CATANet: Efficient Content-Aware Token Aggregation for Lightweight Image Super-Resolution

Supplementary Material

Table A. Ablation Study on the subgroup size g_s. PSNR and SSIM are calculated with a scale factor of 4. All models are trained 250K on DIV2K [6] from scratch.

g_s	Set5		Urban100	
	PSNR	SSIM	PSNR	SSIM
64	32.43	0.8984	26.67	0.8027
128	32.46	0.8985	26.68	0.8025
256	32.51	0.8989	26.68	0.8030
CATA (Ours)	32.54	0.8990	26.70	0.8033

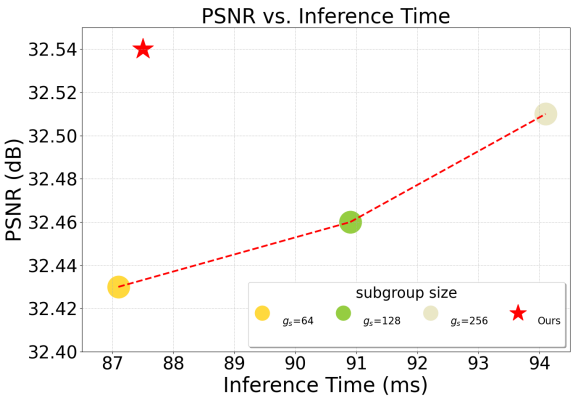

Figure A. Performance and model inference speed comparison on Set5 dataset for upscaling factor $\times 4$. The test output image size is $3 \times 1024 \times 1024$.

The supplementary material is organized as follows. In Sec. A, we provide the implementation details of CATANet. In Sec. B, we conduct more ablation studies on the Token Aggregation Block (TAB) of CATANet and Sec. C provide further analysis to investigate the advantages of our method. Finally, in Sec. D, we present more illustrations of the CATA module and visual examples.

A. Implementation Details

Network hyperparameters. We set the number of Residual Groups K=8, each containing one TAB and one LRSA. For LRSA, the overlapping patch size is set to [16, 20, 24, 28, 16, 20, 24, 28]. For TAB, inspired by the progressive size setting of overlapping patches in LRSA, the number of token centers M and subgroup sizes g_s are set to [16,32,64,128,16,32,64,128],[256,128,64,32,256,128,64,32], respectively. The channel dimension, number of attention heads, and MLP hidden layer dimension of CATANet are set to 40, 4, and 96, respectively. The number of iterations for updating the token centers in each TAB module is set to 5.

Training Details. We train the model with a batch size of 64, where each input image is randomly cropped to a size of 64×64. During the training phase, we applied common data augmentation techniques, including random rotation and horizontal flipping. Following previous work [3, 10], we employ the Adam optimizer [4] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ to minimize the L1 loss. For the case of $\times 2$ zooming factor, we train the model from scratch for 800k iterations. The initial learning rate is set to 2×10^{-4}, and is halved at milestones $[300\text{K}, 500\text{K}, 650\text{K}, 700\text{K}, 750\text{K}]$. For the $\times 3$ and $\times 4$ models, we fine-tuned the well-trained $\times 2$ model for a total of 250K iterations. The initial learning rate is also set to 2×10^{-4}, and is halved at milestones $[125\text{K}, 200\text{K}, 225\text{K}, 237.5\text{K}]$.We use PyTorch [5] to implement our models with 4 Tesla V100 GPUs.

B. More Ablation Studies

Effects of subgroup size g_s. In window attention, a larger window size can provide a larger receptive field, which in turn leads to improved performance. We conduct experiments to explore the influence of varying the subgroup size g_s from 64 to 256 on TAB, as shown in Tab. A. The model performance gradually improves as g_s increases to 256. However, compared to our progressive g_s setting strategy, a fixed g_s of 256 shows inferior performance. This is because our progressive strategy flexibly assigns different g_s to different TAB modules, resulting more effective information capture over different scales. In addition, our progressive g_s strategy results in a faster inference speed compared to setting g_s of all TABs to 256, as illustrated in Fig. A.

Effects of Global Token Centers. In Tab. B, we perform an ablation study regarding whether to learn token centers from each image, and our method achieves better performance. This is because learning the token center for each image separately may lead to drastic changes in the token center during training, and such drastic changes increase the learning difficulty of the model. We use EMA to ensure that the token center will not change drastically to ensure the stability and consistency of model learning. This method is also similar to updating the mean and variance using EMA in BatchNorm. This also makes our model does't need to re-learn the token center in inference, improving inference speed.

Table B. Ablation Study on different types of token centers.

Token Centers	Params	Multi-Adds	Set5	Set14	B100	Urban100	Manga109
Individual	536K	46.8G	32.49	28.75	27.76	26.84	31.27
Global (Ours)	536K	46.8G	32.58	28.90	27.75	26.87	31.31

Table C. Ablation Study on LRSA module.

