SUPPLEMENTARY MATERIAL FOR Efficient Information Modulation Network for Image Super-Resolution

Paper #801

Here we provide more experiment results and analysis for EIMN. The appendix document is arranged as follows:

- In Part 1 and Tables 1, more implementation details can be viewed.
- In Part 2 and Tables 2, 3, 4 and 5, more details about relationship between all configurations in ablation experiments and model performance can be observed.
- In Part 3 and Fig. 1, more visual results, such as "img 012" and "img 076" in Urban100 dataset, "Shimattelkouze vol101" in Manga109 dataset and "Barbara" in Set14 dataset, are displayed.
- In Part 4 and Fig. 2, more LAM results are displayed.
- In Part 5 and Fig. 3, features in the SADFFM are visualized. We can clearly observe that after modulation by the DFFM, channel diversity is notably increased and the network pays more attention to high-frequency edge details, which is consistent with the expected theory and results.

1 Implementation Details

Our proposed method includes 16 basic blocks, each with 64 channels. We only make minor channel adjustment on the image reconstruction part for scaling factors of $\times 2$, $\times 3$, and $\times 4$. We implement our method with Pytorch 1.12.0 and train it on a single NVIDIA RTX 3090 GPU. Further training details are provided in Table 1 for a more clearer understanding.

Table 1: Hyper-parameters of the training process.

Training Config	Settings
Dataset	DF2K (Flick2K [6]+DIV2K [1])
Random rotation	(90°, 180°, 270°)
Random flipping	Horizontal
Patch size	64×64
Batch size	16
Optimizer	Adam [5]
Base learning rate	$5e^{-4}$
Optimizer mementum	β_1=0.9, β_2=0.999
Weight decay	$1e^{-4}$
Learning rate schedule	Cosine decay
Learning rate bound	$1e^{-7}$
Loss function	L_1

2 More details about relationship between all configurations in ablation experiments and model performance

Architecture configuration. We first perform ablation experiments on the number of EIMBs in the nonlinear mapping part to search for the better balance between model complexity and performance, as shown in Table 1. Experiment results indicate that the performance improvements with the increase in the number of stacked blocks until the highest value is reached at 16 blocks. Further increasing the number of blocks would lead to a slight decrease in network performance. Therefore, considering both model complexity and performance, we set the number of blocks to 16.

Subordinate components in the SADFFM layer. We conduct a detailed study on the impact of each component in the SADFFM layer, as presented in Table 2. To compare the effectiveness of the proposed SADFFM layer, we also include the results of original FFN, which is frequently used in Transformer-style model. Remarkably, both DFFM and SAL achieve performance enhancements by a large margin, demonstrating the effectiveness of two modules.

Subordinate components in MOLRCM layer. Finally, we also conduct a detailed research on the impact of each component in MOLRCM layer based on the convolutional modulation technology, as shown in Table 3 and Table 4, including the choice of layer sequences and activation functions. Experiment results demonstrate the efficiency of our approach and the significant performance gains achieved through the carefully designed convolutional modulation module which consists of 5-5-7 layer squence for extracting large-range and multi-order contextual information and SiLU activation function that preserves the mean and variance of the input data and improves the learning process.

3 More visual results

In Fig. 1, we display the $\times 4$ SR results visualization. For the images "img 012" and "img 076" in Urban100 dataset [4], "Shimattelkouze vol101" in Manga109 dataset [8] and "Barbara" in Set14 dataset [10], our method reconstructs the clearest lattice, stripe and text patterns with the minimal blurry effects and artifacts compared to other methods, which validates the usefulness and effectiveness of our method. Take the image "Barbara" as an example, only our method generates stripes with accurate direction and minimal blurry, while the other methods produce incorrect stripes and a large range blurring effects.

4 More LAM results

In Fig. 2, we analyze local attribution maps (LAM) [3] results between AAN [2], EDSR [7], LMAN [9] and our method to investigate pixels utilization range in the input image when reconstructing

Table 2: Ablation study on the architecture configuration. The best performance is in red colors.

