applied sciences

MDPI

Article

Multi-Scale Cross-Attention Fusion Network Based on Image Super-Resolution

Yimin Ma [1,2,†], Yi Xu [1,*,†], Yunqing Liu [1,2], Fei Yan [1,2], Qiong Zhang [1,2], Qi Li [1,2] and Quanyang Liu [1,2]

[1] The College of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China; mayimincust@163.com (Y.M.); liuyunqing@cust.edu.cn (Y.L.); yanf@cust.edu.cn (F.Y.); zhangqiong@cust.edu.cn (Q.Z.); liqicust@163.com (Q.L.); 2018100409@mails.cust.edu.cn (Q.L.)

[2] Jilin Provincial Science and Technology Innovation Center of Intelligent Perception and Information Processing, Changchun 130000, China

[*] Correspondence: xuyicust@163.com; Tel.: +86-177-6771-1633

[†] These authors contributed equally to this work.

Abstract: In recent years, deep convolutional neural networks with multi-scale features have been widely used in image super-resolution reconstruction (ISR), and the quality of the generated images has been significantly improved compared with traditional methods. However, in current image super-resolution network algorithms, these methods need to be further explored in terms of the effective fusion of multi-scale features and cross-domain application of attention mechanisms. To address these issues, we propose a novel multi-scale cross-attention fusion network (MCFN), which optimizes the feature extraction and fusion process in structural design and modular innovation. In order to make better use of the attention mechanism, we propose a Pyramid Multi-scale Module (PMM) to extract multi-scale information by cascading. This PMM is introduced in MCFN and is mainly constructed by multiple multi-scale cross-attention modules (MTMs). To fuse the feature information of PMMs efficiently in both channel and spatial dimensions, we propose the cross-attention fusion module (CFM). In addition, an improved integrated attention enhancement module (IAEM) is inserted at the network's end to enhance the correlation of high-frequency feature information between layers. Experimental results show that the algorithm significantly improves the reconstructed images' edge information and texture details, and the benchmark dataset's performance evaluation shows comparable performance to current state-of-the-art techniques.

Keywords: image super-resolution; pyramid multi-scale features; inter-attention mechanism

check for updates

Citation: Ma, Y.; Xu, Y.; Liu, Y.; Yan, F.; Zhang, Q.; Li, Q.; Liu, Q. Multi-Scale Cross-Attention Fusion Network Based on Image Super-Resolution. *Appl. Sci.* **2024**, *14*, 2634. https://doi.org/10.3390/app14062634

Academic Editors: Giuliana Ramella and Isabella Torcicollo

Received: 30 January 2024
Revised: 12 March 2024
Accepted: 15 March 2024
Published: 21 March 2024

1. Introduction

Image super-resolution (SR) is a fundamental task in computer vision, the primary goal of which is to reconstruct a low-resolution image (LR) into a high-resolution photo (HR). Image super-resolution (ISR) reconstruction is an ill-posed problem because multiple HR images may degrade into the same LR image, and details may be lost in the degradation process. Image super-resolution has been widely studied and applied to medical images, remote sensing images, video surveillance, and other fields needing high-frequency information. In recent years, as deep learning technology has made significant progress in computer vision, this technology has been applied to more tasks. Compared with the image super-resolution methods based on interpolation [1], reconstruction [2], and learning [3,4], the use of deep learning methods can reconstruct high-frequency information more effectively.

SRCNN [5] first applied a Convolutional Neural Network (CNN) to the field of image super-resolution and solved the problem through the mapping function from LR input to HR output. Since then, deep CNN-based methods have been widely used in ISR. Following SRCNN, methods such as FSRCNN [6], ESPCN [7], VDSR [8], EDSR [9], LapSRN [10], and DRRN [11] provide a wider sensory field by deepening the network structure and introducing a residual learning mechanism to alleviate the gradient vanishing problem that increases with network deepening.

Recently, CNN-based methods, such as MSRN [12], MSFRNE [13], and MSAR [14], have demonstrated the ability to further enhance network performance by making full use of multi-scale extracted feature information to increase image texture details. However, despite advances in these methods, more work still needs to be done on the effective fusion of multi-scale features and deep utilization of attention mechanisms. In particular, how to fully use different multi-scale information and enhance the ability of feature information expression while maintaining network efficiency. Therefore, to solve these problems, this paper proposes a multi-scale cross-attention fusion network (MCFN) for the image super-resolution task. The main contributions of this paper are as follows:

(1) A multi-scale cross-attention fusion network (MCFN) is proposed to achieve total extraction and compelling fusion of feature information at different scales and promote high-quality image reconstruction.

(2) A multi-scale Trans-attention module (MTM) is proposed to efficiently extract and fuse multi-scale feature information. MTM utilizes a pyramid multi-scale module (PMM) to extract feature information of various scales, which is then input into a Cross Attention Fusion module (CFM) in a cross-module manner. This approach incorporates a cross-connect strategy that combines channel and spatial attention mechanisms to fuse the multi-scale feature information effectively and capture the correlation dependence between them.

(3) An improved integrated Attention Enhancement module (IAEM) is proposed to extract more feature information from the middle layer through a dense connection strategy. The module learns the correlation between the middle layers and integrates the feature information of each module effectively.

(4) The objective metrics and subjective vision of public datasets show that our method is competitive compared with existing methods. At the same time, we prove the proposed method's effectiveness through many ablation and experimental studies.

This paper is organized as follows: Section 2 will introduce the relevant studies. Section 3 will elaborate on our proposed method and structure. Section 4 will show the experimental results of the method on a public benchmark dataset. The last section will summarize the main conclusions of the paper.

2. Related Works

2.1. Deep CNN-Based Image Super-Resolution

Methods based on deep learning have recently been widely used in image super-resolution [15] and have achieved significant advantages over traditional methods. Dong et al. proposed SRCNN [5], the first article to apply a convolutional neural network to the field of image super-resolution. They used a three-layer convolutional neural network to establish an end-to-end mapping SR method between LR images and their corresponding HR images. Kim et al. proposed the VDSR [8] algorithm, which used a deep convolutional neural network and added residual learning to improve the SRCNN network. At the same time, the DRCN [16] algorithm was proposed, which is the first method to introduce recursive learning to realize parameter sharing in SR. Although the initial application of the CNN method can improve the performance of traditional methods, it will increase the computational cost and produce artifacts. Therefore, Dong et al. proposed the FS-RCNN [6] approach to improve computational efficiency by introducing deconvolution in up-sampling. The ESPCN [7] algorithm was suggested by Shi et al., which presents a sub-pixel convolutional layer to upsample the final LR features as HR output to improve the computational performance to achieve a complete end-to-end mapping. Due to the effectiveness of the sub-pixel convolutional layer, the EDSR [9] algorithm also directly uses it for upsampling and removes the BN layer at the same time to increase the amount of network calculation, reduce the model parameters, and improve image performance. Lai et al. proposed the LapSRN [10] algorithm to reduce the amount of network calculation by using a cascade structure to gradually enlarge image reconstruction. Tai et al. proposed the MemNet [17] algorithm, which uses dense blocks for deep networks. Jiang et al. proposed the HDRN [18] algorithm, which uses hierarchical thick blocks to reconstruct the image to reduce the amount of calculation brought by the dense residual method. These methods show that deep, residual, and dense connections can improve the network's performance. There are other ways to improve network performance.

2.2. Multi-Scale Feature Extraction Based on Image Super-Resolution

Multi-scale feature extraction is widely used in object detection [19] and semantic segmentation tasks [20]. Multi-scale feature extraction can fully use information features at different depths to improve accuracy. The classical scheme for multi-scale feature extraction is the Inception [21] module proposed by Szegedy et al., which uses multiple convolution kernels of different sizes at the same level to extract features, obtain various receptive fields, and improve image quality. Recently, multi-scale feature extraction has also been introduced into image super-resolution. Li et al. proposed an MSRN algorithm [12] that uses multi-scale feature extraction to extract image features of different scales adaptively. He et al. proposed the MRFN [22] algorithm, which uses a multi-receptive field module to remove parts of various receptive fields and proposed a new training loss to reduce reconstruction error. Feng et al. proposed the MSRFN [13] algorithm, which uses a multi-scale extraction module and adds multiple paths for fusion to improve image reconstruction quality. Although these methods are optimized at the network and training levels to enhance the performance of image reconstruction, there is still room for improvement in the extraction and fusion of feature information at different scales.