LRSA	Params	Multi-Adds	Set5	Set14	B100	Urban100	Manga109
✗	403K	37.1G	32.26	28.69	27.62	26.36	30.75
✓	536K	46.8G	32.58	28.90	27.75	26.87	31.31

Effects of LRSA. In Tab. C, we conducted an ablation study on the LSRA module. As shown, local attention also plays a key role in the image SR task, which provides fine local information modeling.

C. Further Analyses

C.1. LAM analyses

In this section, we show more LAM [2] analysis with the state-of-the-art lightweight SR methods, including RCAN [9], SwinIR-light [3] and SRFormer-light [10]. LAM can show the pixels that contribute the most to the reconstruction of a selected region, and the corresponding pixels are marked in red. More marked pixels mean that the model can use more information and achieve better performance. As shown in Fig. B, our CATANet has the most labeled pixels. It indicates that our CATANet has larger receptive fields and utilizes more information to restore image. This is because our method can efficiently aggregate more non-localized pixels via Token Aggregation Block.

C.2. Perceptual Similarity Analyses

The paper [1] reveals that the superiority of PSNR values does not always accord with better visual quality. To further evaluate our method, we introduce metric LPIPS [8]. Compared to PSNR, LPIPS is more alignable with human perception. As shown in Tab. D, our CATANet achieves the best performance (lowest value) on all datasets. This result demonstrates the superiority of our method.

D. More Visual Examples

D.1. Qualitative Comparison

In this section, we show some visual examples of different methods under scaling factor $\times 4$,[1] as shown in Fig. C and Fig. D[2]. These images clearly demonstrate our advantage in recovering sharp edges and clean textures from severely degraded LR input.

D.2. CATA Visualization

In this section, we provide more examples of visualizations from Content-Aware Token Aggregation (CATA) module. We visualize only a few groups for each input image for simplicity. As shown in Fig. E and Fig. F, these token grouping results indicate that our CATA module is capable of grouping multiple tokens based on their content similarity.

[1] No Comparison with SPIN Due to Lack of Pretrained Weights and Test Results.

[2] CRAN Did Not Provide Pretrained Weights and Manga109 Test Results.

Table D. Comparison (LPIPS) with the state-of-the-art methods for image SR. Best and second best results are colored with red and blue.

Method	Scale	Params	Set5	Set14	B100	Urban100	Manga109
SwinIR-light [3]	×2	878K	0.0865	0.1368	0.1561	0.1065	0.0505
SRFormer-light [10]	×2	853K	0.0859	0.1361	0.1555	0.1050	0.0496
OmniSR [7]	×2	785K	0.0900	0.1381	0.1581	0.1082	0.0526
CATANet (Ours)	×2	477K	0.0843	0.1360	0.1545	0.1036	0.0493
SwinIR-light [3]	×3	886K	0.1589	0.2339	0.2681	0.2045	0.1116
SRFormer-light [10]	×3	861K	0.1576	0.2322	0.2655	0.2022	0.1107
OmniSR [7]	×3	793K	0.1612	0.2340	0.2688	0.2048	0.1136
CATANet (Ours)	×3	550K	0.1560	0.2304	0.2649	0.1991	0.1085
SwinIR-light [3]	×4	897K	0.2071	0.3002	0.3459	0.2786	0.1633
SRFormer-light [10]	×4	873K	0.2063	0.3000	0.3439	0.2747	0.1615
OmniSR [7]	×4	805K	0.2136	0.3017	0.3484	0.2783	0.1654
CATANet (Ours)	×4	477K	0.2048	0.2973	0.3418	0.2683	0.1602

