NTIRE 2025 Challenge on Image Super-Resolution ($\times$4): Methods and Results
Supplementary Material

Zheng Chen[†] Kai Liu[†] Jue Gong[†] Jingkai Wang[†] Lei Sun[†] Zongwei Wu[†]
Radu Timofte[†] Yulun Zhang[†*] Xiangyu Kong Xiaoxuan Yu Hyunhee Park
Suejin Han Hakjae Jeon Dafeng Zhang Hyung-Ju Chun Donghun Ryou
Inju Ha Bohyung Han Lu Zhao Yuyi Zhang Pengyu Yan Jiawei Hu
Pengwei Liu Fengjun Guo Hongyuan Yu Pufan Xu Zhijuan Huang
Shuyuan Cui Peng Guo Jiahui Liu Dongkai Zhang Heng Zhang Huiyuan Fu
Huadong Ma Yanhui Guo Sisi Tian Xin Liu Jinwen Liang Jie Liu Jie Tang
Gangshan Wu Zeyu Xiao Zhuoyuan Li Yinxiang Zhang Wenxuan Cai
Vijayalaxmi Ashok Aralikatti Nikhil Akalwadi G Gyaneshwar Rao Chaitra Desai
Ramesh Ashok Tabib Uma Mudenagudi Marcos V. Conde Alejandro Merino
Bruno Longarela Javier Abad Weijun Yuan Zhan Li Zhanglu Chen
Boyang Yao Aagam Jain Milan Kumar Singh Ankit Kumar Shubh Kawa
Divyavardhan Singh Anjali Sarvaiya Kishor Upla Raghavendra Ramachandra
Chia-Ming Lee Yu-Fan Lin Chih-Chung Hsu Risheek V Hiremath
Yashaswini Palani Yuxuan Jiang Qiang Zhu Siyue Teng Fan Zhang
Shuyuan Zhu Bing Zeng David Bull Jingwei Liao Yuqing Yang Wenda Shao
Junyi Zhao Qisheng Xu Kele Xu Sunder Ali Khowaja Ik Hyun Lee
Snehal Singh Tomar Rajarshi Ray Klaus Mueller Sachin Chaudhary
Surya Vashisth Akshay Dudhane Praful Hambarde Satya Naryan Tazi
Prashant Patil Santosh Kumar Vipparthi Subrahmanyam Murala Bilel Benjdira
Anas M. Ali Wadii Boulila Zahra Moammeri Ahmad Mahmoudi-Aznaveh
Ali Karbasi Hossein Motamednia Liangyan Li Guanhua Zhao Kevin Le
Yimo Ning Haoxuan Huang Jun Chen

A. More Challenge Methods and Teams

A.1. NJU_MCG

Description. The NJU_MCG team builds their solution upon the Adaptive Token Dictionary Super-Resolution (ATDSR) model [63]. In addition to the original three-branch architecture, they introduce a new Mamba-based [13] branch. This branch incorporates an OmniShift module [57] before the Mamba unit and transforms the input into four-directional sequences for directional scanning [34]. Furthermore, the team replaces the ConvFFN component of the original ATDSR with a GDFN structure [60], which enhances the model's non-linear representation capacity. The core ATM module structure is illustrated in Figure 1.

Implementation Details. While ATDSR relies primarily on multi-head self-attention (MSA) [35] for feature aggregation, it is limited by window size and subcategory definitions, reducing its ability to convey global pixel-level information. The introduction of a four-directional Mamba branch overcomes this limitation by enabling directional pixel communication with linear computational complexity.

To further boost local feature awareness, the OmniShift module is added prior to Mamba. This ensures each input token encodes spatial context from its neighborhood, thus enhancing local structure recovery. Meanwhile, replacing

[†]Zheng Chen, Kai Liu, Jue Gong, Jingkai Wang, Lei Sun, Zongwei Wu, Yulun Zhang, and Radu Timofte are the challenge organizers, while the other authors participated in the challenge. *Corresponding author: Yulun Zhang.

Section B in the supplementary materials contains the authors' teams and affiliations. NTIRE 2025 webpage: https://cvlai.net/ntire/2025. Code: https://github.com/zhengchen1999/NTIRE2025_ImageSR_x4.

