SMFANet: A Lightweight Self-Modulation Feature Aggregation Network for Efficient Image Super-Resolution - Supplemental Material -

Mingjun Zheng*, Long Sun*, Jiangxin Dong, and Jinshan Pan†

School of Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing, China
{mingjunzheng, cs.longsun, jxdong, jspan}@njust.edu.cn

Overview

In this document, we further implement ablations to demonstrate more analysis on the SMFANet in Section 1, and present the effect of channel number and FMB number in Section 2. Next, we further compare the performance and efficiency metrics with recent state-of-the-art approaches on $\times 4$ SR in Section 3, and show present more comparisons of the LAM visualization in Section 4. To fully demonstrate the non-local modeling capability, we additionally show the comparison results of the effective receptive fields in Section 5. We show more spectral density visualizations of the output features from the EASA and LDE branches in Section 6. Finally, we provide more visual comparison results between our SMFANet series and CNN- or ViT-based methods in Section 7, and give a user study in Section 8.

1 Further Ablation and Analysis

In this section, we first conduct experiments to demonstrate the effect of input projection. As we project the input feature F_{in} into feature X and Y for feature modulation, one may wonder whether using a single projected feature directly would generate better results or not. To answer this question, we replace the channel splitting operation with feature duplication in the proposed network and train it using the same settings as ours for fairness. Table 1 shows that using a single projected feature does not generate better results, indicating that using separate inputs drives the proposed EASA and LDE to learn better feature representations. We then investigate the effect of fusing X_l and Y_d with channel concatenation. Table 1 shows that the performance of using channel concatenation is almost unchanged compared to the baseline model that uses addition for feature fusion. Thus, we use addition to feature X_l and Y_d with less model complexity.

* Equal contribution
† Corresponding author.

Table 1: Ablation experiments for ×4 SMFANet. All measured metrics are calculated in the same way as in Table 1 of the main paper. The experimental setup follows the same settings as those used in Section 5 of the main paper.

	Ablation	#Params(K)	#FLOPs(G)	Set5	Set14	B100	Urban100	Manga109
Baseline	-	197.33	10.88	32.25/0.8956	28.67/0.7825	27.61/0.7371	26.19/0.7861	30.72/0.9097
Input projection	Splitting → Duplication	186.67	10.25	32.23/0.8955	28.64/0.7822	27.60/0.7368	26.13/0.7838	30.63/0.9088
Feature fusion	$X_l + Y_d \rightarrow \mathcal{C}([X_l, Y_d])$	207.69	11.51	32.32/0.8962	28.67/0.7824	27.62/0.7376	26.18/0.7860	30.73/0.9099
Normalization	L2 Norm → LayerNorm	198.48	10.97	32.29/0.8959	28.67/0.7827	27.63/0.7373	26.20/0.7866	30.74/0.9100

Furthermore, we study the effect of L2 norm. The L2 norm in the SMFA layer is used to prevent the training instability problem caused by the attention operator in Equation (5). In a ViT architecture, this step is typically performed by layer normalization (LN). Table 1 shows that replacing L2 norm with LN keeps performance largely unchanged, while introducing additional parameters and slowing down the inference (12.34ms v.s. 9.26ms).

2 Effect of the Channel Number and FMB Number

We show the effect of channel number and FMB number with ×4 SMFANet in Table 2. It is observed that PSNR/SSIM values are positively correlated with these two hyperparameters. For the number of channels, although the performance keeps increasing, the total number of parameters is growing rapidly. As regards the number of FMBs, the performance gain becomes saturated gradually with the increase of its number. To balance performance and model size, we choose 36 as the channel number and 8 as the FMB number.

Table 2: Effect of channel number and FMB number on performance. #Num is the number of channel dimensions or FMBs.

