HAAT: Hybrid Attention Aggregation Transformer for Image Super-Resolution

Song-Jiang Lai[a,b], Tsun-Hin Cheung[a,b], Ka-Chun Fung[a,b], Kai-wen Xue[a,b], and Kin-Man Lam[a,b]

[a]Centre for Advances in Reliability and Safety. New Territories, Hong Kong
[b]Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. Kowloon, Hong Kong

ABSTRACT

In the research area of image super-resolution, Swin-transformer-based models are favored for their global spatial modeling and shifting window attention mechanism. However, existing methods often limit self-attention to non-overlapping windows to cut costs and ignore the useful information that exists across channels. To address this issue, this paper introduces a novel model, the Hybrid Attention Aggregation Transformer (HAAT), designed to better leverage feature information. HAAT is constructed by integrating Swin-Dense-Residual-Connected Blocks (SDRCB) with Hybrid Grid Attention Blocks (HGAB). SDRCB expands the receptive field while maintaining a streamlined architecture, resulting in enhanced performance. HGAB incorporates channel attention, sparse attention, and window attention to improve nonlocal feature fusion and achieve more visually compelling results. Experimental evaluations demonstrate that HAAT surpasses state-of-the-art methods on benchmark datasets.

Keywords: Image super-resolution, Computer vision, Attention mechanism, Transformer

1. INTRODUCTION

Single Image Super-Resolution (SISR) aims to reconstruct a high-quality image from a low-resolution one. The development of effective super-resolution algorithms has emerged as a pivotal research domain in computer vision owing to its extensive applications. Recent research has integrated the self-attention mechanism into computer vision challenges.[1,2]

CNN-based techniques for Single Image Super-Resolution (SISR) have markedly improved the restoration of image texture features. SRCNN[3] was the inaugural model to tackle super-resolution with convolutional neural networks. VDSR[4] implemented residual learning to enhance learning and successfully address the gradient vanishing issue in deep networks. In SRGAN,[5] Ledig et al. employed generative adversarial networks to refine super-resolution image generation, with the generator converting low-resolution images to high-resolution ones and improving quality via adversarial training. ESRGAN[6] included the Residual Dense Block (RRDB) as a fundamental network component, diminishing perceptual loss by utilizing characteristics prior to activation, hence producing images with more authentic textures. Moreover, researchers persist in suggesting novel structures to retrieve progressively realistic information in super-resolution images. CNN-based networks have demonstrated considerable performance efficacy. Nonetheless, the inductive bias inherent in CNNs constrains the ability of SISR models to capture long-range relationships. The constraints arise from the parameter-dependent scaling of the receptive field and the convolution operator's kernel size across many layers, potentially neglecting non-local spatial information in pictures.

To enhance the joint modeling of different hierarchical structures in pictures, researchers have taken advantage of the self-attention mechanism's benefits in multi-scale processing and long-range dependency modeling. Transformer-based SISR models have been developed to overcome the shortcomings of CNN-based networks by utilizing their capacity to simulate long-range dependencies and improving SISR performance. For instance,

Further author information: (Send correspondence to Songjiang-Lai.)
Song-Jiang Lai.: E-mail: songjiang.Lai@connect.polyu.hk
Tsun-Hin Cheung.: E-mail: tsun-hin.cheung@connect.polyu.hk

International Workshop on Advanced Imaging Technology (IWAIT) 2025, edited by Masayuki Nakajima, Chuan-Yu Chang, Chia-Hung Yeh, Jae-Gon Kim, Kemao Qian, Phooi Yee Lau, Proc. of SPIE Vol. 13510, 135101A
 · 0277-786X · doi: 10.1117/12.3058003

super-resolution results have been markedly enhanced by SwinIR,[1] which makes use of the Swin Transformer.[2] Furthermore, state-of-the-art results have been produced with a hybrid attention transformer (HAT)[7] that combines an overlapping cross-attention module with window-based self-attention and channel attention.