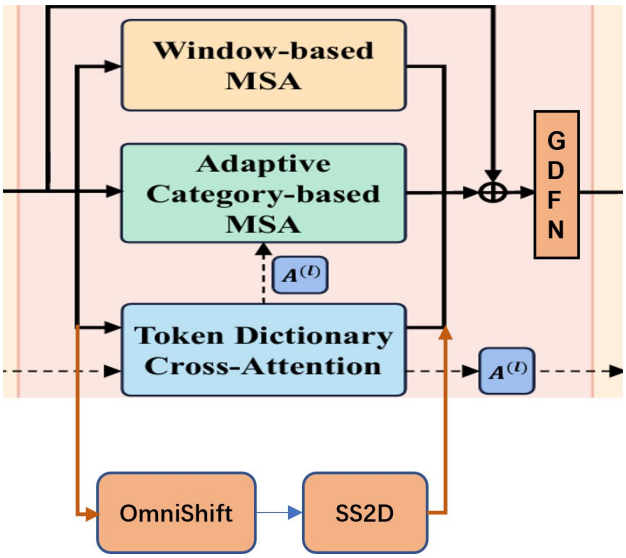


Figure 1. **Team NJU_MCG**

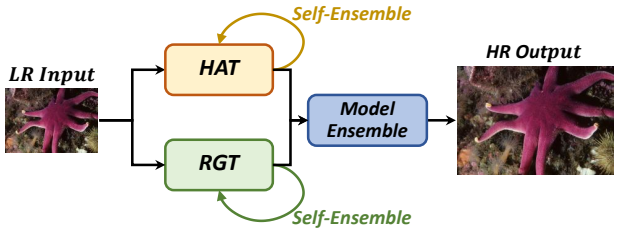


Figure 2. **Team X-L**

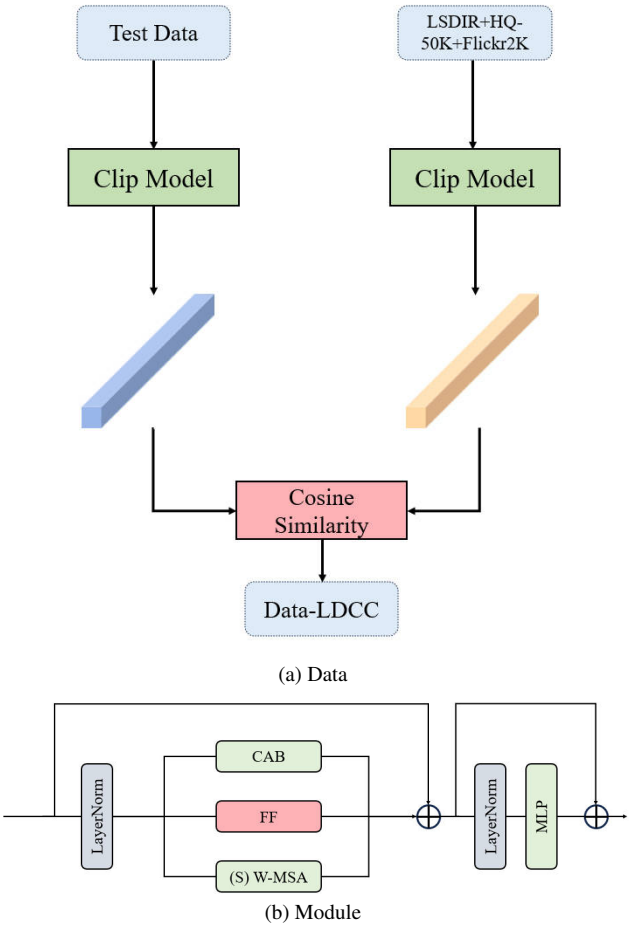


Figure 3. **Team Endeavour**

ConvFFN with GDFN improves both spatial feature modeling and non-linear capacity, leading to stronger pixel-level reasoning performance.

The team trains their ATM model using the Adam optimizer [26] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The training loss is computed using the L1 loss. A MultiStepLR scheduler is adopted, with learning rate milestones set at 300000 and 500000 iterations. The initial learning rate is 2×10^{-4} and is halved at each milestone. The training dataset consists of DIV2K, Flickr2K, and LSDIR.

A.2. X-L

Description. Existing super-resolution methods typically require substantial computational resources. To ensure performance while reducing computational overhead, the team members adopted the following strategy: leveraging two leading approaches, HAT (Hybrid Attention Transformer) [6] and RGT (Recursive Generalization Transformer) [9], the team members directly utilized their pre-trained models to perform self-ensemble, generating two output results. Then, the team members conducted a model ensemble on these two outputs, integrating the results between models to obtain the final reconstruction result. The overall pipeline is shown in Figure 2.

As for the Training strategy, the team members do not require additional training; instead, the team members directly leverage existing methods and their pre-trained models for inference. This approach not only saves significant computational resources and time but also fully utilizes the excellent models and valuable expertise available in the field. By directly employing these pre-trained models, the team members can quickly generate high-quality predictions while avoiding the high costs and complexity associated with training models from scratch.

Implementation Details. The team leverages pretrained HAT [6] and RGT [9] models, which are fine-tuned on DIV2K. During inference, the team members perform self-ensemble and model ensemble to enhance the final results, which do not need training at all.

A.3. Endeavour

Description. The Endeavour team proposes an innovative approach based on the HAT network model [6], enhanced

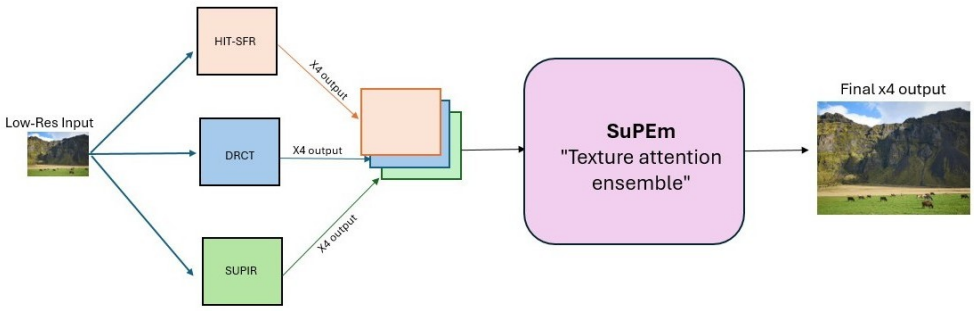


Figure 4. **Team CidautAi**

with a frequency-domain fusion module to improve performance on image restoration tasks. A key feature of the method is the construction of a high-quality dataset called Data-LDCC, specifically designed to bridge the domain gap between training and testing distributions. To enable this, the team uses the CLIP model to extract latent features from both the training candidate data and test data, and then applies cosine similarity [47] to select the most semantically aligned samples. The final Data-LDCC dataset includes selected images from LSDIR [27], HQ-50K, and Flickr2K [29]. The dataset construction pipeline is shown in Figure 3a.

Network Architecture. The proposed model is built upon HAT, which incorporates channel attention and window-based self-attention mechanisms. To further enhance detail reconstruction, a frequency-domain fusion module is added. This module operates in the frequency domain and complements the spatial features extracted by the base network, thereby improving the model's ability to represent high-frequency information. The complete network structure is illustrated in Figure 3b.

Implementation Details. The training process consists of three distinct stages. In the first stage, the model is pretrained on the ImageNet dataset to establish a strong initial representation. In the second stage, it is optimized on the DF2K dataset, composed of DIV2K and Flickr2K images, to enhance generalization. Finally, in the third stage, fine-tuning is performed jointly on DF2K and Data-LDCC, which enables the model to adapt to the target test domain more effectively.

