Contextual Feature Modulation Network for Efficient Super-Resolution

Wandi Zhang[1], Hao Shen[1], Biao Zhang[1], Weidong Tian[1], and Zhong-Qiu Zhao[1,2,3](✉)

[1] College of Computer and Information, Hefei University of Technology, Hefei, China
2021171171@mail.hfut.edu.cn
[2] Intelligent Manufacturing Institute of HFUT, Hefei, China
[3] Guangxi Academy of Sciences, Guangxi, China

Abstract. In recent years, single image super-resolution (SISR) reconstruction models based on convolutional neural networks (CNNs) have shown remarkable visual effects and reconstruction accuracy. However, the abundance number of parameters and relatively slow execution speed make it challenging to deploy these models on devices with limited memory and processing power. To address this challenge, we propose a Contextual Feature Modulation Network, denoted as CFMN, for efficient SISR tasks in this paper. This model successfully reduces model size and computational burden while maintaining high-quality image reconstruction. The proposed CFMN consists of a Multi-scale Feature Spatial Modulation (MFSM) and a Channel Attention Fusion Module (CAFM). Specifically, the MFSM replaces the traditional attention mechanism with a spatial modulation strategy. This module adaptively selects contextual feature representations at various scales and granularities through a multi-scale mechanism and a gated matrix, modulating the input features in the spatial dimension. Another core module CAFM complementarily extracts local contextual information and incorporates the Squeeze-and-Excitation Block to capture inter-channel dependencies. It effectively combines features from various channels through feature fusion, enhancing the network's ability to perceive image details. The performance analysis demonstrates that the proposed CFMN effectively balances model complexity and performance.

Keywords: Efficient Super-Resolution · Spatial Modulation · Multi-scale

1 Introduction

SISR aims to extract feature information from low-resolution (LR) images for reconstructing high-resolution (HR) images. Following the pioneering work of SRCNN [1], CNN-based SISR models have consistently exhibited superior performance. With the continuous development of CNNs, the scale of network models has gradually increased. In VDSR [2], the SR network is first deepened to 20 layers. EDSR [3] model architecture surpasses 60 layers with approximately 43M parameters. Subsequently, RDN [4] and RCAN [5] further expand the network depth to over 100 layers and even over 400 layers. The primary trend presented by these models is to further enhance the SR performance

D.-S. Huang et al. (Eds.): ICIC 2024, LNCS 14867, pp. 15–26, 2024.
https://doi.org/10.1007/978-981-97-5597-4_2

by adding convolutional layers. However, convolutional layers also introduce excessive computational costs and storage requirements, limiting applications on mobile devices. In recent years, the emergence of Visual Transformers (ViTs) [6–8] has had a profound impact on the field of image super-resolution. Thanks to its unique self-attention mechanism, which captures global image dependencies, Transformer has stronger capabilities for recovering complex textures and details. However, deploying a self-attention module on devices is challenging because of the high computational cost.

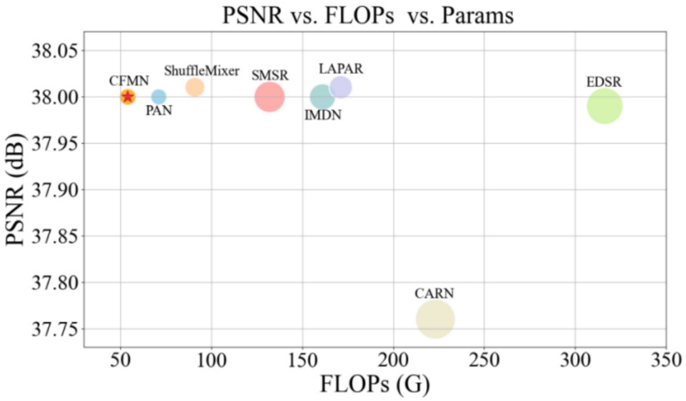

Fig. 1. Comparison of the trade-off between PSNR performance and model complexity on the Set5 dataset for × 2 SR