Ablation	#Num	#Params(K)	#FLOPs(G)	Set5	Set14	B100	Urban100	Manga109
Channel number	20	68.05	3.68	31.95/0.8911	28.44/0.7777	27.47/0.7329	25.74/0.7732	30.07/0.9020
	28	124.44	6.81	32.15/0.8942	28.52/0.7789	27.53/0.7351	25.98/0.7819	30.36/0.9064
	36	197.33	10.88	32.25/0.8956	28.67/0.7825	27.61/0.7371	26.19/0.7861	30.72/0.9097
	44	286.93	15.89	32.35/0.8965	28.71/0.7836	27.65/0.7386	26.25/0.7887	30.83/0.9114
FMB number	4	106.96	5.93	32.00/0.8918	28.49/0.7786	27.51/0.7338	25.83/0.7750	30.18/0.9027
	8	197.33	10.88	32.25/0.8956	28.67/0.7825	27.61/0.7371	26.19/0.7861	30.72/0.9097
	12	287.68	15.82	32.33/0.8964	28.75/0.7845	27.66/0.7386	26.32/0.7907	30.90/0.9118
	16	378.05	20.77	32.37/0.8970	28.76/0.7846	27.68/0.7392	26.40/0.7933	30.96/0.9127

3 Further Comparisons with Recent SOTA Methods

We further compare the reconstruction quality and efficiency metrics with recent state-of-the-art methods on ×4 SR, including VapSR [17], MDRN [11], LKDN [15], Omni-SR [14], DAT-light [1]. Table 3 shows that SMFANet+ obtains a more favorable trade-off between reconstruction results and model efficiency. Our SMFANet+ (DF2K) achieves similar performance to DAT-light (DF2K) on five test sets, but only uses **40**% of the GPU usage and is ×**20** times faster.

Table 3: Comparisons with recent lightweight methods on ×4 SR.

	Methods	#Params [K]	#FLOPs [G]	#GPU Mem. [M]	#Avg. Time [ms]	Set5	Set14	B100	Urban100	Manga109
CNN	VapSR [17] (DF2K)	342	21	90	23.74	32.38/0.8978	28.77/0.7852	27.68/0.7398	26.35/0.7941	30.89/0.9131
	MDRN [11] (DIV2K+LSDIR)	322	17	328	33.92	32.35/0.8970	28.80/0.7861	27.69/0.7404	26.60/0.8005	31.02/0.9146
	LKDN [15] (DF2K)	322	18	260	15.97	32.39/0.8979	28.79/0.7859	27.69/0.7402	26.42/0.7965	30.97/0.9140
ViT	Omni-SR [14] (DIV2K)	792	46	265	49.40	32.49/0.8988	28.78/0.7859	27.71/0.7415	26.64/0.8018	31.02/0.9151
	DAT-light [1] (DF2K)	573	50	648	362.44	32.57/0.8991	28.87/0.7879	27.74/0.7428	26.64/0.8033	31.37/0.9178
Ours	SMFANet+ (DIV2K)	496	28	259	17.39	32.43/0.8979	28.77/0.7849	27.70/0.7400	26.45/0.7943	31.06/0.9138
	SMFANet+ (DF2K)	496	28	259	17.39	32.51/0.8985	28.87/0.7872	27.74/0.7412	26.56/0.7976	31.29/0.9163

4 Comparisons of Local Attribution Maps

The LAM [4] performs attribution analysis on SR networks, finding pixels (red labeled) in the input LR image that contribute to the reconstruction of a certain location/area (**rectangular patch**) in the output image. The Diffusion Index (DI)is used to indicate the range of involved pixels, which is defined by $DI = (1 - G) \times 100$, where G is the Gini coefficient. Figure 1 and 2 show that our proposed approaches obtain larger diffusion index values compared to previous CNN- or ViT-based lightweight SR methods [5, 7, 9, 12, 13, 16], reflecting that the proposed models can exploit more information for accurate image reconstruction.

5 Comparisons of Effective Receptive Field

To fully demonstrate the non-local modeling capability of our proposed method, we also show visualization results of the effective receptive fields [3]. We compare the ERFs of the proposed SMFANet family with EDSR-baseline [8], ShuffleMixer [13], SAFMN [12], SwinIR-light [7], NGswin [2], SRFormer-light [18] on $\times 4$ SR. Figure 6 evidences that our SMFANet family have favourable long-range modeling capabilities.