Block Numbers	#Params(K)	Set5	Set14	BSDS100	Urban100	Manga109
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
12	759	31.90/0.8895	28.49/0.7788	27.53/0.7366	25.75/0.7730	30.09/0.8959
14 (Ours)	881	32.53/0.8993	28.89/0.7882	27.79/0.7447	26.68/0.8027	31.22/0.9148
16 (Ours)	1002	32.63/0.9008	28.94/0.7897	27.82/0.7458	26.88/0.8084	31.52/0.9183
18	1124	32.62/0.9005	28.94/0.7898	28.81/0.7455	26.85/0.8075	31.49/0.9182
20	1246	32.62/0.9007	28.94/0.7896	27.81/0.7455	26.86/0.8078	31.48/0.9181

Table 3: Ablation study on the subordinate components of the proposed SADFFM layer. The best performance is in red colors. Where, CA and SA denote general channel and spatial attention, DFFM denotes the proposed dynamic feature flow modulation, SAL denotes the propoesd spatial awareness layer.

Model	Componets				#Params(K)	Set5	Set14	BSDS100	Urban100	Manga109
	CA	SA	DFFM	SAL		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
FFN	-	-	-	-	896	31.42/0.8824	28.09/0.7693	27.25/0.7271	25.12/0.7502	28.98/0.8835
FFN+CA	√				929	31.73/0.8888	28.36/0.7757	27.43/0.7323	25.55/0.7667	29.68/0.8948
FFN+SA		√			914	31.68/0.8872	28.33/0.7745	27.39/0.7313	25.45/0.7629	29.52/0.8925
FFN+DFFM			√		949	31.84/0.8900	28.42/0.7774	27.47/0.7337	25.67/0.7702	29.82/0.8965
FFN+SAL				√	937	31.94/0.8920	28.52/0.7790	27.52/0.7355	25.81/0.7758	30.01/0.8998
SADFFM(Ours)			√	√	1002	32.63/0.9008	28.94/0.7897	27.82/0.7458	26.88/0.8084	31.52/0.9183

Table 4: Ablation study on the different layer sequences of the convolutional modulation technology within the proposed MOLRCM layer. The best performance is in red colors. where, k_1-k_2-k_3 represents the kernel size of $DWConv$-$DWDConv$-$DWDConv$ squence in the convolutional modulation module.

Layers Sequence	#Params(K)	Set5	Set14	BSDS100	Urban100	Manga109
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
3-3-5	968	32.42/0.8981	28.77/0.7861	27.70/0.7416	26.48/0.7967	31.00/0.9129
5-5-7(Ours)	1002	32.63/0.9008	28.94/0.7897	27.82/0.7458	26.88/0.8084	31.52/0.9183
7-7-9	1053	32.60/0.9002	28.91/0.7890	27.79/0.7450	26.82/0.8069	31.41/0.9177

Table 5: Ablation study on the different activation functions of the convolutional modulation technology within the proposed MOLRCM layer. The best performance is in red colors.

model	Activation				#Params	Set5	Set14	BSDS100	Urban100	Manga109
	No use	Sigmoid	GELU	SiLU		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Model I	√				1002	31.45/0.8837	28.15/0.770	27.29/0.7272	25.22/0.7536	29.11/0.8855
Model II		√			1002	32.59/0.9004	28.91/0.7892	27.79/0.7450	26.80/0.8064	31.36/0.9162
Model III			√		1002	32.62/0.9006	28.93/0.7896	27.81/0.7455	26.84/0.8079	31.49/0.9184
Ours				√	1002	32.63/0.9008	28.94/0.7897	27.82/0.7458	26.88/0.8084	31.52/0.9183

Figure 1: Qualitative comparison of state-of-the-art methods on the Set14 (×4), Urban100 (×4) and Manga109 dataset (×4). Our method achieves better performance with fewer artifacts.

