


Article

# Triple-Attention Mixed-Link Network for Single-Image Super-Resolution

Xi Cheng <sup>1</sup>, Xiang Li <sup>2</sup>  and Jian Yang <sup>1,\*</sup>

<sup>1</sup> Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

<sup>2</sup> Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

\* Correspondence: csjyang@njust.edu.cn

Received: 26 May 2019; Accepted: 23 July 2019; Published: 25 July 2019



**Featured Application:** Single-image super-resolution (SISR) is an important low-level computer-vision task with high practical value in many fields such as industrial inspection, medical imaging and security monitoring.

**Abstract:** Single-image super-resolution is of great importance as a low-level computer-vision task. Recent approaches with deep convolutional neural networks have achieved impressive performance. However, existing architectures have limitations due to the less sophisticated structure along with less strong representational power. In this work, to significantly enhance the feature representation, we proposed triple-attention mixed-link network (TAN), which consists of (1) three different aspects (i.e., kernel, spatial, and channel) of attention mechanisms and (2) fusion of both powerful residual and dense connections (i.e., mixed link). Specifically, the network with multi-kernel learns multi-hierarchical representations under different receptive fields. The features are recalibrated by the effective kernel and channel attention, which filters the information and enables the network to learn more powerful representations. The features finally pass through the spatial attention in the reconstruction network, which generates a fusion of local and global information, lets the network restore more details, and improves the reconstruction quality. The proposed network structure decreases 50% of the parameter growth rate compared with previous approaches. The three attention mechanisms provide 0.49 dB, 0.58 dB, and 0.32 dB performance gain when evaluating on Set5, Set14, and BSD100. Thanks to the diverse feature recalibrations and the advanced information flow topology, our proposed model is strong enough to perform against the state-of-the-art methods on the benchmark evaluations.

**Keywords:** super-resolution; mixed-link networks; triple-attention

## 1. Introduction

Single-image super-resolution (SISR) is an important low-level computer-vision task with high practical value in many fields such as industrial inspection, medical imaging, and security monitoring. SISR aims at recovering a high-resolution image from only one low-resolution image. For this ill-posed inverse problem, widely used interpolation methods cannot achieve visually pleasing results and many learning-based methods [1,2] have been proposed. In recent years, deep learning-based algorithms [3–9] have been developed which have greatly improved the quality, and the detail of the images can be better preserved with these powerful deep networks. The introduction of attention mechanisms further improves the representation power of the neural networks. SENet [10] focuses on

the attention between channels and achieves good results. Attention is not limited to channels; many other kinds of attention (e.g., spatial attention) play important roles in tasks of segmentation, detection, and re-id. However, most of the recent approaches in this task focus only on stacking deep networks to enhance the representation power of the models, which leads to large computation and memory cost. These methods ignore the use of attention mechanisms which can greatly reduce information redundancy and improve network efficiency. Therefore, we propose a powerful deep network with three different types of effective attention mechanisms to enhance the super-resolution task.

The network structure is another important factor that influences the representation power of networks. In recent studies, residual networks with very deep structure [4] or networks with dense connections [8,9] were found effective in super-resolution tasks, both of which have achieved good results. In order to make the networks more powerful, much recent research has focused on taking advantage of both the effective network topologies [7,11]. In this work we introduced an attention-enhanced multi-kernel mixed-link network. In the proposed structure, the network was designed with synchronous residual and dense connection to gain contiguous memory [8] (CM) and make full use of the hierarchical information from the LR space. Also, we used different kernel size to help the network gain different receptive fields and kernel attention was introduced to fuse different outputs between convolutional layers to improve the network performance. Moreover, we add channel attention between the connections of the layers and spatial attention in the reconstruction networks for further filtering and enhancing the flow of gradient and information. When training the network, we introduced multi-supervise, which enables block-output high-resolution images, and we calculate the loss between all of the outputs and target to make TAN stably output high-resolution images with high quality. We summarize our main contributions in the following points:

- We propose a triple-attention mixed-link network (TAN) for image super-resolution. The proposed three attention mechanisms provide 0.49 dB, 0.58 dB, and 0.32 dB performance gain when evaluating on Set5, Set14, and BSD100.
- We propose an attention-enhanced multi-kernel mixed-link block which could help achieve better performance with 50% parameter growth rate compared with previous approaches.
- Our model achieved state-of-the-art performance according to the benchmark.

With the help of mixed-link topology, the model could gain better representation power than other methods. The channel, kernel, and spatial attention enable the network to select and fuse information to make the model more effective. Thanks to the effective attention and network topology, our proposed TAN could not only achieve better performance but also contain fewer parameters, which is of great practical value.

In the next section, we mainly analyze the recent research related to this paper and analyze the strengths and problems of the previous methods. In Section 3, we explain the structure of the TAN and the calculation method for the attention modules. In Section 4 we conduct experiments to compare performance and parameters with the state of the art. We also analyze the performance gain of the network topology fusion and the various attention modules. Section 5 mainly discusses the strengths and weaknesses of the approach we propose. Section 6 is the conclusion of the full text.

## 2. Related Works

Single-image super-resolution has become a research hotspot in recent years. Deep-learning-based methods have shown great improvement compared with conventional methods such as interpolations, anchored neighborhood regression [2], self-exemplars [12], and methods based on sparse encoding [1]. SRCNN [3] firstly used convolutional neural networks to sample images and achieved significant improvements. The performance of SRCNN was limited by its shallow structure. To achieve higher performance, the networks tend to be increasingly deeper. Kim et al. proposed the VDSR [13] model with a deeper structure. In order to make the deep model trainable, recursive supervision and residual models were introduced [8,14,15]. Recently, some very deep models have been proposed such as EDSR

and MDSR [4], which achieves very pleasing performance on super-resolution tasks. In addition, super-resolution models integrated with dense connections are proposed, including SRDenseNet [9] and MemNet [8], which effectively use the hierarchical features. Later, RDN [7] was proposed, which effectively used hierarchical features and controlled parameter growth. RCAN [6] was proposed with channel attention and ultra-deep structure, which showed great improvements of quality. In terms of a reconstruction network, the model also gradually uses deconvolution and effective subpixel shuffle [5,16] to replace the traditional pre-interpolation process, which simplified the computational complexity and preserved more image details. The above methods showed impressive performance; however, their structures were complex and very deep. To achieve higher performance, attention plays another key role in addition to the scale and complexity of network. RCAN [6] only introduced channel attention, which is not sufficient to make full use of the flow of information, and the ultra-large scale (over 400 layers) makes it consume much computation and memory. Moreover, the above methods cannot make full use of the residual and dense features. Although many methods [7–9] have used dense connections to create a contiguous memory, all of them construct the network with separated residual and dense links. We build the network with synchronous residual and dense connections, which enables linking the two connections together and gives a better fusion of residual and dense features. Also, feature extraction usually requires different receptive fields and effective recalibration or fusion for features from different subfields, which were neglected in previous super-resolution methods. In order to solve these problems, we proposed a triple-attention network with multi-kernel mixed-link structure for single-image super-resolution. We will introduce our proposed TAN in detail in Section 3.