1. **Stage 1:** The model is trained on ImageNet for 70,000 iterations with a patch size of 64×64 and a batch size of 256.
2. **Stage 2:** Training continues on DF2K for 50,000 iterations, with a larger patch size of 96×96 and a reduced batch size of 128.
3. **Stage 3:** The model is fine-tuned for 20,000 iterations on DF2K + Data-LDCC using 112×112 patches and a batch size of 64.

This progressive training pipeline allows the model to gradually adapt to more complex data and resolution scales. The training strategy includes fixed learning rates within each stage and checkpoint-based model selection.

A.4. CidautAi

Description. The team members propose a $x4$ SR solution for both tracks, however, the results might be more optimal if the team members consider **track 2) perceptual metrics**.

The model is a learned ensemble of state-of-the-art super-resolution techniques. It takes as input the results of three super-resolution models: (i) two models optimized for fidelity and minimal distortion, MSE (PSNR) [18, 65]. (ii) one perceptual model, SUPIR [58], designed to enhance perceptual metrics and visual quality from a human perspective.

The model is based on DRCT [18]. Using attention mechanisms designed to enhance textures, the results of DRCT are combined with the outputs of the other two models (HIT-SCR [65] and SUPIR [58]). the team members illustrate this in Figure 4.

Implementation Details. For training the model, a combined loss function was used to balance maintaining good PSNR performance while enhancing visual details.
$L = L_2(\hat{x}-x)+(1-SSIM(\hat{x}-x))+0.5\cdot Lpips(\hat{x}-x)$.
The team members used the DIV2K and FLICKR2K datasets for training, which were preprocessed through the three models before being fed into the ensemble model. The images generated by the three models were concatenated and processed through the network with a batch size of 16 and 512 for crop size.

The results obtained have been quite encouraging with respect to the goal. Although the PSNR of SuPEm is slightly lower than that of DRCT, perceptual metrics show a considerable improvement. The results differ from those

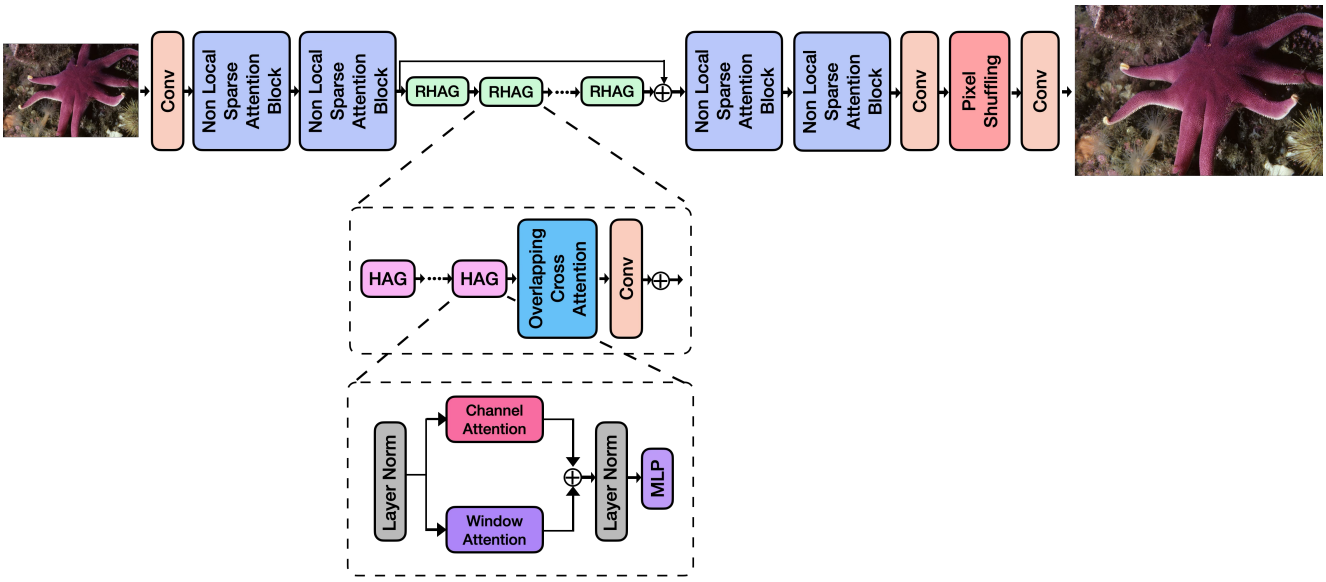


Figure 5. **KLETech-CEVI**

reported as the metrics were calculated using techniques different from those provided in the challenge.

These results are consistent with those obtained from the other test datasets used for the super-resolution problem: DIV2K, LSDIR, Urban100, Manga109, Set5, and Set14. Contact us if you want to know the results.

Despite using three pre-trained models, the approach has a certain degree of novelty, as it aims to leverage the advantages of perceptual-focused models and integrate them into the results of other models through attention mechanisms.

A.5. KLETech-CEVI

Description. The proposed framework, WHAT (Wavelet Hybrid Attention Transformer), is built upon a pre-trained HAT-L model and designed to enhance $\times 4$ image super-resolution by integrating hybrid attention and frequency-aware learning. The architecture introduces Non-Local Sparse Attention (NLSA) blocks to improve long-range dependency modeling. Training is performed on LSDIR and DIV2K with standard augmentations. The framework is further guided by a composite loss function combining pixel-wise, wavelet, perceptual, and MS-SSIM terms. It requires 33 hours of training and achieves an inference speed of 0.9s per image.

Implementation Details. The proposed super-resolution framework integrates hybrid attention mechanisms and wavelet-based loss functions to enhance image quality. The methodology consists of the following key components.

