electronics

MDPI

Article

Closed-Loop Residual Attention Network for Single Image Super-Resolution

Meng Zhu [1] and Wenjie Luo [1,2,*]

[1] School of Cyber Security and Computer, Hebei University, Baoding 071002, China; lwj12111@hbu.edu.cn
[2] Hebei Machine Vision Engineering Research Center, Hebei University, Baoding 071002, China
* Correspondence: luowenjie@hbu.edu.cn;

Abstract: Recent research on single image super-resolution (SISR) using convolutional neural networks (CNNs) with the utilization of residual structures and attention mechanisms to utilize image features has demonstrated excellent performance. However, previous SISR techniques mainly integrated extracted image features within a deep or wide network architecture, ignoring the interaction between multiscale features and the diversity of features. At the same time, SISR is also a typical ill-posed problem in that it allows for several predictions for a given LR image. These problems limit the great learning ability of CNNs. To solve these problems, we propose a closed-loop residual attention network (CLRAN) to extract and interact with all the available diversity of features features efficiently and limit the space of possible function solutions. Specifically, we design an enhanced residual attention block (ERA) to extract features, and it dynamically assigns weight to the internal attention branches. The ERA combines multi-scale block (MSB) and enhanced attention mechanism (EAM) base on the residual module. The MSB adaptively detects multiscale image features of different scales by using different 3 × 3 convolution kernels. The EAM combines multi-spectral channel attention (MSCA) and spatial attention (SA). Therefore, the EAM extracts different frequency component information and spatial information to utilize the diversity features. Furthermore, we apply the progressive network architecture and learn an additional map for model monitoring, which forms a closed-loop with the mapping already learned by the LR to HR function. Extensive experiments demonstrate that our CLRAN outperforms the state-of-the-art SISR methods on public datasets for both ×4 and ×8, proving its accuracy and visual perception.

Keywords: image super-resolution; attention mechanism; convolutional neural networks; deep learning

Citation: Zhu, M.; Luo, W. Closed-Loop Residual Attention Network for Single Image Super-Resolution. *Electronics* **2022**, *11*, 1112. https://doi.org/10.3390/electronics11071112

Academic Editor: Gemma Piella

Received: 7 March 2022
Accepted: 30 March 2022
Published: 31 March 2022

1. Introduction

Single image super-resolution (SISR) refers to the technology of reconstructing an underlying high-resolution (HR) image from a single low-resolution (LR) image of the scene. It is known as a typical ill-posed problem, as several HR outputs may correspond to the input LR image. To tackle this inverse problem, numerous algorithms have been proposed. According to the three tier classification of [1], SISR algorithms can be divided into two types: learning methods [2–4] and reconstruction methods [5,6]. The SISR algorithms based on deep learning try to hallucinate the missing details of the super-resolution (SR) images. The methods based on the reconstruction requires the degradation model and explicit prior information to define constraints for the target HR image.

In recent years, numerous studies based on deep learning methods with utilization of residual structures and attention mechanisms have demonstrated outstanding performance in SISR challenges. Dong et al. [7] proposed a super-resolution convolutional neural network (SRCNN) in 2014, which is the first successful effort at introducing CNN with its three convolution layers into SISR. Subsequently, a number of CNN-based SISR models have been proposed to learn the mapping between LR and HR images. Ledig et al. [8] proposed SRResNet, which introducing residual learning to train deep network in SISR.

Kim et al. [9] proposed VDSR, which inspired by the proposal of the residual network [10] and extended the depth of CNN to twenty layers. Lim et al. [11] proposed a 69-layer model named EDSR to improve the high-frequency details, which was inspired by SRResNet and removed redundant modules and expanded model. The success of EDSR also illustrated the efficiency of network deepening. On this foundation, Zhang et al. [12] proposed a 400-layer model named RCAN, which combined the residual structure with the attention mechanism and achieved state-of-the-art performance. The success of RCAN also illustrated the efficiency of deep network combined residual structure and attention block.

However, there are still some limitations for CNN-based SISR models. First, very deep and very wide SISR networks lead to a huge computational cost, which is difficult to apply in real-world applications. Second, most of the deepened networks with stacked convolution operations neglect the full utilization of the feature information in the LR image.

To tackle these problems, Lai et al. [13] designed a pyramid network in a coarse-to-fine fashion to gradually predict sub-band residuals. Li et al. [14] proposed a multi-scale residual network (MSRN), which is not designed to be very deep and very wide, but employs different kernel sizes (3×3 and 5×5) in two-bypass convolution layers to exploit the multi-scale spatial features. Furthermore, MSRN employed the hierarchical feature fusion (HFF) technique to combine the outputs of all residual blocks, utilizing the intermediate features. MSRN obtained equivalent performance with a 7-times smaller model size than EDSR. Subsequently, Muqeet et al. [15] proposed HRAN, which employed dilated convolution layers with different dilation factors to attain a larger receptive field and exploited the channel and spatial dependencies. HRAN proposed the binarized feature fusion (BFF) structure, considering that the HFF is difficult to integrate the features extracted from the CNN smoothly. Behjati et al. [16] combined channel attention mechanisms with residual blocks following two independent but parallel computing paths to attend to relevant features and preserve higher frequency details. Dense connections were employed in prior work [17], which extended each feature to subsequent features through residual connections. Instead of the residual block, Wang et al. [18] proposed a residual in a residual dense block (RRDB), which combines a multi-layer residual network and a dense connection to improve the perceptual quality of the SR image in deep models. Musunuri et al. [19] employed RRDB to replace the residual block in EDSR, yielding better reconstruction results and achieving perceptual quality. The SISR models based on CNN, which combine multiscale feature extraction and attention mechanisms, have achieved excellent performance. However, most networks do not limit the function space when designing the network. The channel attention may discard relevant details contained in other frequency components, which ignores the diversity of features. Moreover, not all attention mechanisms improve network performance, and attention employed across all levels is inefficient, as also described in [16,20].

In this paper, we propose a novel closed-loop residual attention network (CLRAN) that combines residual structures and attention mechanisms to utilize the multiscale features and the diversity of features. The CLRAN also limits the space of possible functions while learning the mapping from LR to HR. We introduce a progressive framework for the reconstruction from LR to HR. The framework is based on the cascade of deep CNNs to gradually reconstruct the HR image and naturally apply deep supervision simultaneously at each level of CLRAN, and it is easily extended to other upscaling factors. Guo et al. [21] proposed that, ideally, the SR image can be downsampled to obtain the same LR image as the input LR image. With this limitation, it is possible to estimate the underlying downsample kernel and reduce the space of potential functions to learn a more effective map. Therefore, we employ an extra map that the SR image uses to reconstruct the input LR image to limit the potential space. The extra mapping utilizes the features from the process of gradually reconstructing the HR image, which plays a supervising function in our model. Specifically, the CLRAN is trained by the Charbornnier penalty loss function [13] to achieve a better visual SR result.

The framework employs basic architecture block (Basic-CLRAN) to gradually obtain the HR image. To achieve multi-scale SR, we only need to modify the number of Basic-CLRANs. In this way, the parameters are shared between different scales, and the network parameters are reduced in our model. Considering the structure of our model is simple, in Basic-CLRAN, we employ HFF technique rather than BFF technique [15] to combine local multi-scale features and global features.

In Basic-CLRAN, an enhanced residual attention block (ERA) is proposed as the basic building block to interact features between each other and extract the diversity features for more powerful feature representations. The ERA contains a multi-scale feature extraction part and an enhanced attention part. In the multi-scale feature extraction part, we propose the multi-scale block (MSB) to obtain the multiscale image features. Considering the stacking multiple dilated convolutions used in [15] to attain a larger receptive field that caused some pixel information not be utilized in the network, we adopt two 3 × 3 convolutions instead of 5 × 5 convolutions in multi-scale residual block (MSRB) [14] of MSRN and introduce two 1 × 1 convolutions, which not only obtain the same effect, but also reduce parameters and indirectly increase the depth of the network. In the enhanced attention part, motivated by the attention mechanism [22–24], we propose an enhanced attention mechanism (EAM) to improve the interactions of the deep multi-scale features and utilize the diversity features. The EAM mainly contains a multi-spectral channel attention (MSCA) block and spatial attention (SA) block. The MSCA block has the ability to capture other frequency component channel-wise information for more powerful feature representations. The SA block further extracts the spatial information and helps the network discriminate "where" to concentrate the features. Considering the drawback described in [16,20], we design a non-attention branch to concentrate on the information that is ignored by the enhanced attention branch. The weights of the two branches are automatically calculated by introducing an attention dropout module (ADM) [20].