6 More Visualizations of the Power Spectral Density

We present spectral density maps for different scenes to intuitively illustrate the complementary roles of the EASA and LDE. Figure 3 shows the power spectral density map of deep features from SMFANet and SMFANet+. Compared to the input feature F_{in}, after processing by the EASA branch, the feature X_l has significantly stronger energy in the central region of the PSD map, and the spectrum of feature Y_d spreads more widely towards the margins after processing by the LDE branch. These results suggest that EASA effectively exploits non-local information to enable accurate low-frequency representations, while LDE captures local details to enhance high-frequency components.

7 More Visual Results

In this section, we present additional visual comparisons with state-of-the-art methods [2, 5–10, 12, 13, 16] on $\times 4$ benchmark datasets. Figures 4-5 show that the proposed SMFANet family generates clearer images with finer details and structures than those by the evaluated CNN-based or ViT-based SR methods.

8 User study

For a further comprehensive comparison, we conduct a user study to evaluate the $\times 4$ benchmark results. We compare our SMFANet+ with ELAN-light [16], SwinIR-light [7], and SRFormer-light [18]. We invite a total of 10 participants to take part in this user study. Each volunteer was presented with a set of 20 randomly selected images, which included an input image, the results of the listed comparison methods, and our results. Figure 7 shows that the volunteers preferred our results over the other methods.

References

1. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: ICCV (2023)
2. Choi, H., Lee, J., Yang, J.: N-gram in swin transformers for efficient lightweight image super-resolution. In: CVPR (2023)
3. Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., Sun, J.: Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In: CVPR (2022)
4. Gu, J., Dong, C.: Interpreting super-resolution networks with local attribution maps. In: CVPR (2021)
5. Hui, Z., Gao, X., Yang, Y., Wang, X.: Lightweight image super-resolution with information multi-distillation network. In: ACM MM (2019)
6. 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. In: NeurIPS (2020)
7. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: SwinIR: Image restoration using swin transformer. In: ICCV Workshops (2021)
8. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR Workshops (2017)
9. Liu, J., Chen, C., Tang, J., Wu, G.: From coarse to fine: Hierarchical pixel integration for lightweight image super-resolution. In: AAAI (2023)
10. Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T.: Transformer for single image super-resolution. In: CVPR Workshops (2022)
11. Mao, Y., Zhang, N., Wang, Q., Bai, B., Bai, W., Fang, H., Liu, P., Li, M., Yan, S.: Multi-level dispersion residual network for efficient image super-resolution. In: CVPR Workshops (2023)
12. Sun, L., Dong, J., Tang, J., Pan, J.: Spatially-adaptive feature modulation for efficient image super-resolution. In: ICCV (2023)
13. Sun, L., Pan, J., Tang, J.: ShuffleMixer: An efficient convnet for image super-resolution. In: NeurIPS (2022)
14. Wang, H., Chen, X., Ni, B., Liu, Y., Liu, j.: Omni aggregation networks for lightweight image super-resolution. In: CVPR (2023)
15. Xie, C., Zhang, X., Li, L., Meng, H., Zhang, T., Li, T., Zhao, X.: Large kernel distillation network for efficient single image super-resolution. In: CVPR Workshops (2023)
16. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: ECCV (2022)
17. Zhou, L., Cai, H., Gu, J., Li, Z., Liu, Y., Chen, X., Qiao, Y., Dong, C.: Efficient image super-resolution using vast-receptive-field attention. In: ECCV Workshops (2022)
18. Zhou, Y., Li, Z., Guo, C., Bai, S., Cheng, M., Hou, Q.: Srformer: Permuted self-attention for single image super-resolution. In: ICCV (2023)