### 3. Proposed Method

This section describes the construction details and the calculation of the TAN. The organization of this section is as follows: Section 3.1 shows the overall framework of the proposed TAN. Section 3.2 shows the basic principles and calculation methods of mixed-link connection. Section 3.3 shows the three attention modules proposed in this paper, where Section 3.3.1 illustrates the attention among channels, Section 3.3.2 is the attention between convolution kernels, and Section 3.3.3 shows the spatial attention in the reconstruction network. Section 3.4 is our algorithm for multiple supervision and loss function.

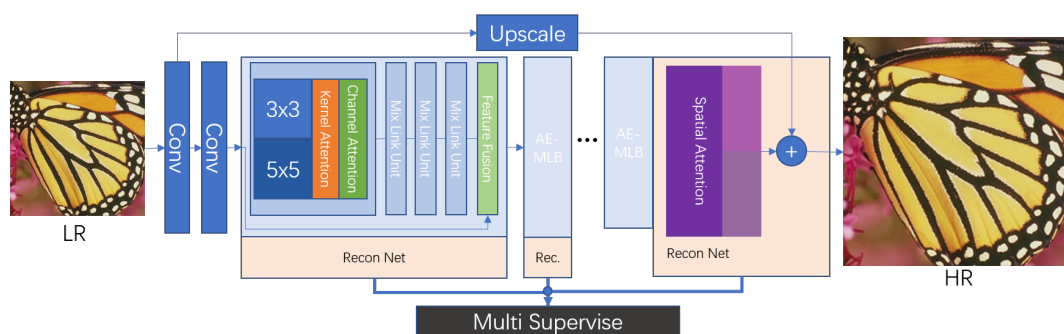


Figure 1. Overall structure of triple-attention mixed-link network.

#### 3.1. Overall Model Structure

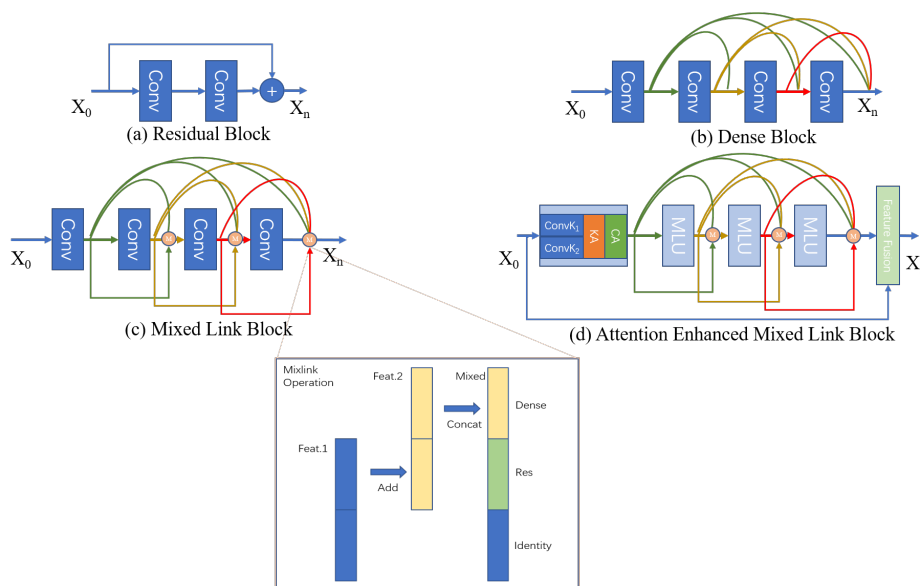
Figure 1 shows our proposed TAN, containing three basic parts, which are shallow feature extractor (SFENet), attention-enhanced mixed-link blocks (AE-MLBs), and reconstruction networks with multi-supervise. The SFENet contains two convolution layers to grab the shallow features through the network. Low-resolution images were fed directly into the network and divided into two branches; one was input to the upscale module after the first convolution layer in SFENet, and the other is then passed through the second convolution layer and input to the AE-MLBs to predict the details.

The reconstruction network uses global residual learning [13] which combines the upscaled image with the predicted details to generate the high-resolution image.

### 3.2. Mixed-Link Connections

Figure 2d shows the inner structure of the mixed-link connections. Operator M in the figure denotes the mixed-link operation, which gives a fusion of residual and dense connections between the current layer and previous layer. The mixed-link operation could be calculated as Formulas (1)–(4). This could be divided into three parts. The first slices the input channels into two parts equally;  $S(\cdot)$  means the slice in Formula (1). Suppose the feature map  $x_i$  has  $2N$  channels, the output feature maps  $x_i^1$  and  $x_i^2$  will contain  $N$  channels after the slicing operation.

$$x_i^1, x_i^2 = S(x_i) \tag{1}$$



**Figure 2.** Comparison of different network structures. (a) residual block in ResNet, (b) dense block in DenseNet, (c) mixed-link block with a topology fusion of residual block and dense block, (d) attention-enhanced mixed-link block. Feat1 and Feat2 are features from current and previous convolution layers. Polyline denotes the residual (skip) connection and the curve denotes dense (concatenate) connection.