2.3. Attention Mechanism Based on Image Super-Resolution

Attention usually means that the human visual system adaptively focuses on salient areas in visual information. Therefore, the attention mechanism can help the network focus on essential details. A non-local neural network for image classification tasks [23] was first proposed by Wang et al. After that, Hu et al. designed a Squeeze and Excitation Network (SENet) [24] to improve image classification performance by introducing a channel attention mechanism. Attention-based networks have also been increasingly applied in image super-resolution (ISR) tasks. Inspired by the SENet network [25], Zhang et al. referred to the channel attention mechanism in SR [26] to improve image quality. The SAN [27] algorithm recently used a second-order channel attention mechanism to refine features adaptively. In the AIDN [28] algorithm, information recognition ability is enhanced using a refined attention mechanism to improve network performance. In the MSAR [14] algorithm, the multi-scale attention residual module of feature refinement is used to refine the edge of parts at each scale to improve performance. Therefore, using a multi-scale attention mechanism for feature correlation learning can achieve a more comprehensive and in-depth improvement in performance. We propose a multi-scale cross-attention fusion network (MCFN) to extract and effectively fuse image feature information fully.

3. Methods

The ISR aims to reconstruct a high-resolution image $I_{HR} \in R^{C\times rH\times rW}$ on top of a low-resolution image $I_{LR} \in R^{C\times H\times W}$. The height and width of the image are denoted as W and H, C is the number of channels in the color space, and r is the scale factor. LR images are usually obtained by down-sampling the HR image.

Firstly, this section shows the overall framework of the multi-scale cross-attention fusion network (MCFN). We will then detail each core component, including the pyramid multi-scale module (PMM) in the multi-scale trans-attention module (MTM), the cross-attention fusion module (CFM), and the optimized, integrated attention enhancement module (IAEM). In addition, we will provide an in-depth analysis and justification of the overall architecture strategy of the network.

3.1. Network Framework

We proposed a multi-scale cross-attention fusion network architecture, as shown in Figure 1, which consists of a shallow feature extraction module (SFM), a deep feature extraction module (DFM), and a feature reconstruction module (FRM).

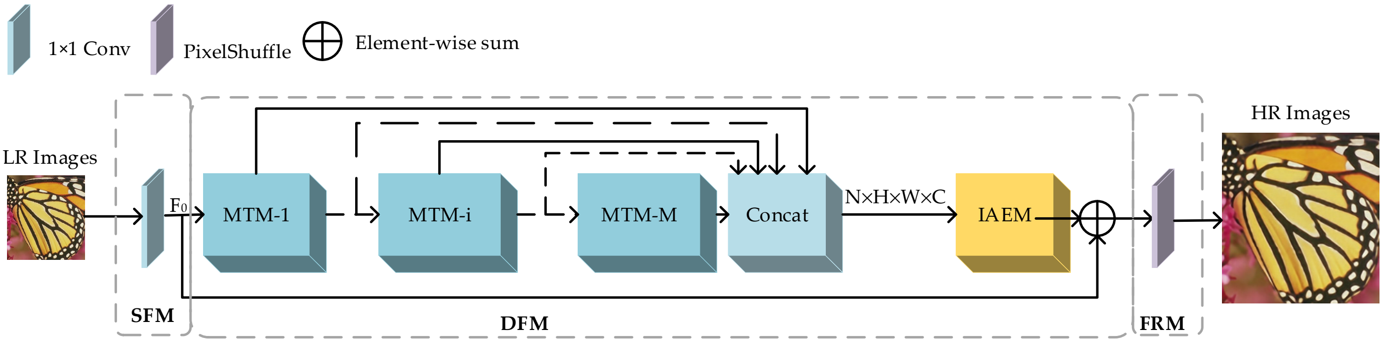


Figure 1. Framework of the multi-scale cross-attention fusion network.

First, the SFM extracts shallow feature information $F_0 \in R^{C\times H\times W}$, including edges and corners, through a single 3×3 convolution function $f_{sf}(\cdot)$, given an input LR image $I_{LR} \in R^{C\times H\times W}$.

$$F_0 = f_{SFM}(I_{LR}), \tag{1}$$

At the same time, F_0 is also the input to the Deep Feature Extraction Module (DFM). Inside the DFM, the F_0 is used as the input of the M multi-scale trans-attention modules and an optimized, Integrated Attention Enhancement Module (IEAM) in order to extract and fuse image feature information. The function of this process is called f_{DFM}. In addition, global skip and dense connections are introduced to make the central part of the network focus on high-frequency information, which can be formally expressed as follows:

$$\begin{aligned} &F_R = f_{DFM}(F_0) = F_0 + f_{IAEM}(Concat[F_1, F_2, \ldots, F_i, \ldots, F_M]), \\ &F_i = f^i_{MTM}(F_{i-1}), \end{aligned} \tag{2}$$

f^i_{MTM} denotes the mapping of the i-th multi-scale trans-attentive module, and $[\cdot]$ denotes concatenation. F_i denotes the output of the i-th MTM, and its input is a concatenation of the outputs of the previous $i-1$ MTM modules. *Concat* denotes the connectivity operator, and f_{IAEM} denotes the mapping in which the module learns feature information from the outputs of the M MTMs, enhancing the feature information for high-frequency information. The IAEM module is designed to enhance the feature layers that are highly informative in their contribution and suppress the feature layers that contain redundant information. Finally, the feature reconstruction module generates a high-resolution image $I_{SR} \in R^{C\times rH\times rW}$ according to the feature information F_R, which is upsampled to the required size by sub-pixel convolution:

$$I_{SR} = f_{PixelShuffle}(F_R), \tag{3}$$

where $f_{PixelShuffle}$ denotes sub-pixel convolution, which aggregates low-resolution feature information to reconstruct the image.

Currently, loss functions such as L_1, L_2, perceptual loss, and adversarial loss are commonly used to train SR models. In this paper, we choose loss L_1 to reduce computational complexity. In a given training set, $\{I^i_{LR}, I^i_{SR}\}^N_{i=1}$, N images, and corresponding images, L_1 loss is defined as:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{n}(f_{MCFN}(I^i_{LR}) - I^i_{HR})_1 = \frac{1}{n}\sum_{i=1}^{n}(I^i_{SR} - I^i_{HR})_1, \tag{4}$$

where f_{MCFN} and Θ denote the proposed functional mapping and its learning parameters, respectively. The configuration of each module will be shown in detail next.

3.2. Multi-Scale Trans-Attention Module

The multi-scale trans-attention module (MTM) is the core of this method, where the extraction and fusion of multi-scale deep feature information are mainly carried out. Figure 2 shows the pyramid multi-scale module (PMM) and the cross-attention fusion module (CFM).