Despite the successful use of Transformer-based approaches to image restoration problems, there are areas for improvement. Current window-based Transformer networks confine self-attention calculations to a concentrated region. This method results in a constrained receptive field and fails to properly use the feature information from the original picture. This research proposes a hybrid multiaxial aggregation network, termed HAAT, to address these issues. HAAT is constructed by integrating Swin-Dense-Residual-Connected Blocks (SDRCB)[8] with Hybrid Grid Attention Blocks (HGAB). HGAB, inspired by GAB,[9] integrates channel attention, sparse attention, and window attention, leveraging the global information perception capabilities of channel attention to address the deficiencies of self-attention. Sparse self-attention is used to enhance global feature interactions while maintaining computational efficiency further enhancing the potential performance of the model.

2. HYBRID ATTENTION AGGREGATION TRANSFORMER

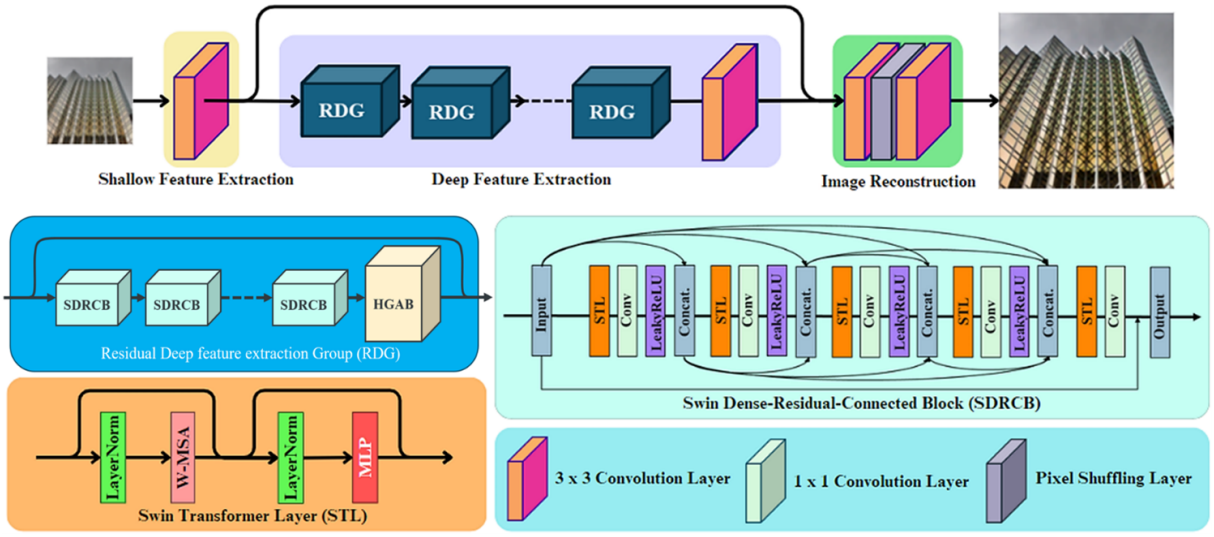

Figure 1. SDRCB Framework.

Figure 1 illustrates the comprehensive structure of HAAT. SDRCB incorporates Swin Transformer Layers and transition layers into each Residual Deep feature extraction Group (RDG), enhancing the receptive field while using fewer parameters and a more streamlined design, resulting in superior performance. Furthermore, we provide HGAB to describe cross-area similarity for enhanced picture reconstruction. The architecture of HGAB, seen in Figure 2, comprises a Mix Attention Layer (MAL) and a Multi-Layer Perceptron (MLP) layer. HGAB employs sparse self-attention to augment global feature interactions while controlling computational complexity, facilitating the joint modeling of analogous features for enhanced picture reconstruction. Furthermore, the employed channel attention mechanism can help the model extract more effective information between different channels.