-

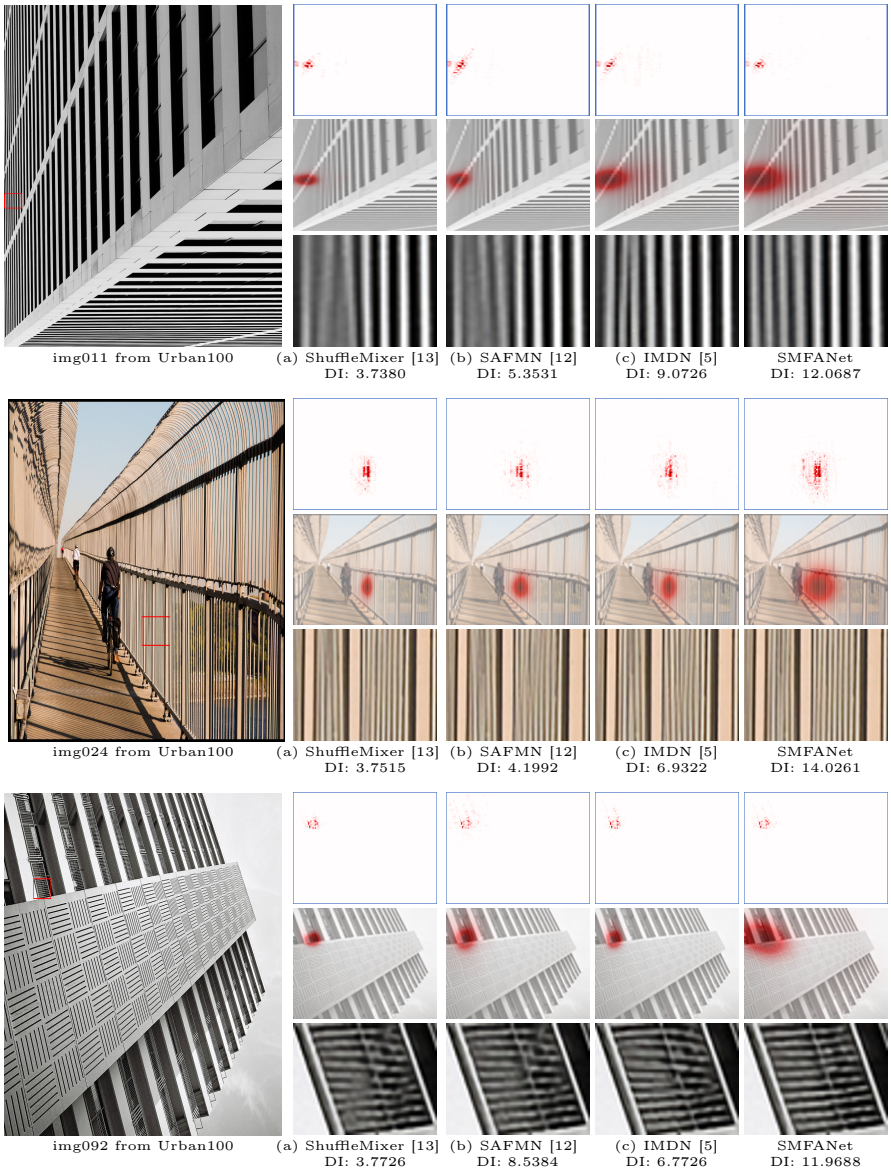


Fig. 1: Comparison of local attribution maps (LAMs) [4] and diffusion indices (DIs) [4] between our SMFANet and other CNN-based lightweight SR models. The proposed model exploits more feature information and achieves better results.