To address the issue above, we propose a Contextual Feature Modulation Network, namely CFMN. This network consists of Multi-scale Feature Spatial Modulation (MFSM) and Channel Attention Fusion Module (CAFM), aiming to maintain the quality of reconstructed images while reducing model parameters and computational costs. As shown in Fig. 1, we find that the CFMN achieves a better balance between SR performance and model complexity. Inspired by FocalNet [9], we adopt a spatial modulation strategy in the design of MFSM to replace the traditional attention mechanism. The module adaptively selects the representations of contextual features at different granularities, enabling dynamic integration of both global and local contextual information. Additionally, it employs a multi-scale approach to deeply explore rich feature information from images at various scales. Another module CAFM complementarily extracts local information and performs feature fusion along the channel dimension. To further model the dependencies among channels, this module incorporates the SE Block [10], enabling the network to focus on the most significant features for the reconstruction task. The following summarizes our contributions:

- We design a lightweight attention module as an alternative to the self-attention module, ensuring comparable performance while utilizing minimal additional parameters and computational resources.
- We propose a channel feature fusion module to encode local contextual information and adjust channel weights.

- Our model is evaluated on multiple benchmark datasets, indicating that it effectively maintains image SR performance with lower computational cost and fewer parameters.

2 Related Work

2.1 Deep Learning-Based Image Super-Resolution

Deep learning-based methods have become the dominant approach in SISR due to their powerful representation and fitting capabilities. The pioneering SRCNN [1] introduces a three-layer CNN to acquire the relationship between HR and LR images and significantly improve performance compared to conventional methods. Subsequently, VDSR [2] and EDSR [3] capitalize on deeper information through a very deep and wide backbone combined with residual learning. RCAN [5] extends the backbone into over 400 layers based on channel attention and residual in residual to further exploit intermediate features. Most recently, more and more researchers have paid attention to Vision Transformers [6, 7] due to their ability to capture long-range dependence of images. Many Transformer-based methods have emerged for image restoration [8, 11, 12], advancing state-of-the-art performance. A significant pre-trained model known as IPT [6], based on the Transformer architecture, is employed for super-resolution tasks. Building upon the Swin Transformer [13], SwinIR [8] executes self-attention within a localized window during feature extraction, yielding remarkable outcomes. Although these networks attain state-of-the-art reconstruction accuracy, the high computational cost and memory footprint limit their applications on resource-constrained devices.

2.2 Efficient Image Super-Resolution

Efficient Image Super-Resolution aims to decrease the computational cost and the parameter count of the SR networks while improving inference speed and maintaining high performance. IMDN [14] employs a progressive feature distillation strategy to gradually compress features, improving the efficiency of the model. PAN [15] obtains the attention map with only a 3 × 3 convolution and also introduces self-calibrating convolutions to exploit long-distance dependencies. LAPAR [16] performs linear coefficient regression tasks on the predefined filter dictionary, achieving state-of-the-art results with fewer model parameters and MultiAdds. ShuffleMixer [17] introduces the large kernel convolution into the lightweight SR network. In recent related research, model compression and acceleration techniques have been introduced to image super-resolution tasks. Wang et al. [18] propose the SMSR network to learn sparse masks to locate and skip redundant computations, accelerating the inference process. Shen et al. [19] present a joint operation and attention block search algorithm for SR, aiming to create an efficient network and enhance feature representation. Zhang et al. [20] propose structural regularization pruning, aligning the positions of pruned filters across different layers, efficiently eliminating redundant filters. Although the aforementioned methods above have progressed in various efficiency aspects, our goal is to further strike a balance between achieving optimal model efficiency and maintaining SR performance.

