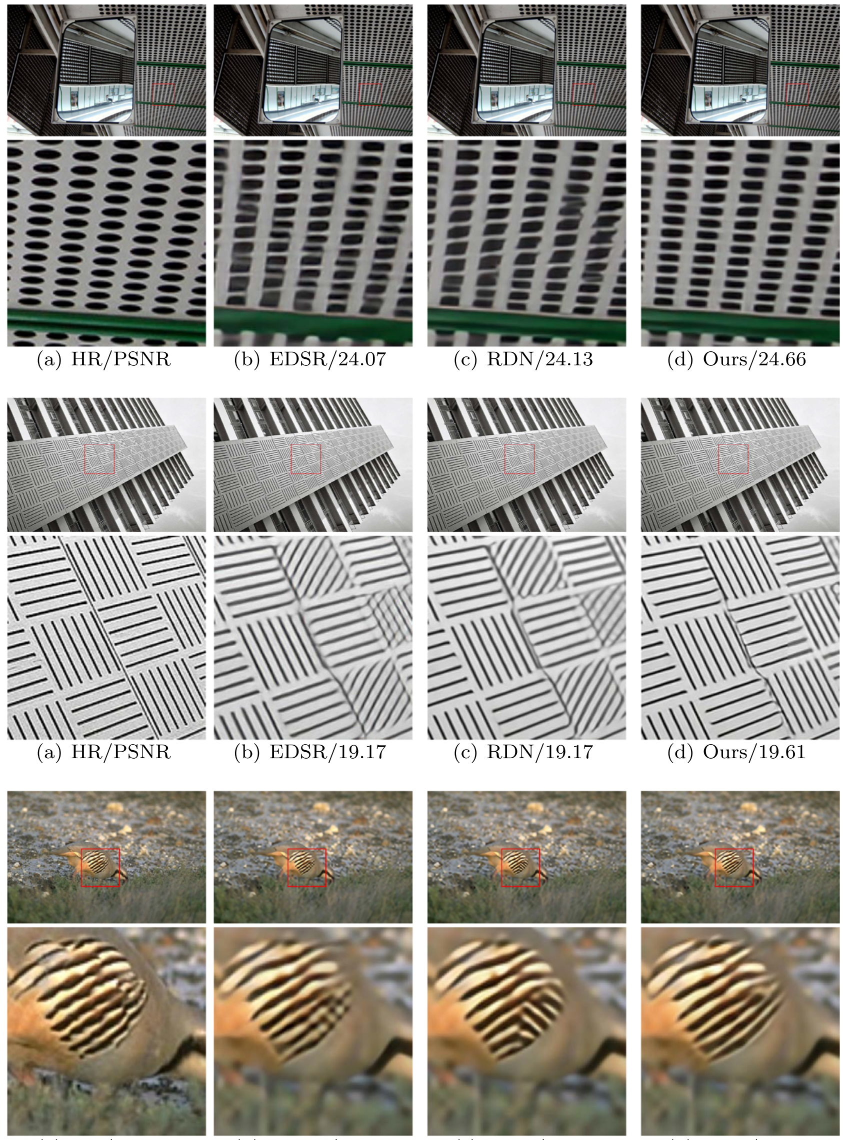
Fig. 9 Visual result comparisons for "img_073" (top) from Urban100, and "YumeiroCooking" (bottom) from Manga109 at scaling factor $\times 4$. (Zoom in for best view)

shapes. That is because the proposed AAMN can take rich information to comprehensively decide which parts should be reserved or neglected, thus not be interfered by other scale features unexpectedly.

In addition, we also display visual comparisons with two mainstream models (EDSR [28] and RDN [25]) in several complicated images from Urban100 and BSD100 in Fig. 10. For the images "img_004" and "img_092", both the other two models bring obvious artifacts and make more ambiguous prediction. Moreover, when dealing with the tiny and intricate stripes, EDSR and RDN tend to restore false direction, as shown on the image "img_092" and "8023". This phenomenon may be due to that EDSR and RDN are not good at processing more subtle structures, or make a proper feature selection from a complicated background. On the contrary, with the multi-scale structure and scale aggregating module, the AAMN can (1) grasp a variety of scale information to deal well with objects at varying sizes in an image; (2) take global information as a guidance to selectively keep correct scale and discard the rest, which provides a potential supervision for effective scale feature aggregation. Consequently, the proposed method is qualified to distinguish error-prone shapes from complex structures and infer accurate and sharper details, which fit to the Ground Truth better.

