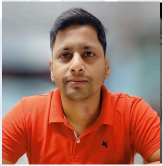
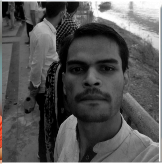
CFAT: Unleashing Triangular Windows for Image Super-resolution


Abhisek Ray[1] Gaurav Kumar[1] Maheshkumar H. Kolekar[1]
[1]Indian Institute of Technology Patna

Motivations & Contributions

- The shifted-rectangular window has limited number of unique shifting modes due to rotational repeatation.
  - We are the first to introduce the triangular window-based self-attention mechanism in the computer vision task that exibits more non-identical shifting modes than the conventional rectangular one.
- The use of rectangular window technique in Image SR also results boundary-level distortion due to insufficient neighboring pixels at rectagular boundaries.
  - We smoothly integrate the proposed triangular window with traditional rectangular windows to employ non-overlapping self-attention in single-image SR.
  - It not only eradicate the boundary-level distortion but also execute the multi-region attention.
- The smaller window reduces the computational complexity with a heavy penalty on performances due restricted receptive field.
  - We introduce two variants of triangular window attention: (i) dense and (ii) sparse. The dense and sparse attention concentrate more on local image features and global image context respectively.

Shift Modes in Triangular & Rectangular Windows

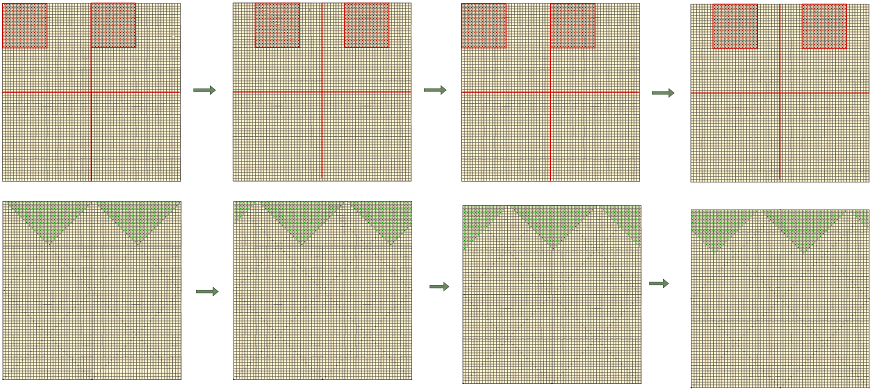


Figure 1: Shifting modes of rectangular and triangular windows in a 64 $\times$ 64 image patch

Design of Triangular & Rectangular Windows

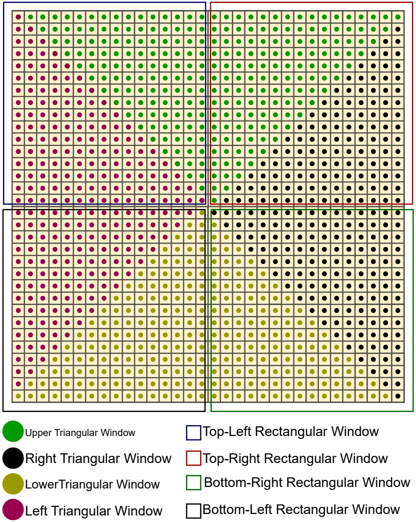


Figure 2: A rectangular and triangular patch in 32×32 window.