In order to verify the effectiveness of the proposed methods, we propose a closed-loop residual attention network (CLRAN), combining the progressive framework with the Basic-CLRAN. In summary, the main contributions of this paper are threefold:

(1) We propose an extra mapping that limits the potential space with the progressive framework in our model, thus forming a closed loop to enhance the performance of the SR model.

(2) We propose an enhanced residual attention block (ERA). This block is based on the residual structure that fuses features at several scales by introducing the multi-scale block (MSB) and utilizes diversity of features by introducing the enhanced attention mechanism (EAM). The MSB and the EAM also can be employed for feature extraction in other computer vision tasks.

(3) We propose a closed-loop residual attention network (CLRAN). The network extracts diversity of features from the input LR image and integrates them with the features throughout the middle process to obtain high accuracy SR images. By employing a progressive framework, the CLRAN gradually obtains the SR result. At the same time, the network is easily extended to certain upscaling factors by modifying the number of Basic-CLRANs in the progressive framework.

The rest of this paper is organized as follows. In Section 2, related work on image-super resolution and attention mechanisms is introduced. In Section 3, the details of the proposed methods are presented. In Section 4, the experimental process, the results, and analysis of the proposed method on different benchmark datasets are presented. Additionally, the ablation study on the proposed network is presented. Model complexity comparisons are also included. In Section 5, the conclusions of the paper are presented.

2. Related Work

In recent years, with the development of neural networks, the image super-resolution algorithms have made remarkable progress. In order to address the ill-posed issue in SISR, researchers continuously widen and deepen the network. However, only broadening and

deepening the network did not achieve the expected significant improvement. Therefore, researchers designed some network structures and learning strategies such as residual networks, recursive networks, dense connections, progressive structure designs, attention mechanisms, and GAN models. In this section, we first describe the related SISR algorithms based on CNNs. We then discuss the attention mechanism.

2.1. CNN-Based Networks

Dong et al. [7] first proposed a shallow three-layer convolutional neural network (SRCNN) for learning a nonlinear mapping function from LR → HR. Subsequently, He et al. [10] proposed a residual learning technique. Ledig et al. [8] proposed SRResNet introducing residual learning to SISR. Kim et al. proposed VDSR [9] with the deep (20 layers) CNN and global residual connection and DRCN [25] with a recursive block to increase the depth without introducing new parameters. Based on DRCN, Tai et al. [26] proposed DRRN combined residual learning and recursive learning. These approaches extract features from an interpolated LR image, which takes much memory and computation time. To address this problem, Dong et al. [27] proposed the FSRCNN, which improves the training speed of SRCNN. Shi et al. [28] proposed ESPCN, designing a sub-pixel convolution layer. Subsequently, numerous networks were proposed to boost the reconstruction performance of HR images. Lim et al. [11] proposed EDSR with an extremely deep and broad network structure that was based on SRResNet and removed unnecessary modules in residual blocks, resulting in considerable promotion. SRDenseNet [17] introduced dense connections [29] in SISR. Tai et al. [30] proposed MemNet, adopting memory blocks consisting of recursive and gate units. RDN [31] employed the dense connections to utilize all the hierarchical features of the convolutional layers. Wang et al. [18] proposed ESRGAN, in which a residual in a residual dense block (RRDB) combined residual blocks, and a dense connection was proposed to improve the perceptual quality of the SR image. Subsequently, Musunuri et al. [19] employed to RRDB replace the residual block in EDSR, yielding better reconstruction results. Recently, some networks have focused on balancing the performance and memory consumption of SISR. Lai et al. [13] proposed LapSRN, which employs the Laplacian pyramid structure to progressively reconstruct the sub-band residuals of the HR image. Ahn et al. [32] proposed CARN, which employs group convolution and learns high-frequency details by locally and globally cascading connections. For multiscale feature extraction techniques, Li et al. [14] proposed MSRN, which employs different kernel size convolution to exploit multiscale spatial features. Muqeet et al. [15] proposed HRAN, which employs different dilation factors dilated convolution layers to exploit the multiscale features.

2.2. Attention-Based Networks

The attention mechanism in deep learning is comparable to the attention mechanism in human vision. It is viewed as a means of biasing the allocation of available computational resources towards the most informative components of a signal [22]. The attention mechanism has recently been widely applied in computer vision tasks such as image classification [33] and image captioning [22]. This mechanism aims to bias the allocation of available resources towards the most informative parts of an input signal [34]. Hu et al. [22] proposed the squeeze-and-excitation (SE) block, which is focused on the channel-to-channel relationship. Woo et al. [24] proposed convolutional block attention module (CBAM), in which channel attention mechanism and spatial attention mechanism are combined. Dai et al. [35] proposed second-order channel attention (SOCA) to adaptively rescale features by considering second-order statistics of features, so the network could focus on more informative features and enhance discriminative learning ability. Qin et al. [23] proposed multi-spectral channel attention by compressing channels in the channel attention mechanism by applying a discrete cosine transform (DCT).

Some researchers have successfully applied attention mechanisms to CNN-based image enhancement methods, especially to SISR. Liu et al. [36] originally proposed employing

non-local operations in a recurrent neural network for image restoration. Zhang et al. [12] considered that if all channels of features were treated equally, the network would lack the ability to discriminate and learn, thus proposed a channel attention (CA) mechanism that employed the residual channel attention network (RCAN), in which the features of each channel were adaptively re-scaled by modeling the interdependence between feature channels. Subsequently, some models [15,34,37] that combined channel attention and spatial attention mechanisms were proposed to learn more discriminative features.

Recently, researchers have started to introduce more sophisticated attention mechanisms to further improve the performance of SISR. Liu et al. [38] proposed enhanced spatial attention (ESA), which reduces the number of channels and adopts a larger stride convolution to shrink spatial dimensions, effectively enlarging the receptive field. Inspired by ESA, Muqeet et al. [39] proposed a cost-efficient attention mechanism (CEA) with dilated convolutions to refine the features. Zhao et al. [40] designed PAN, introducing a pixel-wise channel attention to SISR. Mei et al. [41] designed PANet to capture multi-scale feature. Behjati et al. [16] combined channel attention mechanisms with residual blocks following two independent but parallel computational paths, in which features and attention are processed simultaneously.

3. Proposed Method

3.1. Network Architecture

The complete framework of the proposed network is shown in Figure 1. As we have discussed in Section 1, the CLRAN employs a progressive framework by Basic-CLRAN to reconstruct the HR image from the LR image step by step. For 4× SR task, we employ two Basic-CLRANs, in which we obtain 2× SR for each input image. The Basic-CLRAN in Figure 1 is composed of two parts: feature extraction and reconstruction. We set the original LR image (I_{LR}) as the input of the Basic-CLRAN; the shallow feature E_0 is obtained through initial feature extraction with a 3 × 3 convolutional layer

$$E_0 = H^3_{HF}(I_{LR}) \tag{1}$$

where $H^i_{HF}(\cdot)$ denotes the convolution operation and i denotes the size of convolution kernel.

The extracted feature E_0 is sent to the enhanced residual attention feature extraction part with several ERA modules. We denote the proposed the ERA module as $H_{ERA_i}(\cdot)$, given by

$$E_i = H_{ERA_i}(E_0) \tag{2}$$

where $E_i (i \neq 0)$ is the output feature map of the ith ERA module. After enhanced residual attention feature extraction, we introduce HFF structure expressed as follows:

$$F_{DFS} = \omega * [E_0, E_1, E_2, \ldots, E_n] + b \tag{3}$$

where $[E_0, E_1, E_2, \ldots, E_n]$ denotes the connection operation and denotes the input features of reconstruction part.

The extracted features F_{DFS} from the feature extraction are sent to the reconstruction part; the configuration information for the reconstruction module is shown in Table 1. We employ a PixelShuffle [28] layer upsampled to the same dimensions as HR. We use $I_{HR'}$ to denote the final output from the reconstruction module. Therefore, the final output SR image I_{SR} from Basic-CLRAN is expressed as follows:

$$I_{SR} = H_{UP}(I_{LR}) + I_{HR'} \tag{4}$$

where $H_{UP}(\cdot)$ and I_{SR} denote an upsampled module that contains a pixelshuffle layer.