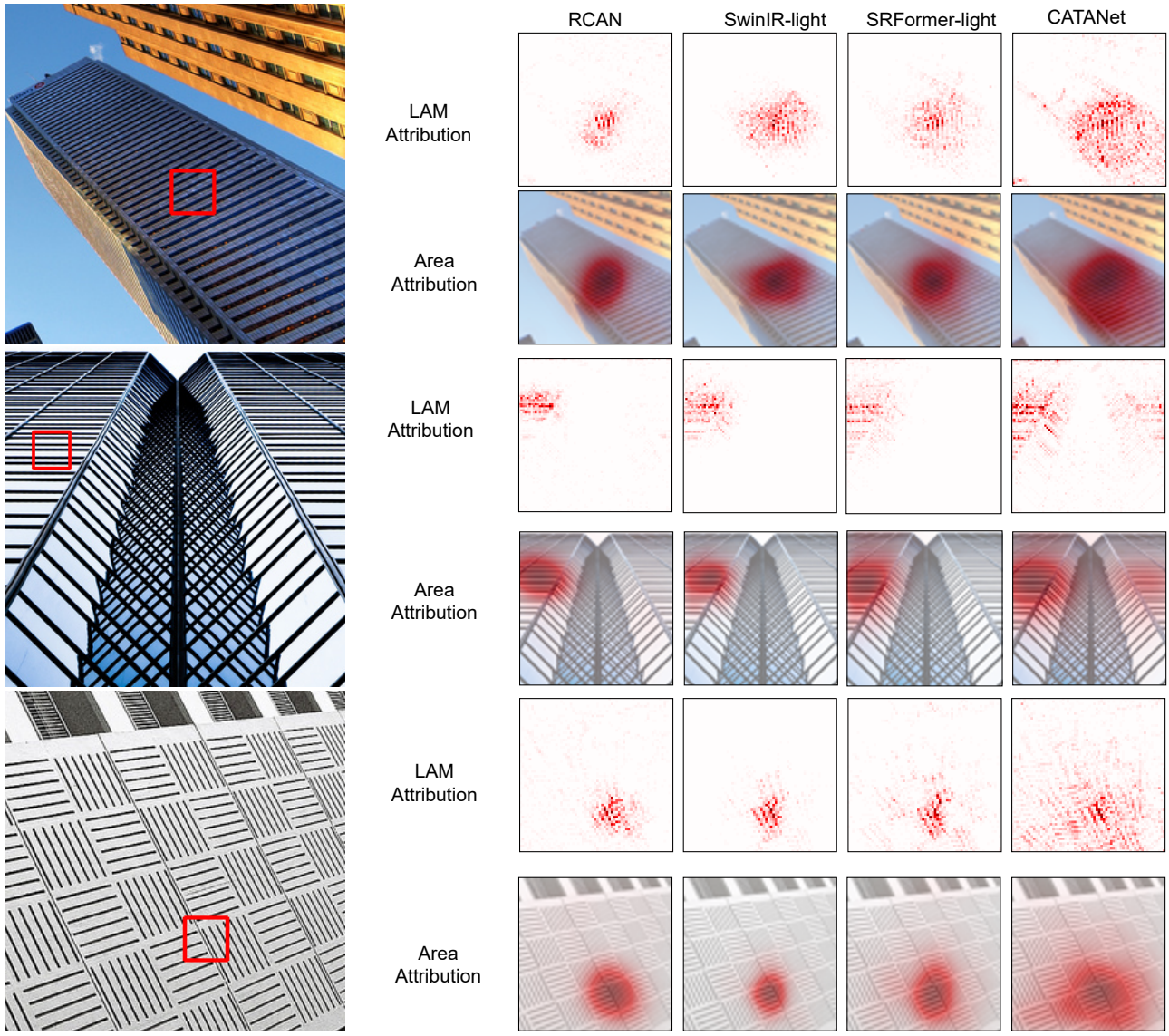

Figure B. LAM Comparison: RCAN [9], SwinIR-light [3], SRFormer-light [10] and CATANet (Ours) for $\times 4$ SR.

Figure C. Visual comparisons of CATANet and other state-of-the-art lightweight SR methods. Metrics (PSNR/SSIM) are calculated on each patch. Best and second best results are colored with red and blue, respectively.