CFAT vs SOTA Models

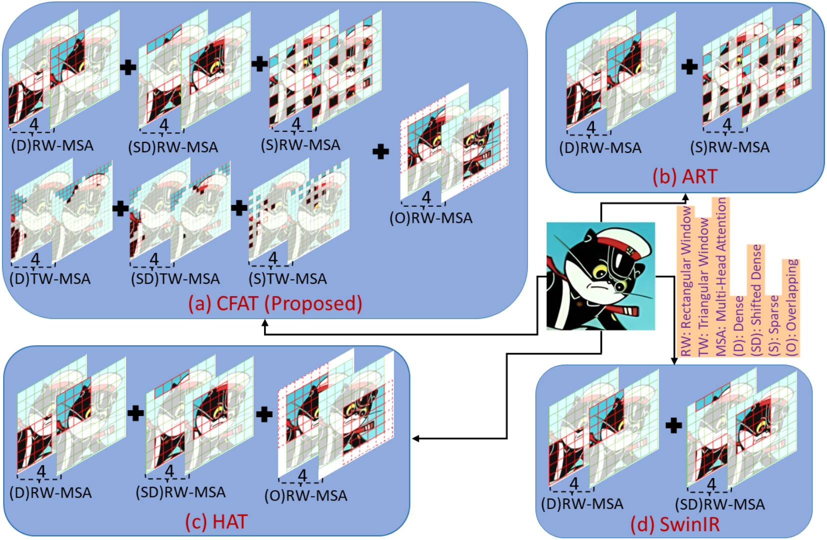


Figure 3: Proposed CFAT vs other SOTA models

Model Architecture

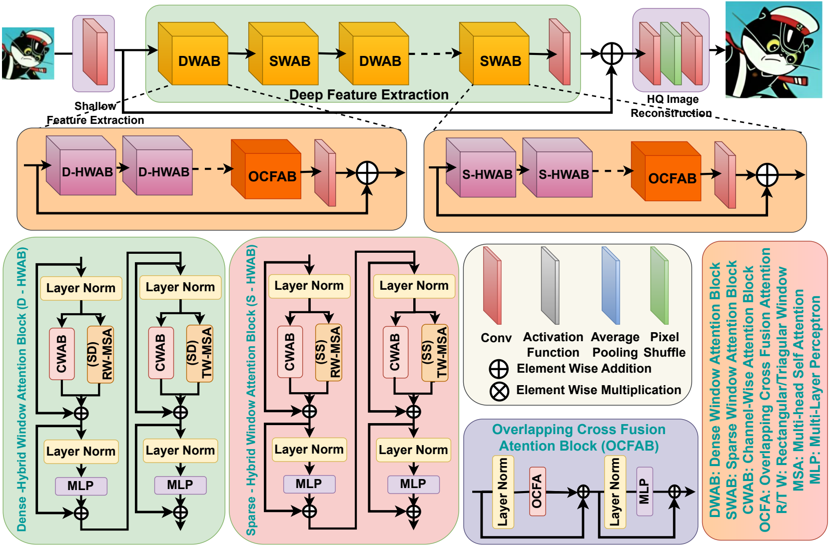


Figure 4: The overall architecture of CFAT with all internal modules

Model Equations

① Dense Window Attention Blocks [DWAB]:

$$
\begin{aligned}
F_{dp} &= f_{conv}(f_{op}(f_{nop}^{n}...(f_{nop}^{2}(f_{nop}^{1}(F_{sh}))))) + F_{sh}, \\
F_{DA} &= f_{nop}^{x}(F_{sh}) = f_{tri}^{n} f_{rect}^{n}...(f_{tri}^{1}(f_{rect}^{1}(F_{sh})))), \\
F_{int} &= f_{MSA}^{rect}(f_{LN}^{1}(F_{in})) + \alpha f_{CA}(f_{LN}^{1}(F_{in})) + F_{in}, \\
F_{out} &= f_{MLP}(f_{LN}^{2}(F_{int})) + F_{int}, \\
F_{int} &= f_{MSA}^{tri}(f_{LN}^{1}(F_{in})) + \beta f_{CA}(f_{LN}^{1}(F_{in})) + F_{in}, \\
F_{out} &= f_{MLP}(f_{LN}^{2}(F_{int})) + F_{int}.
\end{aligned}
\tag{1}
$$

② Overlaping Cross Fusion Attention Block [OCFAB]:

$$
R_0 = (1 + k)R. \tag{2}
$$

③ Computational Complexity for Triangular-MSA:

$$
\begin{aligned}
O(MSA) &= 4[HW]C^2 + 2[HW]^2 C, \\
O(\text{D-MSA}) &= (4HWC^2 + 2HWL^2 C), \\
O(\text{S-MSA}) &= (4HWC^2 + 2(\frac{HW}{S})^2 C).
\end{aligned}
\tag{3}
$$

Environment & Training Settings

① Environment Settings:

1. **GPU:** NVIDIA GTX 1080 ti, CUDA 10.1.243, CuDNN 8.1.0,
2. **Language:** Python 3.10.11.
3. **Framework:** PyTorch 2.0.1
4. **Library:** Torch, Numpy...