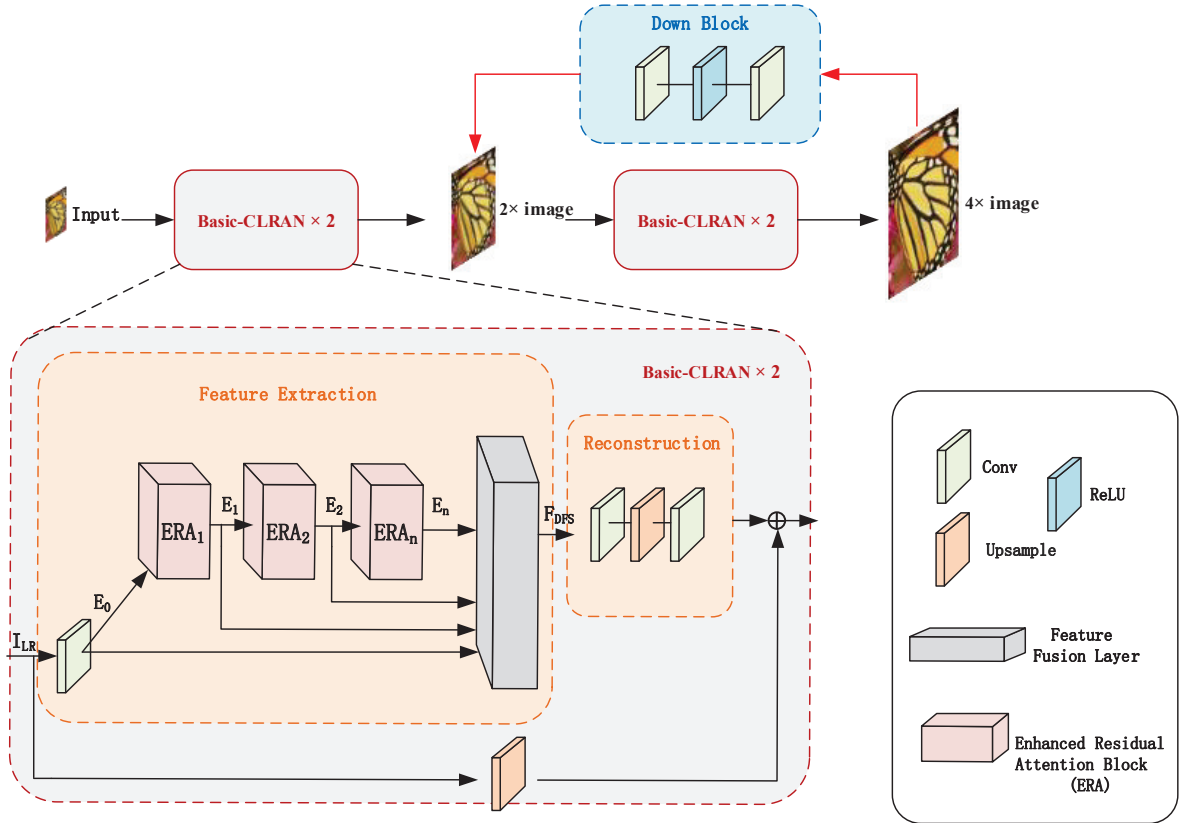


Figure 1. The complete architecture of the closed-loop residual attention network (CLRAN) for 4× SR. The CLRAN contains a primal network (marked with black lines) and a dual regression network (marked with red lines).

Table 1. Detailed configuration information for the reconstruction structure.

Layer	Input Channel	Output Channel	Kernel Size
Input conv	64	64 × 2 × 2	3 × 3
PixelShuffle (×2)	64 × 2 × 2	64	/
Input conv	64	1	3 × 3

In CLRAN, we incorporate progressive architecture into our network. Therefore, for different upscaling factors, we only need to change the number of Basic-CLRANs. The details of our network for different SR tasks are shown in Table 2.

Table 2. The design details for different upscaling factors in our network.

Upscaling Factor	Number of Basic-CLRANs	Upscaling Factor in PixelShuffle	Number of ERAs
×4	2	×2	2
×8	3	×2	2

Loss Function: Different from most networks that have used L1 loss function, we choose the Charbornnier penalty function [13] to train our model. Our ultimate goal is to learn an end-to-end mapping function f from LR $\rightarrow$ HR. However, the space of the possible mapping functions is extremely large, making the function training difficult. Guo et al. [21] provided the derivation of the generalization error bound for the dual regression scheme to prove that introducing dual regression mapping (DRM) to limit the space of the possible mapping functions is effective. Inspired by Guo et al. [21], we learn

the primary mapping P for HR reconstruction and the dual regression mapping D for LR reconstruction simultaneously. Given a training dataset $\{I^i_{LR}, I^i_{HR}\}^N_{i=1}$, we address the following problem in our network:

$$\sum_{i=1}^{N} L_P(P(I^i_{LR}), I^i_{HR}) + \lambda L_D(D(P(I^i_{LR}), I^i_{LR})) \tag{5}$$

where L_P and L_D denote the loss function for the primal mapping and DRM tasks, respectively. The weight of the DRM loss is controlled by λ. Guo et al. [21] discussed the sensitivity of λ; according to the analysis, we set $\lambda = 0.1$ during our training.

In CLRAN, we input an LR image, and then the SR image is progressively predicted at $\log_2 S$ levels, where S is the scale factor. The expression $I_{HR'}$ denotes the output SR image at level s. We denote the desired output SR image at level s by y_s. The overall loss function is defined as:

$$L_{tP} = \sum_{S}^{\log_2 S} L_P(y_s, I^S_{HR}) \tag{6}$$

$$L_{tD} = \sum_{S}^{\log_2 S-1} L_D(D(y_{s+1}), y_s) \tag{7}$$

$$L_T = L_{tP} + L_{tD} \tag{8}$$

where L_{tP} and L_{tD} denote the total loss for the primal mapping and DRM tasks in our network, respectively, and L_T represents the overall loss of our work.

3.2. Enhanced Residual Attention Block (ERA)

The enhanced residual attention block (ERA) of the proposed network, shown in Figure 2, is composed of two parts: the multi-scale part and the enhanced attention part. The multi-scale part contains the MSB, and the enhanced attention part consists of the enhanced attention branch and the non-attention branch.

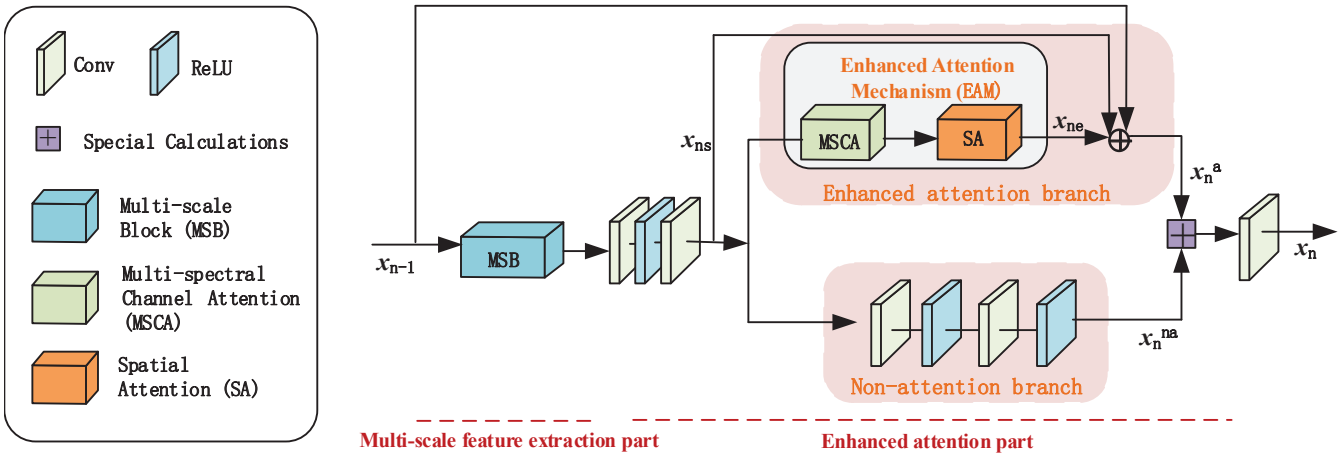


Figure 2. The structure of the enhanced residual attention block (ERA). The purple box denotes special calculations where each added component is multiplied by an automatically generated trainable scalar parameter by the ADM.