Hybrid Attention Architecture. To improve the receptive field and enhance feature learning, the model builds upon the Hybrid Attention Transformer (HAT) [6] framework while incorporating Non-Local Sparse Attention (NLSA) blocks. The architecture follows these enhancements. NLSA blocks are added before and after the core HAT model to strengthen global feature aggregation and long-range dependencies. The transformer backbone uses a pre-trained HAT-L model with a default configuration where the embedding dimension is set to 180, the patch embedding size is 4, and the total number of trainable parameters is 41.3 M. This structure helps in capturing both local and global features efficiently, making it suitable for super-resolution $\times 4$ tasks. The overall architecture is shown in Figure 5.

Wavelet-Based Loss Function. To improve the recovery of high-frequency details, the model incorporates wavelet-domain losses alongside conventional RGB pixel-wise losses. A Symlet filter is used to compute wavelet coefficients. In the wavelet loss function, the weights for sub-bands λ_j are set to 0.05 to prevent chroma artifacts. The loss balances low- and high-frequency details, preserving textures and edges. The loss function is inspired by [24] and is defined as:

$$L_G = L_{\text{RGB}} + L_{\text{Wavelet}} + L_{\text{MS-SSIM}} + L_{\text{VGG}} \quad (1)$$

where L_{RGB} is the pixel-wise loss in the RGB domain, and L_{Wavelet} is the wavelet-based loss that ensures sharp details in the reconstructed image.

A.6. JNU620

Description. Inspired by the recent success of Transformer- and Mamba-based architectures, the JNU620 team proposes a framework for image super-resolution based on pre-trained models and ensemble learning, named PMELSR. As shown in Figure 6, the pipeline consists of two stages and follows an ensemble learning paradigm. In the first stage, the team employs three powerful pre-trained back-

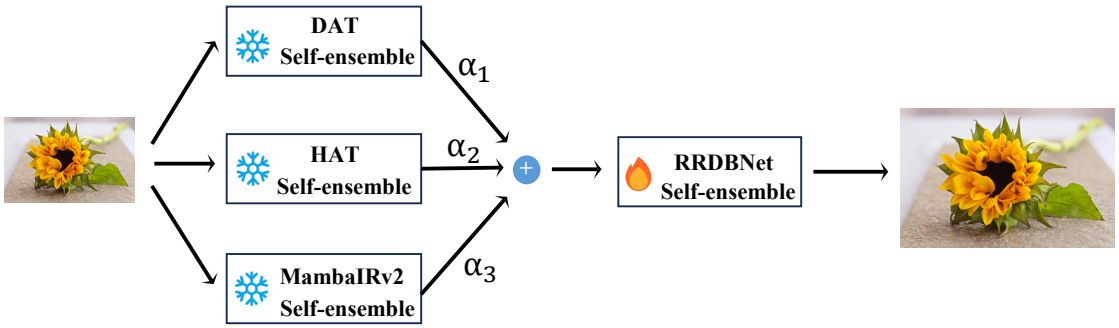


Figure 6. **Team JNU620**

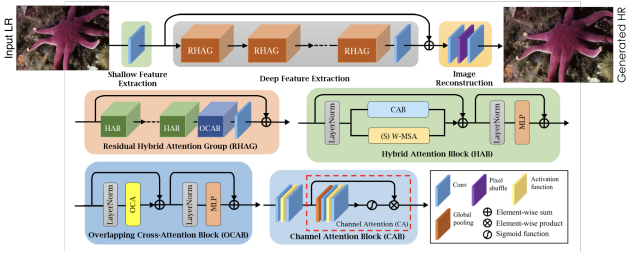


Figure 7. **Team ACVLAB**

bones—DAT [7], HAT-L [6], and MambaIRv2 [14]—to process the low-resolution input independently. The outputs are then fused to generate a preliminary super-resolved result. In the second stage, a RRDBNet [51] refinement module is used to further optimize the fused results and enhance the final image quality. Although this design increases complexity, it is intended to maximize performance.

Implementation Details. During training, the parameters of the pre-trained models (DAT, HAT-L, and MambaIRv2) remain fixed, and only the RRDBNet refinement module is updated. The model is optimized using the Adam optimizer with the *L1* loss function, and data augmentations such as random flips and rotations are applied. Training is conducted on the DIV2K dataset, with HR patches of size 512×512 randomly cropped from the full-resolution images. The batch size is set to 4, and training runs for a total of 400K iterations. The initial learning rate is set to 2e-4 and decayed using a cosine annealing schedule.

In the testing phase, the team adopts two ensemble strategies to reduce prediction bias: *self-ensemble* and *model ensemble*. First, the self-ensemble technique is applied to each pre-trained model to enhance individual performance. Then, the outputs of all models are combined through model ensemble. Finally, the fused results are further refined using RRDBNet, again applying self-ensemble to improve the final output quality.

A.7. ACVLAB

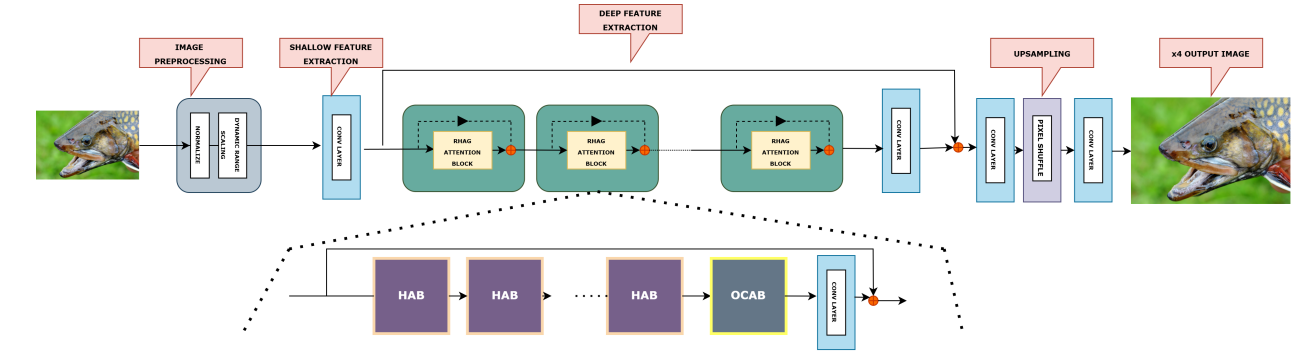
Description. The proposed method is built upon the Hybrid Attention Transformer (HAT) [6] and enhanced through self-ensemble fusion strategies. The approach aims to improve reconstruction performance by leveraging multiple training stages and combining diverse model checkpoints. The training is conducted on the LSDIR and DIV2K datasets, with data augmentations including random flips and rotations. The method is designed to progressively refine model accuracy through careful learning rate scheduling and loss function transitions.

