CFAT: Unleashing Triangular Windows for Image Super-resolution -:Supplementary Material:-

Abhisek Ray[1] Gaurav Kumar[1] Maheshkumar H. Kolekar[1]
[1]Indian Institute of Technology Patna, India
{rayabhisek0610, gaurav19.iitp}@gmail.com, mahesh@iitp.ac.in

In this supplementary material, we provide additional information about model analysis and experimental results of Composite Fusion Attention Transformer (CFAT). We discuss the performance of our architecture and its variants along with their complexity in Sec. 1. In Sec. 2, we present the extensive results of an ablation study related to CFAT. We also compare the proposed model with various transformer-based state-of-the-art architectures graphically and based on LAM score [2] in the Sec. 3.

1. Extensive Model Analysis

We examine how the performance of our model varies alongside changes in complexity under two different settings: (i) channel size variation and (ii) model size variation. All the performances are evaluated on the BSD100 dataset for a scale of $\times 4$. In the first setting, we evaluate the complexity and the performance by increasing the value of channel counts from $96 \rightarrow 144 \rightarrow 180 \rightarrow 192$ as shown in Tab. 1. We observe a significant rise in performances (both PSNR and SSIM) when we widen the channel counts from $96 \rightarrow 144 \rightarrow 180$. However, it comes with a computational burden in terms of parameter counts and Multiply-Add operations. After 180, we observe a saturation in performance, e.g., only a 0.01dB improvement in PSNR when the channel is increased from 180 to 192.

After the channel-centric evaluation of CFAT, we set the channel count to 180 in our final model and compared it with various state-of-the-art architectures of the same channel count, as shown in Fig. 1. CFAT achieves the best performance and also maintains an excellent trade-off between parameter count and number of multiply-add operations. The Multiply-Add operations in HAT [1] and ART [6] are too high with moderate parameter counts, whereas SwinIR [4], EDT [3], and ACT [5] show the opposite trends. We observe an identical performance exhibited by ACT, ART, and HAT with little variations whereas the outcomes from EDT and SwinIR are comparively lesser.

We consider Dense Window Attention Blocks (DWAB) and Sparse Window Attention Blocks (SWAB) to be the basic units of CFAT and termed Window Attention Blocks (WAB). We compose three architectures, CFAT-l (large), CFAT (medium), and CFAT-r (reduced), based on model depth, i.e., the number of WAB units present in CFAT. We take (8, 8, 8, 8, 8, 8, 8, 8) WAB units for CFAT-l, (8, 8, 8, 8, 8) for CFAT, and (8, 8, 8) for CFAT-r. Here, '8'

Table 1. Analysis of CFAT based on channel counts.

Channels	Params (M)	Multi-Adds (G)	PSNR/SSIM
192	25.01	102.6	28.18dB/0.7524
180	22.07	90.59	28.17dB/0.7524
144	14.35	59.22	27.99dB/0.7504
96	6.74	28.18	27.78dB/0.7469

Table 2. Analysis of CFAT based on model size.

Models	Params (M)	Multi-Adds (G)	PSNR/SSIM
CFAT-l	34.89	142.08	28.25dB/0.7531
CFAT	22.07	90.59	28.17dB/0.7524
CFAT-s	14.35	59.22	27.99dB/0.7504
CFAT-r	13.52	56.27	27.93dB/0.7498

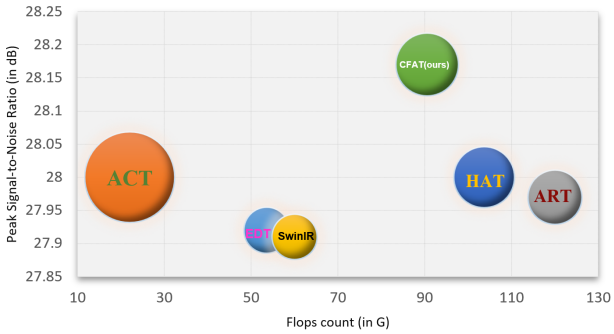


Figure 1. Performance vs Complexity plot of CFAT compare to other state-of-the-art models.
Performance: PSNR (on X-axis) in dB. **Complexity:** Flops (on Y-axis) in G and Parameters (area of the circle) in M