Inspired by [16,20], we design the non-attention branch to learn the information that is ignored by the enhanced attention branch. The two branches enable CNNs to make the best use of existing feature information and fully explore the correlation and dependence between the features.

We also introduce the ADM [20] into ERA to balance the enhanced attention branch and non-attention branch. Formally, we have:

$$x_n = f_{1\times1}(\pi^{na}_n \times x^{na}_n + \pi^a_n \times x^a_n) \tag{9}$$

where x_n^{na} is the output feature of the non-attention branch, and x_n^a is the output feature of the enhanced attention branch; π_{na} and π_a are weights of the non-attention branch and the enhanced attention branch, respectively. The dynamic weights are computed by the ADM block; $f_{1\times1}(\cdot)$ denotes the convolution function of 1 × 1 kernel convolution, and x_n is the output feature of the ERA.

Local Residual Learning Structure: The residual learning and shortcut connections alleviate the difficulty of learning between the LR and HR images. We adopt residual structure in the enhanced attention branch to maximize the utilization of the local residual features and enable the network to be more efficient. The utilization of local residual learning in our network significantly reduces the computational complexity, and the performance of the network is enhanced.

As shown in Figure 2, we use x_{n-1} to describe the input feature maps sent to the ERA, x_{ns} to describe the input feature maps sent to the enhanced attention part, and x_{ne} to describe the output feature maps from the EAB. Formally, we describe the output of the enhanced attention branch x_n^a as

$$x_n^a = x_{n-1} + x_{ns} + x_{ne} \tag{10}$$

where the operation $x_{n-1} + x_{ns} + x_{ne}$ is performed by a shortcut connection and element-wise addition.

3.3. Multi-Scale Block (MSB)

Several studies [14,15] have proposed a block to extract the multiscale spatial features. Although the dilated convolution used in [15] achieves much larger receptive fields, not all pixels are used for calculation, resulting in the loss of extracted information details. Therefore, we still use the conventional convolution layers to extract features. As shown in Figure 3a, the multi-scale residual block (MSRB) is used in MSRN [14] to extract the multiscale spatial features. Inspired by the successful application of MSRB, we propose the multi-scale block (MSB) to detect image features at different scales. As shown in Figure 3b, we adopt two 3 × 3 convolutions instead of 5 × 5 convolutions and introduce two 1 × 1 convolutions to reduce parameters and accelerate calculation. In addition, we remove the local shortcut connection (LSC) in MSB and directly follow the attention enhanced attention part to extract diversity of features. In this way, redundancy is reduced in feature utilization and the cost of computational complexity is reduced. The whole operation is defined as

$$S_1 = \sigma_3^1(\sigma_1^1(E_{n-1})) \tag{11}$$

$$P_1 = \sigma_3^3(\sigma_3^2(\sigma_1^2(E_{n-1}))) \tag{12}$$

$$S_2 = \sigma_3^4(\sigma_1^3([S_1, P_1])) \tag{13}$$

$$P_2 = \sigma_3^6(\sigma_3^5(\sigma_1^4([P_1, S_1]))) \tag{14}$$

$$E_n = \sigma_1^5([S_2, P_2]) \tag{15}$$

where E_{n-1} represents the feature maps sent to the MSB, and E_n represents the output feature maps of MSB; σ_i^j denotes a fusion function that combines the convolution function and the ReLU function, where i denotes the size of the convolution kernel and j denotes the number of σ_i; $[S_1, P_1]$, $[S_2, P_2]$, and $[P_1, S_1]$ denote the concatenation operation.

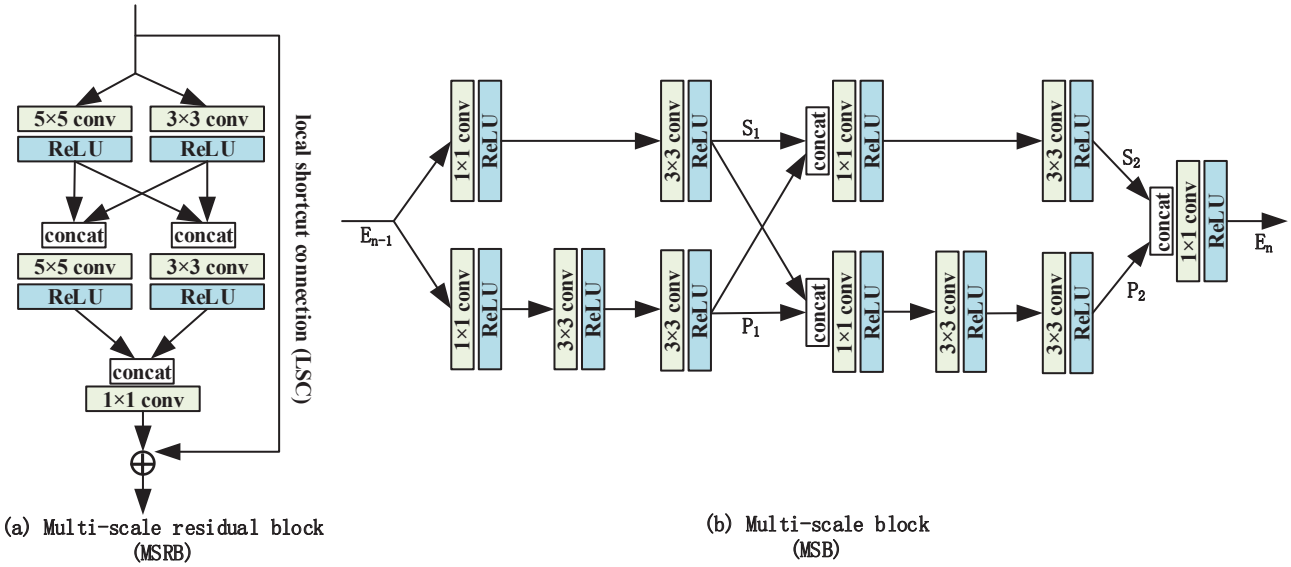


Figure 3. The structure of multi-scale residual block (MSRB) and multi-scale block (MSB), respectively.

3.4. Enhanced Attention Mechanism

Enhanced Attention Mechanism (EAM): In the enhanced attention part, we introduce multi-spectral channel attention (MSCA) and spatial attention (SA) mechanisms into our network. Convolution operations extract meaningful features by combining channel and spatial information together. However, the MSCA block depicted in Figure 4 only utilizes the inter-channel relationship, which neglects spatial information. SA is critical in determining "where" to concentrate. In our work, we propose the EAM that focuses on features in both channel and spatial dimensions. As shown in Figure 5, the EAM infers attention feature maps sequentially along two distinct dimensions, channel and spatial, and attention feature maps multiply with the input feature maps for adaptive feature refinement. Our module contributes significantly to the efficient flow of information within a network. The EAM is expressed as

$$F' = M_f(F) \otimes F \tag{16}$$

$$F'' = M_s(F') \otimes F' \tag{17}$$

where $F \in R^{C \times H \times W}$ denotes input feature maps, M_f denotes the MSCA block, M_s denotes the SA block, $\otimes$ denotes element-wise multiplication, and F'' is the final refined output features.

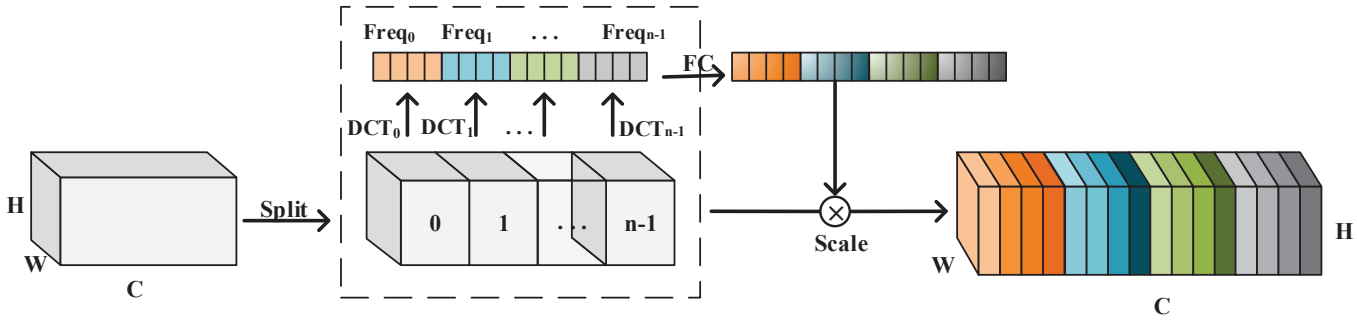


Figure 4. The structure of the multi-spectral channel attention (MSCA) block.