3 Proposed Method

3.1 Network Architecture

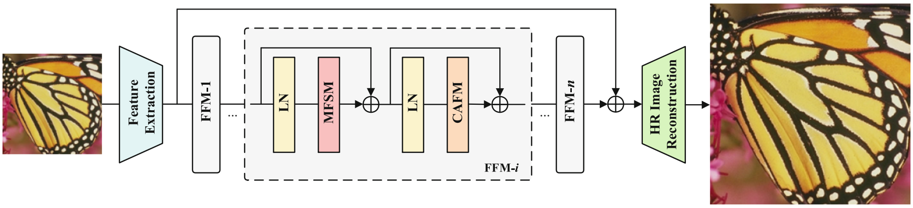

Fig. 2. The overall architecture of the CFMN

As depicted in Fig. 2, the whole architecture contains three components including a shallow feature extraction module M_s, a deep feature extraction module M_d utilizing Feature Fusion Modules (FFMs), and an efficient high-resolution image reconstruction module M_r. Given an input image $I_{LR} \in \mathbb{R}^{H \times W \times 3}$, with dimensions denoted by H and W, the M_s employs a simple 3×3 convolution layer to convert I_{LR} into a shallow feature map. Then, this feature map is sent to cascading FFMs for generating deep features. Finally, the M_r aggregates both low-frequency and high-frequency features through a global residual connection to predict the HR image $I_{SR} \in \mathbb{R}^{rH \times rW \times 3}$, where r is a scale factor. The M_r consists of a 3×3 convolution layer and a sub-pixel convolution [21].

3.2 Multi-scale Feature Spatial Modulation

Most existing models [5, 8, 11, 12] typically employ attention mechanisms to enable convolutional neural networks to focus on critical information. Commonly used attention mechanisms include Channel Attention (CA) [10] and Self-attention (SA) [7, 8, 13]. The former is limited in capturing long-range dependencies, while the latter supports global interactions within an image and plays a crucial role in the success of the Vision Transformer [7]. However, the high algorithmic complexity of the self-attention module has consistently been a crucial concern for researchers. Therefore, we present a lightweight module as an alternative for modeling long-range dependencies by learning global contextual features. This model combines a multi-scale mechanism and a gating mechanism to dynamically fuse feature information at different scales and granularities. This fusion strategy enhances the network's comprehension of complex image structures by increasing feature diversity.

Multi-scale Mechanism. As shown in Fig. 3, given the input feature $X \in \mathbb{R}^{H \times W \times C}$, we initially divide the input features evenly into four parts based on the channel dimension to reduce model complexity and learn attention maps at multiple scales. These four parts undergo downsampling through a max pooling layer and are then fed into Spatial Modulations to generate output. The generated outputs are subsequently fused and normalized, resulting in the final estimated attention map $\widehat{X}$ that focuses on important regions and

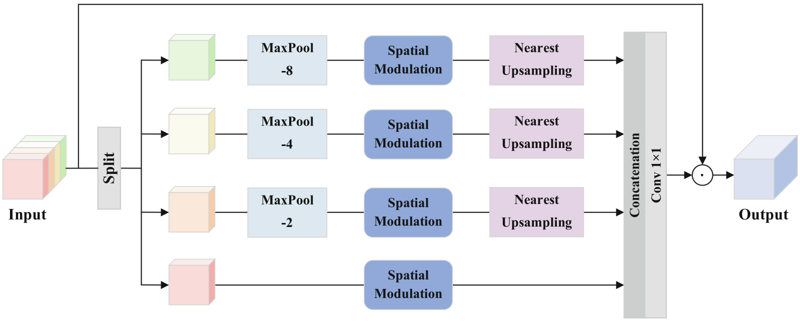

Fig. 3. The architecture of the MSFM

details of the original feature map. In the final step, $\widehat{X}$ is multiplied element-wise with the input X to obtain the modulated feature map X_{out}:

$$\begin{aligned} X_0, X_1, X_2, X_3 &= \mathrm{Split}(X) \\ \hat{X}_0 &= \mathrm{SM}(X_0) \\ \hat{X}_i &= \uparrow_{2^i}\big(\,\mathrm{SM}\,\big(\downarrow_{2^i}(X_i)\big)\big), 1 \le i \le 3 \\ \hat{X} &= \mathrm{Conv}_{1\times1}\big(\,\mathrm{Concat}\big(\big[\hat{X}_0, \hat{X}_1, \hat{X}_2, \hat{X}_3\big]\big)\big) \\ X_{out} &= \mathrm{GELU}\big(\hat{X}\big) \odot \hat{X} \end{aligned} \tag{1}$$