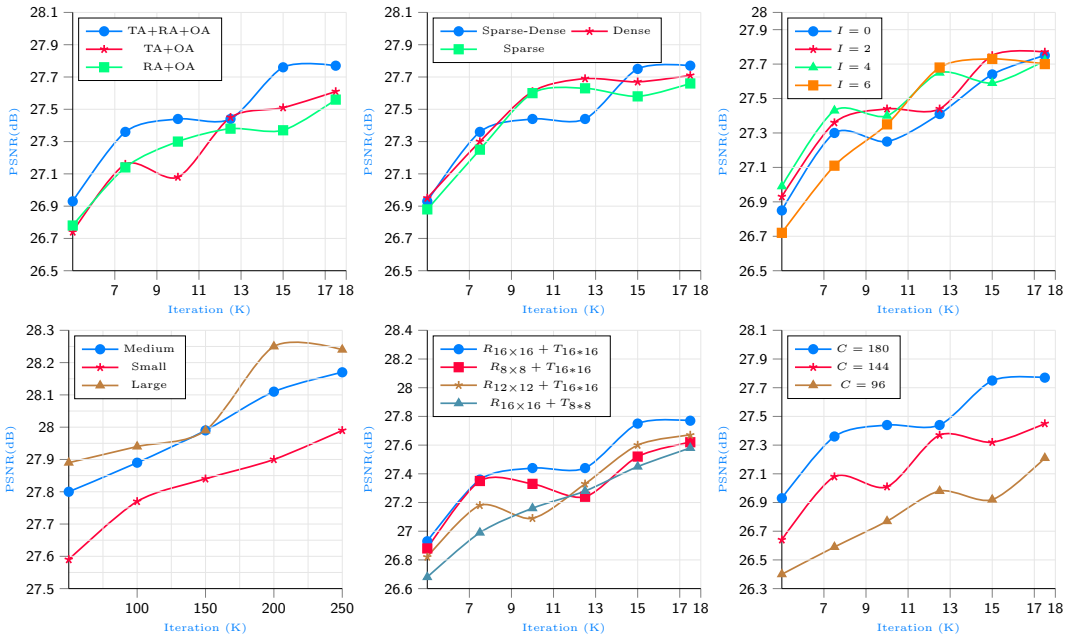
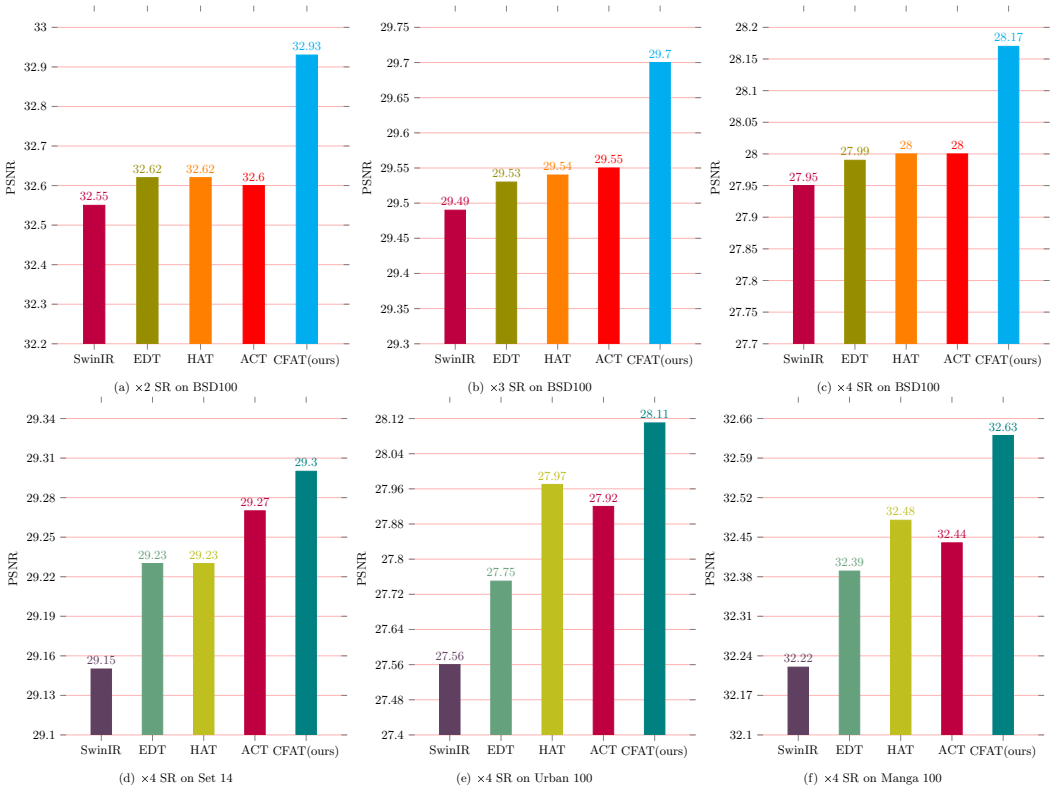