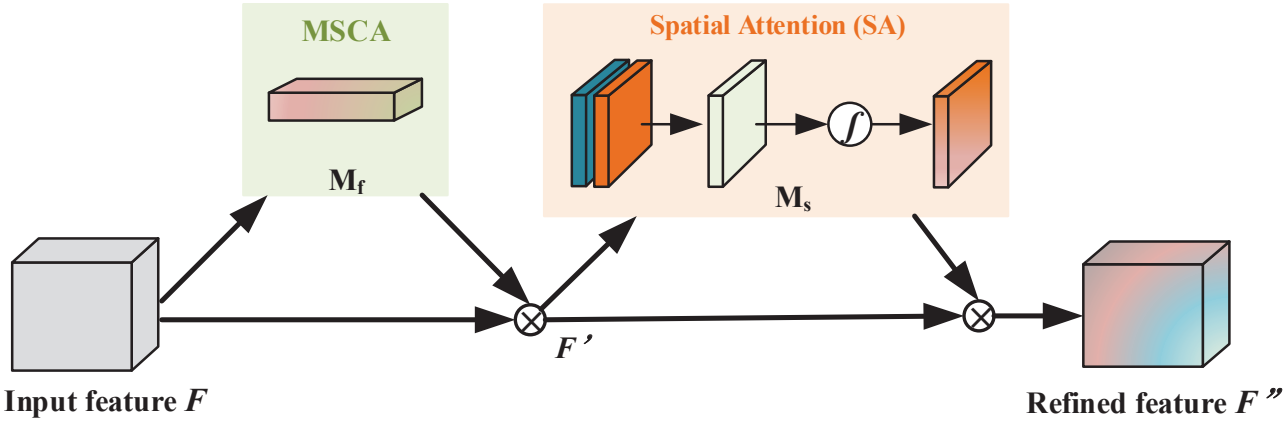


Figure 5. The structure of the enhanced attention mechanism (EAM).

Multi-spectral Channel Attention (MSCA) Module: The channel attention (CA) mechanism uses a scalar to represent and evaluate the importance of each channel and automatically distributes weights to different channels so as to extract critical and important information so that the model makes accurate judgments and will not incur greater overhead in the calculation and storage of the model.

Low-level and mid-level features, in addition to high-level features, are important for reconstructing an SR image. Due to massive information loss, the channel attention mechanism that uses a scalar to represent a channel is difficult. Qin et al. [23] proposed that using global average pooling (GAP) in the channel attention mechanism means only preserving the lowest frequency information and discarding the useful information in representing the channels from other frequencies. Their proposed MSCA mechanism generalizes GAP to more frequency components of 2D discrete cosine transform (DCT).

As shown in Figure 4, the input features $F \in R^{C\times H\times W}$ are split along the channel dimension into several parts. For each part, a corresponding 2D DCT frequency component $Freq^i$ is assigned by employing selection criterion. Finally, the multi-spectral vector $Freq \in R^C$ is obtained by concatenation:

$$Freq = cat([Freq^0, Freq^1, \ldots, Freq^{n-1}]) \tag{18}$$

The feature maps from MSCA module is then expressed as

$$M_f(F) = sigmoid(f_c(Freq)) \tag{19}$$

where *sigmoid* denotes the sigmoid function, and f_c represents fully connected layer.

Spatial Attention (SA) Module: SA tells the network on which informative part it should be focused. As shown in Figure 5, in the SA block, the input features $F^{'} \in R^{C\times H\times W}$ first apply average-pooling and max-pooling operations along the channel axis and then concatenate the outputs to generate an efficient feature map. The combined output is convolved with the convolution function of 7 × 7 kernel convolution, producing our 2D spatial attention map. In short, the spatial attention weight is expressed as follows:

$$M_s(F^{'}) = sigmoid(f^{7\times 7}([AvgPool(F^{'}), MaxPool(F^{'})])) \tag{20}$$

where *sigmoid* denotes the sigmoid function, and $f^{7\times 7}$ represents the convolutional layer with the filter size 7 × 7.

4. Experiments

In this section, we evaluate the performance of our model on several benchmark test datasets. The datasets used for training and testing are introduced first, and next the implementation details are discussed. Following that, we compare our model to several

other methods. Finally, we conducted an ablation study to validate and evaluate the effectiveness of our proposed methods. Specially, we employed the PyTorch framework to all of the implementations.

4.1. Datasets and Metrics

We trained on the DIV2K dataset [42], which contains 800 training images. Bicubic downsampling is employed to obtain the LR images. We evaluated our model using the standard and publicly available benchmark datasets Set5 [43], Set14 [44], B100 [45], Urban100 [46], and Manga109 [47]. Set5 [43], Set14 [44], and B100 [45] contain animals, humans, and natural settings, whereas Urban100 [46] focuses only on urban settings. Urban100 contains rich structure contents. The PSNR and SSIM metrics are employed to evaluate the SR results on the Y channel of the transformed YCbCr color space.

4.2. Implementation Details

In this section, we specify the implementation details of our proposed model. We provided two models, namely a small model CLRAN-S and a large model CLRAN-L for 4× and 8× SR. In our model, we employed two enhanced residual attention blocks (ERA, N = 2) in each Basic-CLRAN, and the output from each ERA was 64 feature maps. We chose the Charbornnier penalty [37] function to train our model.

In each training batch, we randomly extracted 16 LR patches with a size of 128 × 128 and 1500 epochs. We trained our model with ADAM optimizer [48] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate was initialized as 1×10^{-4}. We employed the PyTorch framework to implement our models with GeForce RTX 2080 GPU.

4.3. Results

We compared our model with several state-of-the-art methods in terms of quantitative results and visual results. For quantitative comparison, we compared the PSNR and SSIM values of different methods for 4× and 8× SR. The results of all comparison approaches were derived from their pre-trained models, publicly available code, or original papers.

The results of the PSNR and SSIM values are presented in Table 3. It was found that CLRAN yielded promising performance. CLRAN achieved comparable or superior results compared with all the other methods, including the extremely competitive MSRN. CLRAN-S has the best PSNR on Set5, Set14, B100 and best SSIM on Set5, Set14, B100, and Manga109 for scale ×4. Our CLRAN-L also has excellent SSIM performance on Set5, B100, and Manga109 for scale ×8. Compared with other methods, we found that CLRAN-S and CLRAN-L had achieved almost the best SSIM performance on all benchmark datasets. This confirms that CLRAN is able to gradually aggregate, select, and save relevant details throughout the network. That was mainly because we employed the Charbornnier penalty function [13], thus our model was capable of aggregating rich structured information to generate more representative features. Our model employed YCbCr color space.

For quality comparison, we provided visual comparisons between our method and the considered methods (see Figure 6). We observe that the majority of the approaches were unable to properly recover the tiniest details and so lost the structures, as well as a hazy effect in the majority of the methods. Our model was capable of reconstructing clear and natural images and outperformed other approaches evaluated.

In order to fully utilize the features from the input LR image, our network combined residual structures and attention mechanisms to extract multiscale and diversity of features. Inspired by [14], we proposed the MSB to extract multiscale features. Muqeet et al. [15] was also inspired by [14], which used different dilated convolution layers and channel and spatial attention mechanisms. However, dilated convolution is not friendly to pixel level prediction, and a network based on dilated convolution to design needs some skills, which makes it difficult to migrate directly to other tasks. Moreover, not all attention mechanisms improve network performance, and attention mechanisms may discard relevant details. Behjati et al. [16] designed the network to integrate channel attention mechanisms with

residual blocks via two independent but parallel processing routes. However, in [16], the features of the attention branch and residual branch connect directly and neglect spatial information. The residual blocks are simple and neglect to extract the multiscale features. Wang et al. [18] proposed RRDB combined residual network and dense connection. Musunuri et al. [19] employed RRDB to replace the residual block in EDSR, improving the perceptual quality of the SR image. However, as EDSR is a deep and wide network, training this model will cost more memory, space, and datasets. In short, these models do not limit the space of the possible functions and neglect to extract the diversity of features. Moreover, our model employing the loss function is different from these models.