where $\mathrm{Split}(\cdot)$ represents the channel split operation, $\mathrm{SM}(\cdot)$ is Spatial Attention Modulation, $\downarrow_{2^i}(\cdot)$ and $\uparrow_{2^i}(\cdot)$ represents upsampling and downsampling of the input features, $\mathrm{Concat}(\cdot)$ denotes the concatenation operation, $\mathrm{Conv}_{1\times1}$ is a 1×1 convolution and $\mathrm{GELU}(\cdot)$ stands for the activation function [22].

Spatial Modulation. As shown in Fig. 4, spatial modulation aggregates the attention map at various granularities to generate a modulator. In detail, given the input feature $X \in \mathbb{R}^{H\times W\times C}$ and the number of layers L, the feature undergoes a linear layer to expand the channel count to $2C+L+2$. Subsequently, it is partitioned along the channel dimension into $Query \in \mathbb{R}^{H\times W\times C}$,$Context \in \mathbb{R}^{H\times W\times C}$, and $Gate \in \mathbb{R}^{H\times W\times (L+2)}$. It can be expressed by:

$$Query, Context, Gate = \mathrm{Split}(\mathrm{Linear}(X)) \tag{2}$$

where $\mathrm{Linear}(\cdot)$ represents a linear layer. The $Context$ gradually acquires broader contextual features through a sequence of L depth-wise convolutions and a global average pooling layer. At level $l \in \{1, \cdots, L\}$, the $Context_l$ is obtained by:

$$Context_l = \mathrm{GELU}(\mathrm{DWConv}_{k\times k}(Context_{l-1})) \tag{3}$$

where $k = 2 \times l + 1$, $\mathrm{DWConv}_{k\times k}(\cdot)$ represents a depth-wise convolution with a kernel size of $k \times k$. Then, the $Context_L$ is fed into an average pooling layer to capture global contextual dependencies $Context_g$ by:

$$Context_g = \mathrm{GELU}(\ \mathrm{AvgPool}(Context_L)) \tag{4}$$

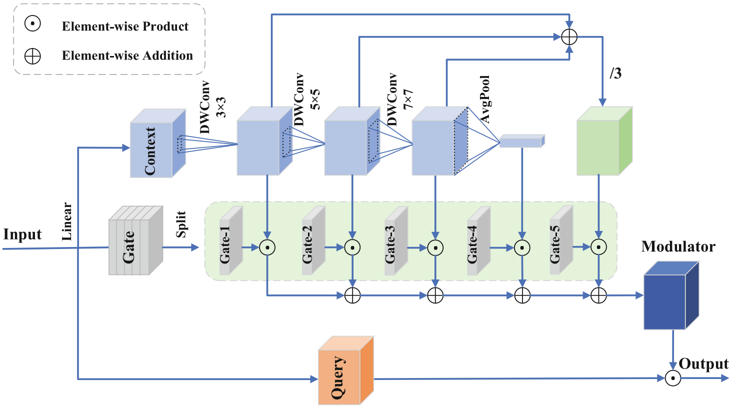

Fig. 4. The architecture of spatial modulation. The level is set to 3.

where AvgPool(·) denotes a global average pooling layer. However, effectively integrating these contextual features at different granularities poses a challenge. To address this, an average contextual feature $Context_{avg}$ is introduced to serve as a buffer during the fusion process. It can be written as follows:

$$Context_{avg} = \frac{1}{L}\sum_{l=1}^{L} Context_l \tag{5}$$

In addition, a gating mechanism is introduced to dynamically integrate features, resulting in the contextual feature denoted as $\widehat{X}$:

$$\hat{X} = \sum_{l=1}^{L}(Gate_l \odot Context_l) + Gate_{L+1} \odot Context_g + Gate_{L+2} \odot Context_{avg} \tag{6}$$

$$\hat{X} = \text{Conv}_{1\times1}\left(\hat{X}\right) \tag{7}$$

where $Gate_l \in \mathbb{R}^{H\times W\times 1}$ refers to the l-th layer of the $Gate$, $\odot$ is the element-wise product. After obtaining the modulator $\widehat{X}$, it is then applied to adjust $Query$ through element-wise multiplication. Then a 1×1 convolution is employed for cross-channel interaction. The final output X_{out} is derived by:

$$X_{out} = \text{Dropout}\left(\text{Conv}_{1\times1}\left(\hat{X} \odot Query\right)\right) \tag{8}$$

where Dropout(·) is a regularization layer.