Implementation Details. The training process is divided into two phases, each consisting of 800,000 iterations. The Adam optimizer is used with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the initial learning rate is set to 2×10^{-4}. A multi-step learning rate scheduler reduces the learning rate at iterations 300,000, 500,000, 650,000, 700,000, and 750,000. No weight decay is applied. High-resolution training patches of size 256×256 are extracted and augmented using horizontal flips and rotations. The first training phase optimizes the model using L1 loss with a batch size of 16, while the second phase switches to MSE loss for further refinement. The method is implemented in PyTorch 1.13.1 and trained on two NVIDIA GeForce RTX 3090 GPUs. The final model integrates multiple trained checkpoints for ensemble inference.

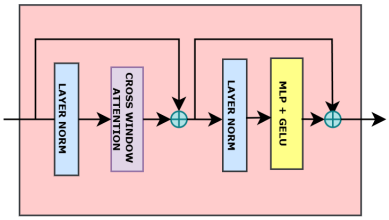
A.8. CV_SVNIT

Description. The single-image super-resolution model utilizes the Hybrid Attention Transformer (HAT) [6] architecture, incorporating pretrained weights to enhance overall performance. The network is divided into three primary stages: shallow feature extraction, deep feature extraction, and image reconstruction. Each stage plays a crucial role in capturing relevant features from low-resolution input images to reconstruct HR outputs with high fidelity.

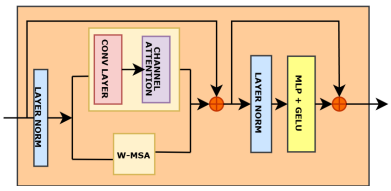
The model is designed for single-image super-resolution with a scaling factor of $\times 4$. It consists of 6 attention blocks and a depth of 6 layers. Pretrained weights were used to



(a) The block schematic of the proposed architecture for scaling factors $\times 4$



(b) OVERLAPPING CROSS ATTENTION BLOCK (OCAB)



(c) HYBRID ATTENTION BLOCK (HAB)

Figure 8. **Team CV_SVNIT**

initialize the model, which allows for faster convergence and improved accuracy during the training process. The input low-resolution (LR) image is processed through shallow feature extraction layers that capture important low-frequency information. Deep feature extraction layers are used to extract high-frequency details essential for producing high-quality super-resolved images.

The model uses the Gaussian Error Linear Unit (GELU) activation function, which helps maintain smooth gradients and enhances learning performance. Global residual connections are employed to combine shallow and deep features, enabling a comprehensive understanding of the image across multiple layers. The reconstructed image is produced using the gathered feature information in the final stage.

Hybrid attention blocks (HAB) are applied to preserve high-frequency details by selectively emphasizing important features. Each residual attention block includes attention mechanism blocks (AMB), an overlapping cross attention block (OCAB), and a 3×3 convolution layer with residual connections. Channel attention blocks in the AMB perform adaptive re-scaling of features on a per-channel basis to refine the reconstruction. The pixel shuffle operation is used to upscale the feature maps to the desired $\times 4$ factor.

To ensure smoothness and reduce artifacts in the output image, total variation (TV) loss is combined with Charbonnier loss. Additionally, the team members incorporated L1 loss and SSIM (Structural Similarity Index Measure) loss to improve the image's structural quality.

The overall loss function is defined as:

$$\text{Loss} = B_1 \cdot \text{L1\_Loss} + B_2 \cdot \text{SSIM} + B_3 \cdot \text{TV\_Loss} \quad (2)$$

where,

$$B_1 = 1, \quad B_2 = 0.5, \quad B_3 = 10^{-1} \quad (3)$$

Implementation Details. The model was trained using the DIV2K and LSDIR datasets. The architecture consists of 6 attention heads with a window size of 16, while the number of hybrid attention blocks (HAB) and attention mechanism blocks (AMB) is set to 6. The number of channels is set to 64.

The learning rate is initialized at 5×10^{-5} and decays by half every 10k iterations. The model was trained for a total of 100k iterations with a batch size of 2. The Adam optimizer is used to minimize the combination of L1 loss, SSIM, and TV loss, optimizing performance [6].

A.9. HyperPix

Description. The proposed solution integrates a Transformer-CNN hybrid architecture with advanced attention mechanisms to address single-image super-resolution. The network is designed to efficiently extract both low-frequency and high-frequency features, leveraging Hybrid Attention Blocks (HABs) and Overlapping Cross-Attention Blocks (OCABs). This approach significantly improves feature representation and super-resolution performance. The model is further enhanced through specialized pre-training on large-scale datasets, improving generalization across diverse low-resolution inputs.