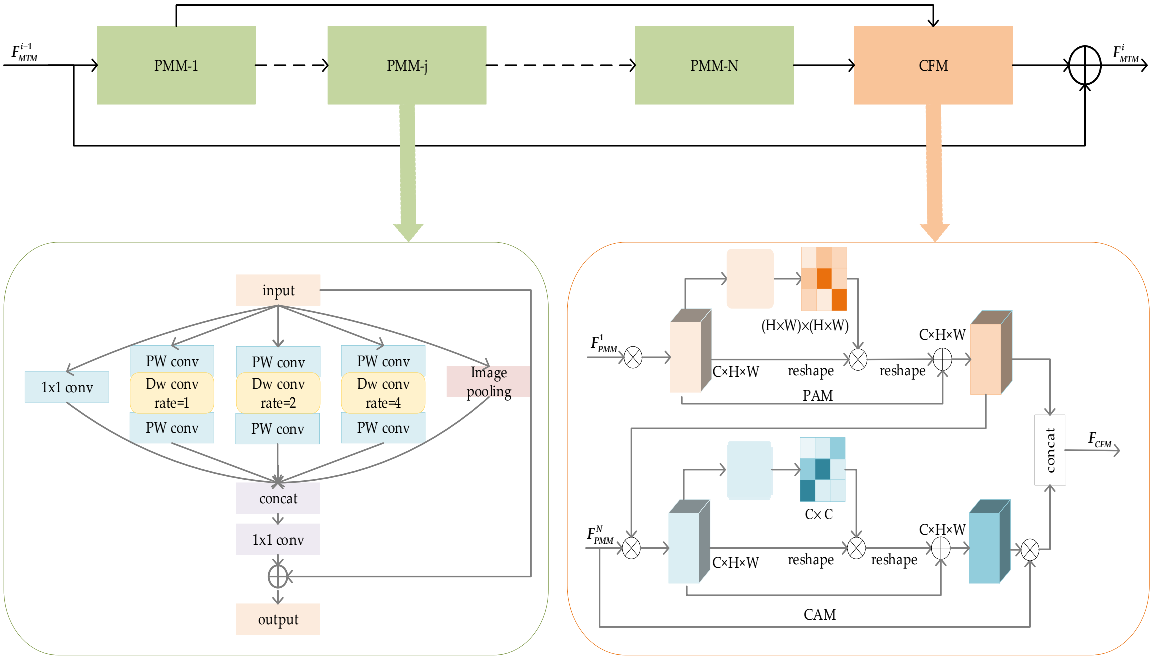


Figure 2. Architecture of the multi-scale trans-attentive module. The core consists of the PMM as a multi-scale pyramid module, which extracts feature information at different scales by incorporating depth-separable convolution to improve efficiency. In addition, CFM is the cross-attention fusion module, which fully fuses feature information by cross-learning the correlation of shallow and deep PMM output feature information.

We constructed a pyramid multi-scale module (PMM) to fully extract feature information and a cross-attention fusion module (CFM) for feature information fusion. We adopt the global residual to minimize loss in the feature information extraction process. The pyramid multi-scale module (PMM) we designed extracts features, such as detail texture and contour area, to extract feature information comprehensively. Then, the heads and tails of multiple modules are fed into the cross-attention fusion module as cross-module outputs for related learning. The specific process is as follows:

$$F^{j}_{MTM} = F^{j-1}_{MTM} + f_{CFM}(F^{1}_{PMM}, F^{N}_{PMM}), \tag{5}$$

where F^{1}_{PMM} and F^{N}_{PMM} denote the outputs of the 1st and Nth pyramid multi-scale modules, and f_{IFM} denotes the mapping of the cross-attention fusion module.

3.2.1. Pyramid Multi-Scale Module

The multi-scale CNN can provide more informative features and help generate high-quality super-resolution images. In order to extract the informative part of all scales more comprehensively, we designed a pyramid multi-scale module for feature lifting, as shown in Figure 2.

In feature extraction, the shallower convolutional layers contain more global information, so extracting more than one detailed texture information feature is crucial. Inspired by DEEP Lab V3 [29] and Mobile Net V2 [30], the ASPP module is improved to extract detailed texture feature information. ASPP uses multiple cavity convolutions with different expansion rates to extract sensory fields of different sizes and then uses standard convolutions to achieve multi-scale feature information fusion. In order to improve the efficiency and performance of ASPP as well as reduce its computational overhead, this paper improves ASPP. It proposes the pyramid multi-scale module to extract the feature information at different scales more effectively. A comparison is shown in Figure 3.

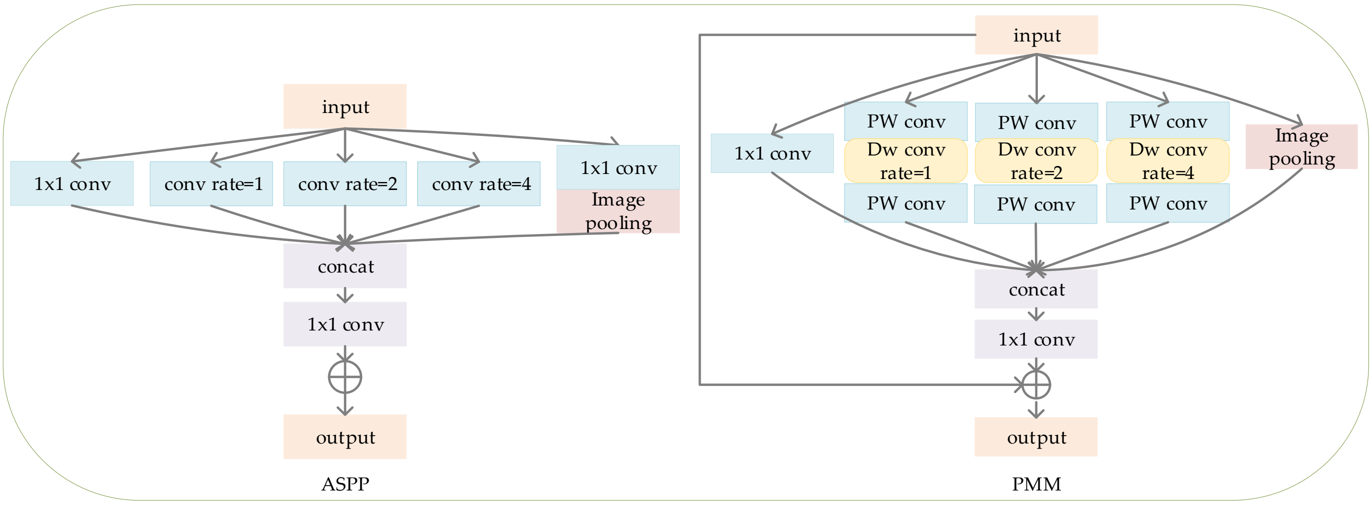


Figure 3. Comparative structural diagram of PMM and ASPP.

We replaced the dilated convolution module with depthwise separable and point convolutions to improve computational efficiency. Experiments [30] show that the number of channels has an essential impact on the overall performance. The comparison shows that the performance of the dilated dimension is better than that of the compressed dimension. This paper uses point convolution to expand and restore the dimension and control the number of channels. Point convolution can effectively promote the information exchange between different channels, and depthwise separable convolution can also focus on extracting multi-scale feature information on each channel independently. The leaky Relu function, which has smaller parameters and better feature extraction ability than Relu6, is selected in this paper. The representation process is as follows:

$$F^{PMM}_{pwconv} = f_{lrelu}(f^{\exp and}_{1\times 1conv}(F^{j-1}_{MTM})), \tag{6}$$

$$F^{PMM}_{dwconv} = f_{lrelu}(f_{dwconv,rate=n}(F^{PMM}_{pwconv})), \tag{7}$$

$$F^{PMM}_{pwconv,rate=n} = f_{lrelu}(f^{regain}_{1\times 1conv}(F^{PMM}_{dwconv})), \tag{8}$$

where F^{i-1}_{MTM} denotes the output of the j-1st MTM, $f^{\exp and}_{1\times 1conv}$ denotes the convolution function of the expanded dimension, F^{PMM}_{pwconv} denotes the output after the expanded dimension, $f_{dwconv,rate=n}$ denotes the Depthwise Convolution function with expansion rate n, F_{dwconv} denotes the output after the expanded rate, $f^{regain}_{1\times 1conv}$ denotes the convolution function of the recovered dimension, f_{lrelu} denotes the Leaky Relu function, and $F_{pwconv,rate=n}$ denotes the output after the recovered dimension. Thus, the process concludes with the introduction of global residual connectivity in this paper in order to increase the stability of the module. Formally, the process is described as follows:

$$\begin{aligned} F^{j}_{PMM} &= F^{j-1}_{PMM} + f^{j}_{PMM}\left(F^{\mathrm{j}-1}_{PMM}\right), \\ f^{j}_{PMM}\left(F^{\mathrm{j}-1}_{PMM}\right) &= Concat\left(F_{conv} + F_{global} + F^{PMM}_{pwconv,rate=1} + F^{PMM}_{pwconv,rate=2} + F^{PMM}_{pwconv,rate=4}\right) \end{aligned} \tag{9}$$

where f^{j}_{PMM} denotes the mapping of the PMM, F_{conv} denotes the feature mapping obtained after convolutional layer processing, and F_{global} denotes the feature information after pooling. Compared with the previous improvement, the module's parameters and computational overhead are reduced, and more detailed texture information features can be extracted.

