NTIRE 2025 Image Super-Resolution($\times$4) Chanllenge

Junyi Zhao, Qisheng Xu, Kele Xu*

March 22, 2025

1 Team details

- Team name: Junyi
- Team leader name: Junyi Zhao
- Team leader institution and email: Key Laboratory for Parallel and Distributed Processing, Changsha, China, z15236936309@gmail.com
- Rest of the team members: Qisheng Xu, Kele Xu
- Team website URL: None
- Affiliations: Key Laboratory for Parallel and Distributed Processing, Changsha, China
- User names and entries on the NTIRE 2025 Codalab competitions (development/validation and testing phases): wentheart, 1 entry in development/validation phases, 3 entries in testing phases
- Link to the codes/executables of the solution(s): https://github.com/wentheart/cvpr2025sr

2 Contribution details

2.1 Method Overview

Deep learning models currently applied in the field of image super-resolution have achieved significant advancements, with increasingly diversified network architectures such as the Dual Aggregation Transformer (DAT) model[1] and the Shifted Window Transformer (SwinIR) model[2].

*Key Laboratory for Parallel and Distributed Processing, Changsha, China

To investigate whether these models can achieve enhanced super-resolution performance through fusion techniques, we conducted a series of model fusion experiments based on three foundational architectures: DAT, SwinIR, and the Residual Feature Distillation Network (RFDN)[3]. This study systematically evaluates the synergistic efficacy of integrating these state-of-the-art models, aiming to explore potential performance improvements in super-resolution reconstruction tasks through architectural hybridization and parameter optimization strategies.

Beyond the aforementioned fusion framework constituting our primary investigation, we further implemented a secondary fusion paradigm by integrating the Real-ESRGAN[4][5] architecture with the Dual Aggregation Transformer (DAT) model. This comparative experiment was specifically designed to quantify performance discrepancies arising from distinct fusion methodologies. Subsequent analysis of experimental results, rigorously evaluated through standardized metrics (Peak Signal-to-Noise Ratio [PSNR] and Structural Similarity Index [SSIM]), demonstrated superior quantitative performance in the initial fusion strategy. Based on this empirical evidence, we conclusively adopted the first fusion configuration as our optimal solution, thereby establishing its technical predominance in balancing reconstruction fidelity and computational efficiency within our experimental framework.

2.2 Network

Our fusion methodology employs an output-level model fusion approach. Specifically, we first fine-tuned three baseline models: DAT, SwinIR and the RFDN. Subsequently, we devised a lightweight attention-based weight allocation network to dynamically optimize the fusion coefficients among these three super-resolution outputs.

This architectural design enables adaptive spatial weighting across different reconstruction results, where the attention mechanism automatically prioritizes region-specific contributions from each model based on local texture complexity and edge preservation characteristics. The weight optimization process was conducted through end-to-end training using perceptual loss constraints, ensuring coordinated enhancement of both quantitative metrics and visual quality in the fused super-resolution output.

The detailed architectural configuration of the proposed model is illustrated in Figure 1.

3 Global Method Description

3.1 Training strategy

All experimental procedures in this study were conducted exclusively on the training set of the DIV2K dataset, which served as the foundational data source for model development and evaluation.

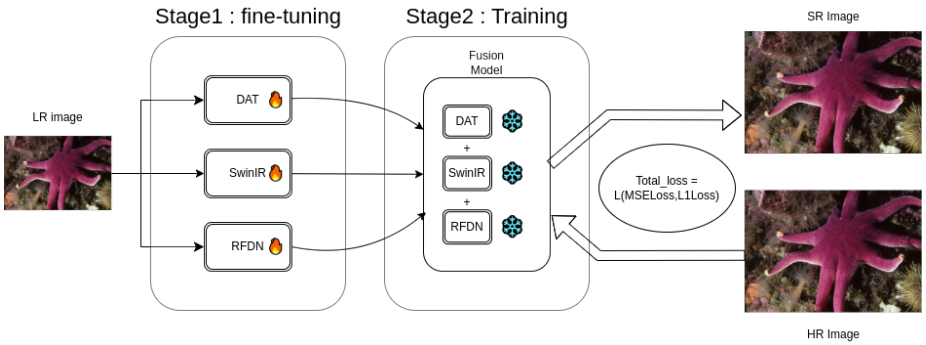

Figure 1: Traing pipeline

The experimental procedure was initiated by fine-tuning three baseline models, with the critical hyperparameters governing the fine-tuning process being systematically documented in Table 1 for reproducibility analysis.