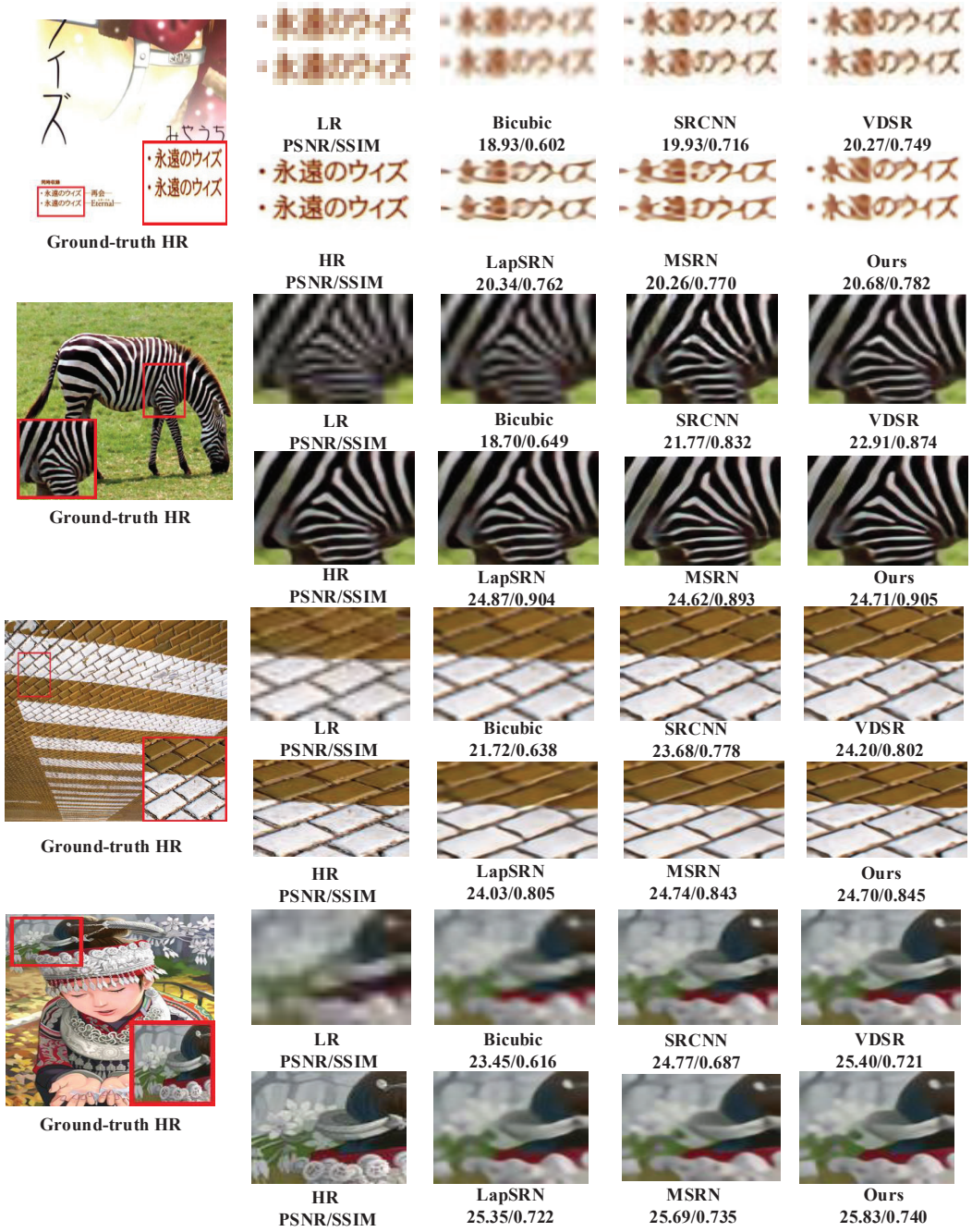


Figure 6. Visual comparison of different methods for 4× image SR.

Table 3. Quantitative results with the BI degradation model for all upscaling factors ×4 and ×8. The **red** number indicates the best result, and the **blue** number indicates the second best result. "-" denotes the results that are not reported.

Algorithms	Scale	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM	Manga109 PSNR/SSIM
Bicubic	4	28.42/0.810	26.10/0.702	25.96/0.667	23.15/0.657	24.92/0.789
SRCNN [7]		30.48/0.863	27.50/0.751	26.90/0.710	24.52/0.722	27.58/0.856
FSRCNN [27]		30.72/0.866	27.61/0.775	26.98/0.715	24.62/0.728	27.90/0.861
VDSR [9]		31.35/0.883	28.02/0.768	27.29/0.726	25.18/0.754	28.83/0.887
SRDenseNet [29]		32.02/0.893	28.50/0.778	27.53/0.733	26.05/0.781	29.49/0.899
DRCN [25]		31.56/0.881	28.15/0.763	27.24/0.715	25.15/0.753	28.98/0.882
LapSRN [13]		31.54/0.881	28.19/0.772	27.32/0.728	25.21/0.756	29.09/0.890
DCSR [49]		31.58/0.887	28.21/0.772	27.32/0.726	**27.24**/**0.831**	-/-
MemNet [30]		31.74/0.889	28.26/0.772	27.40/0.728	25.50/0.763	29.42/0.894
SRMDNF [50]		31.96/0.893	28.35/**0.779**	27.49/**0.734**	25.68/0.773	30.09/0.902
MSRN [14]		32.07/0.890	**28.60**/0.775	27.52/0.727	26.04/**0.790**	30.17/**0.903**
CARN [32]		32.13/0.894	**28.60**/**0.781**	**27.58**//**0.735**	**26.07**/0.784	-/-
IMDN [51]		**32.21**/**0.895**	28.58/**0.781**	27.56/**0.735**	26.04/0.784	**30.45**/**0.908**
CLRAN-S(Ours)		**32.24**/**0.898**	**28.65**/**0.781**	**27.59**/**0.735**	26.05/0.785	**30.37**/**0.908**
Bicubic	8	24.39/0.657	23.19/0.568	23.67/0.547	20.74/0.515	21.47/0.649
SRCNN [7]		25.34/0.647	23.86/0.544	24.14/0.504	21.29/0.513	22.46/0.661
FSRCNN [27]		20.13/0.552	19.75/0.482	24.21/0.568	21.32/0.538	22.39/0.673
SCN [52]		25.59/0.707	24.02/0.603	24.30/0.570	21.22/0.557	22.68/0.696
VDSR [9]		25.73/0.674	23.20/0.511	24.34/0.517	21.48/0.529	22.73/0.669
SRDenseNet [29]		25.99/0.704	24.23/0.581	24.45/0.530	21.67/0.562	23.09/0.712
DRCN [25]		25.93/0.674	24.25/0.551	24.49/0.517	21.71/0.529	23.20/0.669
LapSRN [13]		26.14/0.737	24.35/0.620	24.54/0.585	21.81/0.580	23.39/0.734
MemNet [30]		26.16/0.741	24.38/0.620	24.58/0.584	21.89/0.583	23.56/0.739
MSLapSRN [53]		26.34/**0.756**	24.57/**0.627**	24.65/**0.590**	22.06/0.596	23.90/**0.756**
MSRN [14]		**26.59**/0.725	**24.88**/0.596	**24.70**/0.541	**22.37**/**0.598**	**24.28**/**0.752**
CLRAN-L(Ours)		**26.97**/**0.776**	**24.85**/**0.637**	**24.76**/**0.593**	**22.35**/**0.610**	**24.35**/**0.773**

4.4. Discussion

To validate the effectiveness of our work, we conduct a set of experiments to compare the performance of the MSB, DRM, ADM, and attention mechanisms [22–24], and the number of ERAs in SISR tasks. The results are displayed in Tables 4–6. In Table 4, we conduct the ablation study to validate the effectiveness of MSB, DRM, and ADM. All comparative experiments employ attention mechanisms with the MSCA and SA. In Table 5, we conduct the ablation study to validate the effectiveness of different attention mechanisms in the enhanced attention branch of ERA, and all comparative experiments employ ADM.