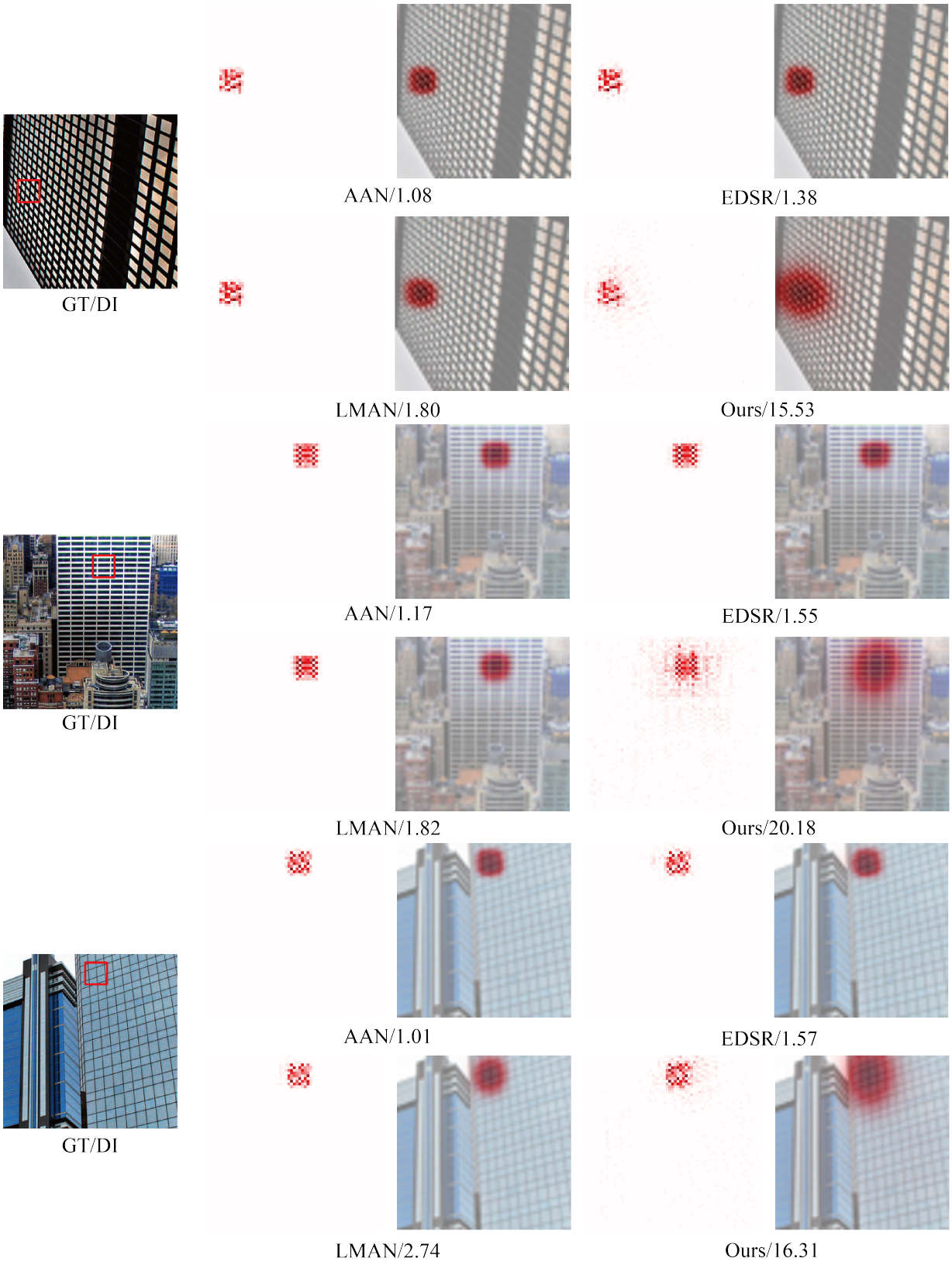
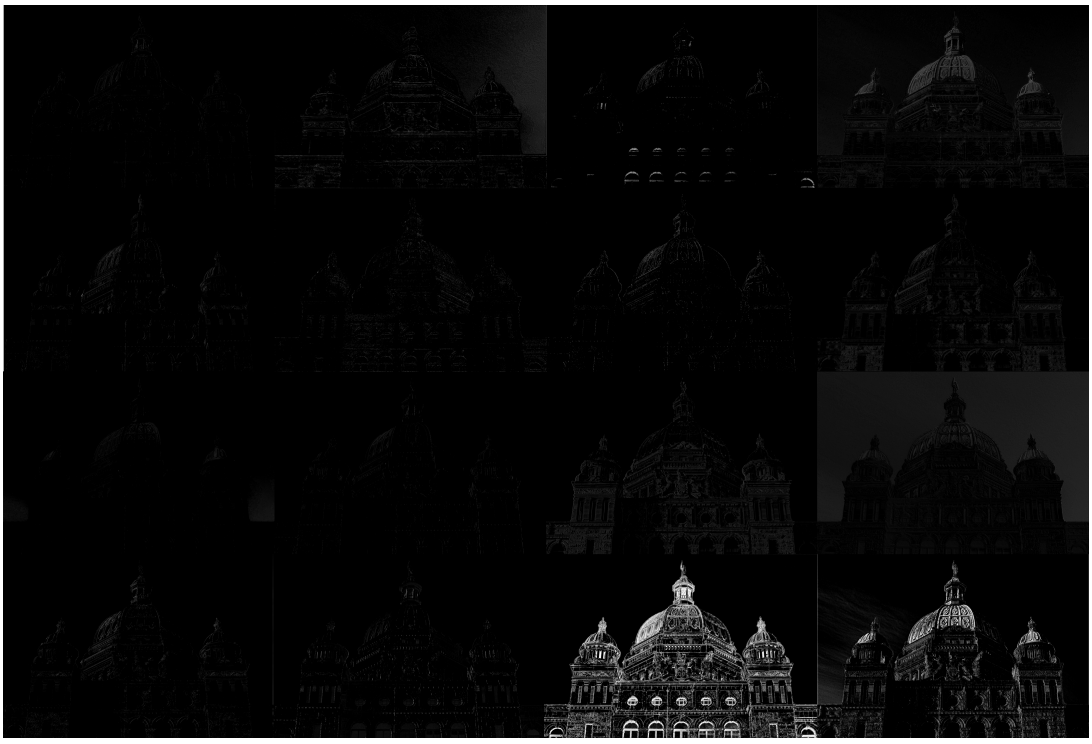