2.1 Swin-Dense-Residual-Connected Block

We use the shifting window self-attention mechanism of the Swin-Transformer Layer (STL)[1,2] to capture long-range dependencies via adaptive receptive fields. STL modifies the model's emphasis according to global content, enhancing feature extraction. This technique maintains global details as the network deepens, enlarging the receptive area without degradation. Integrating STL with dense-residual connections expands the receptive area and improves emphasis on critical information, hence increasing performance in SISR tasks that need thorough, context-sensitive processing. The SDRCB for input feature maps $\boldsymbol{Z}$ inside RDG is delineated as follows.

$$\boldsymbol{Z_j} = H_{trans}(\boldsymbol{STL}([\boldsymbol{Z}, ...\boldsymbol{Z_{j-1}}]), j = 1, 2, 3, 4, 5, \tag{1}$$

$$SDRCB(\boldsymbol{Z}) = \alpha \cdot \boldsymbol{Z_5} + \boldsymbol{Z}, \tag{2}$$

where $[\cdot]$ denotes the concatenation of multi-level feature maps produced by the previous layers. $H_{trans}(\cdot)$ refers to the convolution layer with a LeakyReLU activation function for feature transition. The negative slope of LeakyReLU is set to 0.2. Conv1 is the 1×1 convolution layer, which is used to adaptively fuse a range of features with different levels.[10] α represents residual scaling factor, which is set to 0.2 for stabilizing the training process.[6]

2.2 Hybrid Grid Attention Block(HGAB)

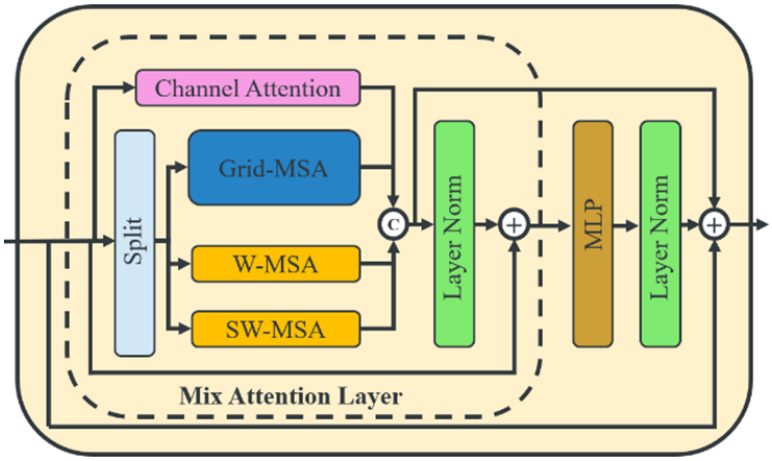

Figure 2. The structure of HGAB.

The structure of Hybrid Grid Attention Block (HGAB) is illustrated in Figure 2. This advanced hybrid attention block integrates channel, sparse, and self-attention to enhance feature modeling and representation learning beyond traditional hybrid attention approaches. Within the Hybrid Global Attention Block (HGAB), the Multi-Attention Layer (MAL) enables parallel processing of divided input features, boosting efficiency. This design improves global feature interaction and utilizes non-local spatial information, overcoming self-attention limitations while controlling computational complexity. Consequently, it significantly enhances performance in image super-resolution tasks through more sophisticated feature extraction and learning. The HGAB consists of a Mix Attention Layer (MAL) and an MLP layer. Regarding the MAL, we first split the input feature F_{in} into two parts by channel: $F_G \in R^{(H \times W \times C/2)}$ and $F_W \in R^{(H \times W \times C/2)}$. Besides, the input is fed to another branch, where channel attention s applied. Subsequently, we split F_W into two parts by channel again and input them into W-MSA and SW-MSA, respectively. Meanwhile, F_G is input into Grid-MSA.[9]The computation process of MAL is as follows:

$$X_{W_1} = W - MSA(F_{W_1}), \tag{3}$$

$$X_{W_2} = SW - MSA(F_{W_2}), \tag{4}$$

$$X_G = Grid - MSA(F_G), \tag{5}$$

$$X_C = CA(F_{in}), \tag{6}$$

Table 1. Quantitative comparison with SOTA methods

Method	Scale	Training Dataset	Set5		Set14	
			PSNR	SSIM	PSNR	SSIM
EDSR	×2	DIV2K	38.11	0.9602	33.92	0.9195
RCAN	×2	DIV2K	38.27	0.9614	34.12	0.9216
SAN	×2	DIV2K	38.31	0.9620	34.07	0.9213
IGNN	×2	DIV2K	38.24	0.9613	34.07	0.9217
HAN	×2	DIV2K	38.27	0.9614	34.16	0.9217
NLSN	×2	DIV2K	38.34	0.9618	34.08	0.9231
SwinIR	×2	DIFK	38.42	0.9623	34.46	0.9250
CAT-A	×2	DIFK	38.51	0.9626	34.78	0.9265
HAT	×2	DIFK	38.63	0.9630	34.86	0.9274
DAT	×2	DIFK	38.58	0.9629	34.81	0.9272
DRCT	×2	DIFK	38.72	0.9646	34.96	0.9287
HAAT (Ours)	×2	DIFK	38.74	0.9645	34.97	0.9287
EDSR	×3	DIV2K	34.65	0.9280	30.52	0.8462
RCAN	×3	DIV2K	34.74	0.9299	30.65	0.8482
SAN	×3	DIV2K	34.75	0.9300	30.59	0.8476
IGNN	×3	DIV2K	34.72	0.9298	30.66	0.8484
HAN	×3	DIV2K	34.75	0.9299	30.67	0.8483
NLSN	×3	DIV2K	34.85	0.9306	30.70	0.8485
SwinIR	×3	DIFK	34.97	0.9318	30.93	0.8534
CAT-A	×3	DIFK	35.06	0.9326	31.04	0.8538
HAT	×3	DIFK	35.07	0.9329	31.08	0.8555
DAT	×3	DIFK	35.16	0.9331	31.11	0.8550
DRCT	×3	DIFK	35.15	0.9333	31.22	0.8569
HAAT (Ours)	×3	DIFK	35.17	0.9336	31.23	0.8569
EDSR	×4	DIV2K	32.46	0.8968	28.80	0.7876
RCAN	×4	DIV2K	32.63	0.9002	28.87	0.7889
SAN	×4	DIV2K	32.64	0.9003	28.92	0.7888
IGNN	×4	DIV2K	32.57	0.8998	28.85	0.7891
HAN	×4	DIV2K	32.64	0.9002	28.90	0.7890
NLSN	×4	DIV2K	32.59	0.9000	28.87	0.7891
SwinIR	×4	DIFK	32.92	0.9044	29.09	0.7950
CAT-A	×4	DIFK	33.08	0.9052	29.18	0.7960
HAT	×4	DIFK	33.04	0.9056	29.23	0.7973
DAT	×4	DIFK	33.08	0.9055	29.23	0.7973
DRCT	×4	DIFK	33.09	0.9061	29.32	0.7982
HAAT (Ours)	×4	DIFK	33.12	0.9062	29.32	0.7983

$$X_{MAL} = LN(Cat(X_{W_1}, X_{W_2}, X_{W_G}) + X_C) + F_{in}, \tag{7}$$

where X_{W_1}, X_{W_2}, X_G and X_C are the output features of W-MSA, SW-MSA, Grid-MSA, and CA, respectively. Furthermore, it is worth noting that we adopt the post-norm method in HGAB to enhance the network training stability. For a given input feature F_{in}, the computation process of HGAB is as follows:

$$F_M = LN(MAL(F_{in})) + F_{in}, \tag{8}$$

$$F_M = LN(MAL(F_M)) + F_M, \tag{9}$$

3. EXPERIMENTAL RESULTS

Our HAAT model is trained on DF2K, a substantial aggregated dataset that includes DIV2K[11] and Flickr2K.[12] DIV2K provides 800 images for training, while Flickr2K contributes 2650 images. For the training input, we

generate LR versions of these images by applying the bicubic down sampling method with scaling factors of 2, 3, and 4. To assess the effectiveness of our model, we conduct performance evaluations using well-known SISR benchmark datasets such as Set5[13] and Set14.[14]

In the DRCT architecture, the depth and width configuration mirrors that of HAT. Specifically, both models have 6 RDG and SDRCB units, with 180 channels for intermediate feature maps. For window-based multi-head self-attention (W-MSA), the number of attention heads is set to 6, and the window size is 16. In the HGAB block, the channel squeeze factor is 16, with 180 channels for intermediate features. The Grid MSA and (S)W-MSA use 3 and 2 attention heads, respectively. HR patches of 256 × 256 pixels were extracted from HR images, with random horizontal flips and rotation for data augmentation. As shown in Table 1, our method outperforms state-of-the-art techniques in both PSNR and SSIM.