3.2.2. Cross-Attention Fusion Module

The CNN convolution module is usually used to extract features and perform simple feature fusion. In order to comprehensively fuse information features, this paper proposes a cross-attention fusion module (CFM) to learn the correlation of feature information and fuse them. As shown in Figure 2, the PAM and CAM [31] modules are imported.

In the feature extraction process, the deeper layers can extract more advanced feature information, such as shape feature information, and reduce the deformation during image reconstruction. However, there will be a loss of feature information. Considering such a problem, we designed the cross-module input method, which focuses on shallow and deep feature information to complement the feature information. We designed a cross-attention fusion module (CFM), containing a channel attention module (CAM) [31] and a position attention module (PAM) [31]. Shallow feature information contains more comprehensive and rich spatial location information. After extracting feature information through the location attention module, spatial location features are weighted and selectively aggregated for each location. Deeply extracted feature information often contains rich semantic context, so the information is cross-processed after the output of the location attention module is combined with the deeply extracted feature information. Then, through the channel attention module, the correlation feature information between all channel mappings is learned to achieve the purpose of selectively emphasizing the interdependence. This information is then multiplied with the input features to refine the feature boundaries and finally cross-fertilized with spatial location feature information and semantic feature information. Formally, the process is described as:

$$F_{CFM} = Concat\left(f_{CAM}\left(F^{N}_{PMM} \otimes f_{PAM}\left(F^{1}_{PMM}\right)\right) \otimes F^{N}_{PMM}, f_{PAM}\left(F^{1}_{PMM}\right)\right), \tag{10}$$

where f_{CAM} and f_{PAM} denote spatial attention and position attention function mapping, F_{CFM} denotes cross-attention fusion function mapping, and $\otimes$ denotes element-wise multiplication. The module we designed adopts the strategy of cross-module learning and cross-learning to fuse the correlation of spatial location and semantic context of feature information, making the learning process more comprehensive and detailed.

3.3. Integrating the Attention Enhancement Module

Currently, most SR networks usually use standard convolutional connections and ultimately perform deep feature extraction. Adding an extra module enhances the feature learning capability and thus improves the network's overall performance. Therefore, we designed the Integrated Attention Enhancement Module (IAEM) according to this assumption. We continue with the attention mechanism, inspired by DANet [31], and optimize it. We treat the mapping of each deep feature extraction module as a specific response; different module responses correlate. The interdependence between module mappings is used to enhance the interdependent feature information mapping and the feature representation ability of modules, as shown in Figure 4.

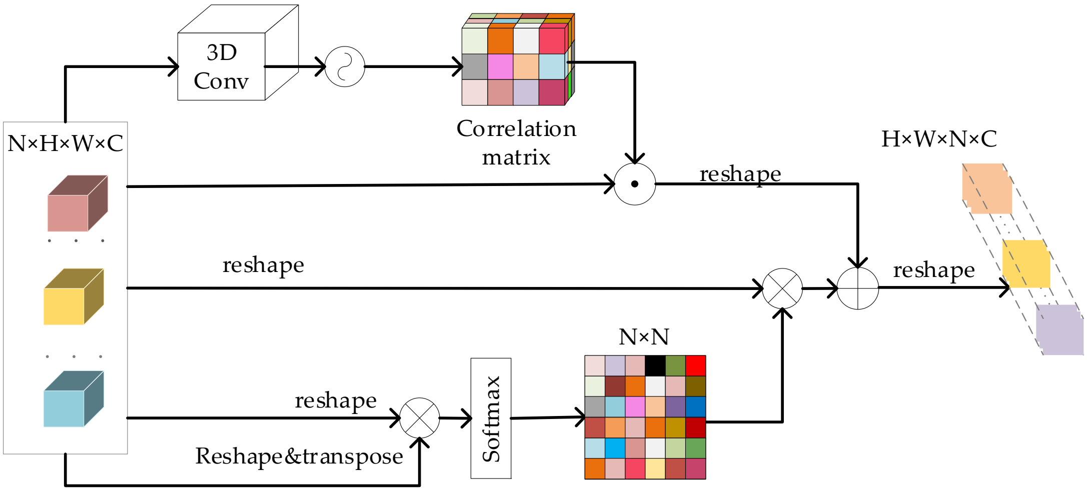


Figure 4. Integration of the Attention Enhancement Module architecture.

Different from the above CAM module, the input is the deep feature group $F_{IFG} \in R^{N \times H \times W \times C}$ output from N multi-scale cross-attention modules, and the dimension is $N \times H \times W \times C$. Through the change of dimension, the weight of feature information is re-learned to strengthen the attention of high-frequency information. Firstly, in this paper, the feature group is convolved by 3D convolution to strengthen the representation of local context features. Then, the sigmoid function is used to extract the feature information of the deep feature group and generate the corresponding attention map.

According to the dimensions of the feature groups, we chose a 3D convolution with kernel size three and step size 1 to generate the attention maps of the three feature groups. Then, in this paper, we multiply it element-by-element with the original input depth-extracted feature layer and multiply it by the scale parameter C to generate the attention map B. Formally, the process is described as:

$$F_{IAEM}^{CHW} = \mu(\sigma(f_{3dconv}(F_{IFG})) \cdot F_{IFG}), \tag{11}$$

where f_{3dconv} represents the 3D convolution function, σ represents the softmax function, $\cdot$ represents element-wise multiplication, and μ learns weights starting from initialization 0.

Secondly, this paper reshapes these deep extracted feature groups IFGs into a two-dimensional matrix of $N \times HWC$. After that, the reshaped feature group is matrix multiplied with its transpose, and then, softmax is applied to obtain the attention map $S \in R^{N \times N}$ that strengthens the correlation between modules. Formally, the process is described as follows:

$$s_{ji} = \frac{\exp(IFGs_i, IFGs_j)}{\sum\limits_{i=1}^{N} \exp(IFGs_i, IFGs_j)}, i, j = 1, 2, \ldots, N, \tag{12}$$

where s_{ji} represents the influence between the i-th module and the j-th module, the attention map of the depth extraction feature layer is obtained by multiplying the reshaped depth extraction feature set with the original feature set matrix and then multiplying the result with the scale parameter λ. Finally, the two attention maps are summed element-wise to obtain the output $F_{IAEM} \in H \times W \times NC$. Formally, the process is described as follows:

$$F_{IAEM_j} = \lambda \sum_{i=1}^{N} (s_{ji} IFGs_i) + F_{CHW_j}^{IAEM}, \tag{13}$$

where λ learns the weights from the initialization of 0, and the final feature of each module represents a weighted sum of all the parts of the module that are related to the original quality and models the long-range semantic dependencies of the entire feature graph. Thus, integrating the attention enhancement modules by learning the interdependencies between the modules is a way to enhance and optimize the overall network's performance effectively.

4. Results

4.1. Datasets and Metrics

In this paper, DIV2K [32] is used as the training set of the model, and the DIV2K dataset contains 800 training images, 100 validation images, and 100 test images. Five standard test sets: Set5 [33], Set14 [34], B100 [35], Urban100 [36], and Manga109 [37] are used. According to the current work, all training and testing are performed based on the luminance channel of the YCbCr color space, and only the Y-channel is processed. This paper uses bicubic down-sampling (BI) to obtain the low-resolution image (LR). The commonly used evaluation metrics PSNR and SSIM are selected for quantitative comparison with other SR methods. Visualization results are also provided for a more intuitive comparison with other methods.

4.2. Implementation Details

In this paper, the LR image is randomly cropped into blocks of size 48 × 48 as training input, and the corresponding patch size of the HR image is 48r × 48r, where r is the scale factor. The minibatch is set to 16, and data enhancement such as horizontal flipping and random rotation of 90° are performed on the training set. This paper sets the number of MTMs M = 5 and the number of PMMs N = 7 for hyper-parameter settings. The model in this paper is trained using the ADAM optimizer [3–6] with $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 10^{-8}$, L1 loss function, the number of channels (number of filters) C = 64, and sets the learning rate to 10^{-4} every 200 backpropagation iterations to reduce the learning rate to 0.5 per 100 iterations. Backpropagation iterations were reduced by half. In increasing the image resolution to 3× and 4× for model training, we adopt the trained 2× image upsampling model as a pre-trained model to further train the ×3 and ×4 models. This approach captures the underlying upsampling mechanism and features by learning with a small (×2) upsampling time. When this pre-trained model is trained on the task of upsampling to higher magnifications (×3 and ×4), it can learn the complex details required for the task more efficiently, accelerating training time and improving model performance at higher resolutions. This paper uses the PyTorch framework and NVIDIA GeForce RTX 3090 GPU for training and testing.