Figure 2. Iterative performance (PSNR in dB) comparison of the proposed CFAT for **Top-Left:** triangular vs rectangular vs overlapping attention, **Top-Middle:** sparse vs dense attention, **Top-Right:** various interval size, **Bottom-Left:** small vs medium vs large CFAT model, **Bottom-Middle:** various combinations of rectangular (8×8, 12×12, 16×16) with triangular ($8 * 8$, $16 * 16$) windows, and **Bottom-Right:** various channel lengths. [on BSD100($\times$4) for epoch 70]



(a) ×2 SR on BSD100 (b) ×3 SR on BSD100 (c) ×4 SR on BSD100

(d) ×4 SR on Set 14 (e) ×4 SR on Urban 100 (f) ×4 SR on Manga 100

Figure 3. Comparing performance (PSNR in dB) of various state-of-the-art models with CFAT on **Top-Left:** BSD100 for scale 2, **Top-Middle:** BSD100 for scale 3, **Top-Right:** BSD100 for scale 4, **Bottom-Left:** Set14 for scale 4, **Bottom-Middle:** Urban100 for scale 4, and **Bottom-Right:** Manga109 for scale 4.

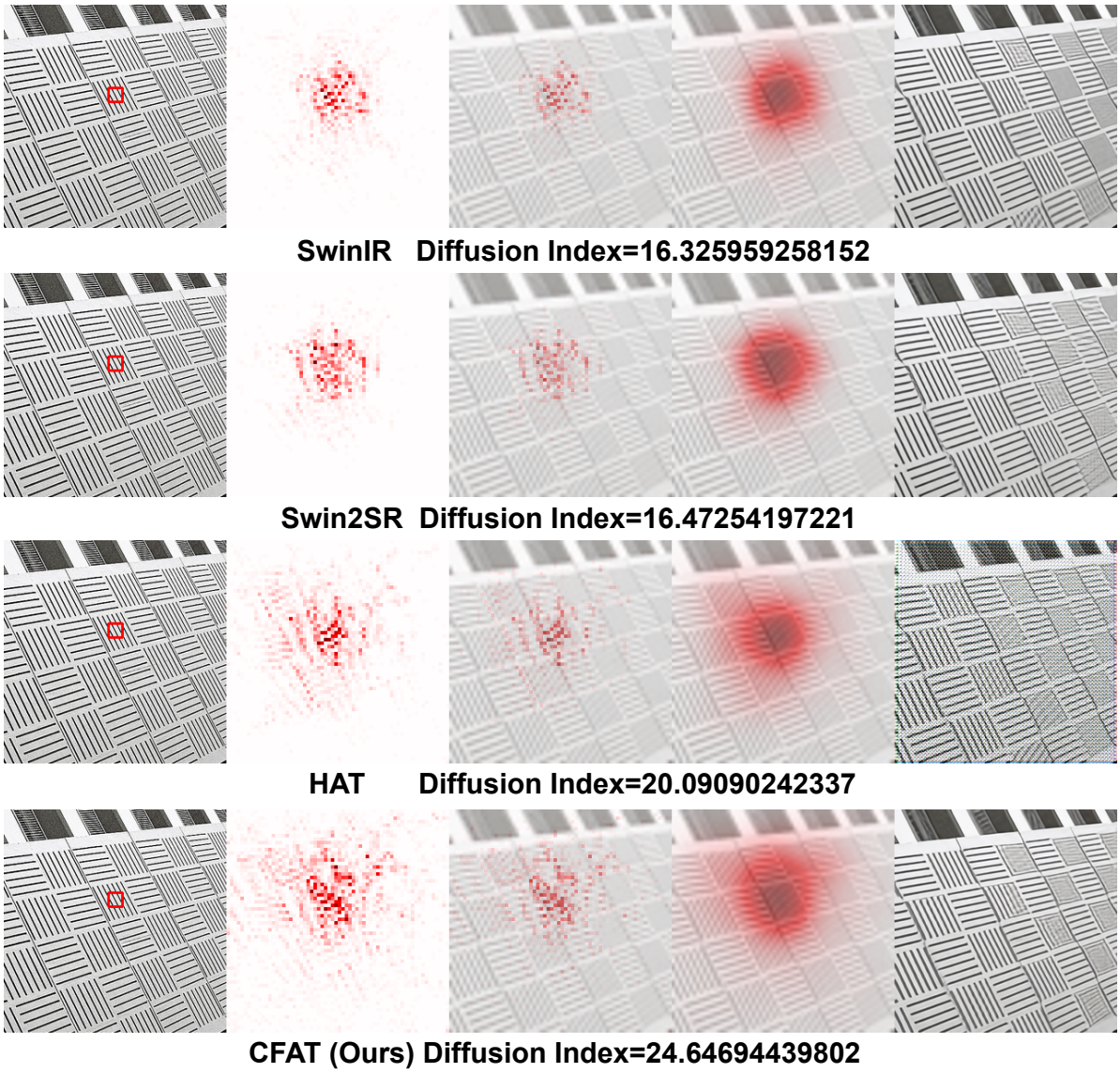


Figure 4. LAM results and corresponding Diffusion Index for CFAT and various SOTA methods.

signifies '4' pairs of Shifted-Dense Rectangular Window MSA ((SD)RW-MSA) and Shifted-Dense Triangular Window MSA ((SD)TW-MSA) arranged in an alternative fashion. We also consider another CFAT variant with 144 channels to compete with CFAT-r while finalizing our small variant. The corresponding parameters, Multiply-Add operations, and performances are displayed in Tab. 2. To finalize the small version, we prioritize more on performance over complexity. This table shows that CFAT with 144 channels yields higher performance than CFAT-r, while both possess identical complexity levels. Therefore, we designate CFAT with 144 channels as our small variant, CFAT-s.

2. Extensive Ablation Study

In this section, we investigate the performance variation of CFAT under the influence of different hyperparameters and model units. We evaluate all the model variants on the BSD100 dataset for scale factor $\times 4$ at distinct training iterations. We plot these outcomes in the X-axis along with their respective iterations in the Y-axis as displayed in Fig. 2. A large deviation in performance is observed when we evaluate the model within the first 17.5k iterations. Therefore, we adopt an averaging technique to keep the results steady at iterations of 5k, 7.5k, 10k, 12.5k, 15k, and 17.5k. The averaging technique is implemented within the iteration or epoch range of $\pm 2.5k$ iteration or ± 5 epoch, respectively. The Top-Left plot shows the outcomes for models taking rectangular with overlapping attention, triangular with overlapping attention and rectangular, triangular with overlapping attention. We find that the last configuration yields the best results. The Top-Middle plot displays the significance of the combined dense-sparse attention-based model