3.3 Channel Attention Fusion Module

The proposed MFSM mainly focuses on the extraction and utilization of global contextual features, effectively integrating global information of the image through spatial modulation. However, a significant limitation of this module is the lack of an attention mechanism in the channel dimension, which partly restricts the comprehensive utilization of image features. To address this limitation and extract complementary local features, we propose a Channel Attention Fusion Module (CAFM) based on FMBConv [23]. Since BatchNorm [24] is not suitable for image super-resolution tasks. Therefore, we substitute BatchNorm with LayerNorm [25] and move it to the front of the module, which improves the stability of model training. As illustrated in Fig. 5, the proposed CAFM comprises a 3 × 3 convolution, an SE Block [10], and a 1 × 1 convolution. Specifically, the role of the 3 × 3 convolution is to extract local spatial context and quadruple the number of channels for channel mixing. Subsequently, the SE block is employed to model inter-channel dependencies, adaptively adjusting the importance of features among channels. Finally, the 1 × 1 convolution compresses the channel count back to its original size.

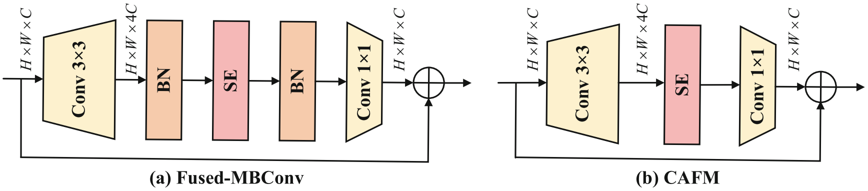

Fig. 5. The architectures of FMBConv and CAFM.

3.4 Feature Fusion Module

In general, people widely believe that the outstanding performance of Visual Transformer (ViT) [7] is primarily attributed to the token mixer. However, recent research has shown that replacing this token mixer with spatial MLP still maintains impressive results. This leads to a conclusion: the success of ViT is attributed to the adoption of a general architecture known as MetaFormer [26], which consists of a token mixer and a feed-forward neural network. This paper similarly adopts this structure, integrating MFSM and CAFM into a feature fusion module, formulated as:

$$Y = \mathrm{MFSM}(\mathrm{LN}(X)) + X \tag{9}$$

$$Z = \mathrm{CAFM}(\mathrm{LN}(Y)) + Y \tag{10}$$

where $\mathrm{LN}(\cdot)$ refers to the LayerNorm [25] layer and X, Y and Z represent the intermediate features.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. Similar to prior works [15–17], our training and validation datasets consist of 3450 and 100 images respectively, sourced from the DIV2K [27] and Flickr2K [3] datasets. The LR images were derived from HR images using bicubic downsampling. For evaluation, we employ five commonly used benchmark datasets [28–32]. To assess the quality of the reconstructed images, two standard evaluation metrics are utilized: PSNR and SSIM [33]. These measurements are computed using the Y channel within the YCbCr color space, obtained through conversion from the RGB color space.

Implementation Details. The configuration of the CFMN involves setting the number of FFMs to 8 and the level of SM to 3. During the training phase, we extract 64 patches of 64 × 64 pixels at random from the LR images for each mini-batch. To enhance the diversity of the input patches, we perform data augmentation by randomly flipping it horizontally and vertically and then rotating it. The CFMN is trained by the Adam [34] optimizer with the momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The training starts with the initial learning rate of 1×10^{-3} and progressively decreases to a minimum of 1×10^{-5}, utilizing the Cosine Annealing strategy [35] for updates. The network is implemented by the PyTorch framework and trained with 2 Nvidia RTX 2080Ti GPUs.