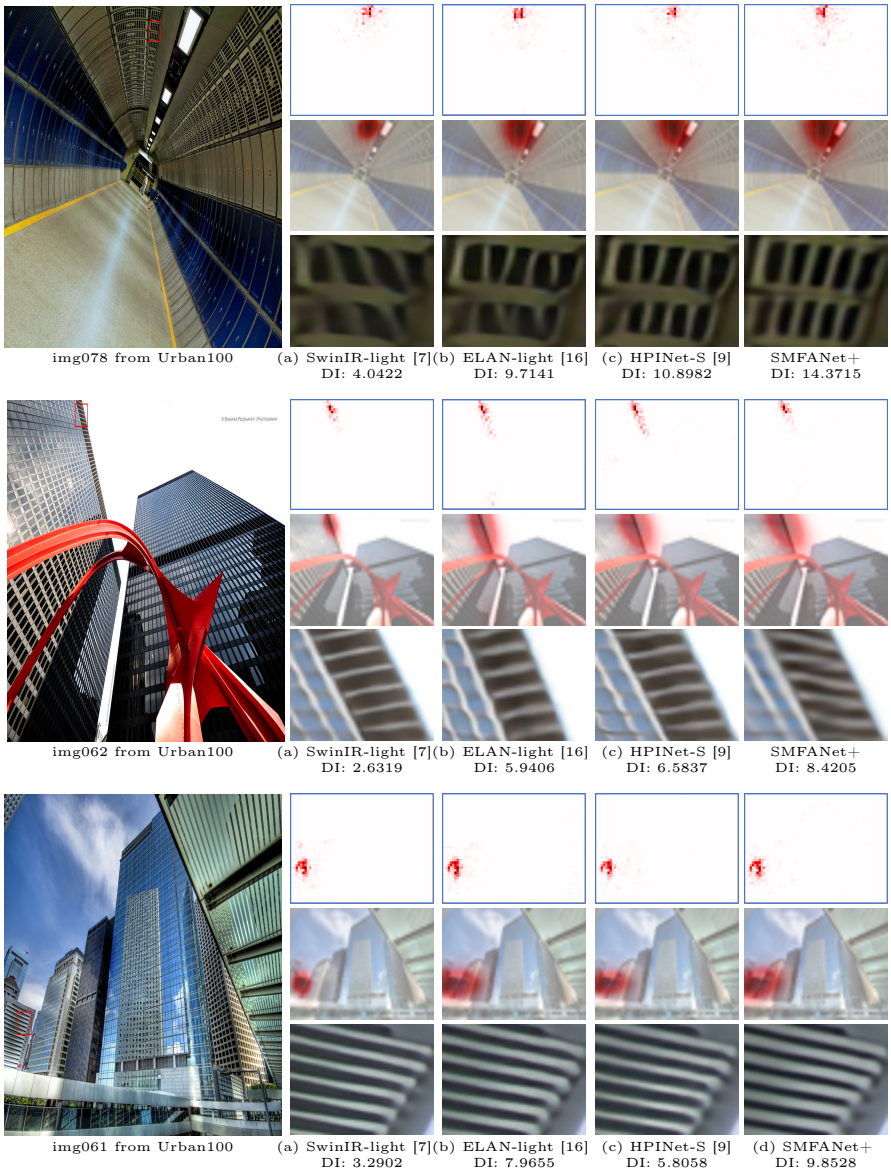


Fig. 2: Comparison of local attribution maps (LAMs) [4] and diffusion indices (DIs) [4] between our SMFANet+ and other ViT-based lightweight SR models. The proposed SMFANet+ exploits more feature information and reconstructs a more accurate image structure.