4.3. Comparison with State-of-the-Art Methods

In this section, we compare the performance of the MCFN network in detail with several state-of-the-art network models. The comparison covers the following network models: double cubic interpolation, A+ [38], SRCNN [5], VDSR [8], EDSR-baseline [9], Lap-SRN [10], CARN [39], IDN [40], MSRN [12], MSFRN [13], MIPN [41], MSCIF [42], and MSAR [14]. Through quantitative analysis and subjective visual evaluation methods, we aim to objectively assess the performance metrics of each model in order to comprehensively demonstrate the performance of the MCFN network in various aspects. This study performed detailed comparisons on different scaling factors, i.e., c2, ×3, ×4. The specific comparison results are shown in Table 1.

Table 1. Comparison of PSNR and SSIM values on standard datasets. In this table, the bolded numbers indicate the optimal values in each dataset, while the slanted numbers represent the suboptimal values.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
BICUBIC	×2	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9339
A+	×2	36.60	0.9542	32.42	0.9059	31.24	0.8870	29.25	0.8955	35.37	0.9663
SRCNN [5]	×2	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946	35.60	0.9663
VDSR [8]	×2	37.53	0.9590	33.05	0.9130	31.90	0.8960	30.77	0.9140	37.22	0.9750
Lap-SRN [10]	×2	37.52	0.9591	33.08	0.9130	31.08	0.8950	30.41	0.9101	37.27	0.9740
EDSR-baseline [9]	×2	37.99	0.9604	33.57	0.9175	32.16	0.8994	31.98	0.9272	38.54	0.9769
CARN [39]	×2	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
IDN [40]	×2	37.83	0.9600	33.30	0.9148	32.08	0.8985	31.27	0.9196	38.01	0.9749
MSRN [12]	×2	38.08	0.9605	33.74	0.9170	32.23	0.9002	32.22	*0.9326*	38.82	0.9772
MSCIF [42]	×2	37.89	0.9605	33.41	0.9153	32.15	0.8892	*31.47*	0.9220	--------	---------
MSFRN [13]	×2	38.02	0.9606	33.68	0.9184	32.19	0.8998	32.17	0.9287	38.59	0.9770
MIPN [41]	×2	38.12	*0.9610*	33.73	0.9180	32.25	0.9006	32.42	0.9310	*38.88*	*0.9773*
MSAR [14]	×2	**38.22**	**0.9616**	*33.79*	*0.9189*	*32.27*	**0.9108**	32.46	0.9322	--------	---------
MCFN	×2	*38.17*	*0.9610*	**33.95**	**0.9211**	**32.29**	*0.9011*	**32.80**	**0.9345**	**38.93**	**0.9775**
BICUBIC	×3	30.39	0.8682	27.55	0.7742	27.21	0.7385	24.46	0.7349	26.95	0.8556
A+	×3	32.63	0.9085	29.25	0.8194	28.31	0.7828	26.05	0.8019	29.93	0.9089
SRCNN [5]	×3	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989	30.48	0.9117
VDSR [8]	×3	33.67	0.9210	29.78	0.8320	28.83	0.7990	27.14	0.8290	32.01	0.9340
Lap-SRN [10]	×3	33.82	0.9227	29.87	0.8320	28.82	0.7980	27.07	0.8280	32.21	0.9350
EDSR-baseline [9]	×3	34.37	0.9270	30.28	0.8417	29.09	0.8052	28.15	0.8527	33.45	0.9439
CARN [39]	×3	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IDN [40]	×3	34.11	0.9253	29.99	0.8354	28.95	0.8013	27.42	0.8359	32.71	0.9381
MSRN [12]	×3	34.38	0.9262	30.34	0.8395	29.08	0.8041	28.08	0.8554	33.44	0.9427
MSCIF [42]	×3	34.24	0.9266	30.09	0.8371	29.01	0.8024	27.69	0.8411	-------	-------
MSFRN [13]	×3	34.40	0.9272	30.34	0.8423	29.10	0.8052	28.19	0.8530	33.59	0.9447
MIPN [41]	×3	34.53	0.9280	*30.43*	0.8440	29.15	*0.8060*	28.38	0.8570	33.86	0.9460
MSAR [14]	×3	*34.59*	*0.9285*	**30.53**	*0.8446*	**29.35**	0.8044	*28.44*	*0.8586*	**33.98**	**0.9470**
MCFN	×3	**34.63**	**0.9289**	**30.53**	**0.8458**	*29.29*	**0.8087**	**28.73**	**0.8635**	**33.98**	*0.9469*
BICUBIC	×4	28.43	0.8020	26.10	0.6940	25.96	0.6600	23.150	0.6590	21.460	0.6138
A+	×4	30.33	0.8560	27.44	0.7450	26.83	0.7000	24.340	0.7210	22.390	0.6454
SRCNN [5]	×4	30.48	0.8630	27.49	0.7500	26.90	0.7100	24.520	0.7260	27.580	0.8555
VDSR [8]	×4	31.35	0.8840	28.01	0.7670	27.29	0.7250	25.18/	0.7520	28.830	0.8870
Lap-SRN [10]	×4	31.54	0.8850	28.19	0.7720	27.32	0.7280	25.210	0.7560	29.090	0.8900
EDSR-baseline [9]	×4	32.09	0.8938	28.58	0.7813	27.50	0.7357	26.040	0.7849	30.350	0.9067
CARN [39]	×4	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.070	0.7837	30.470	0.9084
IDN [40]	×4	31.82	0.8900	28.25	0.7730	27.41	0.7300	25.410	0.7630	29.410	0.8942
MSRN [12]	×4	32.26	0.8960	28.63	0.7836	27.61	0.7380	26.220	0.7911	30.570	0.9103
MSCIF [42]	×4	31.91	0.8923	28.35	0.7751	27.46	0.7308	25.640	0.7692	-------	-------
MSFRN [13]	×4	32.16	0.8947	28.62	0.7823	27.57	0.7362	26.090	0.7868	30.470	0.9082
MIPN [41]	×4	32.31	0.8970	28.65	0.7830	27.61	0.7370	26.230	0.7900	*30.670*	*0.9107*
MSAR [14]	×4	*32.29*	**0.8989**	*28.67*	*0.7841*	**27.95**	**0.7410**	*26.250*	*0.7907*	30.660	0.9100
MCFN	×4	**32.43**	*0.8976*	**28.78**	**0.7858**	*27.71*	*0.7405*	**26.637**	**0.8012**	**31.008**	**0.9133**

It can be observed in these results that the MCFN network shows a significant advantage in most of the performance metrics compared to the recently proposed methods. In particular, compared to the more extensive network MSRN proposed by ECCV, the MCFN network shows higher PSNR and SSIM values by 0.21dB and 0.0041, respectively, on the Set14 test set with a scaling factor of 2. On Set5, with a scaling factor of 3, compared to the MIPN, the MCFN also improves its PSNR and SSIM values by 0.1 dB and 0.0009. As the scaling factor increases, the low-resolution image loses more high-frequency information, limiting the high-quality reconstruction of super-resolution images. In the Urban100 dataset, which is rich in detailed information, MCFN outperforms the following highest method, MSAR, by 0.387 dB and 0.0105 in PSNR and SSIM metrics, respectively, when the scaling factor is four. In summary, our network exhibits recognizable performance, which initially proves the validity of the network that we designed.

In order to present a more comprehensive picture of the performance of our model, we selected several representative detail parts from different super-resolution images. We reconstructed the images with ×2, ×3, and ×4 for these detail parts to show and compare these key details more obviously. As shown in Figures 5–8, the selected details were marked with rectangular boxes and enlarged three times to show and contrast these key details more obviously.