4.2 Comparison to Other Methods

Quantitative Evaluation. To comprehensively assess the performance of the CFMN, we contrast it with a series of state-of-the-art lightweight image super-resolution models. These models include IMDN [14], PAN [15], LAPAR [16], SMSR [18], and ShuffleMixer [17]. Presents the quantitative evaluation results on the commonly used SR image benchmarks. Apart from using PSNR and SSIM metrics, we have provided the parameter count and computational complexity to assess the performance and complexity of the model. These calculations are based on upsampling a LR image to a HR of 1280 × 720 pixels. Thanks to its simple and efficient structural design, the proposed CFMN model exhibits significant advantages in terms of both parameter count and computational complexity. Taking × 4 SR as an example, CFMN reduces the number of parameters by up to 59% (295K vs. 715K) and significantly reduces computational complexity by 66% (14G vs. 41G) compared to IMDN. However, CFMN maintains comparable performance to IMDN, fully demonstrating the effectiveness and efficiency of its structural design. In addition, compared to ShuffleMixer, which is also a lightweight model, CFMN reduces the number of parameters by 28% (295K vs. 411K) and reduces computational complexity by 50% (14G vs. 28G). The results prove that CFMN effectively reduces model complexity and computational cost while maintaining high performance (Table 1).

Qualitative Evaluation. Besides quantitative evaluation, we also conduct a qualitative comparison of the proposed method. Figure 6 presents the visual comparison results on the Urban100 dataset for × 4 SR. It can be clearly observed that, compared to other lightweight models, the CFMN exhibits higher accuracy in generating parallel lines and grid patterns. The accurate reproduction of these fine structures is crucial for super-resolution tasks as they directly impact the clarity and overall visual effect of the image. Therefore, it can be concluded that the CFMN possesses significant advantages in detail recovery and structure preservation.

Table 1. Comparison of efficient SR models. '-' represents unreported results

Model	Scale	Params	FLOPs	Set5	Set14	BSD100	Urban100	Manga109
IMDN [14]	×2	694K	161G	38.00/0.9605	33.63/0.9177	**32.19**/0.8996	32.17/0.9283	**38.88/0.9774**
PAN [15]		**261K**	71G	38.00/0.9605	33.59/0.9181	32.18/0.8997	32.01/0.9273	38.70/0.9773
LAPAR-A [16]		548K	171G	**38.01**/0.9605	33.62/**0.9183**	**32.19/0.8999**	32.10/0.9283	38.67/0.9772
SMSR [18]		985K	132G	38.00/0.9601	**33.64**/0.9179	32.17/0.8990	**32.19/0.9284**	38.76/0.9771
ShuffleMixer [17]		394K	91G	**38.01/0.9606**	33.63/0.9180	32.17/0.8995	31.89/0.9257	38.83/**0.9774**
CFMN(Ours)		283K	**54G**	38.00/0.9604	33.57/0.9179	32.17/0.8994	31.82/0.9254	38.69/0.9770
IMDN [14]	×3	703K	72G	34.36/0.9270	30.32/0.8417	29.09/0.8046	28.17/0.8519	33.61/0.9445
PAN [15]		**261K**	39G	**34.40**/0.9271	30.36/0.8423	29.11/0.8050	28.11/0.8511	33.61/0.9448
LAPAR-A [16]		594K	114G	34.36/0.9267	30.34/0.8421	29.11/0.**8054**	28.15/0.8523	33.51/0.9441
SMSR [18]		993K	68G	**34.40**/0.9270	30.33/0.8412	29 10/0 8050	**28.25/0.8536**	33.68/0.9445
ShuffleMixer [17]		415K	43G	**34.40/0.9272**	**30.37/0.8423**	**29.12**/0.8051	28.08/0.8498	**33.69/0.9448**
CFMN(Ours)		288K	**24G**	**34.35**/0.9268	30.35/0.8413	29.09/0.8047	27.97/0.8482	33.50/0.9439
IMDN [14]	×4	715K	41G	**32.21**/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838	30.45/0.9075
PAN [15]		**261K**	22G	32.13/0.8948	28.61/0.7822	27.59/0.7363	26.11/0.7854	30.51/0.9095
LAPAR-A [16]		659K	94G	32.15/0.8944	28.61/0.7818	**27.61/0.7366**	**26.14/0.7871**	30.42/0.9074
SMSR [18]		1006K	42G	32.12/0.8932	28.55/0.7808	27.55/0.7351	26.11/0.7868	30.54/0.9085
ShuffleMixer [17]		411K	28G	**32.21/0.8953**	**28.66/0.7827**	**27.61/0.7366**	26.08/0.7835	**30.65/0.9093**
CFMN(Ours)		295K	**14G**	32.18/0.8946	28.61/0.7815	27.56/0.7354	26.01/0.7826	30.42/0.9071