For evaluation, we use all RGB channels and exclude the outermost (2 × scale) border pixels. PSNR and SSIM metrics are employed for assessment. Table 1 shows a quantitative comparison of our method with state-of-the-art approaches such as EDSR,[15] RCAN,[16] SAN,[17] IGN,[18] HAN,[19] NLSN,[20] SwinIR,[1] CATA,[21] DAT[22] and CDRT.[8] Our method consistently outperforms these methods across all benchmark datasets. Therefore, it can be clearly summarized HAAT achieves significantly better results than the other state-of-the-art models.

4. CONCLUSION

This work introduces a unique Hybrid Attention Aggregation Transformer (HAAT) for single-image super-resolution. HAAT enhances the DRCT architecture, emphasizing the stabilization of information flow and the expansion of receptive fields via dense connections in residual blocks, in conjunction with the shift-window attention mechanism to adaptively acquire global information. This enables the model to enhance its emphasis on global geographical information, optimizing its capabilities and circumventing information bottlenecks. Furthermore, motivated by the hierarchical structural resemblance in images, we provide HGAB to represent long-range relationships. The network improves multi-level structural similarity via the integration of channel attention, sparse attention, and window attention. The model was trained on the DF2K dataset and verified using the Set5 and Set14 datasets. Experimental findings indicate that our strategy surpasses SOTA techniques on benchmark datasets for single-image super-resolution tasks.

REFERENCES

[1] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R., "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844 (2021).

[2] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

[3] Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., and Liao, Q., "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia* **21**(12), 3106–3121 (2019).

[4] Kim, J., Lee, J. K., and Lee, K. M., "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654 (2016).

[5] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690 (2017).

[6] Hanson, K. M., "Introduction to Bayesian image analysis," in *Medical Imaging: Image Processing*, Loew, M. H., ed., *Proc. SPIE* **1898**, 716–731 (1993).

[7] Chen, X., Wang, X., Zhou, J., Qiao, Y., and Dong, C., "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22367–22377 (2023).

[8] Hsu, C.-C., Lee, C.-M., and Chou, Y.-S., "Drct: Saving image super-resolution away from information bottleneck," *arXiv preprint arXiv:2404.00722* (2024).

[9] Chu, S.-C., Dou, Z.-C., Pan, J.-S., Weng, S., and Li, J., "Hmanet: Hybrid multi-axis aggregation network for image super-resolution," *arXiv preprint arXiv:2405.05001* (2024).

[10] Tai, Y., Yang, J., Liu, X., and Xu, C., "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, 4539–4547 (2017).

[11] Agustsson, E. and Timofte, R., "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135 (2017).

[12] Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., and Zhang, L., "Ntire 2017 challenge on single image super-resolution: Methods and results," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114–125 (2017).

[13] Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L., "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," (2012).

[14] Zeyde, R., Elad, M., and Protter, M., "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, 711–730, Springer (2012).

[15] Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K., "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144 (2017).

[16] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y., "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 286–301 (2018).

[17] Dai, T., Cai, J., Zhang, Y., Xia, S.-T., and Zhang, L., "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11065–11074 (2019).

[18] Zhou, S., Zhang, J., Zuo, W., and Loy, C. C., "Cross-scale internal graph neural network for image super-resolution," *Advances in neural information processing systems* **33**, 3499–3509 (2020).

[19] Niu, B., Wen, W., Ren, W., Zhang, X., Yang, L., Wang, S., Zhang, K., Cao, X., and Shen, H., "Single image super-resolution via a holistic attention network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 191–207, Springer (2020).

[20] Mei, Y., Fan, Y., and Zhou, Y., "Image super-resolution with non-local sparse attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3517–3526 (2021).

[21] Chen, Z., Zhang, Y., Gu, J., Kong, L., Yuan, X., et al., "Cross aggregation transformer for image restoration," *Advances in Neural Information Processing Systems* **35**, 25478–25490 (2022).

[22] Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., and Yu, F., "Dual aggregation transformer for image super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 12312–12321 (2023).