6 Conclusions

In this paper, we have proposed an attention augmented multi-scale network (AAMN), which can sufficiently and efficiently merge features of different perceptual fields from multiple paralleled branches. Specifically, an attention-driven aggregating strategy is designed to fully aggregate scale features in the discriminatively scale aggregating block (DSAB). This way allows to synthesize multi-cale information to allocate weights for features at different scales, which guides subsequent feature aggregation. Also,



(a) HR/PSNR (b) EDSR/24.07 (c) RDN/24.13 (d) Ours/24.66

(a) HR/PSNR (b) EDSR/19.17 (c) RDN/19.17 (d) Ours/19.61

(a) HR/PSNR (b) EDSR/30.93 (c) RDN/30.34 (d) Ours/31.05

Fig. 10 Visual result comparisons for "img_004" (top) and "img_092" (middle) from Urban 100, and "8023" from BSD100 (bottom) at scaling factor ×4. (Zoom in for best view)

a two-level fusion technique is proposed to further boost the reconstruction power of the network, which involves intra-group local hierarchical feature fusion (LHFF) and inter-group global hierarchical feature fusion (GHFF). The technique not only fully harnesses scale features of hierarchy within a local information collection group(LICG) but also integrates information from all LICGs, to strengthen information transmission and ease training difficulty. With these advantages, AAMN boost discriminative learning ability and multi-scale joint characterization power of the network. Experimental results on benchmark datasets demonstrate the superiority of our AAMN over the state-of-the-art methods, in terms of quantitative and qualitative performance, model complexity and reconstruction efficiency, which is critically important for real-world applications. In the future, we will extend the proposed attention-driven aggregating strategy to the related image restoration tasks, and explore more effective algorithms for accurate and efficient SISR reconstruction.

Funding Information This work was supported by the National Natural Science Foundation of China under Grant 61471400 and"the Fundamental Research Funds for the Central Universities", South-Central University for Nationalities(CZY19016).

Author Contributions Chengyi Xiong designed and annalyzed the algorithm, wrote the paper. Xiaodi Shi designeded algorithm, performed experiment and wrote the paper. Zhirong Gao and Ge Wang contributed to improve the experiment, and revised the manuscript. All authors read and approved the final manuscript.

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

Availability of data and material We declare that all data generated or analysed during this study are included in this article. And the datasets are used during the current study are available online.

Code availability We declare all code generated or used during the study is available from the corresponding github repository of authors.

References

1. Agustsson E, Timofte R (2017) Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 126–135
2. Aquino G, Zacarias A, Rubio JDJ, Pacheco J, Gutierrez GJ, Ochoa G, Balcazar R, Cruz DR, Garcia E, Novoa JF (2020) Novel nonlinear hypothesis for the delta parallel robot modeling. IEEE Access 8:46324–46334
3. Ashfahani A, Pratama M, Lughofer E, Ong Y (2019) Devdan: Deep evolving denoising autoencoder. Neurocomputing
4. Bevilacqua M, Roumy A, Guillemot C, Alberi-morel ML (2012) Low-complexity single-image super-resolution based on nonnegative neighbor embedding
5. Cao F, Liu H (2019) Single image super-resolution via multi-scale residual channel attention network. Neurocomputing 358:424–436
6. Chiang H, Chen M, Huang Y (2019) Wavelet-based eeg processing for epilepsy detection using fuzzy entropy and associative petri net. IEEE Access 7:103255–103262
7. Dai S, Han M, Xu W, Wu Y, Gong Y (2007) Soft edge smoothness prior for alpha channel super resolution. In: 2007 IEEE Conference on computer vision and pattern recognition. IEEE, pp 1–8
8. Dai T, Cai J, Zhang Y, Xia ST, Zhang L (2019) Second-order attention network for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 11065–11074
9. Dong C, Loy CC, He K, Tang X (2014) Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer, pp 184–199
10. Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: European conference on computer vision. Springer, pp 391–407
11. Elias I, Rubio JDJ, Cruz DR, Ochoa G, Novoa JF, Martinez DI, Muniz S, Balcazar R, Garcia E, Juarez CF (2020) Hessian with mini-batches for electrical demand prediction. Appl Sci 10(6):2036
12. Emami H, Aliabadi MM, Dong M, Chinnam RB (2019) Spa-gan: Spatial attention gan for image-to-image translation. arXiv:1908.06616
13. Fan X, Yang Y, Deng C, Xu J, Gao X (2018) Compressed multi-scale feature fusion network for single image super-resolution. Signal Process 146:50–60
14. Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1664–1673
15. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
16. Huang G, Chen D, Li T, Wu F, van der Maaten L, Weinberger KQ (2017) Multi-scale dense networks for resource efficient image classification. arXiv:1703.09844
17. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
18. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
19. Jin X, Xiong Q, Xiong C, Li Z, Gao Z (2019) Single image super-resolution with multi-level feature fusion recursive network. Neurocomputing 370:166–173
20. Kim J, Kwon Lee J, Mu Lee K (2016) Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1646–1654
21. Kim J, Kwon Lee J, Mu Lee K (2016) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1637–1645
22. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
23. Kuang P, Ma T, Chen Z, Li F (2019) Image super-resolution with densely connected convolutional networks. Appl Intell 49(1):125–136
24. Kuen J, Wang Z, Wang G (2016) Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pp 3668–3677
25. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 624–632
26. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
27. Li F, Jia X, Fraser D (2008) Universal hmt based super resolution for remote sensing images. In: 2008 15Th IEEE international conference on image processing. IEEE, pp 333–336
28. Li J, Fang F, Mei K, Zhang G (2018) Multi-scale residual network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 517–532
29. Li X, Orchard MT (2001) New edge-directed interpolation. IEEE Trans Image Process 10(10):1521–1527
30. Lim B, Son S, Kim H, Nah S, Mu Lee K (2017) Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 136–144
31. Ma J, Wang X, Jiang J (2019) Image super-resolution via dense discriminative network. IEEE Transactions on Industrial Electronics
32. Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol 2. IEEE, pp 416–423
33. Matsui Y, Ito K, Aramaki Y, Fujimoto A, Ogawa T, Yamasaki T, Aizawa K (2017) Sketch-based manga retrieval using manga109 dataset. Multimed Tools Appl 76(20):21811–21838
34. Medacampana JA (2018) On the estimation and control of nonlinear systems with parametric uncertainties and noisy outputs. IEEE Access 6:31968–31973
35. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
36. Qin J, Huang Y, Wen W (2019) Multi-scale feature fusion residual network for single image super-resolution. Neurocomputing