Effects of MSB: We propose MSB, which is an efficient multiscale feature extraction structure. This module adaptively detects image features at different scales and fully utilizes the potential features of images. To validate the effectiveness of MSB, we visualize the output feature maps of MSB. The result is shown in Figure 7. With the deepening of the number of network layers, the features extracted by the module become more and more abstract, which is not conducive to our observation. Therefore, we visualize the features extracted by the first application of MSB in the network. From Figure 7, we can observe that the output of MSB retains almost all the information of the original image.

When we employed MSB in our network, 32.47 dB PSNR was obtained with 3.33 M parameters; when we employed without MSB, and the performance of our network with 0.88 parameters decreased by 0.32 dB. Although employing our proposed module increases memory consumption, the effect on performance is obvious, so employing this MSB block in our network is necessary.

Effects of ADM and DRM: In order to evaluate the effects of ADM and DRM, we conducted the comparative experiments. As shown in Table 2, the experiments without

ADM and DRM have lower PSNR than the experiments that employed the ADM and DRM. Therefore, our modules are designed reasonably.

Effects of Different Attention Mechanisms: As shown in Figure 8, in the same way as the visualization of the application of MSB, we visualized the MSCA block heatmaps and the CA block heatmaps. As can be seen, for the MSCA employed in our model, the image structure is clear and the high-frequency and low-frequency regions of the feature map are correctly detected, but the CA employed is incapable of precisely locating them.

From Table 5, we also displayed the comparative experiment results to evaluate the performance of different attention mechanisms in our model. As can be seen, the combination of the MSCA and the SA in our model achieved the best performance. Therefore, we apply the MSCA block and the SA block in the enhanced attention branch of ERA. For case 1, our model did not have the attention mechanism, and the performance was much lower than those cases combined with the attention mechanism. Therefore, the attention mechanism applied to our model is necessary.

Effects of Increasing the Number of ERAs: It is well established that increasing the depth of the network may effectively increase network performance. In our work, increasing the number of ERAs is the easiest way to obtain better SR results. In order to verify the influence of the number of ERAs on the network, we conducted a series of experiments. As shown in Table 6, our network performance improved quickly with increasing ERAs.

In order to gain a more intuitive sense of the effect of the number of ERAs on our model, we plotted the changes in the model metrics during the first 50 epochs, with every 5 epochs as a sample. Given the parameter size of the ERA module itself, we increased the number of ERA from 1 to 5. As shown in Figure 9, the improvement of our model was obvious with the growing number of ERAs, although increasing the number of ERAs in our model will lead to a more complex network. Considering balancing network performance and complexity, we employed two ERAs (N = 2) in our network, which resulted in the optimal balance of performance and model parameters.

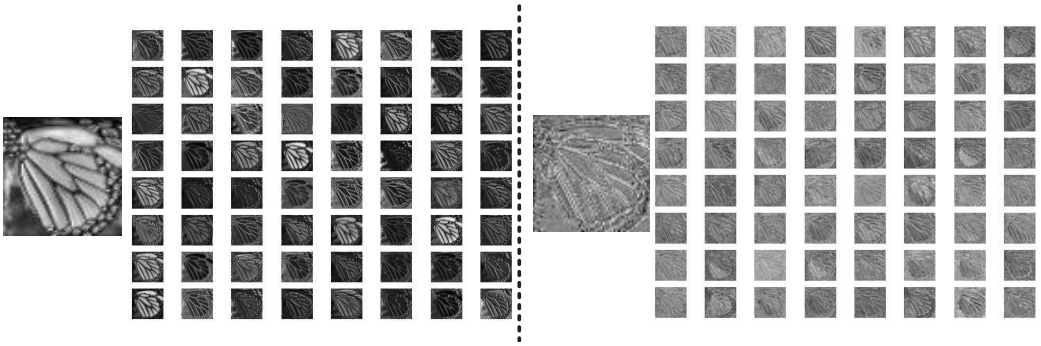


Figure 7. Feature map visualization. On the left: the input feature map of the MSB. On the right: the output feature map of MSB. The 64-channel summation feature map and each channel feature map are shown, respectively.

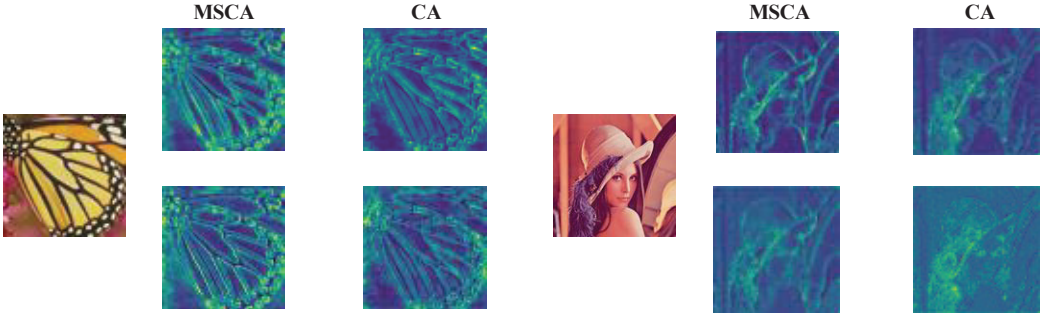


Figure 8. Attention block heatmaps for the MSCA block and the CA block. **The first row:** averaged input feature map of attention layers. **The second row:** averaged output feature map of attention layers.

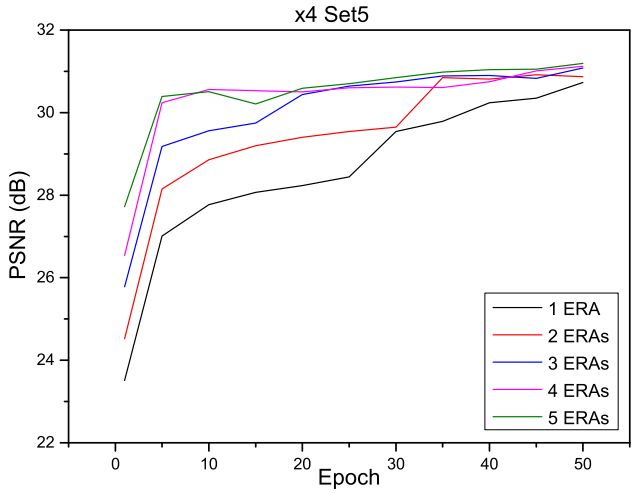


Figure 9. Performance comparison of CLRAN-S with a different number of ERAs.

Table 4. Ablation study: effect of different components of CLRAN-S. Test on Set5 (×4).

Case Index	1	2	3	4
MSB	×	✓	✓	✓
DRM	✓	×	✓	✓
ADM	✓	✓	×	✓
Parameter (M)	0.88	3.32	3.32	3.33
PSNR (dB)	31.92	32.20	32.12	32.24

Table 5. Ablation study: effect of different attention mechanisms of CLRAN-S. Test on Set5 (×4).

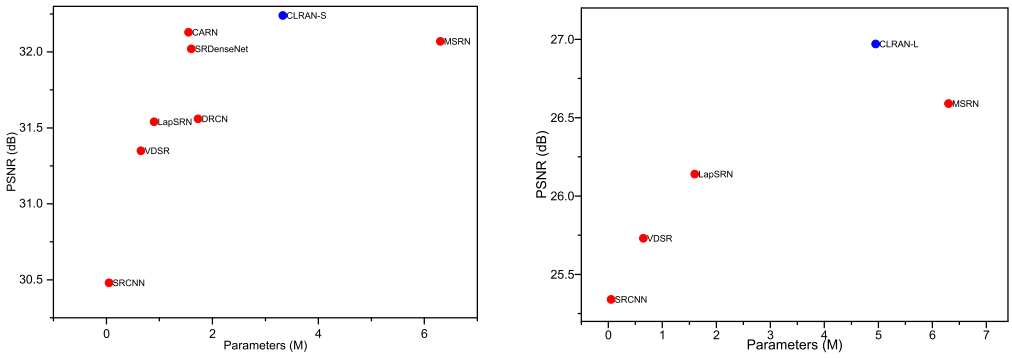
Case Index	1	2	3	4	5
SA	×	✓	×	×	×
CA+SA	×	×	×	✓	×
MSCA	×	×	✓	×	×
MSCA+SA	×	×	×	×	✓
Parameter (M)	3.15	3.32	3.32	3.32	3.33
PSNR (dB)	32.04	32.14	32.15	32.17	32.24

Table 6. Effect of the number of ERAs on the performance of CLRAN-S (testing on Set5) for 4× SR.

