HMANet: Hybrid Multi-Axis Aggregation Network for Image Super-Resolution

Supplementary Material

1. Training Details

1.1. Study on the pre-training strategy

We calculate the interlayer CKA [3] similarity in $\times 2$ SR, $\times 3$ SR, and $\times 4$ SR, except for the shallow feature extraction and image reconstruction modules. In Fig. 1, we can see that Fig. 1(a) and Fig. 1(c) show high similarity on the diagonal, while Fig. 1(b) has a low similarity score on the diagonal. Therefore, we train the $\times 3$ SR model after training the $\times 2$ SR model as the initial parameter and then use the $\times 3$ SR model as the initial parameter of the $\times 2$ SR model and the $\times 4$ SR model.

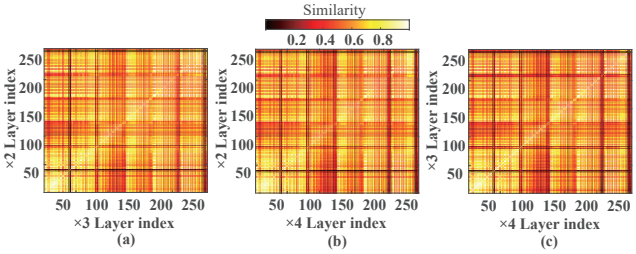

Figure 1. (a) CKA similarity map between layers of the $\times 2$ SR model and the $\times 3$ SR model, (b) CKA similarity map between layers of the $\times 2$ SR model and the $\times 4$ SR model, (c) CKA similarity map between layers of the $\times 3$ SR model and the $\times 4$ SR model.

We train the model using nine pre-training strategies to test the impact of different pre-training strategies on performance. Tab. 1 shows the training results, which are evaluated on the Set5 [1] dataset. We can find that our proposed pre-training strategies can effectively improve the model performance (0.05dB∼0.09dB). It can also be observed that using models with different degradation levels as model initialization parameters has different effects on motivating the model potential. Using the $\times 3$ SR model as the initialization parameter for the $\times 2$ and the $\times 4$ SR models maximizes the model performance. Whereas using the $\times 2$ SR model as the initialization parameter of the $\times 4$ model, on the contrary, reduces the model performance. This suggests that a suitable pre-training strategy can lead to better performance gains for HMA.

Scale	Initialization parameters							
	w/o		×2		×3		×4	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	38.84	0.9642	38.86	0.9644	**38.95**	**0.9647**	38.78	0.964
×3	35.25	0.9342	**35.35**	**0.9346**	35.27	0.9343	**35.30**	0.9345
×4	33.26	0.9083	33.24	0.9081	**33.38**	**0.9086**	33.25	0.9083

Table 1. Quantitative results of HMA PSNR (dB) on $\times 4$ SR using different pre-training strategies.

2. Analysis of Model Complexity

We experiments to analyze Grid Attention Block (GAB) and Fused Attention Block (FAB). We also compare our method with the Transformer-based method SwinIR. The $\times 4$ SR performance on Urban100 is reported and the number of Multiply-Add operations is computed when the input size is 64×64. Note that the pre-training technique is not used for all models in this section.

we use SwinIR with a window size of 16 as a baseline to study the computational complexity of the proposed GAB and FAB. As shown in Tab. 2, our GAB obtains performance gains by finitely increasing parameters and Multi-Adds. It proves the effectiveness and efficiency of the proposed modules. In addition, FAB brings better performance at the same time although it brings more parameters and Multi-Adds.

Method	#Params.	#Multi-Adds.	PSNR
SwinIR	12.1M	63.8G	27.81dB
w/GAB	24.4M	76.9G	28.37dB
w/FCB	57.6M	157.0G	28.30dB
Ours	69.9M	170.1G	28.42dB

Table 2. Model complexity comparison of GAB and FAB.

3. Visual Comparisons with LAM

We provide visual comparisons with the LAM [2] results to compare SwinIR, HAT, and our proposed HMA. The red dots in the LAM results represent the pixels used for reconstructing the patches marked with red boxes in the HR images, and we give the Diffusion Indices (DI) in Fig. 2 to reflect the range of pixels involved. In this case, the more pixels are used to recover a specific input block, the wider the distribution of red dots in LAM, and the higher the DI. As shown in Fig. 2, both HAT and HMA can effectively extend the effective pixel range compared to the baseline SwinIR, where the pixel range is only clustered in a limited area. Compared to HAT, HMA can extend the range of utilized pixels more widely due to the introduction of the

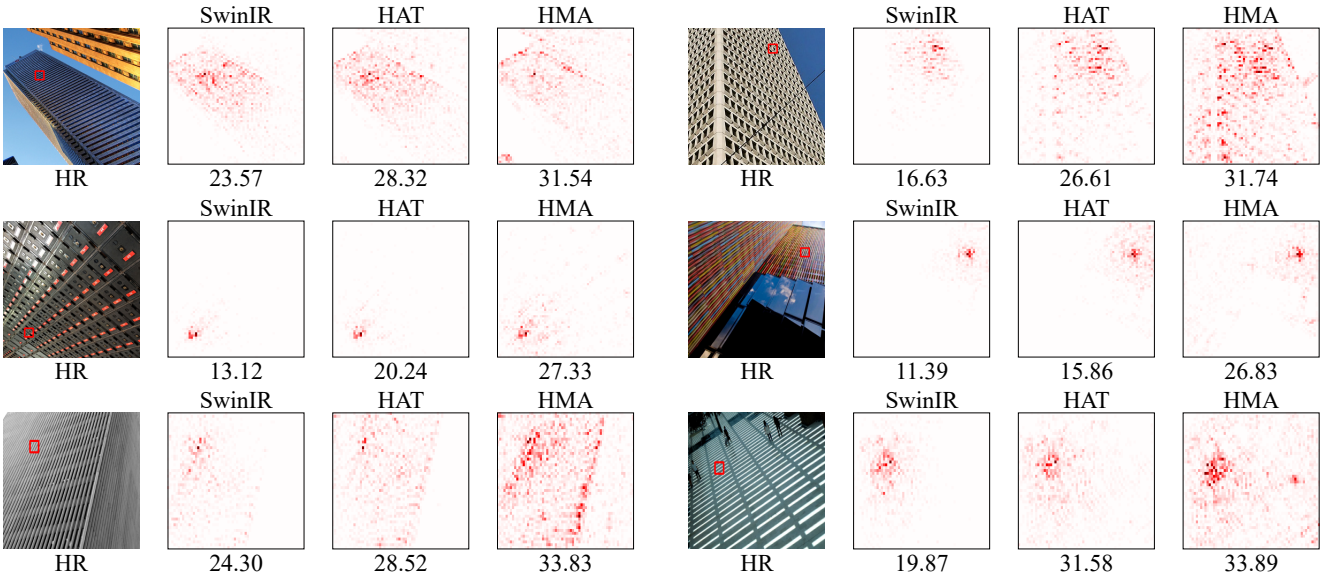

Figure 2. Comparison of LAM results between SwinIR, HAT and HMA.

GAB module. Also, for quantitative metrics, HMA obtains much higher DI values than SwinIR and HAT. The visualization results and quantitative evaluation metrics show that HMA can better utilize global information for local area reconstruction. As a result, the method generated by HMA is more capable of generating high-resolution images with better visualization.

References

[1] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[2] 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*, pages 9199–9208, 2021.

[3] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.