Second, the output of one layer or unit should also be sliced into two equal parts in the channel dimension. In Figure 2, structure (c) and structure (d) could be calculated as Formulas (2) and (3).

$$\hat{x}_i^1, \hat{x}_i^2 = S(\sigma(W_1(x_i))) \tag{2}$$

$$\hat{x}_i^1, \hat{x}_i^2 = S(A_c(A_k(\sigma(W_1(x_i)) + \sigma(W_2(x_i)))))) \tag{3}$$

where  $A_c$  and  $A_k$  denote the channel and kernel attention,  $\sigma$  means the PReLU [17] activation function.  $W_1$  and  $W_2$  means the weight of convolution layer  $ConvK_1$  with  $3 \times 3$  kernel size and  $ConvK_2$   $5 \times 5$  kernel size in Figure 2d.

$$x_{i+1} = C(x_i^1, C((\hat{x}_i^1 + x_i^2), \hat{x}_i^2)) \tag{4}$$

The last step is shown as Formula (4).  $C(\cdot)$  denotes the concatenate operation,  $x_i^1$  and  $x_i^2$  means the sliced parts of the feature from the previous layer while  $\hat{x}_i^1$  and  $\hat{x}_i^2$  means the parts from the current layer. The addition between  $\hat{x}_i^1$  and  $x_i^2$  makes the topology residual and the concatenation among

$\hat{x}_i^1 + x_i^2$ ,  $\hat{x}_i^2$  and  $x_i^1$  makes the topology dense. The above formulas enable the network to become a partial residual network [18] and partial dense network [19].

There is also a local feature fusion in a mixed-link block, which filters the information from proceeding state and used a  $1 \times 1$  convolution to fuse the information from the current block. The above operations consist of the basic mixed-link unit (MLU) in our proposed network and each AE-MLB contains 4 MLUs in our designed topology. In the final step of AE-MLBs we add the following two parts as the final output for block-level feature selection and fusion. The process could be seen as Formula (5):

$$F_{i+1} = W_F(x_i) + A_c(F_i) \tag{5}$$

where  $F_{i-1}$  denotes the feature from the proceeding attention-enhanced mixed-link block and  $A_c$  is a channel attention module for block-feature filtering.  $F_{i+1}$  is the output feature of the current attention-enhanced mixed-link block.  $x_i$  is the feature from the last mixed-link units in the current block,  $W_F$  means the weight of a  $1 \times 1$  convolution for block-feature fusion. With the help of mixed-link mechanism, the network could synchronously do both residual and dense connections, which decreased half of the growth parameters and improved the network performance.

### 3.3. Triple Attention

This section shows three different attention modules that we proposed, with Section 3.3.1 for channel attention, Section 3.3.2 for kernel attention, and Section 3.3.3 for spatial attention.

#### 3.3.1. Channel Attention

Channel attention could help the network gain the ability of modeling and selecting the information among channels. As shown in Figure 3a, the channel attention module consists of a global average pooling layer [20] which squeezes the features spatially to grab the global information among channels. Then two  $1 \times 1$  convolutions named ConvD and ConvU generate a bottleneck. Finally, a Sigmoid layer used to normalize the information and the output is used to reweight the original output to generate a self-learned channel-wise attention. The process could be calculated as the following formulas:

$$S_q(x_c) = \frac{1}{HW} \sum_i^H \sum_j^W x_c(i, j) \tag{6}$$

where  $S_q(\cdot)$  represents the spatial squeeze with global average pooling. H and W denote the height and width of the feature map.  $x_c$  means channel c of the input feature map x.

$$A_c(x) = \sigma^s(W_u \sigma^p(W_d S(x))) * x \tag{7}$$

where  $A_c(\cdot)$  denotes the channel attention,  $\sigma^s$  means the sigmoid activation function and  $\sigma^p$  means the PReLU activation function,  $W_u$  and  $W_d$  means the two convolution layers with  $1 \times 1$  kernel size.

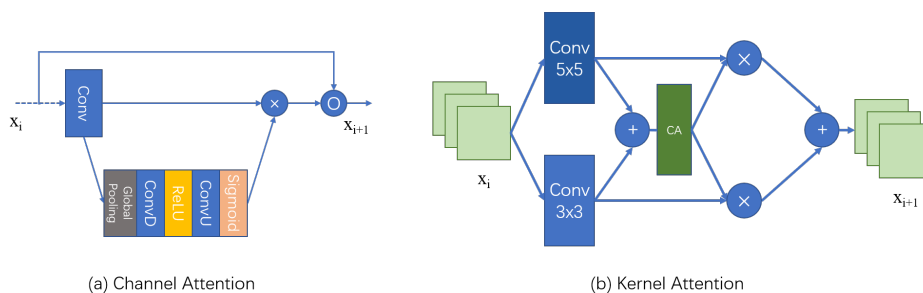


Figure 3. Channel attention and kernel attention, ‘O’ denotes add or concatenate operation.

### 3.3.2. Kernel Attention

Different sizes of convolution kernels can provide different receptive fields, extracting different features [21]. However, many previous methods only used  $3 \times 3$  convolution, which cannot fully extract the information from the LR space. Therefore, in order to improve the network capability, we use  $3 \times 3$  and  $5 \times 5$  convolution kernels to extract different information pieces and  $1 \times 1$  convolution for extraction and transition. In addition, we use kernel attention for feature recalibration on the channels output from layers with different kernels and learn the fusion for features from different receptive fields. The structure of kernel attention could be seen as Figure 3b and can be derived from formulas (8) and (9).

$$W_k = A_c(\sigma W_1(x_i) + \sigma W_2(x_i)) \tag{8}$$

$$A_k(x_i) = W_k * (\sigma W_1(x_i) + \sigma W_2(x_i)) \tag{9}$$

### 3.3.3. Spatial Attention

To further enhance the details in reconstructed images, a spatial attention module for fusing global and local information is added in the reconstruction network, as shown in Figure 4. The reconstruction network also adopts the strategy of global residual learning. The global residual features are increased to twice the original. Half of the channels are weighted by global information, and the other half retains the local information. Then, the two are summed and averaged as the fusion of global and local information. The network generates the final high-resolution image from all the blocks with spatial attention according to the following equations. Figure 5 gives a visualization on how spatial attention affect the global residual, which directly related to the high frequency details in the images.

$$R_1, R_2 = S(M(x)) \tag{10}$$

$$I_{HR} = \sum_b^B w^b \frac{1}{2} (R_1 + R_2 G(M(x))) + U(I_{LR}) \tag{11}$$

where B denotes the AE-MLBs,  $w^b$  denotes the weight for the images.  $R_1$  and  $R_2$  mean the features, G means the channel squeeze convolution, M means the channel multiplier convolution. U denotes the subpixel shuffler [16] for upscaling which increases the width and height of  $I_{LR}$  to the desired size.