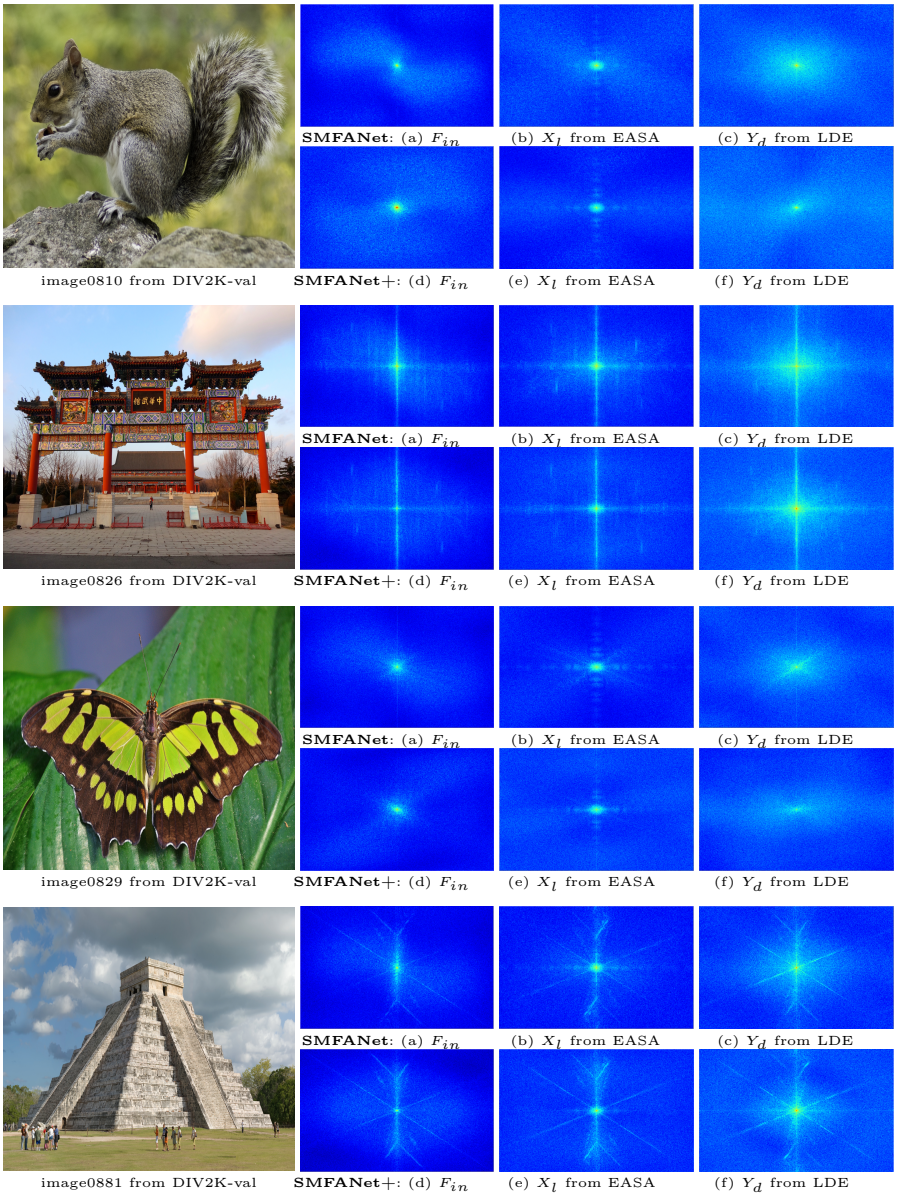


Fig. 3: The power spectral density (PSD) visualizations of feature F_{in}, X_l, Y_d. We perform a periodic shift of the spectrum map such that the low-frequency component is moved to the center. The EASA activates more low-frequency components for feature X_l, and the LDE enhances high-frequency representations for feature Y_d.