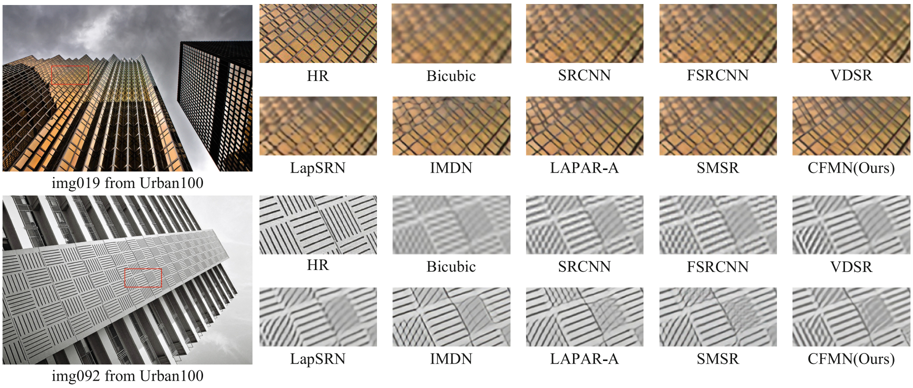

Fig. 6. Visual comparison of other SR models on the Urban100 dataset for × 4 SR

4.3 Ablation

To validate the influence of each module and its internal design, we conduct more detailed ablation experiments. The results of the experiments are presented in Table 2. In all experiments, we consistently adopt the same settings and measure performance on the DIV2K validation dataset for × 4 SR. The parameters and computational complexity are evaluated when enlarging a 320 × 180 image to 1280 × 720 to ensure the accuracy and consistency of the evaluation.

Multi-Scale Feature Spatial Modulation. Removing the MFSM module leads to a 0.17 dB decrease in PSNR on the DIV2K-val dataset, emphasizing the effectiveness of this module in enhancing super-resolution performance. Additional ablation experiments reveal that utilizing a single spatial modulation without multi-scale fusion reduces PSNR by 0.04 dB while increasing parameters and computational cost. Furthermore, removing the gating mechanism for adaptive adjustment also results in a 0.07 dB PSNR drop. Meanwhile, the introduction of average contextual features can also bring a 0.03 dB improvement in performance. These findings underscore the contributions of both the multi-scale and gated mechanisms in improving model performance while maintaining efficiency.

Channel Attention Fusion Module. The absence of the CAFM module causes a notable 0.7 dB decrease in PSNR, confirming its effectiveness in super-resolution tasks. When only channel fusion is performed without considering channel attention, there was a slight 0.06 dB drop in PSNR, indicating the importance of attention mechanisms. Furthermore, when we use MLP instead of CAFM, the PSNR decreases by 0.23 dB, proving the advantage of CAFM in improving SR performance.

Normalization. Through experiments comparison, the results indicate that LayerNorm [25] is more suitable for super-resolution tasks than BatchNorm [24].

Table 2. Ablation experiments of CFMN on the DIV2K-val dataset for × 4 SR.