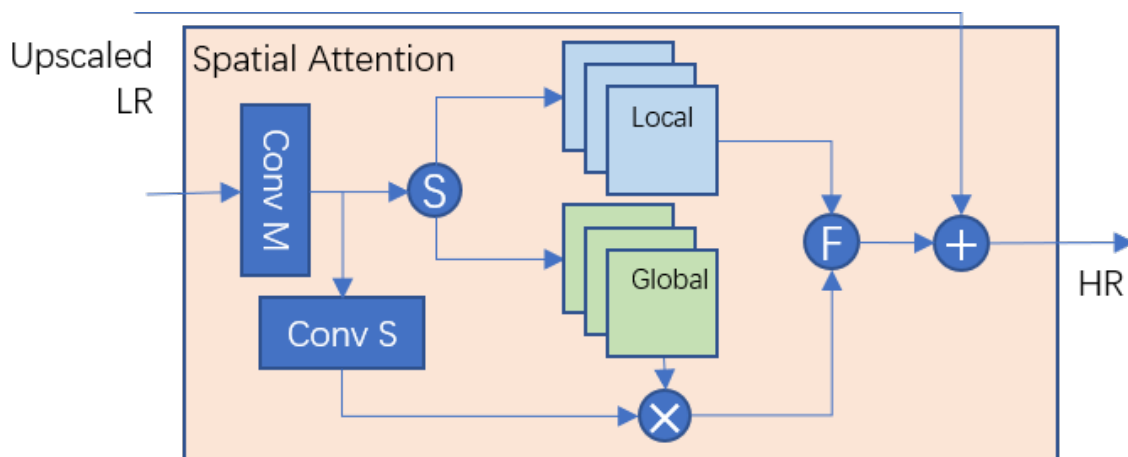
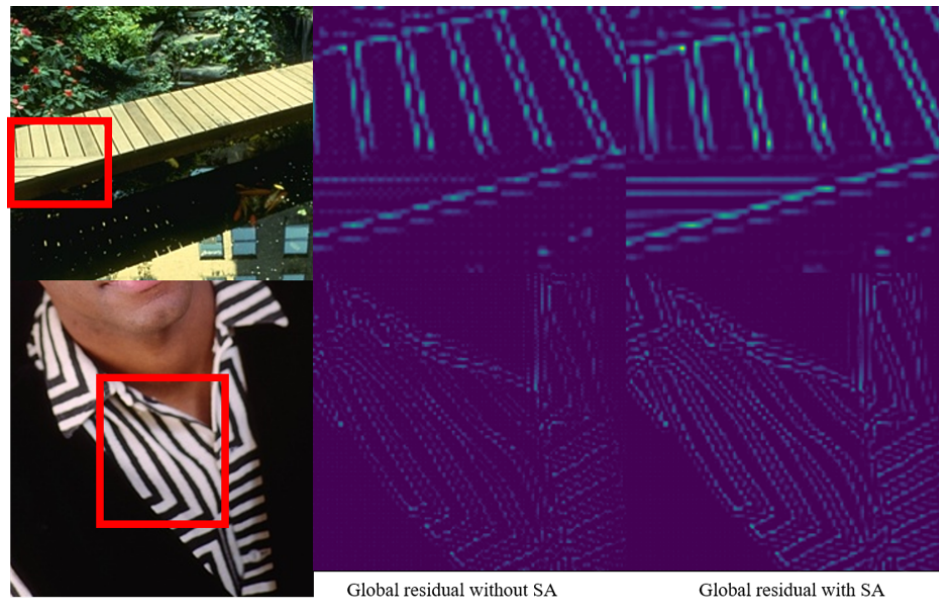


Figure 4. Reconstruction net with spatial attention.



**Figure 5.** Reconstruction net with spatial attention.

### 3.4. Multi-Supervise and Loss Function

Multi-supervise is used during training process. For each AE-MLB, a HR image is generated and the Charbonnier loss [5] is calculated. Finally, the arithmetic mean of the loss from each block is used as overall loss as Formula 12 shows.

$$Loss = \frac{1}{NB} \sum_{i=1}^N \sum_{b=1}^B p(\hat{y}_b^i - y_b^i) \quad (12)$$

where  $N$  denotes the batch size,  $B$  means the number of blocks,  $\hat{y}_b^i$  and  $y_b^i$  means the ground truth and the HR image generated from the network.  $p(x)$  is the Charbonnier penalty function and this can be calculated as  $p(x) = \sqrt{x^2 + \epsilon^2}$ , where  $\epsilon$  is set to 0.001.