Implementation Details. The proposed super-resolution

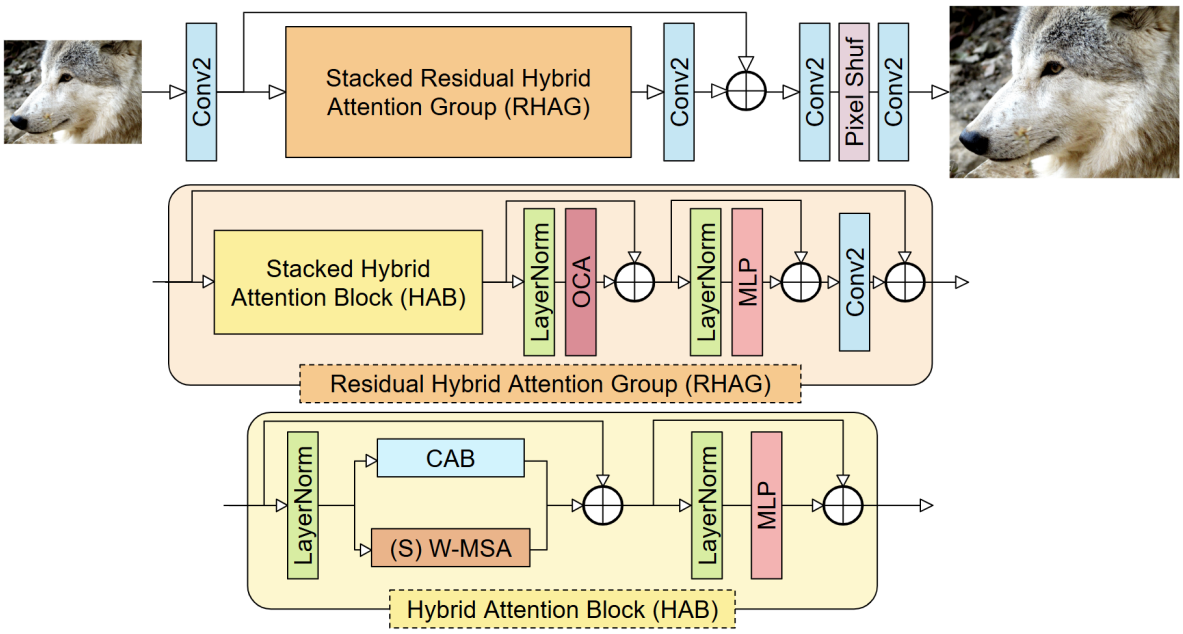


Figure 9. **Team HyperPix**

framework consists of three main components: shallow feature extraction, deep feature extraction, and image reconstruction. The low-resolution (LR) input I_{LR} is passed through a convolutional layer for shallow feature extraction, followed by a series of Residual Hybrid Attention Groups (RHAG) and a 3x3 convolution layer for deep feature extraction. A global residual connection fuses shallow and deep features, which are then upsampled using a pixel shuffle operation.

Hybrid Attention Block (HAB). The HAB is designed to enhance pixel activation through channel attention and self-attention. The HAB works in tandem with the Swin Transformer block to improve the network's feature representation. Each HAB contains a channel attention block (CAB) and a window-based multi-head self-attention (W-MSA) module, with self-attention computed within local windows of size $M \times M$.

Overlapping Cross-Attention Block (OCAB). The OCAB further enhances representation capability by establishing direct cross-window connections. Unlike traditional self-attention, OCAB employs an overlapping cross-attention mechanism to compute attention within pixel tokens across windows of varying sizes, enabling richer feature extraction and stronger inter-window connections.

For training, the model uses the DIV2K and LSDIR datasets. The depth and width of the network are comparable to SwinIR, with 6 Hybrid Attention Blocks (HABs) and 6 DAT Blocks (DATBs). The model is trained for 20,000 iterations with a batch size of 8, using a learning rate of 1×10^{-4}, which decays by 0.5 every 10,000 iterations. The Adam optimizer is used, and the model is implemented in PyTorch. The pre-training strategy leverages large-scale datasets to improve generalization, and the network parameters are optimized using the L1 loss function.

A.10. BVIVSR

Description. The BVIVSR team proposes HIMambaSR, a novel architecture that integrates the strengths of MambaIRv2-B [14] and HIIF [22] to enhance super-resolution performance. As shown in Figure 10, the team adopts the MambaIRv2-B model—excluding its upsampling modules—as a latent encoder E_{φ}. The core of MambaIRv2 consists of a sequence of Attentive State Space Groups (ASSG), each comprising multiple Attentive State Space Blocks (ASSBs). Within each ASSB, a local-to-global modeling strategy is applied using Window Multi-Head Self-Attention (MHSA) for local feature capture and an Attentive State Space Model (ASSM) for global dependencies. Each block follows a "Norm $\rightarrow$ Token Mixer $\rightarrow$ Norm $\rightarrow$ FFN" design, incorporating two residual connections with learnable scaling. This encoder is responsible for extracting latent features from the low-resolution image.

The decoder D_{ϱ} is built on HIIF, which reconstructs high-resolution images from the extracted features. The HIIF decoder includes a multi-scale hierarchical encoding module, multiple multi-head linear attention blocks, and MLPs. It uses hierarchical positional encoding to capture the local implicit image function across multiple scales. These encodings are progressively injected throughout the network, promoting effective feature propagation and en-