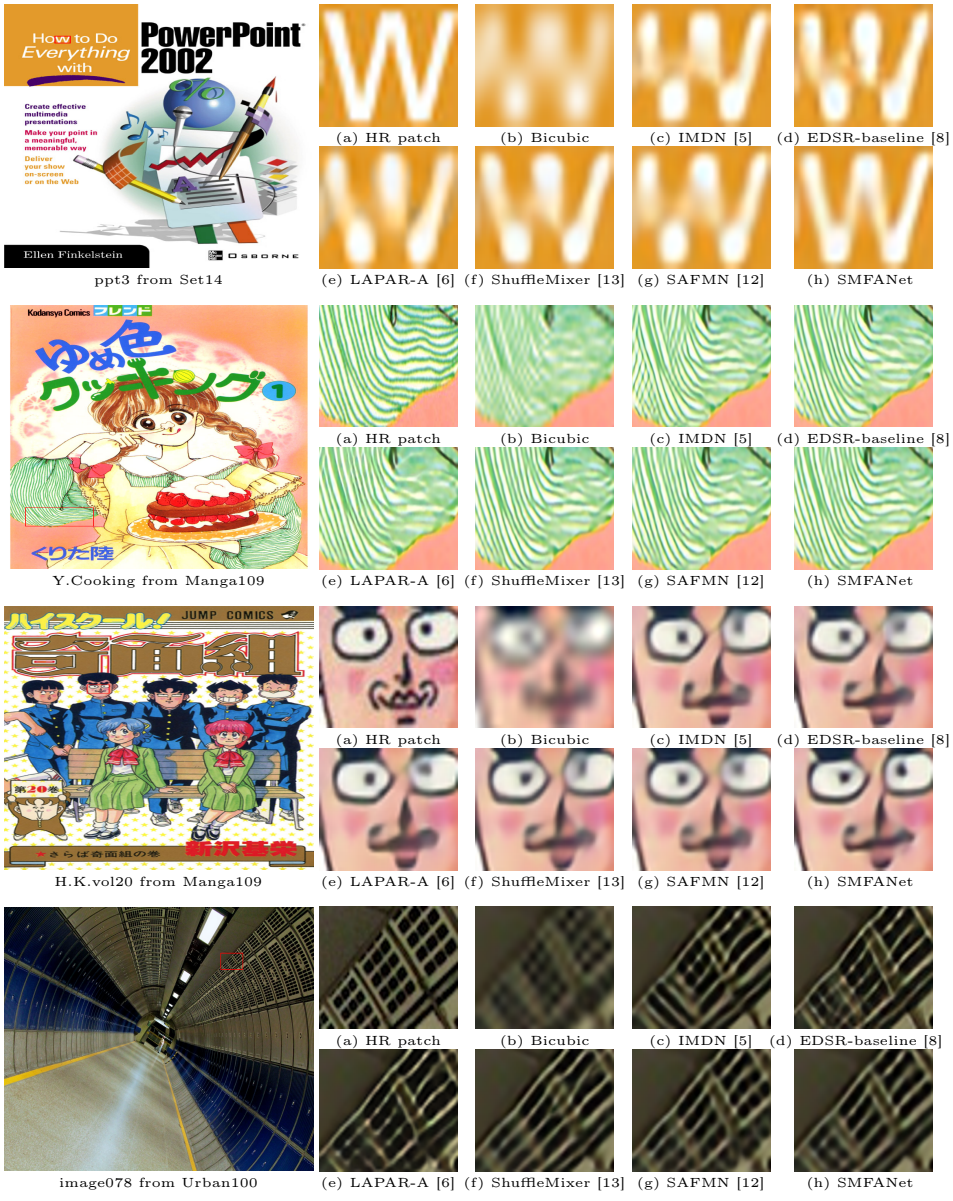


Fig. 4: Visual comparisons with CNN-based Efficient SR models for $\times 4$ SR on benchmark datasets. The proposed SMFANet recovers more accurate results with clearer character, lines and face, as well as finer stripe patterns.