## 4. Experiment

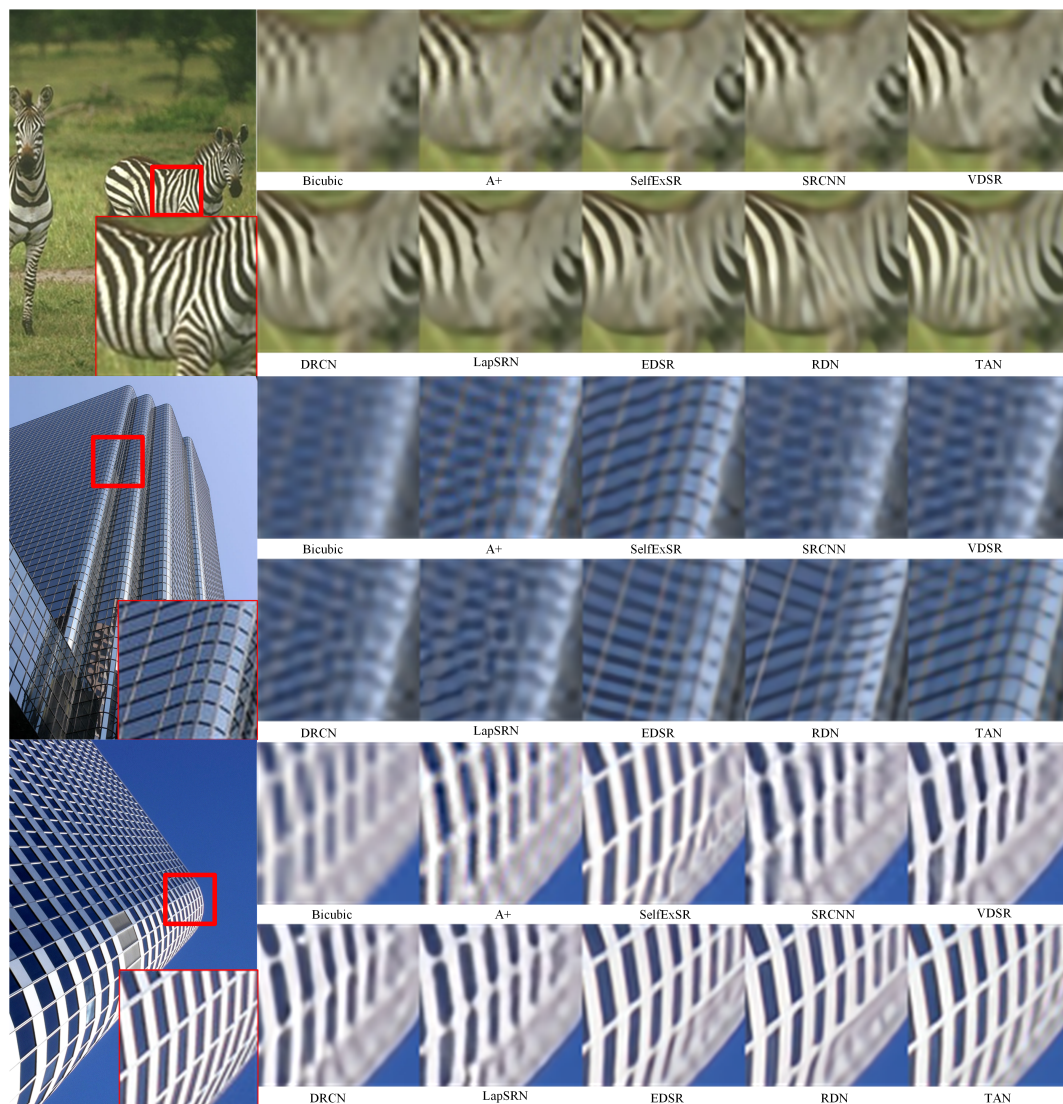
This section mainly shows the design and results of the experiment. Section 4.1 shows the data set used for the experiment and the hyperparameter settings for the training. Section 4.2 compares our proposed TAN and state-of-the-art methods in terms of performance, visual effects, and number of parameters. Section 4.3 is a comparison of self-ensemble performance. Section 4.4 is a study of the performance gains brought about by the connection within the model. Section 4.5 is a study of various attention modules. Section 4.6 is a study of the scale and trainability of the network structure.

### 4.1. Datasets and Training Details

We used DIV2K [22] for training, which contains 800 high-resolution images and we used Set5 [23], Set14 [24], BSD100 [25], Urban100 [12] and Manga109 [26] for testing. All RGB images were converted to YCbCr color space and we selected the Y channel (luminance) for training and testing the super-resolution model. Adam optimizer with the initial learning rate 0.0001 was used and we decrease the learning rate with factor 0.1 every 20 epochs and we totally trained for 80 epochs. We used PyTorch 0.40 as the deep-learning framework to build the network. To further enhance the performance, we add self-ensemble [4] (marked with '+') to gain higher performance.

#### 4.2. Compare with State-of-the-Art Methods

Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) are used as the image-quality metrics. We compare the result generated by our proposed TAN with Bicubic, A+ [2], and deep learning-based methods including SRCNN [3], LapSRN [5], MemNet [8], EDSR [4], and RDN [7]. Table 1 shows the quantitative results on Set5 [23], Set14 [24], BSD100 [25], Urban100 [12], and Manga109 [26] with scale factor  $\times 2$ ,  $\times 4$ , and  $\times 8$ . We also illustrate a visual quality comparison to show the reconstructed details on BSD100 and Urban100 datasets. As shown in Figure 6, the small red rectangle represents where the sub-image was taken from and the zoomed-in ground truth is located on the bottom-right marked with a larger red rectangle. The example images shown in Figure 6 are img\_052 from BSD100 [25], img\_074, and img\_005 from Urban100 [12]. The textures of these images were reconstructed clear with our proposed TAN, while images generated from other methods were blurred, distorted, and have errors with the image details. This illustrates that our proposed method could generate high-resolution images with more accurate details. In addition, as shown in Figure 7, our proposed TAN only contains 7.2 M parameters which are 28% fewer than DBPN(10 M) [27] 67.3% fewer than RDN(22 M) [7], and 83.3% fewer than EDSR(43 M) [4].

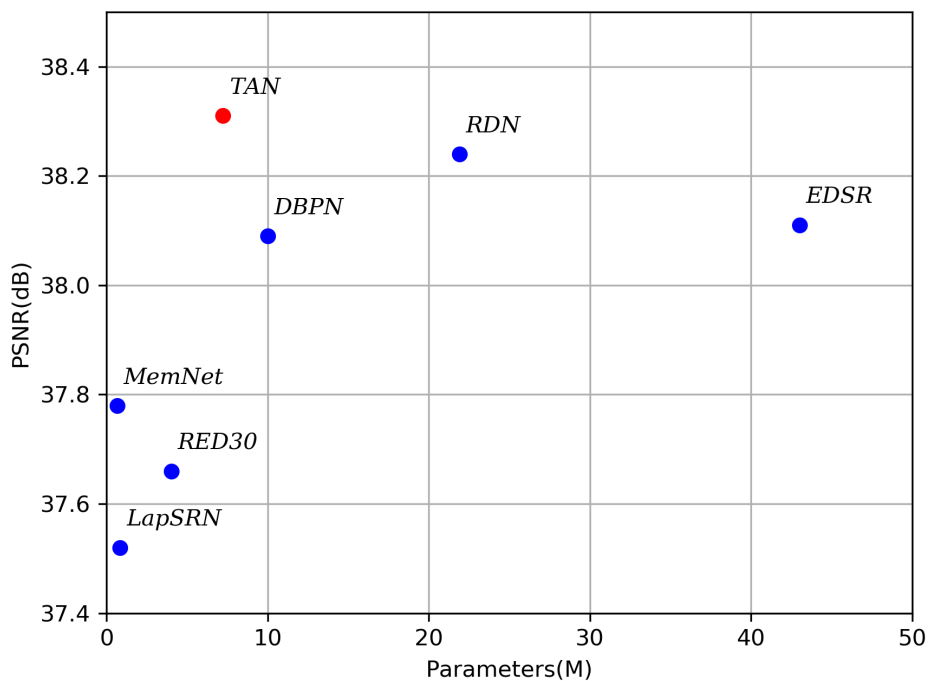


**Figure 6.** Visual comparison for details of reconstructed images. Compared with Bicubic, A+ [2], SelfExSR [12], SRCNN [3], VDSR [13], DRCN [14], LapSRN [5], EDSR [4] and RDN [7], our proposed TAN shows the best visual quality.