② Training Settings:

1. **Iterations:** 250K
2. **Batch Size:** 32
3. **Obejective Function:** L_1 Loss
4. **Optimizer:** Adam.
5. **Leaning Rate:** 0.0002
6. **lr Decay:** 0.5
7. **Step Size:** [112.5K, 175K, 200K, 225K]

Ablation Study

CVPR
JUNE 17-21, 2024
SEATTLE, WA

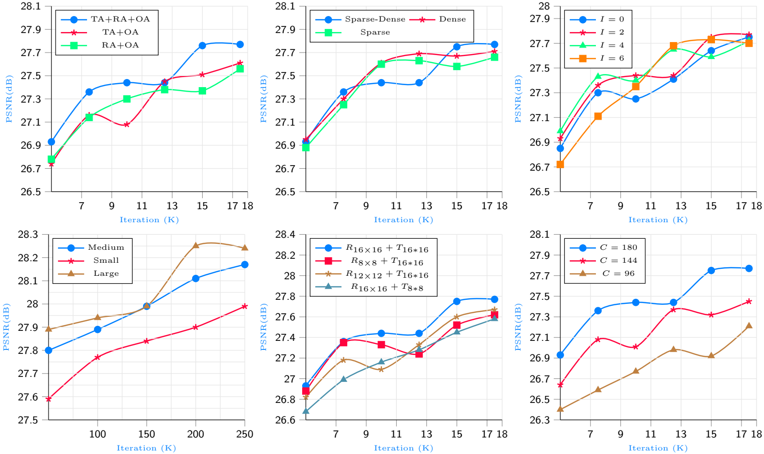


Figure 5: Iterative performance (PSNR in dB) comparison of the proposed CFAT for **Top-Left:** triangular vs rectangular vs overlapping attention, **Top-Middle:** sparse vs dense attention, **Top-Right:** various interval size, **Bottom-Left:** small vs medium vs large CFAT model, **Bottom-Middle:** various combinations of rectangular (8×8, 12×12, 16×16) with triangular ($8 * 8$, $16 * 16$) windows, and **Bottom-Right:** various channel lengths. [BSD100($\times$4) epoch 70]

Quantitative Comparison

Table 1: Quantitative comparison of the proposed CFAT. red and green color indicate the best and second best respectively