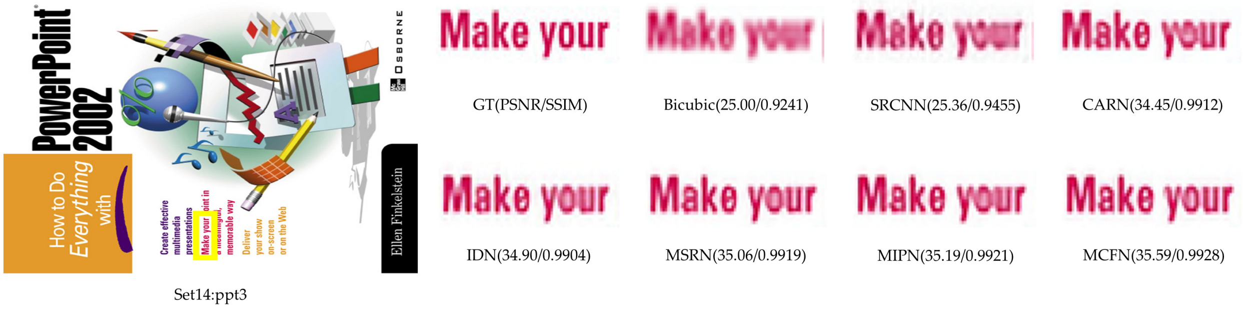


Figure 5. Visual comparison of our method with other methods (×2).

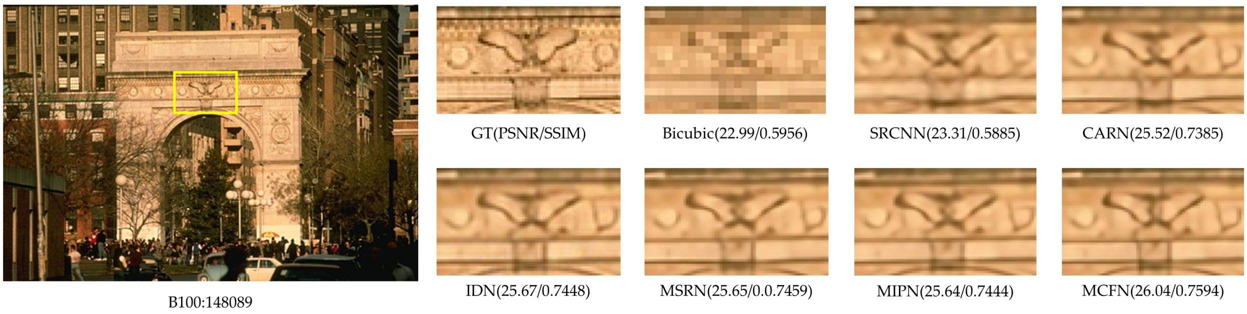


Figure 6. Visual comparison of our method with other methods (×3).

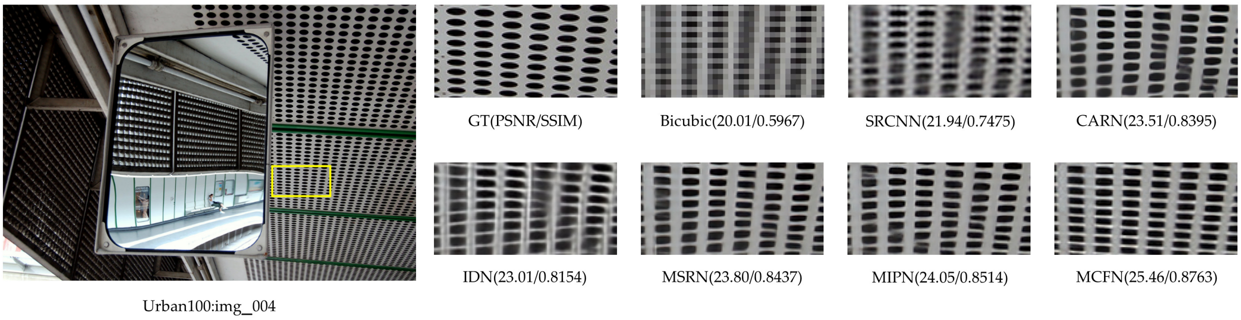
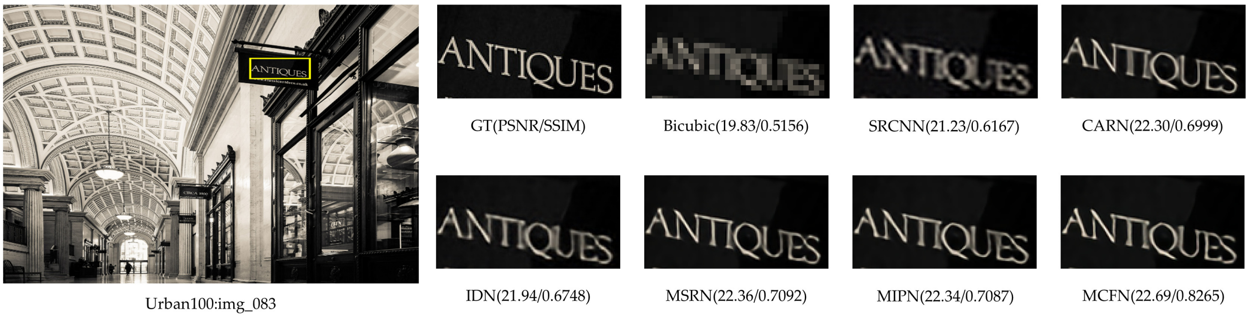


Figure 7. Visual comparison of our method with other methods (×4).



GT(PSNR/SSIM) Bicubic(19.83/0.5156) SRCNN(21.23/0.6167) CARN(22.30/0.6999)

Urban100:img_083 IDN(21.94/0.6748) MSRN(22.36/0.7092) MIPN(22.34/0.7087) MCFN(22.69/0.8265)

Figure 8. Visual comparison of the proposed method with other methods in terms of letters (×4).

In Figure 5, a significant difference in the clarity of the letters reconstructed by the different algorithms can be observed when the magnification factor is two times. For example, the letters reconstructed by SRCNN and CARN could be more precise and quieter. Although IDN, MSRN, and MIPN methods have improved the clarity, some details of the letter shape still need to be recovered. In contrast, the letters reconstructed by MCFN are more transparent and less noisy.

In Figure 6, a building at sunset at 3× magnification shows that MCFN performs better in preserving the edge texture and reducing the artifacts. Figure 7 shows a car roof image at 4× magnification. MCFN demonstrates less distortion and effectively reduces ringing effects, with richer information on the edge contours.

In addition, in Figure 8, the selected sign text in the scene is displayed under a magnification factor of 4, and our method improves the edge clarity while also improving the brightness to obtain a better visual effect. In general, our network performs well in objective indicators and shows significant advantages in subjective visual effects.