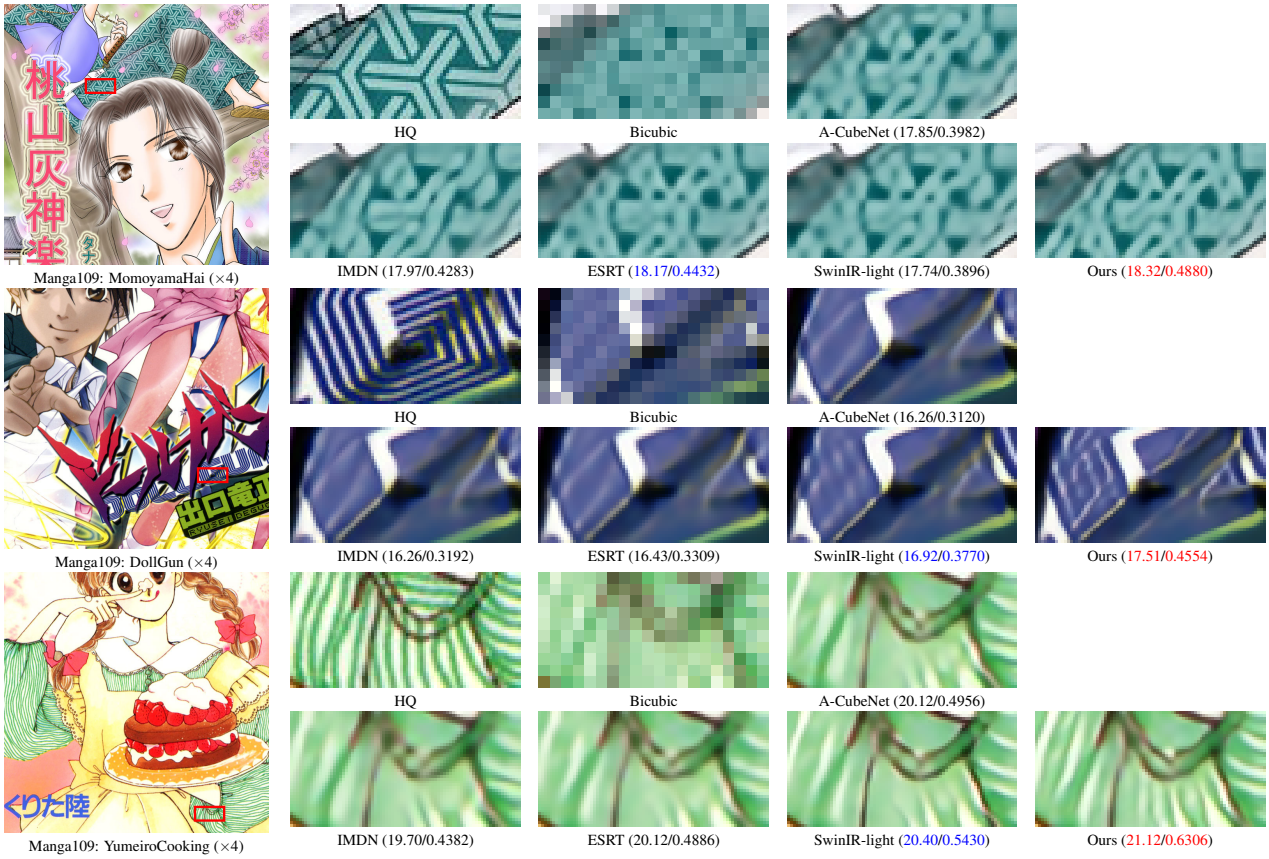

Figure D. Visual comparisons of CATANet and other state-of-the-art lightweight SR methods. Metrics (PSNR/SSIM) are calculated on each patch. Best and second best results are colored with red and blue, respectively.

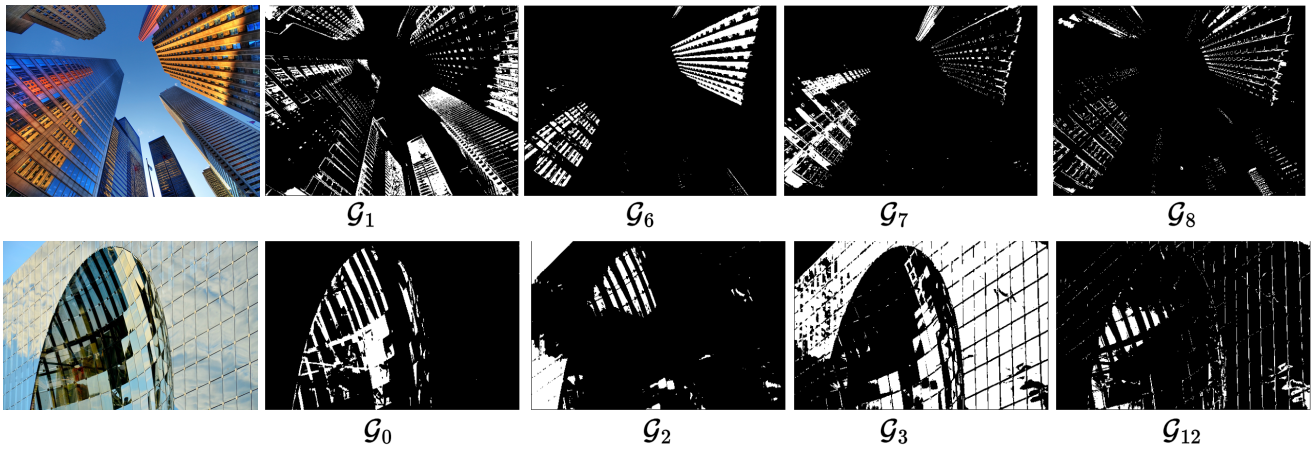

Figure E. Visualization of token grouping examples on Urban100. The white area in each binarized image denotes a single group.

Figure F. Visualization of token grouping examples on Manga109. The white area in each binarized image denotes a single group.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 2

[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. 2

[3] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 2, 3, 4

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[5] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[6] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 1

[7] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387, 2023. 3

[8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2

[9] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2, 4

[10] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12780–12791, 2023. 1, 2, 3, 4