Upon completing the training of the three base models, we proceeded to optimize the weight prediction network under the following computational configuration: 150 training epochs with a batch size of 72, requiring approximately 5 hours per training session utilizing an NVIDIA GeForce RTX 4090 GPU. The optimization employed the Adam optimizer with combined MSE and L1 loss functions. Detailed hyperparameter specifications are systematically presented in Table 2.

Models	Training Time	Epochs	Params. (M)	GPU
DAT	1.5h	100	11.21	Nvidia GTX 4090
SwinIR	1.7h	100	11.8	Nvidia GTX 4090
RFDN	0.3h	100	0.55	Nvidia GTX 4090

Table 1: Baseline Model Details

3.2 Testing description

For testing, we applied the trained baseline models and weight network to generate super-resolution images on the test set, incorporating a self-ensemble strategy to enhance output stability.

Epoch	Training Time	Params
150	5h	28.4K
Extra Data	Batch Size	GPU
No	72	Nvidia GTX 4090

Table 2: Fusion Model Details

Fusion Model	DAT	SwinIR	RFDN
31.42 ↑	31.23	31.18	30.86

Table 3: Validation Psnr Comparison between Fusion Model and Baseline Models

3.3 Valid and Test results

3.3.1 Valid results

During the validation phase, we conducted systematic comparisons between the fused model and three baseline models on the DIV2K validation set through PSNR evaluation. As quantitatively documented in Table 3, the fused model consistently surpassed all baseline counterparts in metric performance. This empirical outcome substantiates that the model fusion paradigm effectively integrates constituent sub-models' complementary strengths, thereby achieving enhanced reconstruction capability.

3.3.2 Test results

During the testing phase, both our original fusion framework and the comparative fusion architecture incorporating Real-ESRGAN were rigorously evaluated through parallel submissions on the CodaLab platform. Quantitative analysis of test set metrics revealed statistically significant superiority of the original fusion strategy over its Real-ESRGAN-enhanced counterpart. Based on this empirical validation, we conclusively determined to retain the original fusion configuration as the optimal solution for final deployment. We can see the details in Table 4.

Fusion Model(ours)	Single DAT	DAT and Real-ESTGAN
30.90 ↑	30.85	28.63

Table 4: test Psnr Comparison

4 Other details

In addition to the executable source code, we have also made essential pretrained weights available for public download in our GitHub repository to facilitate experimental reproducibility.

The non-deterministic nature of our implementation, resulting from unfixed random seed initialization, precludes strict reproducibility guarantees across different computational instances. Output variations may emerge due to inherent stochastic variations in parallel computing architectures and numerical instabilities inherent in floating-point operations.

References

[1] Zheng Chen et al. “Dual aggregation transformer for image super-resolution”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 12312–12321.

[2] Jingyun Liang et al. “Swinir: Image restoration using swin transformer”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 1833–1844.

[3] Jie Liu, Jie Tang, and Gangshan Wu. “Residual feature distillation network for lightweight image super-resolution”. In: *Computer vision–ECCV 2020 workshops: Glasgow, UK, August 23–28, 2020, proceedings, part III 16*. Springer. 2020, pp. 41–55.

[4] Xintao Wang et al. “Esrgan: Enhanced super-resolution generative adversarial networks”. In: *Proceedings of the European conference on computer vision (ECCV) workshops*. 2018, pp. 0–0.

[5] Xintao Wang et al. “Real-esrgan: Training real-world blind super-resolution with pure synthetic data”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 1905–1914.