37. Rubio JDJ (2009) Sofmls: Online self-organizing fuzzy modified least-squares network. IEEE Trans Fuzzy Syst 17(6):1296–1309
38. Schulter S, Leistner C, Bischof H (2015) Fast and accurate image upscaling with super-resolution forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3791–3799
39. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1874–1883
40. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
41. Tai Y, Yang J, Liu X (2017) Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3147–3155
42. Tai Y, Yang J, Liu X, Xu C (2017) Memnet: a persistent memory network for image restoration. In: Proceedings of the IEEE international conference on computer vision, pp 4539–4547
43. Tong T, Li G, Liu X, Gao Q (2017) Image super-resolution using dense skip connections. In: Proceedings of the IEEE International Conference on Computer Vision, pp 4799–4807
44. Van Reeth E, Tham IW, Tan CH, Poh CL (2012) Super-resolution in magnetic resonance imaging: a review. Concept Magn Reson Part A 40(6):306–325
45. Wan J, Yin H, Chong AX, Liu ZH (2020) Progressive residual networks for image super-resolution. Appl Intell:1–13
46. Wang C, Li Z, Shi J (2019) Lightweight image super-resolution with adaptive weighted learning network. arXiv:1904.02358
47. Yamanaka J, Kuwashima S, Kurita T (2017) Fast and accurate image super resolution by deep cnn with skip connection and network in network. In: International conference on neural information processing. Springer, pp 217–225
48. Yang J (2019) Densely convolutional attention network for image super-resolution. Neurocomputing:pp 368
49. Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. IEEE Trans Image Process 19(11):2861–2873
50. Yu F, Wang D, Shelhamer E, Darrell T (2018) Deep layer aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2403–2412
51. Zeng K, Ding S, Jia W (2019) Single image super-resolution using a polymorphic parallel cnn. Appl Intell 49(1):292–300
52. Zeyde R, Elad M, Protter M (2010) On single image scale-up using sparse-representations. In: International conference on curves and surfaces. Springer, pp 711–730
53. Zhang K, Gao X, Tao D, Li X (2012) Single image super-resolution with non-local means and steering kernel regression. IEEE Trans Image Process 21(11):4544–4556
54. Zhang Y, Li K, Li K, Wang L, Zhong B, Fu Y (2018) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 286–301
55. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y (2018) Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2472–2481
56. Zou WW, Yuen PC (2011) Very low resolution face recognition problem. IEEE Trans Image Process 21(1):327–340

Chengyi Xiong received his B.S. degree in radio technology from University of Electronic Science and Technology of China, in 1992, and Ph.D. degree in control science and engineering from School of Automation, Huazhong University of Science and Technology, in 2006. Now he is a Professor in South-Central University for Nationalities. His research interests include image restoration, image compression, compressive sensing, machine learning, and deep learning.

Xiaodi Shi received the B.S. degree from South-Central University for Nationalities, Wuhan, China, in 2018. She is currently pursuing the Master degree in school of electronic and information engineering at South-Central University for Nationalities, Wuhan, China. Her research interests include super resolution, image restoration, and deep learning.

Zhirong Gao received the Ph.D degree from school of Computer, Wuhan University, in 2019. She is now also an associate professor in College of Computer Science, South-Central University for Nationalities. Her research interests include image super resolution, compressive sensing, and machine learning.

Ge Wang is currently a Clark and Crossan Chair Professor and the Director of the Biomedical Imaging Center, Rensselaer Polytechnic Institute, USA. He authored the papers on the first spiral/helical cone-beam/multi-slice CT algorithm. Currently, there are over 100 million medical CT scans yearly with a majority in the spiral conebeam mode. He pioneered bioluminescence tomography. His-group published the first papers on interior tomography and omni-tomography (all-in-one) to acquire diverse data sets simultaneously (all-at-once). His-results were featured in Nature, Science, and PNAS, and recognized with awards. He has written more than 430 peerreviewed journal publications. He is a Fellow of SPIE, OSA, AIMBE, AAPM, and AAAS.