**Table 1.** Compare with state-of-the-art methods. The best result is marked with red and the second is marked with blue.

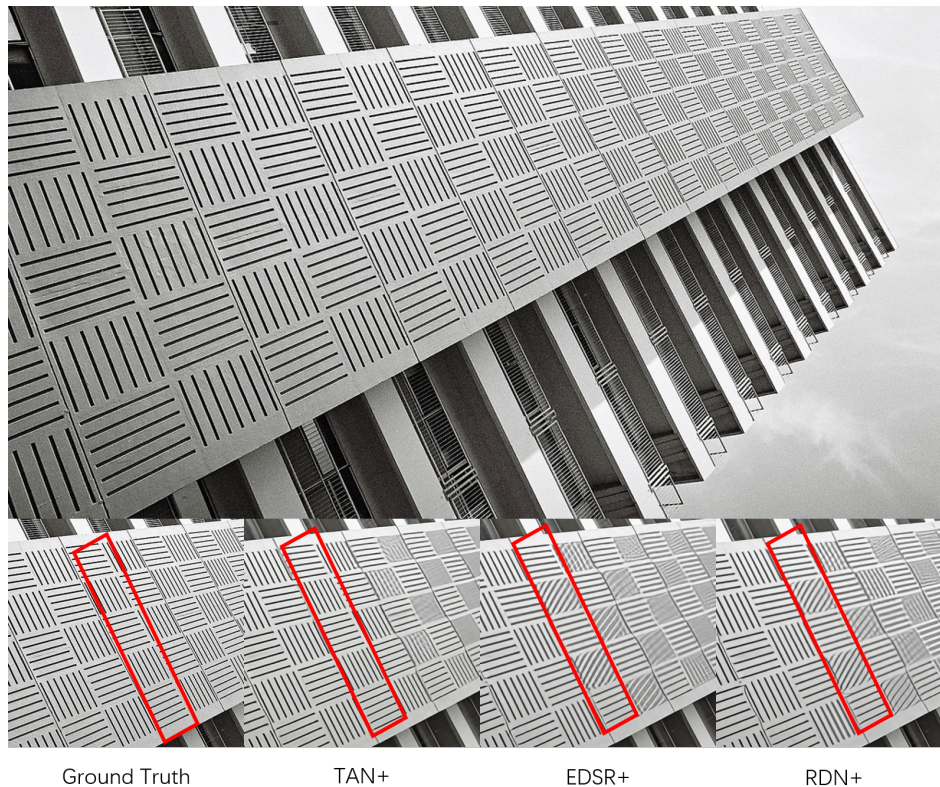
Dataset	Scale	SRCNN	LapSRN	MemNet	EDSR	RDN	TAN (Ours)	TAN + (Ours)
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Set5	×2	36.65/0.955	37.52/0.959	37.78/0.960	38.11/0.960	38.24/0.962	<b>38.30/0.964</b>	<b>38.31/0.965</b>
	×4	30.50/0.863	31.54/0.885	31.74/0.889	<b>32.46/0.897</b>	<b>32.47/0.899</b>	32.34/0.901	32.41/0.896
	×8	25.33/0.690	26.15/0.738	26.16/0.741	<b>26.96/0.776</b>	—/—	26.79/0.772	<b>26.84/0.773</b>
Set14	×2	32.29/0.908	33.08/0.913	33.28/0.914	33.92/0.920	<b>34.01/0.921</b>	33.98/0.924	<b>34.07/0.925</b>
	×4	27.52/0.753	28.19/0.772	28.26/0.772	28.80/0.786	<b>28.81/0.787</b>	28.78/0.795	<b>28.82/0.796</b>
	×8	23.76/0.591	24.35/0.620	24.38/0.620	24.91/0.642	—/—	<b>25.04/0.659</b>	<b>25.08/0.661</b>
BSD 100	×2	31.36/0.887	31.80/0.895	32.08/0.898	32.32/0.901	32.34/0.901	<b>32.37/0.906</b>	<b>32.41/0.907</b>
	×4	26.91/0.712	27.32/0.727	27.40/0.728	<b>27.71/0.742</b>	<b>27.72/0.742</b>	27.62/0.748	27.66/0.741
	×8	24.13/0.566	24.54/0.586	24.58/0.584	<b>24.81/0.598</b>	—/—	24.80/0.614	<b>24.84/0.615</b>
Urban 100	×2	29.23/0.895	30.41/0.910	31.31/0.920	32.93/0.935	<b>33.09/0.937</b>	33.04/0.938	<b>33.13/0.939</b>
	×4	24.53/0.725	25.44/0.756	25.50/0.763	<b>26.64/0.804</b>	26.62/0.803	26.61/0.808	<b>26.73/0.809</b>
	×8	21.29/0.544	21.81/0.581	21.89/0.583	22.51/0.622	—/—	<b>22.74/0.645</b>	<b>22.83/0.648</b>
Manga 109	×2	35.82/0.969	37.27/0.974	37.72/0.974	39.10/0.977	39.18/0.978	<b>39.42/0.981</b>	<b>39.59/0.981</b>
	×4	27.83/0.866	29.09/0.890	29.42/0.894	<b>31.02/0.915</b>	31.00/0.915	30.88/0.918	<b>31.04/0.919</b>
	×8	22.46/0.695	23.39/0.735	23.56/0.738	<b>24.69/0.784</b>	—/—	24.67/0.796	<b>24.80/0.799</b>



**Figure 7.** Performance and parameter comparison of scale × 2 reconstructed for images in Set5. TAN shows the best PSNR with relatively lower model parameters compared with those deep learning-based methods LapSRN [5], MemNet [8], RED [28], DBPN [27], EDSR [4] and RDN [7].