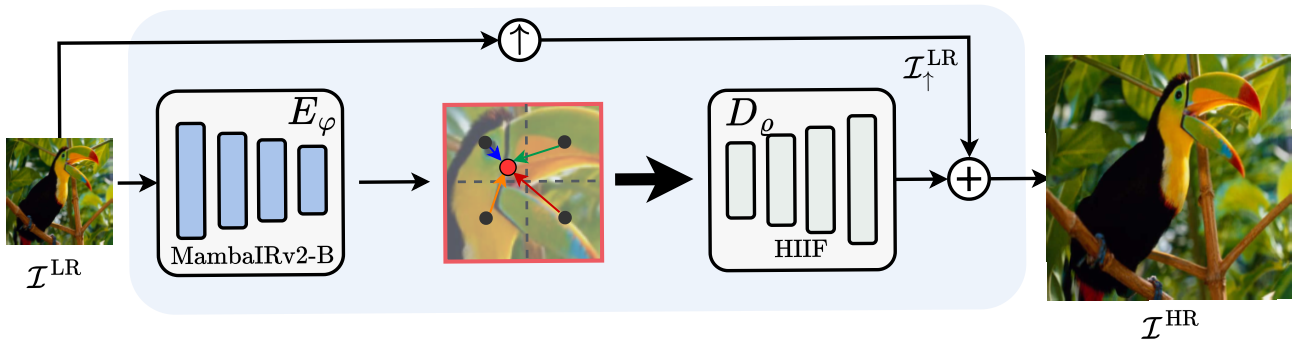


Figure 10. **Team BVIVSR**

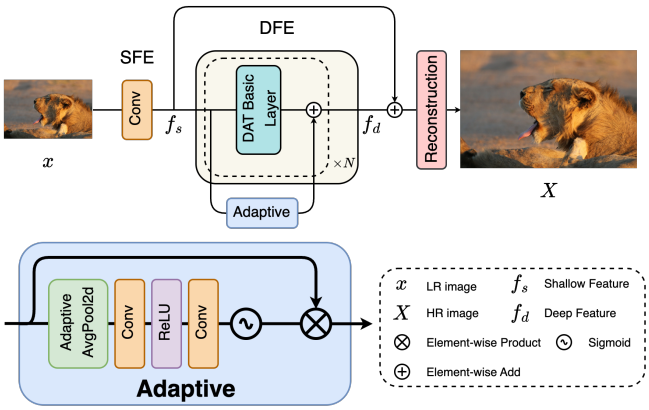


Figure 11. **Team AdaDat**

hancing the ability to recover high-frequency details.

By leveraging the representation capacity of MambaIRv2-B in latent feature encoding and the continuous-scale decoding capability of HIIF, the proposed HIMambaSR model demonstrates flexible and high-quality super-resolution performance.

Implementation Details. The team adopts the original configurations of both MambaIRv2-B and HIIF. The training dataset includes DIV2K [48], 1,000 2K-resolution images from BVI-AOM [40], Flickr2K [29], and 5,000 images from LSDIR [27]. For evaluation, the team follows the standard protocol [22, 23] in continuous super-resolution and uses the DIV2K validation set containing 100 images. The learning rate is set to a maximum of 4×10^{-4} and follows a cosine annealing schedule with a warm-up period of 50 epochs. The model is optimized using the L1 loss and the Adam optimizer [26]. Training and testing are conducted on four A100 GPUs. The final model contains approximately **24.5M** parameters and is trained with resolution of $64 \times 64 \times 3$, using a batch size of 48 for 1,000 epochs.

A.11. AdaDAT

Description. Adaptive DAT (AdaDAT) is an enhanced version of DAT that can adapt to new datasets with the pretrained DAT model [7] and the adaptive layer. The motivation of AdaDAT is that the team members observe that most existing SR models have achieved great performance, and the team members aim to leverage the power of these pretrained models to adapt to new datasets or tasks. Figure 11 illustrates the overall pipeline of the proposed AdaDAT. Given a low-resolution image x, the team members first use the shallow feature extractor (SFE) to extract the shallow features f_s, where the SFE consists of a single convolutional layer. Then, the team members use the deep feature extractor (DFE) to extract the deep features f_d. The DFE consists of N blocks, where each block consists of a pretrained DAT basic layer and an adaptive layer. The pretrained DAT basic layer is used to extract the basic deep features, while the adaptive layer is used to introduce the adaptive information. Given the input features, in the adaptive layer, the team members first use a 2d adaptive average pooling layer (Adaptive AvgPool2d) to process the input features and then use 2 convolutional layers with a ReLU activation function in between to process the output of the pooling layer. The sigmoid function is used to generate the attention map, which is then multiplied with the input features to generate the output of the adaptive layer. The output of the adaptive layer is then added to the output of the pretrained DAT basic layer to generate f_d. The f_d is then added with f_s to serve as the input of the reconstruction layer. The reconstruction layer is used to reconstruct the high-resolution image X.

Implementation Details. the proposed AdaDAT is trained on the LSDIR dataset [27]. The team members employed the Adam optimizer with learning rate of 1e-5, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, running for 250K iterations. The learning rate is reduced by half at iterations [125K, 200K, 225K, 237K]. The batch size is set to 32. The input image is set to 64×64. The L1 loss is used as the loss function to op-