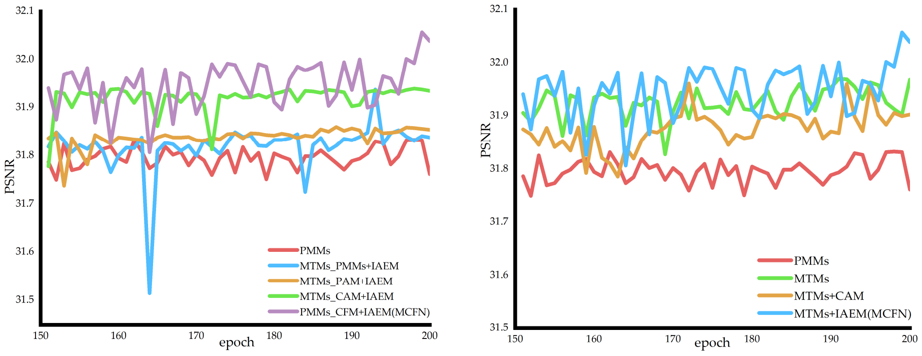
4.4. Ablation Study

4.4.1. Study of Ablation of Network Structures

In this part of the study, we demonstrate the effectiveness of each module in the proposed MCFN and their contribution to the network performance. We design a series of ablation experiments, as shown in Table 2. We evaluate their contribution to network performance by adding or replacing critical modules in the network. Firstly, we construct a base network consisting of a series of PMMs, called the PMMs network. The base network adopts a multi-scale mechanism of depth-separable convolution and pointwise convolution, improving computational efficiency while ensuring adequate feature information extraction at different scales. Then, IAEM was added to our study to evaluate the network performance of PMMs, denoted as MTMs_PMMs + IAEM. Subsequently, CFMs were added to the PMMs to assess the effect of the addition on the network's performance, denoted as MTMs (PMMs + CFM). It is worth noting that we did not perform ablation experiments on the combination of PMMs and CFM alone. Instead, we chose to perform ablation experiments on MTMs (a combination of PMMs and CFMs) together with IAEM, aiming to assess the impact of CFM on performance in the presence of IAEM. Therefore, we used the strategy of replacing CFMs with PAMs and CAMs, denoted as MTMs _PAM + IAEM and MTMs_CAM + IAEM, and similarly, in order to assess the performance of IAEMs, we replaced IAEMs with CAMs in the MCFN structure, denoted as MTMs + CAM. Although this design scheme for ablation experiments may be different from traditional ablation methods, it provides us with an effective way to assess the interactions of the individual modules. In addition, this design approach aligns more with our experimental resource realities, allowing us to perform the most effective performance evaluation under limited conditions. We select most of the modeling methods, PSNR, and SSIM values on Set5, Set14, and B100 test sets for 200 cycles of comparison to ensure the necessity and validity of the experiments. In order to show the experimental results more intuitively, we plotted the experimental data of the last 50 cycles as a line graph-Figure 9.

Table 2. Ablation study of MCFN properties.

Method	Scale	Set5		Set14		B100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
PMMs	×4	31.797	0.889	28.362	0.776	27.424	0.731
MTMs_PMMs + IAEM		31.828	0.889	28.413	0.776	27.444	0.731
MTMs_PAM + IAEM		31.861	0.890	28.391	0.776	27.459	0.732
MTMs_CAM + IAEM		31.943	0.891	28.447	0.777	27.487	0.732
MTMs (PMMs + CFM)		31.995	0.891	28.517	0.780	27.529	0.734
MTMs + CAM		31.997	0.890	28.372	0.776	27.430	0.731
MCFN (MTMs + IAEM)		32.047	0.893	28.520	0.780	27.543	0.735



(**a**) (**b**)

Figure 9. Line plots of the training process: (**a**) plot comparing the results of the fusion module network and (**b**) plot comparing the results of integrating the augmentation module and CAM.

Table 2 and Figure 9a demonstrate a clear trend: adding the fusion and enhancement networks to the base network significantly improves the network's performance metrics, proving the effectiveness of the individual modules and indicating that better results can be obtained. The performance improvement is pronounced in the MTMs_CAM + IAEM network. This network effectively focuses on critical feature information by learning the relevance between different channels, which demonstrates the importance of correlation learning after deep extraction of high-frequency information. In particular, in the MCFN network, we design an innovative cross-attention fusion module. This network not only effectively learns the spatial locations of shallow feature information through the cross-module learning approach but also combines this spatial location feature information with deep feature information through the cross-connection strategy to deeply learn the relevance of the information in the channel. This hierarchical approach improves the comprehensiveness of information utilization. In CFM, by integrating spatial and channel features, we achieve a more comprehensive fusion of information, enabling the network to achieve the best results in several performance metrics.

When analyzing the performance of IAEM, we used CAM as a control group to learn the difference in performance between the two. As shown in Table 2 and Figure 9b, our network performs better in PSNR and SSIM than the control group in the above test set experimental results. The results of the above analyses demonstrate the effectiveness of our module in performing relevant learning. In contrast to accessing channel attention only at the tail, our integrated attention-enhanced network employs a dimensionality transformation technique to fuse feature information at different stages. This strategy enhances the learning of feature information weights and effectively helps the network's performance during the fusion reconstruction process.

4.4.2. Study of Multi-Scale Trans-Module Synthesis

In this part of this study, we analyze the influence of MTM and the number of PMMs in MTM on the network performance and conduct a series of ablation experiments. As shown in Table 3, we set the number of MTMS M to 4, 5, and 6 and evaluate its impact on the number of parameters and network performance in the test set Set 5. The results show that with the increase in M, the PSNR value of the network improves, and the network performance improves, but the growth rate becomes gradually smaller. In addition, we analyze the number N of PMMs, setting them to 6, 7, and 8, respectively, and record the comparative experimental results, as shown in Table 4. The experimental results show that when N increases from 6 to 7, the PSNR value increases by 0.034. However, when N grows to 8, the increase in PSNR value is only 0.01. Therefore, to effectively balance the reconstruction quality and the number of parameters, we set the number of MTM and PMM to 5 and 7.

Table 3. Analysis of the number of MTMs.

Number	Scale	Parameters	Set5	
			PSNR	SSIM
M = 4	×2	1.43	37.108	0.960
M = 5	×2	1.70	38.166	0.961
M = 6	×2	1.98	38.233	0.962

Table 4. Analysis of the number of PMMs in MTM.

Number	Scale	Parameters	Set5	
			PSNR	SSIM
N = 6	×2	1.62	38.132	0.960
N = 7	×2	1.70	38.166	0.961
N = 8	×2	1.79	38.167	0.961

4.4.3. Study of Parameters and FLOPs

As shown in Figure 10, this study compares the number of parameters, the number of floating-point operations (FLOPs), and the average peak Signal-to-Noise Ratio (PSNR) (Avg. PSNR) between MCFN and other advanced methods when 4× magnification (output image resolution is 1280 × 720) is performed on the Set 5 dataset. In order to provide a more intuitive comparison perspective, the relevant data are summarized in Table 5. Compared with other methods, MCFN achieves superior performance with low computational overhead. Although not optimal regarding the number of parameters, MCFN has half the number of parameters compared to MSRN. In summary, MCFN performs well in model efficiency and objective evaluation indicators.

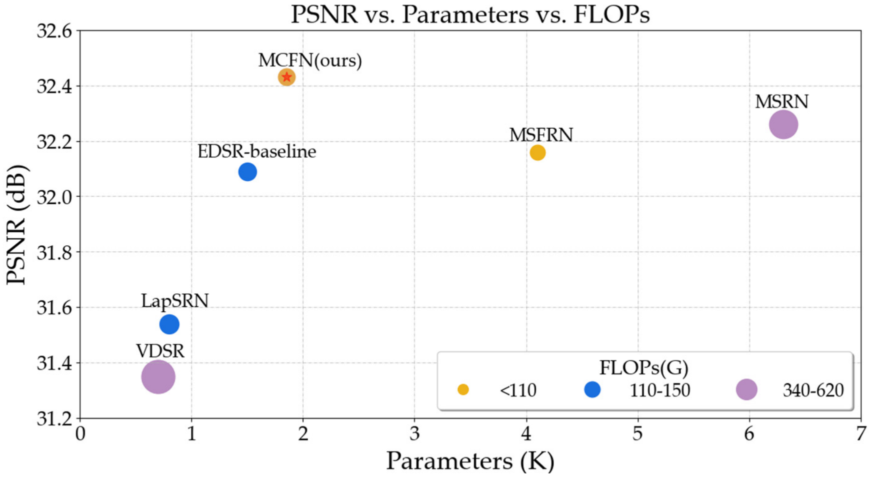


Figure 10. Visualization of PSNR, parameters, and FLOPs. PSNR values were evaluated in Set 5 with scaling factor × 4.

Table 5. Comparison of performance, parameters, and FLOPs with some state-of-the-art ISR methods under a scaling factor of 4 in the Set 5 dataset. Comparison results of the number of parameters, FLOPS, and average PSNR values of the SR method on the Set 5 test set. FLOPs are calculated based on 320 × 180 input features.