### 4.3. Comparison on Self-Ensemble Results

Self-ensemble is an effectively way to take advantage of the augmented data. Since EDSR [4] and RDN [7] also support self-ensemble, we make a visual comparison among these methods. As shown in Figure 8 the red boxes mark the detail of the ground truth and the results of img\_092 from Urban100 [12] with TAN+, EDSR+ [4], and RDN+ [7]. It can be seen that under data augmentation, since TAN+ can make better use of more information under different receptive fields, better results can be obtained than EDSR+ and RDN+. Both EDSR+ and RDN+ had serious errors in reconstructing high-resolution images, even though they were able to get relatively sharp lines. In addition, the distortion of TAN+ on smaller pixel scales is also better controlled.



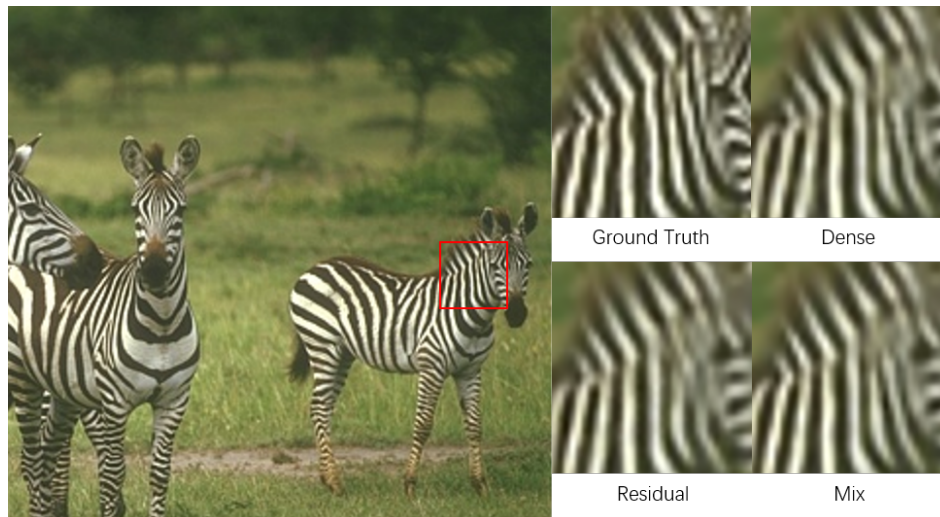
**Figure 8.** Visual comparison of TAN+, EDSR+, and RDN+.

#### 4.4. Study in Model Structure

In this section, we study the effect of network structure and the model performance. For fair comparison, no attention modules or self-ensemble are included in these networks. We replaced the mixed-link connections in the units with dense and residual connections to create a pure dense network and a residual network. We trained the models on DIV2k dataset and we compare the performance and model size of these network structures. The results for scale  $\times 2$  on Set5, Set14, and BSD100 are shown in Table 2. The experiment result shows that mixed-link structure could gain at least 0.09 dB performance with a decrease of 0.17 M model parameters compared with residual structure and gain at least 0.07 dB performance with a decrease of 0.69 M model parameters compared with dense structure. Figure 9 also shows that our proposed mixed-link topology could achieve better performance than residual and dense structure with the least model parameters.

**Table 2.** Model parameter and performance on Set5, Set14 and BSD100 with different network structures.

Structure	MixNet	ResNet	DenseNet
Parameter	1.86 M	2.03 M (+0.17)	2.55 M (+0.69)
Set5	37.81	37.69 (−0.12)	37.74 (−0.07)
Set14	33.40	33.28 (−0.12)	33.32 (−0.08)
BSD100	32.05	31.96 (−0.09)	31.98 (−0.07)



**Figure 9.** Visual comparison of Mix, Dense and Residual network structure.

#### 4.5. Study in Attention Modules

In this section, we study the attention modules we introduced in our network. We define the baseline model as a mixed-link network with 6 blocks. We then conduct an experiment on the effects of these attention modules. We first train three models which only contain one of the three attention modules. Then we train networks with two of the three modules. Finally, we use all these modules for training. There are 7 different combinations and the results for scale factor  $\times 2$  on Set5, Set14, and BSD100 are shown in Table 3. The first line shows the PSNR score and the second line shows the improvement over the baseline. CA denotes the channel attention, KA means the kernel attention, SA means the spatial attention. There is an obvious increase of performance when adding these attention modules to the baseline network among the experiment results. When using one attention module, there will be at most 0.14 dB gain in Set14, 0.42 dB with two module, and 0.58 dB with all three attention modules.

**Table 3.** Performance comparison of networks with different attention modules.

<b>Module</b>	<b>Set5</b> <b>Scale <math>\times 2</math></b>	<b>Set14</b> <b>Scale <math>\times 2</math></b>	<b>BSD100</b> <b>Scale <math>\times 2</math></b>
Baseline	37.81	33.40	32.05
CA	37.89 (+0.08)	33.45 (+0.05)	32.08 (+0.03)
KA	37.95 (+0.14)	33.54 (+0.14)	32.14 (+0.09)
SA	37.91 (+0.10)	33.49 (+0.09)	32.12 (+0.07)
CA + KA	38.19 (+0.38)	33.82 (+0.42)	32.27 (+0.22)
CA + SA	38.15 (+0.34)	33.54 (+0.14)	32.17 (+0.12)
KA + SA	38.09 (+0.28)	33.66 (+0.26)	32.22 (+0.17)
CA + KA + SA	38.30 (+0.49)	33.98 (+0.58)	32.37 (+0.32)

#### 4.6. Study in the Number of AE-MLBs

The number of multi-kernel attention-enhanced mixed-link blocks (AE-MLBs) could directly affect layers and model parameters of the network and determines the super-resolution performance. We study the number of blocks, and the PSNR result for scale  $\times 2$  on Set5, Set14, and BSD100 is shown in Table 4. The experiment result shows the performance is better with more blocks. This illustrates that our proposed TAN allows the training of deeper networks, which enables our model to grab more information from the images and predict more accurate details to reconstruct a high-resolution image with favorable quality.

