Mini Project Report - CS-2-14(PO)

Even Semester - January 2025 to May 2025

Image Super Resolution 4x

Submitted By

D Barghav (2023UME0253)
Purushartha Gupta (2023UCE0062)
Aman Nagar (2023UME0242)
Misti D Shah (2023UCE0055)

Supervised By

Dr. Vinit Jakhetiya
Associate Professor
Department of Computer Science & Engineering

Department of Computer Science & Engineering
Indian Institute of Technology, Jammu
Jammu & Kashmir 181221, India

April 2025

Introduction

Single Image Super-Resolution (SISR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input. The 4× super-resolution task is particularly challenging due to the severe loss of textures and structural details during downsampling, which makes the reconstruction of fine structures a highly ill-posed problem.

In this mini-project, we participated in the NTIRE 2025 Challenge to upscale natural images by a factor of 4 using deep learning. We investigated a variety of state-of-the-art architectures and selected an ensemble-based approach to combine the unique strengths of **HAT**, **SwinIR**, and **RCAN**

Motivation

Super-resolution has critical applications in medical imaging, surveillance systems, satellite photography, and real-time vision tasks. The 4× SR task is particularly difficult due to the high degree of ambiguity and information loss in LR images. Moreover, with higher scaling factors, traditional interpolation-based methods fail to preserve edge information and perceptual clarity.

While standalone models can perform well, each one has certain limitations. RCAN excels in local detail recovery, SwinIR balances accuracy and computational efficiency, and HAT excels at global context learning. By combining their outputs, we aim to create a more robust and accurate super-resolution pipeline. An ensemble mitigates weaknesses of individual models and promotes generalization across diverse textures.

Literature Review

Hybrid Attention Transformer (HAT): Chen et al. introduced HAT in ”Activating More Pixels in Image Super-Resolution Transformer” (CVPR 2023), proposing Hybrid Attention Blocks that fuse channel and spatial attention. These blocks effectively model global-local dependencies and boost reconstruction quality across SR benchmarks [1].

SwinIR: Liang et al. presented SwinIR in ”SwinIR: Image Restoration Using Swin Transformer” (ICCVW 2021) [2]. Built upon the Swin Transformer backbone, SwinIR employs residual Swin Transformer blocks and shifted windows to extract hierarchical spatial features efficiently.

RCAN: Proposed by Zhang et al. in ”Image Super-Resolution Using Very Deep Residual Channel Attention Networks” (ECCV 2018) [3], RCAN introduces Residual Channel Attention Blocks (RCABs) to emphasize informative features across channels and achieves state-of-the-art performance in PSNR-driven tasks.

Additional models explored include **EDSR** [4], which enhances deep ResNet architectures by removing batch normalization layers, enabling stable training of deeper networks. **ESRGAN** [5] improves perceptual quality by incorporating both perceptual and adversarial losses, producing more realistic and detailed textures. **SRResNet** [6] is one of the earliest GAN-based super-resolution methods, setting a foundation for future adversarial approaches. **RGT** [7] introduces a recursive transformer design that generalizes well across different scales, offering efficient and flexible performance in high-resolution image generation.

Methodology and Implementation

Model Architectures

HAT: Built on Hybrid Attention Blocks (HABs), HAT combines channel and spatial attention to capture both local and global features. Each HAB includes layer normalization, multi-layer perceptrons (MLPs), and multi-head cross-attention for inter-window communication. The attention mechanism is formulated as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are projections of the input features. This structure enables HAT to effectively model long-range dependencies and multi-scale representations in high-resolution restoration.

SwinIR: A lightweight, hierarchical network based on Swin Transformers, SwinIR uses window-based multi-head self-attention (W-MSA) to model local dependencies efficiently. The architecture consists of shallow feature extraction, stacked Residual Swin Transformer Blocks (RSTBs), and pixel-shuffle upsampling. Each attention block computes:

$$W{-}MSA(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

applied within non-overlapping windows to reduce complexity. Shifted windowing enhances cross-window interaction while maintaining linear scalability. The final convolutional layer aggregates features for high-resolution reconstruction.

RCAN: RCAN is structured around Residual Groups containing Residual Channel Attention Blocks (RCABs). Each RCAB applies channel attention via global average pooling followed by two dense layers and a sigmoid activation to generate weights:

$$\mathbf{M}_c = \sigma(W_2\,\delta(W_1\,GAP(F)))$$

where F is the feature map, GAP is global average pooling, W_1, W_2 are learnable weights, and δ is ReLU. This mechanism emphasizes informative channels. The residual-in-residual design further stabilizes training of deep networks by enhancing gradient flow.