Figure 2: More results of LAM. A more widely distributed red area and higher DI represent a larger range pixels utilization.

(a) Visualization of some feature maps generated by the first linear layer in the SADFFM.



(b) Visualization of some feature maps after passing through the DFFM.

Figure 3: Feature visualization in the SADFFM. After modulation by the DFFM, channel diversity is notably increased and the network pays more attention to high-frequency edge details. Zoom in for a clearer view.

the selected area. Diffusion index (DI), an evaluation metric tool, reflects the ability of model to extract feature and utilizate effective information. As shown in Fig. 2, our method uses larger range pixel information to reconstruct area drawed with a red box, which demonstrates our method attains a larger receptive field by an efficient large kernel size convolutional modulation manner.

5 Feature visualization in the SADFFM

We first visualize the features generated by the first linear layer in the SADFFM, which exhibit a high degree of similarity obviously. This is due to the first layer expanding the channel dimensions from C to rC, leading to channel redundancy in the intermediate layers. As a result, different channels within a layer can carry similar or redundant information, resulting in increased computational cost without improving performance. Additionally, we visualize the features after passing through the DFFM layer, where we observe that channel diversity is notably increased. Different feature layers now carry distinct high frequency edge details, and readers can zoom in for a clearer view. This observation aligns with our expectations and theory, as image super-resolution is a classic example of a high-frequency information reconstruction task focused on recovering lost or degraded high-frequency components in the LR image. After modulation by the DFFM, the network can adaptively emphasize important regions and suppress irrelevant regions by sufficiently mining the underlying relevance of feature representations from both the channel and spatial perspectives.

References

[1] Eirikur Agustsson and Radu Timofte, 'Ntire 2017 challenge on single image super-resolution: Dataset and study', in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, (2017).

[2] Haoyu Chen, Jinjin Gu, and Zhi Zhang, 'Attention in attention network for image super-resolution', *arXiv preprint arXiv:2104.09497*, (2021).

[3] Jinjin Gu and Chao Dong, 'Interpreting super-resolution networks with local attribution maps', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9199–9208, (2021).

[4] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, 'Single image super-resolution from transformed self-exemplars', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5197–5206, (2015).

[5] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[6] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, 'Enhanced deep residual networks for single image super-resolution', in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, (2017).

[7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, 'Enhanced deep residual networks for single image super-resolution', in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, (2017).

[8] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa, 'Sketch-based manga retrieval using manga109 dataset', *Multimedia Tools and Applications*, **76**, 21811–21838, (2017).

[9] Jin Wan, Hui Yin, Zhihao Liu, Aixin Chong, and Yanting Liu, 'Lightweight image super-resolution by multi-scale aggregation', *IEEE Transactions on Broadcasting*, **67**(2), 372–382, (2020).

[10] Roman Zeyde, Michael Elad, and Matan Protter, 'On single image scale-up using sparse-representations', in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pp. 711–730. Springer, (2012).