Method	Scale	Training Dataset	Set5		Set14		BSD100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR [1]	×2	DIV2K	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
HAN [2]	×2	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
SAN [3]	×2	DIV2K	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
IPT [4]	×2	ImageNet	38.37	-	34.43	-	32.48	-	33.76	-	-	-
SwinIR [5]	×2	DIV2K+Flickr2K	38.46	0.9624	34.61	0.9260	32.55	0.9043	33.95	0.9433	40.02	0.9800
Swin2SR [6]	×2	DIV2K+Flickr2K	38.43	0.9623	34.48	0.9256	32.54	0.905	33.89	0.9431	39.88	0.9798
ACT [7]	×2	DIV2K+Flickr2K	38.53	0.9629	34.68	0.9260	32.60	0.9052	34.25	0.9453	40.11	0.9807
ART [8]	×2	DIV2K+Flickr2K	38.56	0.9629	34.59	0.9267	32.58	0.9048	34.30	0.9452	40.24	0.9808
EDT [9]	×2	DIV2K+Flickr2K	38.63	0.9632	34.80	0.9273	32.62	0.9052	34.27	0.9456	40.37	0.9811
HAT [10]	×2	DIV2K+Flickr2K	38.63	0.9630	34.86	0.9274	32.62	0.9053	34.45	0.9466	40.26	0.9809
CFAT-s (ours)	×2	DIV2K+Flickr2K	38.59	0.9621	34.81	0.92872	32.58	0.9044	34.42	0.9453	40.24	0.9799
CFAT (ours)	×2	DIV2K+Flickr2K	39.09	0.9653	35.25	0.9296	32.93	0.9072	35.01	0.9498	41.00	0.9838
EDSR [1]	×3	DIV2K	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
HAN [2]	×3	DIV2K	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
SAN [3]	×3	DIV2K	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
IPT [4]	×3	ImageNet	34.81	-	30.85	-	29.38	-	29.49	-	-	-
SwinIR [5]	×3	DIV2K+Flickr2K	35.04	0.9322	31.00	0.8542	29.49	0.8150	29.90	0.8841	35.28	0.9543
ACT [7]	×3	DIV2K+Flickr2K	35.09	0.9325	31.17	0.8549	29.55	0.8171	30.26	0.8876	35.47	0.9548
ART [8]	×3	DIV2K+Flickr2K	35.07	0.9325	31.02	0.8541	29.51	0.8159	30.10	0.8871	35.39	0.9548
EDT [9]	×3	DIV2K+Flickr2K	35.13	0.9328	31.09	0.8553	29.53	0.8165	30.07	0.8863	35.47	0.9550
HAT [10]	×3	DIV2K+Flickr2K	35.07	0.9329	31.08	0.8555	29.54	0.8167	30.23	0.8896	35.53	0.9552
CFAT-s (ours)	×3	DIV2K+Flickr2K	34.03	0.9323	31.06	0.8551	29.57	0.8158	30.18	0.8889	35.48	0.9547
CFAT (ours)	×3	DIV2K+Flickr2K	35.31	0.9340	31.32	0.8569	29.70	0.8180	30.43	0.8928	35.82	0.9574
EDSR [1]	×4	DIV2K	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
HAN [2]	×4	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
SAN [3]	×4	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
IPT [4]	×4	ImageNet	32.64	-	29.01	-	27.82	-	27.26	-	-	-
SwinIR [5]	×4	DIV2K+Flickr2K	32.93	0.9043	29.15	0.7958	27.95	0.7494	27.56	0.8273	32.22	0.9273
Swin2SR [6]	×4	DIV2K+Flickr2K	32.92	0.9039	29.06	0.7946	27.92	0.7505	27.51	0.8271	31.03	0.9256
ACT [7]	×4	DIV2K+Flickr2K	33.04	0.9041	29.27	0.7968	28.00	0.7516	27.92	0.8332	32.44	0.9282
ART [8]	×4	DIV2K+Flickr2K	33.04	0.9051	29.16	0.7958	27.97	0.7510	27.77	0.8321	32.31	0.9283
EDT [9]	×4	DIV2K+Flickr2K	33.06	0.9055	29.23	0.7971	27.99	0.7510	27.75	0.8317	32.39	0.9283
HAT [10]	×4	DIV2K+Flickr2K	33.04	0.9056	29.23	0.7973	28.00	0.7517	27.97	0.8368	32.48	0.9292
CFAT-s (ours)	×4	DIV2K+Flickr2K	32.01	0.9045	29.25	0.7972	27.99	0.7504	27.86	0.8358	32.45	0.9279
CFAT (ours)	×4	DIV2K+Flickr2K	33.19	0.9068	29.30	0.7985	28.17	0.7524	28.11	0.8380	32.63	0.9305

Visual Analysis

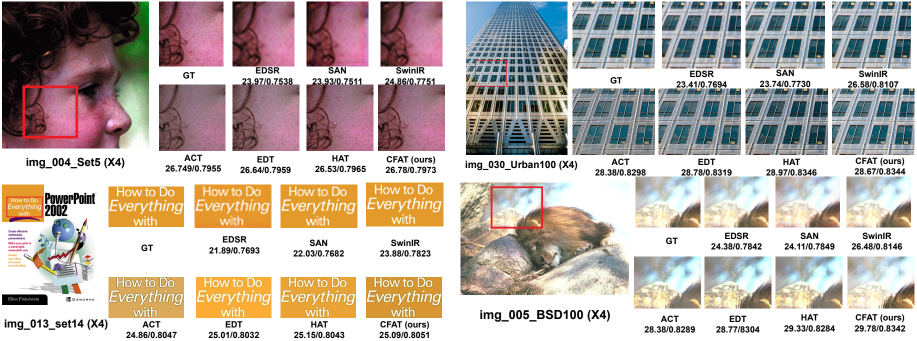


Figure 6: Visual Comparison of CFAT with other state-of-the-art methods.