	VDSR	Lap-SRN	EDSR-Baseline	MSFRN	MSRN	MCFN
Para (M)	0.665	0.812	1.5	4.1	6.3	1.85
FLOPs (G)	612.6	149.9	114.2	94.99	349.8	110.12
PSNR	31.35	31.54	32.09	32.16	32.26	32.426

5. Conclusions

This paper proposes a multi-scale cross-attention fusion network (MCFN) to improve the image quality of image super-resolution tasks. The network combines the advantages of the multi-scale and attention mechanisms, aiming to extract and fuse the feature information of the image more thoroughly. The multi-scale trans-attention module (MTM) we designed includes the pyramid multi-scale module (PMM) and the cross-attention fusion module (CFM). In the pyramid multi-scale module (PMM), to extract feature information of each scale while maintaining the operation efficiency, depth separable convolution and point convolution are introduced using a residual strategy. In the cross-attention fusion module (CFM), the image feature information extracted by cross-fusion is designed to reconstruct the high-frequency information of the image. At the same time, to effectively fuse the cascaded multiple pyramid multi-scale modules (PMMs), a cross-module learning method is designed to learn the multi-scale information extracted by different deep features. In addition, an improved integrated attention enhancement module (IAEM) is inserted in the tail, which fuses the deep parts of different stages through dense connection, enhances the learning feature weight by changing the dimension, and introduces 3D convolution to learn context features to realize the effective fusion of image feature information to improve the quality of image reconstruction more accurately. Finally, experimental results show that MCFN has a certain competitiveness in key performance indicators compared with existing leading methods on public benchmark datasets. In particular, when quadrupled upsampling was performed on the Set 5 dataset, MCFN reached a PNSR of 32.43 dB, 0.14 dB higher than MSAR. In addition, through visual contrast, MCFN has rich texture details and a high level of high-frequency information in the reconstructed images, further proving the method's effectiveness. Although MCFN has shown some competitive performance in the experiment, we also recognize its limitations. Future work plans include training with more realistic datasets to enhance the generalization and practicality of the model. In addition, it includes the introduction of subjective evaluation and other methods to evaluate image quality more comprehensively.

Author Contributions: Conceptualization, Y.M. and Y.X.; methodology, Y.M. and Y.X.; data curation, Y.M.; writing—original draft preparation, Y.M. and Y.L.; writing—review and editing, Y.X., F.Y. and Q.Z.; supervision, Q.L. (Qi Li) and Q.L. (Quanyang Liu); project administration, Y.X.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science and Technology Department Project of Jilin Province, China (20210203039SF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The set Div2K is available online at: https://data.vision.ee.ethz.ch/cvl/DIV2K/, (accessed on 5 January 2023). The Set 5 set is available online at http://people.rennes.inria.fr/Aline.Roumy/results/SR_BMVC12.html, (accessed on 5 January 2023). The Set 14 set is available online at https://doi.org/10.1007/978-3-642-27413-8_47, (accessed on 5 January 2023). The set BSD10 is available online at https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/, (accessed on 5 January

2023). The Urban100 set is available online at https://drive.google.com/drive/folders/1B3DJGQKB6eNdwuQIhdskA64qUuVKLZ9u, (accessed on 5 January 2023). The set Mange109 is available online at http://www.manga109.org/en/, (accessed on 5 January 2023).

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

References

1. Lu, W.; Tan, Y.P. Color filter array demosaicking: New method and performance measures. *IEEE Trans. Image Process.* **2003**, *12*, 1194–1210. [CrossRef] [PubMed]
2. Irani, M.; Peleg, S. Improving resolution by image registration. CVGIP: Graph. *Model. Image Process.* **1991**, *53*, 231–239. [CrossRef]
3. Dong, B.Z.; Yu, M.C.; Zhao, P. Image super-resolution reconstruction method based on wavelet domain. *Liquid Cryst. Displays* **2021**, *36*, 10. [CrossRef]
4. Chen, Z.; Hu, H.; Yao, J.; Yan, Q.; Lin, Z. Single frame image super-resolution reconstruction based on improved generative adversarial network. *Liquid Cryst. Displays* **2021**, *36*, 8. [CrossRef]
5. 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]
6. Dong, C.; Loy, C.C.; Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part II*; Springer: Cham, Switzerland, 2016; Volume 14, pp. 391–407. [CrossRef]
7. 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 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883. [CrossRef]
8. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654. [CrossRef]
9. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 136–144. [CrossRef]
10. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 624–632. [CrossRef]
11. Tai, Y.; Yang, J.; Liu, X. Image Super-Resolution via Deep Recursive Residual Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155. [CrossRef]
12. 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. [CrossRef]
13. Feng, X.; Li, X.; Li, J. Multi-Scale Fractal Residual Network for Image Super-Resolution. *Appl. Intell.* **2021**, *51*, 1845–1856. [CrossRef]
14. Mehta, N.; Murala, S. MSAR-Net: Multi-Scale Attention Based Light-Weight Image Super-Resolution. *Pattern Recognit. Lett.* **2021**, *151*, 215–221. [CrossRef]
15. Hou, J.; Si, Y.; Yu, X. A Novel and Effective Image Super-Resolution Reconstruction Technique via Fast Global and Local Residual Learning Model. *Appl. Sci.* **2020**, *10*, 1856. [CrossRef]
16. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645. [CrossRef]
17. Tai, Y.; Yang, J.; Liu, X.; Xu, C. MemNet: A Persistent Memory Network for Image Restoration. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4539–4547. [CrossRef]
18. Jiang, K.; Wang, Z.; Yi, P.; Jiang, J. Hierarchical Dense Recursive Network for Image Super-Resolution. *Pattern Recognit.* **2020**, *107*, 107475. [CrossRef]
19. Cheng, R.; He, X.; Zheng, Z.; Wang, Z. Multi-Scale Safety Helmet Detection Based on SAS-YOLOv3-Tiny. *Appl. Sci.* **2021**, *11*, 3652. [CrossRef]
20. Wu, Y.; Liu, Z.; Chen, Y.; Zheng, X.; Zhang, Q.; Yang, M.; Tang, G. FCNet: Stereo 3D Object Detection with Feature Correlation Networks. *Entropy* **2022**, *24*, 1121. [CrossRef] [PubMed]
21. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
22. He, Z.; Cao, Y.; Du, L.; Xu, B.; Yang, J.; Cao, Y.; Zhuang, Y. MRFN: Multi-Receptive-Field Network for Fast and Accurate Single Image Super-Resolution. *IEEE Trans. Multimed.* **2019**, *22*, 1042–1054. [CrossRef]

23. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Tang, X. Residual Attention Network for Image Classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164. [CrossRef]
24. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
25. 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. [CrossRef]
26. Guo, T.; Dai, T.; Liu, L.; Zhu, Z.; Xia, S.-T. S2A:Scale Attention-Aware Networks for Video Super-Resolution. *Entropy* **2021**, *23*, 1398. [CrossRef] [PubMed]
27. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-Order Attention Network for Single Image Super-Resolution. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074. [CrossRef]
28. Zang, H.; Zhao, Y.; Niu, C.; Zhang, H.; Zhan, S. Attention Network with Information Distillation for Super-Resolution. *Entropy* **2022**, *24*, 1226. [CrossRef] [PubMed]
29. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587. [CrossRef]
30. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520. [CrossRef]
31. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154. [CrossRef]
32. Agustsson, E.; Timofte, R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 126–135. [CrossRef]
33. 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, BMVC 2012, Surrey, UK, 3–7 September 2012. [CrossRef]
34. Zeyde, R.; Elad, M.; Protter, M. On Single Image Scale-Up Using Sparse Representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, 24–30 June 2010, Revised Selected Papers*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7, pp. 711–730. [CrossRef]
35. 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, ICCV 2001, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 416–423. [CrossRef]
36. Huang, J.B.; Singh, A.; Ahuja, N. Single Image Super-Resolution from Transformed Self-Exemplars. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5197–5206. [CrossRef]
37. 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]
38. Timofte, R.; De Smet, V.; Van Gool, L. A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution. In *Computer Vision—ACCV 2014. Lecture Notes in Computer Science*; Cremers, D., Reid, I., Saito, H., Yang, M.H., Eds.; Springer: Cham, Switzerland, 2015; Volume 9006. [CrossRef]
39. 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. [CrossRef]
40. Hui, Z.; Wang, X.; Gao, X. Fast and Accurate Single Image Super-Resolution via Information Distillation Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 723–731. [CrossRef]
41. Lu, T.; Wang, Y.; Wang, J.; Liu, W.; Zhang, Y. Single Image Super-Resolution via Multi-Scale Information Polymerization Network. *IEEE Signal Process. Lett.* **2021**, *28*, 1305–1309. [CrossRef]
42. Hu, Y.; Gao, X.; Li, J.; Huang, Y.; Wang, H. Single Image Super-Resolution with Multi-Scale Information Cross-Fusion Network. *Signal Process.* **2021**, *179*, 107831. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.