N	1	2	3	4	5
PSNR	32.04	32.24	32.26	32.30	32.34

4.5. Model Complexity Analysis

As shown in Figure 10, we visualize a cost effectiveness analysis between PSNR and model size. CLRAN-S comparisons were done with seven state-of-the-art methods: SRCNN [7], VDSR [9], LapSRN [13], DRCN [25], SRDenseNet [29], MSRN [14], and CARN [32]. CLRAN-S with approximately 3.33M parameters obtained the best performance, which verifies the effectiveness of our model. CLRAN-L comparisons were made with four state-of-the-art methods: SRCNN [7], VDSR [9], LapSRN [13], and MSRN [14]. CLRAN-L with approximately 4.89M parameters obtains best performance, which verifies the effectiveness of our model. In comparison to these methods, CLRAN-S and CLRAN-L achieve higher PSNR with a slightly larger model, demonstrating that the trade-off between performance and model complexity is reasonable.



(a) Results on Set5 (4×) (b) Results on Set5 (8×)

Figure 10. PSNR vs. parameters on Set5.

5. Conclusions

In this paper, we proposed a new closed-loop residual attention network (CLRAN) for single image super-resolution. Specifically, the basic architecture block of closed-loop residual attention network (Basic-CLRAN) allowed CLRAN to fully utilize both local and hierarchical diversity of features and easily migrated to achieve other upscaling factor SR tasks. Additionally, the enhanced residual attention block (ERA) extracted the multiscale and diversity image features. The multi-scale block (MSB) was proposed to fuse features at several scales, and the enhanced attention mechanism (EAM) combined a multi-spectral channel mechanism and a spatial attention mechanism proposed to utilize different frequency components channel features and spatial information. Furthermore, we proposed additional mapping and a progressive framework in our model, restricting the space of possible functions and obtaining the SR result step-by-step, taking into account the ill-posed SR problem and limiting the generation of distinct SR images. Comprehensive experiments and ablation studies on benchmark datasets demonstrate the effectiveness of each proposed module, which suggests our model is reasonable.

Author Contributions: Funding acquisition, W.L.; Resources, W.L.; Supervision, W.L.; Writing—original draft, M.Z.; Writing—review & editing, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Hebei Province (F2019201451).

Data Availability Statement: The datasets used in this paper are public datasets. The DIV2K could be found from https://data.vision.ee.ethz.ch/cvl/DIV2K/ (accessed on 1 March 2022).

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

References

1. Bevilacqua, M. Algorithms for Super-Resolution of Images and Videos Based On Learning Methods. Ph.D. Thesis, Université Rennes 1, Rennes, France, 2014.
2. Gao, X.; Zhang, K.; Tao, D.; Li, X. Image super-resolution with sparse neighbor embedding. *IEEE Trans. Image Process.* **2012**, *21*, 3194–3205. [PubMed]
3. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef] [PubMed]
4. Timofte, R.; De Smet, V.; Van Gool, L. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Computer Vision–ACCV 2014, Proceedings of the 12th Asian Conference on Computer Vision, Singapore, 1–5 November 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 111–126.
5. Dai, S.; Han, M.; Xu, W.; Wu, Y.; Gong, Y. Soft edge smoothness prior for alpha channel super resolution. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
6. Yang, X.; Zhang, Y.; Zhou, D.; Yang, R. An improved iterative back projection algorithm based on ringing artifacts suppression. *Neurocomputing* **2015**, *162*, 171–179. [CrossRef]

7. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Berlin/Heidelberg, Germany, 2014, pp. 184–199.
8. 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, Honolulu, HA, USA, 21–26 July 2017; pp. 4681–4690.
9. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1637–1645.
10. 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, 26 June–1 July 2016; pp. 770–778.
11. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HA, USA, 21–26 July 2017; pp. 136–144.
12. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
13. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 624–632.
14. Li, J.; Fang, F.; Mei, K.; Zhang, G. Multi-scale residual network for image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 517–532.
15. Muqeet, A.; Iqbal, M.T.B.; Bae, S.H. HRAN: Hybrid residual attention network for single image super-resolution. *IEEE Access* **2019**, *7*, 137020–137029. [CrossRef]
16. Behjati, P.; Rodriguez, P.; Mehri, A.; Hupont, I.; Tena, C.F.; Gonzalez, J. Hierarchical Residual Attention Network for Single Image Super-Resolution. *arXiv* **2020**, arXiv:2012.04578.
17. 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. 4799–4807.
18. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops,Munich, Germany, 8–14 September 2018 pp. 0–0.
19. Musunuri, Y.R.; Kwon, O.S. Deep residual dense network for single image super-resolution. *Electronics* **2021**, *10*, 555. [CrossRef]
20. Chen, H.; Gu, J.; Zhang, Z. Attention in Attention Network for Image Super-Resolution. *arXiv* **2021** arXiv:2104.09497.
21. Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; Tan, M. Closed-loop matters: Dual regression networks for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5407–5416.
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
23. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 783–792.
24. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
25. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HA, USA, 21-26 July 2016; pp. 1646–1654.
26. 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, Honolulu, HA, USA, 21–26 July 2017; pp. 3147–3155.
27. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 391–407.
28. 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, Honolulu, HA, USA, 21–26 July 2016; pp. 1874–1883.
29. 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, Honolulu, HA, USA, 21–26 July 2017; pp. 4700–4708.
30. Tai, Y.; Yang, J.; Liu, X.; Xu, C. Memnet: A persistent memory network for image restoration. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4539–4547.
31. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2472–2481.
32. Ahn, N.; Kang, B.; Sohn, K.A. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 252–268.

33. Show, A. Tell: Neural Image Caption Generation with Visual Attention Kelvin Xu. Available online: https://kelvinxu.github.io/projects/capgen.html (accessed on 1 March 2022).
34. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Learning enriched features for real image restoration and enhancement. In *Computer Vision—ECCV 2020, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020* ; Springer: Berlin/Heidelberg, Germany, 2020; pp. 492–511.
35. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
36. Liu, D.; Wen, B.; Fan, Y.; Loy, C.C.; Huang, T.S. Non-local recurrent network for image restoration. *arXiv* **2018**, arXiv:1806.02919.
37. Hu, Y.; Li, J.; Huang, Y.; Gao, X. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3911–3927. [CrossRef]
38. Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; Wu, G. Residual feature aggregation network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2359–2368.
39. Muqeet, A.; Hwang, J.; Yang, S.; Kang, J.; Kim, Y.; Bae, S.H. Multi-attention based ultra lightweight image super-resolution. In *Computer Vision—ECCV 2020 Workshops, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 103–118.
40. Zhao, H.; Kong, X.; He, J.; Qiao, Y.; Dong, C. Efficient image super-resolution using pixel attention. In *Computer Vision—ECCV 2020 Workshops, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 56–72.
41. Mei, Y.; Fan, Y.; Zhang, Y.; Yu, J.; Zhou, Y.; Liu, D.; Fu, Y.; Huang, T.S.; Shi, H. Pyramid attention networks for image restoration. *arXiv* **2020**, arXiv:2004.13824.
42. 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 Workshops, Honolulu, HA, USA, 21–26 July 2016; pp. 126–135.
43. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Alberi-Morel, M.L. *Low-Complexity Single-Image Super-Resolution Based On Nonnegative Neighbor Embedding*; BMVA Press: Swansea, UK, 2012.
44. 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]
45. 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 Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 416–423.
46. 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 (CVPR), Boston, MA, USA, 7–12 June 2015.
47. 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]
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014** arXiv:1412.6980.
49. Zhang, Z.; Wang, X.; Jung, C. DCSR: Dilated convolutions for single image super-resolution. *IEEE Trans. Image Process.* **2018**, *28*, 1625–1635. [CrossRef]
50. Zhang, K.; Zuo, W.; Zhang, L. Learning a single convolutional super-resolution network for multiple degradations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3262–3271.
51. Hui, Z.; Gao, X.; Yang, Y.; Wang, X. Lightweight image super-resolution with information multi-distillation network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2024–2032.
52. Wang, Z.; Liu, D.; Yang, J.; Han, W.; Huang, T. Deep networks for image super-resolution with sparse prior. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 370–378.
53. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2599–2613. [CrossRef] [PubMed]