over isolated attention-based models. The Top-Right plot justifies selecting the interval size as '2' over others. The Bottom-Left plot maps out the performance of three CFAT-variants: CFAT-l, CFAT, and CFAT-s. Based on performances in the Bottom-Middle plot, we decide the best combination of window sizes for rectangular- and triangular-window attention. We map the model outcomes for three-channel counts (180, 144, and 96) in the Bottom-Right plot.

3. Extended Comparison with SOTA Architectures

In this section, we evaluate and compare the performance of the proposed architecture with other transformer-based state-of-the-art models. The top three graphs (Top-Left, Top-Middle, and Top-Right) of Fig. 3 validate the supremacy of CFAT over SwinIR [4], EDT [3], ACT [5], and HAT [1] super-resolution (SR) architectures on BSD100 testing dataset for scales of $\times 2$, $\times 3$, and $\times 4$. These graphs also justify that our model possesses strong expressive power for every scale of super-resolution. As displayed in the bottom three graphs (Bottom-Left, Bottom-Middle, and Bottom-Right), we also verify the generalizability of CFAT by evaluating the performance on different testing datasets for a fixed scale ($\times 4$). All these performances are expressed in terms of peak-signal-to-noise ratio (PSNR). CFAT yields the highest PSNR values for all the above settings, as shown in this figure.

As visualized in Fig. 4, the LAM attributes [2] and Diffusion Index (DI) [2] of the proposed triangular window-based CFAT yield superior results than other rectangular window-based SOTA methods. To check the model's scalability in a low-data environment, all models are trained on the DIV2K dataset with batch size 16.

References

[1] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377. IEEE/CVF, 2023. 1, 4

[2] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9199–9208. IEEE/CVF, 2021. 1, 4

[3] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 3(7):8, 2021. 1, 4

[4] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *International Conference on Computer Vision (ICCV)*, pages 1833–1844. IEEE, 2021. 1, 4

[5] Jinsu Yoo, Taehoon Kim, Sihaeng Lee, Seung Hwan Kim, Honglak Lee, and Tae Hyun Kim. Enriched cnn-transformer feature aggregation networks for super-resolution. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 4956–4965. IEEE/CVF, 2023. 1, 4

[6] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022. 1