Local Attribute Map & Diffusion Index

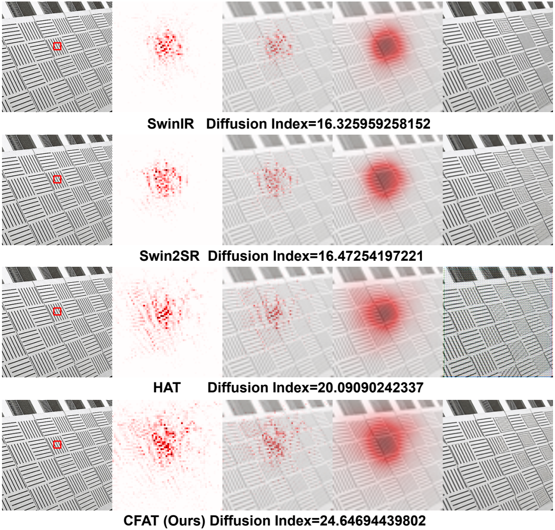


Figure 7: LAM results and corresponding Diffusion Index for CFAT and various SOTA methods.

Model Variants & Complexity

Table 2: 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 3: 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

Model Variants & Complexity

Table 4: Model comparison based on computational cost

Methods	Params (M)	Multi-adds(GMac)	PSNR/SSIM
SwinIR	11.9	53.6	27.92dB/0.7489
ACT	46	22	28.00dB/0.7516
ART	16.55	120	27.97dB/0.7510
EDT	11.6	-	27.91/0.7483
HAT	20.8	103.7	28.00dB/0.7517
CFAT	22.07	90.59	28.17dB/0.7524

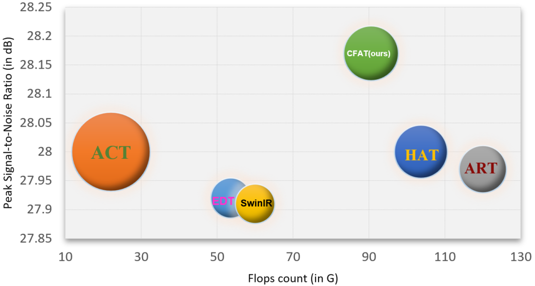


Figure 8: Performance vs Complexity plot of CFAT compare to other state-of-the-art models. **Performance:** PSNR (on X-axis) in dB. **Complexity:** Flops (on Y-axis) in G and Parameters (area of the circle) in M

Conclusion & Future Scope

Conclusion

- ✔ We propose a triangular window attention technique that smoothly integrates with rectangular windows to eliminate boundary-level distortion and allows additional non-identical shifting modes for activating more input pixels that participated in the computer vision task.
- ✔ By incorporating the novel triangular window attention in dense, sparse, and shifted configuration, CFAT outperforms the other state-of-the-art models qualitatively and quantitatively.

Future Scope

- Designing lightweight models for SISR using triangular window attention.
- Exploring the model for other computer vision tasks.

References

[1] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Computer Vision and Pattern Recognition workshops (CVPR-W)*, pp. 136–144, IEEE/CVF, 2017.

[2] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *European Conference on Computer Vision (ECCV)*, pp. 191–207, Springer, 2020.

[3] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 11065–11074, IEEE/CVF, 2019.

[4] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 12299–12310, IEEE/CVF, 2021.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *International Conference on Computer Vision(ICCV)*, pp. 1833–1844, IEEE/CVF, 2021.

[6] M. V. Conde, U.-J. Choi, M. Burchi, and R. Timofte, "Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration," *arXiv preprint arXiv:2209.11345*, 2022.

References (cont.)

[7] J. Yoo, T. Kim, S. Lee, S. H. Kim, H. Lee, and T. H. Kim, "Enriched cnn-transformer feature aggregation networks for super-resolution," in *Winter Conference on Applications of Computer Vision (WACV)*, pp. 4956–4965, IEEE/CVF, 2023.

[8] J. Zhang, Y. Zhang, J. Gu, Y. Zhang, L. Kong, and X. Yuan, "Accurate image restoration with attention retractable transformer," *arXiv preprint arXiv:2210.01427*, 2022.

[9] W. Li, X. Lu, J. Lu, X. Zhang, and J. Jia, "On efficient transformer and image pre-training for low-level vision," *arXiv preprint arXiv:2112.10175*, 2021.

[10] X. Chen, X. Wang, J. Zhou, and C. Dong, "Activating more pixels in image super-resolution transformer. arxiv 2022," *arXiv preprint arXiv:2205.04437*, vol. 1, 2022.

Thank You

Find our Paper Here

Find Our Code Here