**Table 4.** Performance comparison of network with different numbers of AE-MLBs.

Number Blocks	Set5 Scale $\times 2$	Set14 Scale $\times 2$	BSD100 Scale $\times 2$
1	37.97	33.53	32.15
2	38.12	33.69	32.26
3	38.23	33.79	32.32
4	38.28	33.87	32.34
5	38.29	33.96	32.34
6	38.30	33.98	32.37

## 5. Discussion

Many recent approaches have proposed very deep networks for handling the single-image super-resolution task. However, these very large models have relatively low practical value since most personal computers and smart terminals do not have enough GPU memory for these models. Even though stacking very deep networks could bring performance gain, from a practical point of view improving model efficiency is a more important research point. Compared with the previous methods, we propose an efficient network topology fusion, which enables the model to absorb the advantages of ResNet [18] and DenseNet [19]. This enhances the flow of gradient and information in the network. Compared to the asynchronous fusion in RDN [7] and MemNet [8], our proposed model combines the structure of ResNet and DenseNet synchronously at each layer. This enable our TAN to reduce the growth rate of parameters by 50%. The previous methods did not make full use of the attention mechanism, and our TAN proposed three kinds of different attention structures, which enable the network to self-learn the feature selection and fusion among channels, convolution kernels, and spatial pixels. Under these attention mechanisms, our model has higher performance with fewer parameters. Although the model we proposed has many improvements over the previous method, there are still some limitations. Compared with the adversarial training methods [29,30], although current TAN can achieve higher peak signal-to-noise ratio and structural similarity, the visual quality is slightly smoothed, and not as sharp as the Generative Adversarial Networks (GAN) [31] results. This is also a trade-off in visual effects and reconstruction accuracy. However, our proposed network framework can be easily ported to the GAN generator. In addition, our proposed TAN is a data-driven supervisory training method. Although acceptable reconstruction effects can still be achieved for a new task, if a better performance is required, new datasets need to be collected for fine-tuning and migration.

## 6. Conclusions

In this work, we propose a novel single-image super-resolution method which used mixed-link connections and three different attentions, including channel attention, fuse kernel attention, and global spatial attention, and we name the model triple-attention mixed-link network (TAN). The mixed-link structure helps the network gain stronger representation ability and proved to be more powerful than residual or dense networks. Moreover, the attention mechanisms give an impressive improvement on performance. The channel attention (CA) could recalibrate the information among channels.

The kernel attention (KA) can fuse the features from convolution layers with  $3 \times 3$  and  $5 \times 5$  kernels, which helps the model learn information from different receptive fields. The spatial attention (SA) mixes the information from local parts and global parts among channels which greatly enhances the reconstruction network. With these attention modules, the proposed attention-enhanced mixed-link block (AE-MLB) is used as the basic part to build the whole network. According to the experiments, the mixed-link structure can decrease up to 27.06% model parameters and increase up to 0.12 dB in PSNR. The three attention modules help the network gain 0.49 dB, 0.58 dB, and 0.32 dB in Set5 [23], Set14 [24] and BSD100 [25]. Thanks to the sophisticated network structure and effective attention mechanisms, our proposed TAN could not only achieve better performance but also contain fewer parameters, which is of great practical value.

**Author Contributions:** Conceptualization, X.C. and X.L.; Methodology, X.C.; Software, X.C.; Validation, X.C.; Formal analysis, X.C.; Investigation, X.C.; Resources, X.C.; Data curation, X.C.; Writing original draft, X.C.; Writing review and editing, X.L.; Visualization, X.C.; Supervision, J.Y.; Project administration, J.Y.; Funding acquisition, J.Y.

**Funding:** This work was supported by the National Science Fund of China under Grant No. U1713208, Program for Changjiang Scholars.

**Acknowledgments:** The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

TAN	Triple-attention mixed-link network
CA	Channel attention
KA	Kernel attention
SA	Spatial attention
AE-MLB	Attention-enhanced mixed-link block

## References

1. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [[CrossRef](#)] [[PubMed](#)]
2. Timofte, R.; De Smet, V.; Van Gool, L. A+: Adjusted anchored neighborhood regression for fast super-resolution. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 111–126.
3. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [[CrossRef](#)] [[PubMed](#)]
4. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017; Volume 1, p. 4.
5. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate superresolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 2, p. 5.
6. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. *arXiv* **2018**, arXiv:1807.02758.
7. Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; Fu, Y. Residual dense network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
8. Tai, Y.; Yang, J.; Liu, X.; Xu, C. Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4539–4547.

9. Tong, T.; Li, G.; Liu, X.; Gao, Q. Image super-resolution using dense skip connections. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4809–4817.
10. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *arXiv* **2017**, arXiv:1709.01507.
11. Wang, W.; Li, X.; Yang, J.; Lu, T. Mixed Link Networks. *arXiv* **2018**, arXiv:1802.01808.
12. Huang, J.B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5197–5206.
13. Kim, J.; Kwon Lee, J.; Mu Lee, K. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
14. Kim, J.; Kwon Lee, J.; Mu Lee, K. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
15. Tai, Y.; Yang, J.; Liu, X. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; Volume 1, p. 5.
16. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
19. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 4700–4708.
20. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
21. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–10 February 2017; Volume 4, p. 12.
22. Agustsson, E.; Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Venice, Italy, 22–29 October 2017; Volume 3, p. 2.
23. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Alberi-Morel, M.L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the British Machine Vision Conference, Surrey, UK, 3–7 September 2012.
24. Zeyde, R.; Elad, M.; Protter, M. On single image scale-up using sparse-representations. In Proceedings of the International Conference on Curves and Surfaces, Avignon, France, 24–30 June 2010; pp. 711–730.
25. 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 Proceedings of the Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 416–423.
26. Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; Aizawa, K. Sketch-based manga retrieval using manga109 dataset. *Multimed. Tools Appl.* **2017**, *76*, 21811–21838. [[CrossRef](#)]
27. Haris, M.; Shakhnarovich, G.; Ukita, N. Deep backprojection networks for super-resolution. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
28. Mao, X.; Shen, C.; Yang, Y.B. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2802–2810.

29. 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 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 4681–4690.
30. Sajjadi, M.S.; Scholkopf, B.; Hirsch, M. Enhancenet: Single image super-resolution through automated texture synthesis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4491–4500.
31. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).