Ensemble Approach

To leverage the strengths of all models, we average their outputs during inference:

$$SR_{final} = \frac{1}{3}(SR_{HAT} + SR_{RCAN} + SR_{SwinIR})$$

This simple ensemble strategy enhances generalization and image quality while minimizing individual model biases. Weighted averaging could be introduced in future work, based on model-specific confidence maps.

Implementation Details

The networks were implemented in PyTorch. Each model was trained independently using the LSDIR and Flickr2K datasets containing 87641 images. Training involved bicubic downsampled LR inputs and HR targets. The Adam optimizer with cosine decay scheduling was used, along with L1 loss. Models were trained on GPUs with a batch size of 32 for approximately 200 epochs. For validation, we used PSNR and SSIM metrics. Checkpoints were saved based on the best PSNR performance.

Results

Model	**DIV2K**		**Set14**		**Set5**		**Urban100**		**BSD100**		**Overall**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EDSR	23.26	0.664	23.13	0.670	24.75	0.744	23.31	0.675	**32.82**	**0.912**	25.45	0.733
ESRGAN	23.56	0.680	23.56	0.691	25.47	0.776	21.83	0.608	25.46	0.712	23.98	0.693
RGT-S	22.27	0.632	22.82	0.643	25.73	0.756	20.25	0.556	23.94	0.648	23.00	0.647
RGTNet	22.48	0.629	23.34	0.644	26.20	0.755	20.40	0.551	23.99	0.641	23.28	0.644
SRResNet	24.09	0.703	25.06	0.700	28.29	0.817	21.61	0.644	25.62	0.703	24.93	0.713
HAT	32.87	0.786	31.26	0.709	31.99	0.807	30.84	0.685	31.46	0.708	31.69	0.739
SwinIR	28.16	0.348	28.15	0.326	28.26	0.403	28.12	0.336	28.13	0.305	28.16	0.343
RCAN	33.20	0.809	31.58	0.737	32.31	0.842	**31.24**	**0.739**	31.69	0.731	32.00	**0.772**
Ensemble (Ours)	**33.21**	**0.810**	**31.60**	**0.738**	**32.35**	**0.844**	31.23	0.738	31.69	0.730	**32.02**	**0.772**

Quantitative comparison of PSNR and SSIM on multiple datasets for 4×SR

To evaluate the performance of our individual models and the ensemble, we tested them on benchmark datasets including DIV2K, Set14, Set5, Urban100, and BSD100. We computed two standard image quality metrics — PSNR and SSIM — to assess the fidelity of the reconstructed high-resolution images. The ensemble model consistently outperformed individual ones across all datasets, achieving an average PSNR of **32.02** and SSIM of **0.772**, the highest among all models evaluated.

Future Work

Our ensemble-based solution demonstrated strong results, yet several directions remain for future improvement:

- **Model Optimization:** Reduce inference time via pruning and quantization.
- **Dynamic Weighting:** Adaptively combine outputs instead of averaging.
- **Real-Time Use:** Prepare the ensemble for deployment on low-latency edge devices.
- **Web Platform:** Build a lightweight site for user-friendly image enhancement.
- **Generalization:** Train on varied domains (e.g., satellite, medical) to test robustness.

Conclusion

In this project, we addressed the challenge of 4×image super-resolution by evaluating a range of state-of-the-art deep learning models and ultimately designing an ensemble strategy that combines the strengths of HAT, SwinIR, and RCAN. Through extensive experimentation on benchmark datasets, our ensemble model consistently achieved the highest average PSNR and SSIM scores, outperforming individual models. The results demonstrate the effectiveness of ensemble-based methods in enhancing visual quality and reconstruction accuracy. This work lays a strong foundation for future improvements in performance, usability, and real-time deployment of super-resolution systems.

References

[1] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong, "HAT: Hybrid Attention Transformer for Image Restoration," *arXiv preprint arXiv:2309.05239*, Sep. 2023. [Online]. Available: https://arxiv.org/abs/2309.05239

[2] J. Liang, X. Ren, Y. Zhang, Z. Gu, and C. Dong, "SwinIR: Image Restoration Using Swin Transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2021. [Online]. Available: https://arxiv.org/abs/2108.10257

[3] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301. [Online]. Available: https://arxiv.org/abs/1807.02758

[4] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 136–144. [Online]. Available: https://arxiv.org/abs/1707.02921

[5] X. Wang, K. Yu, C. Dong, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, 2018. [Online]. Available: https://arxiv.org/abs/1809.00219

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4681–4690. [Online]. Available: https://arxiv.org/abs/1609.04802

[7] J. Zheng, Y. Chen, K. Li, K. Li, and Y. Fu, "RGT: Recursive Generalist Transformer for Efficient Image Super-Resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024. [Online]. Available: https://arxiv.org/abs/2403.05594