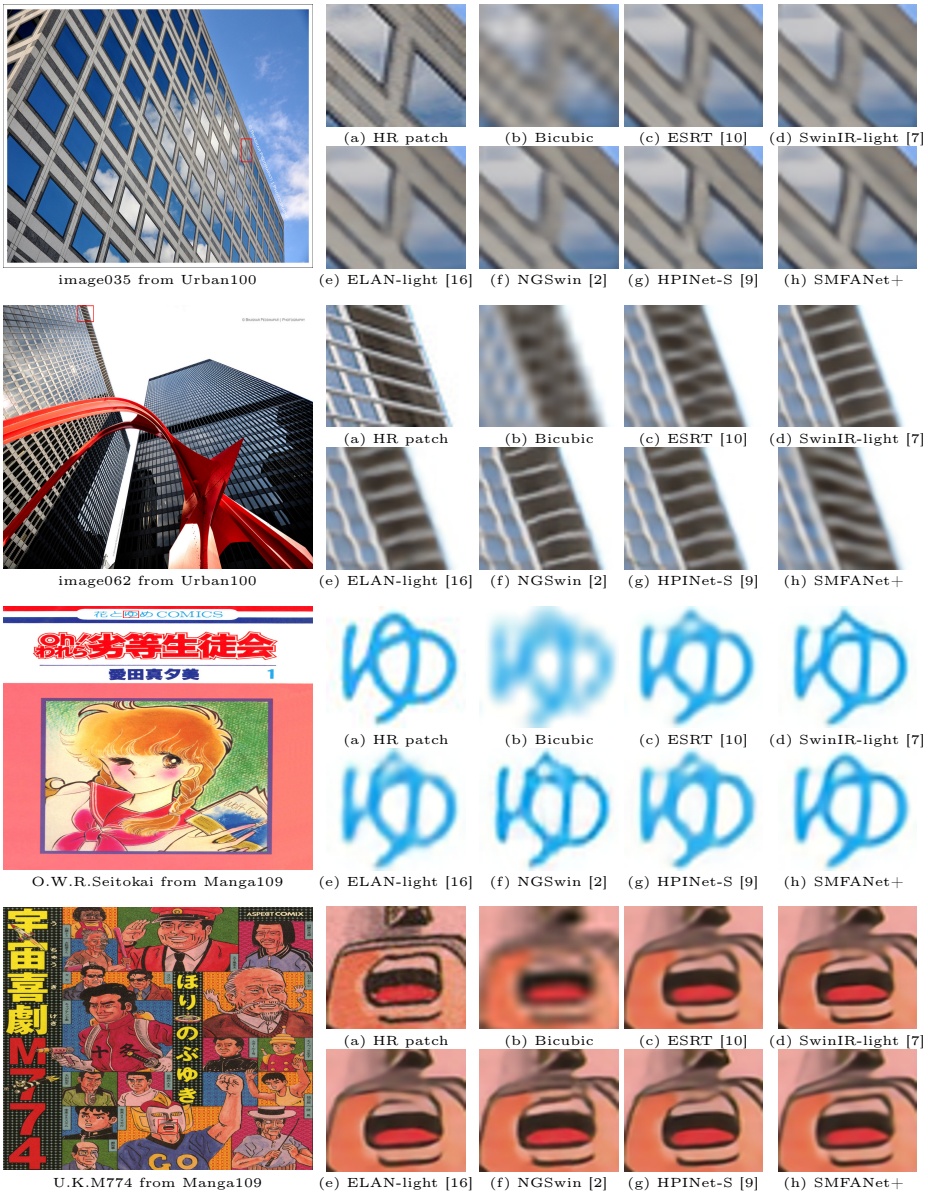
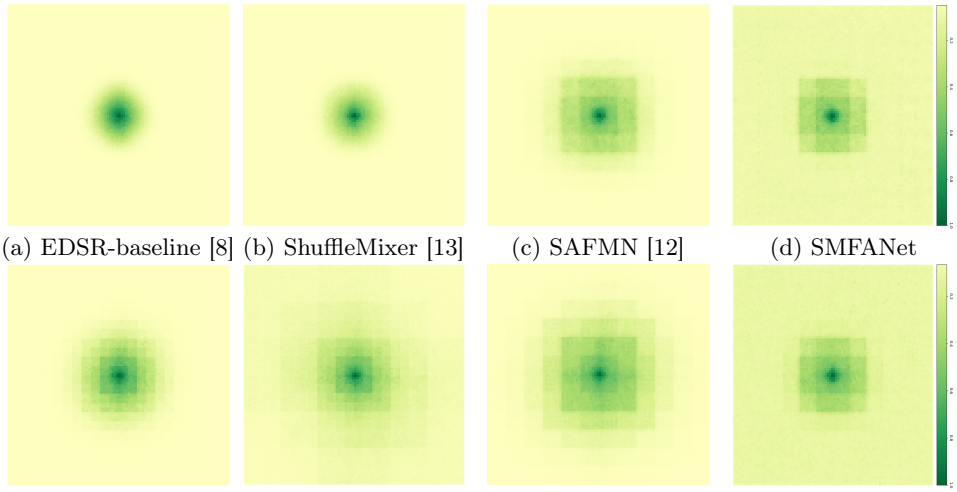


Fig. 5: Visual comparisons with ViT-based lightweight SR models for $\times 4$ SR on benchmark datasets. The proposed SMFANet+ generates more accurate results with clearer lines and mouth contour.



(a) EDSR-baseline [8] (b) ShuffleMixer [13] (c) SAFMN [12] (d) SMFANet

(e) SwinIR-light [7] (f) NGswin [2] (g) SRFormer-light [18] (h) SMFANet+

Fig. 6: Effective receptive field (ERF) [3] visualizations. A more widely distributed dark area represents a larger ERF. Our proposed SMFANet family obtain large ERFs.

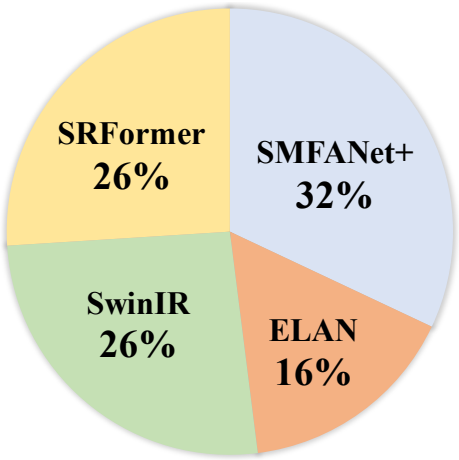


Fig. 7: User study results. Our proposed SMFANet+ is more favored by human voters over other methods.