Ablation	Variant	Params	FLOPs	DIV2K-val
Main Module	MFSM → None	247K	12.93G	30.28/0.8351
	CAFM → None	65K	2.24G	29.75/0.8254
MFSM	w/o Multi-scale Fusion	315K	16.78G	30.41/0.8367
	w/o Gated Matrix	294K	14.19G	30.38/0.8359
	w/o Average Context	295K	14.31G	30.42/0.8365
CAFM	w/o SE Block	274K	14.19G	30.39/0.8362
	CAFM → MLP	150K	7.02G	30.22/0.8328
Normalization	LN → BN	295K	14.22G	30.29/0.8345
CFMN(Ours)	-	295K	14.36G	**30.45/0.8375**

5 Conclusion

In this paper, we propose the Contextual Feature Modulation Network for efficient SR reconstruction. Specifically, MFSM captures richer attention information through a multi-scale mechanism. Additionally, it extracts contextual features at various granularities using depth-wise convolutions and employs gated convolution to select notable features. Furthermore, CAFM complements the extraction of local contextual features and further introduces the SE Block for channel attention and fusion. Experiments on commonly used benchmark datasets demonstrate that the CFMN maintains the performance of existing methods while having lower parameters and computational costs.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 61976079, in part by Guangxi Key Research and Development Program under Grant AB22035022, and in part by Anhui Key Research and Development Program under Grant 202004a05020039.

References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI **38**(2), 295–307 (2015)
2. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR, pp. 1646−1654 (2016)
3. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW, pp. 136–144 (2017)
4. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR, pp. 2472–2481 (2018)
5. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV, pp. 286–301 (2018)
6. Chen, H., et al.: Pre-trained image processing transformer. In: CVPR, pp. 12299–12310 (2021)
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: image restoration using swin transformer. In: ICCV, pp. 1833–1844 (2021)
9. Yang, J., Li, C., Dai, X., Gao, J.: Focal modulation networks. NeurIPS 35, 4203–4217 (2022)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
11. Li, W., Lu, X., Qian, S., Lu, J., Zhang, X., Jia, J.: On efficient transformer-based image pre-training for low-level vision. arXiv preprint arXiv:2112.10175 (2021)
12. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: CVPR. pp. 18278–18289 (2023)
13. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
14. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: ACM MM, pp. 2024–2032 (2019)
15. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: ECCVW. pp. 56–72. Springer (2020). https://doi.org/10.1007/978-3-030-67070-2_3

16. Li, W., Zhou, K., Qi, L., Jiang, N., Lu, J., Jia, J.: Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. NeurIPS **33**, 20343–20355 (2020)
17. Sun, L., Pan, J., Tang, J.: Shufflemixer: an efficient convnet for image super-resolution. NeurIPS **35**, 17314–17326 (2022)
18. Wang, L., et al.: Exploring sparsity in image super-resolution for efficient inference. In: CVPR, pp. 4917–4926 (2021)
19. Shen, H., Zhao, Z.Q., Liao, W., Tian, W., Huang, D.S.: Joint operation and attention block search for lightweight image restoration. PR **132**, 108909 (2022)
20. Zhang, Y., Wang, H., Qin, C., Fu, Y.: Aligned structured sparsity learning for efficient image super-resolution. NeurIPS **34**, 2695–2706 (2021)
21. Shi, W.,et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR, pp. 1874–1883 (2016)
22. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
23. Tan, M., Le, Q.: Efficientnetv2: smaller models and faster training. In: ICML, pp. 10096–10106. PMLR (2021)
24. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456. PMLR (2015)
25. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
26. Yu, W., et al.: Metaformer is actually what you need for vision. In: CVPR, pp. 10819–10829 (2022)
27. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: methods and results. In: CVPRW, pp. 114–125 (2017)
28. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC (2012)
29. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: ICCS, pp. 711–730. Springer (2012). https://doi.org/10.1007/978-3-642-27413-8_47
30. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. TPAMI **33**(5), 898–916 (2010)
31. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR, pp. 5197–5206 (2015)
32. Matsui, Y., et al.: Sketch-based manga retrieval using manga109 dataset. Multimed Tools Appl. **76**, 21811–21838 (2017)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)