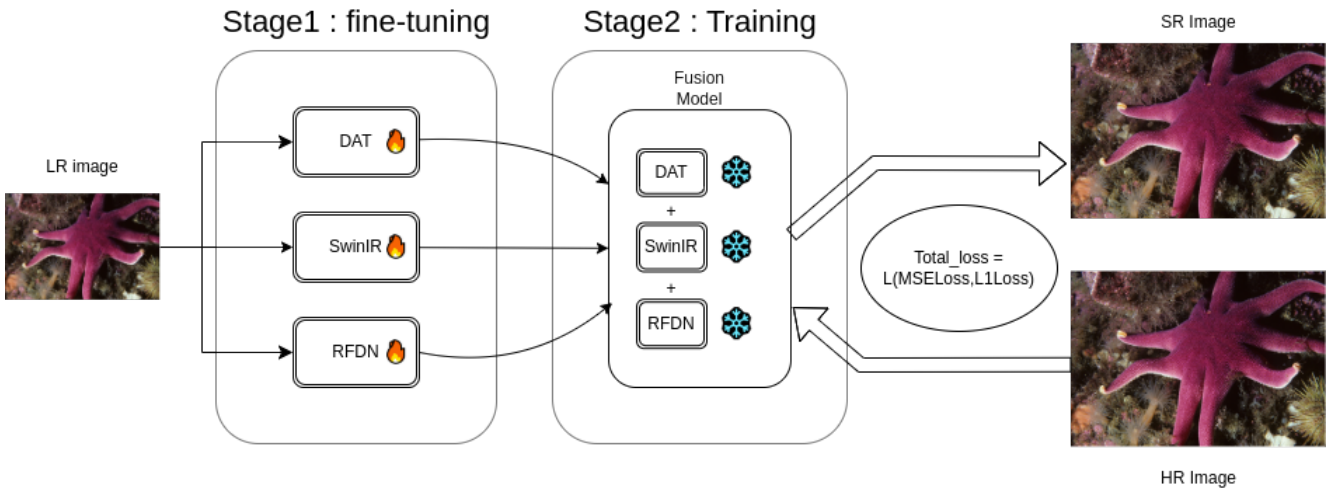


Figure 12. **Team Junyi**

timize the model. The proposed AdaDAT is trained in two phases. In the first phase, the team members optimized only the adaptive layer and fixed the pretrained DAT basic layers. In the second phase, the team members optimized both the adaptive layer and the pretrained DAT basic layers. The first phase is used to adapt the model to the new dataset, while the second phase is used to fine-tune the whole model. In the first 125K iterations, the team members trained the first phase, and then the team members trained the second phase for the rest of the iterations.

A.12. Junyi

Description. 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 [7] and the Shifted Window Transformer (SwinIR) model [28].

To investigate whether these models can achieve enhanced super-resolution performance through fusion techniques, a series of model fusion experiments were conducted based on three foundational architectures: DAT, SwinIR, and the Residual Feature Distillation Network (RFDN) [32]. 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 the primary investigation, a secondary fusion paradigm was further implemented by integrating the Real-ESRGAN [51, 52] 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, the first fusion configuration was conclusively adopted as the optimal solution, thereby establishing its technical predominance in balancing reconstruction fidelity and computational efficiency within the experimental framework.

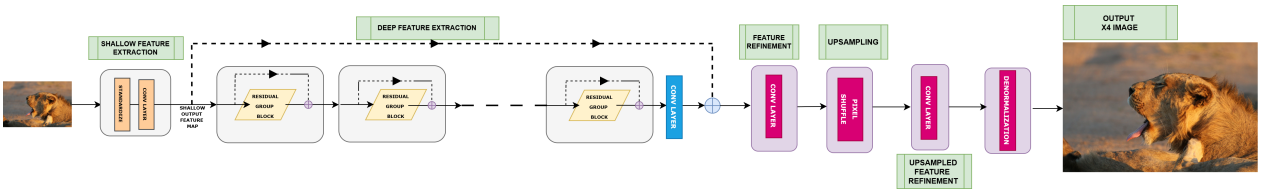
Implementation Details. The fusion methodology employs an output-level model fusion approach. Specifically, three baseline models—DAT, SwinIR, and RFDN—were fine-tuned. A lightweight attention-based weight allocation network was then devised to dynamically optimize the fusion coefficients among the three 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 both quantitative metrics and visual quality in the fused super-resolution output.

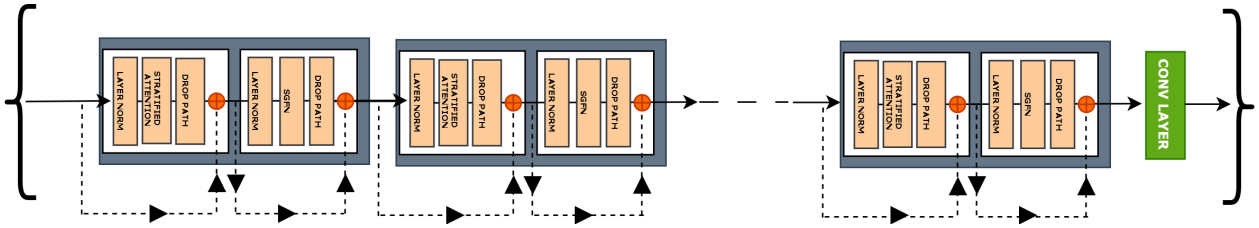
The architectural configuration of the proposed model is illustrated in Figure 12.

Training Strategy. All experimental procedures were conducted using the DIV2K dataset for model development and evaluation. The fine-tuning of three baseline models was the starting point for the training process, with key hyper-parameters documented for reproducibility.

Once the baseline models were trained, the weight prediction network was optimized over 150 epochs with a batch



(a) The block schematic of the proposed architecture for scaling factor $\times 4$.



(b) The schematic of the DAT block with triple attention.

Figure 13. **Team ML_SVNIT**

size of 72, requiring approximately 5 hours of training using an NVIDIA GeForce RTX 4090 GPU. The Adam optimizer was used with combined MSE and L1 loss functions. The final fusion model consisted of 28.4K parameters and was trained without extra data.

A.13. ML_SVNIT

Description. In order to design single-image super-resolution, the proposed solution employs a Transformer and CNN-based network approach. As shown in Figure 13, the overall network consists of three parts, including shallow feature extraction, deep feature extraction, and image reconstruction.

Implementation Details. The proposed pyramid attention and dual aggregate transformer (DAT) architecture for single-image super-resolution with a scaling factor of $\times 4$ is depicted in Figure 13 (a). The low-resolution (LR) image is passed through the network to extract both low-frequency and high-frequency features. Initially, a shallow feature extraction module employs a convolutional layer to capture essential low-frequency details.

The extracted features are then processed through a deep feature extraction module, which incorporates multiple residual group blocks (RBG) and dual aggregate transformer (DATB) blocks. A global residual connection fuses shallow and deep features, ensuring efficient information propagation.

To reconstruct the high-resolution (HR) image, the reconstruction module refines the extracted features and upsamples them using a pixel shuffle operation. This preserves high-frequency details and enhances super-resolution performance. The adaptive self-attention and pyramid attention mechanisms further improve feature representation, making the model highly effective for real-world super-resolution tasks [6, 7].

The DIV2K and LSDIR datasets are used for training. The depth and width of the model are kept the same as SwinIR. Specifically, the RGB number and DATB number are both set to 6. The model is trained for up to 2×10^4 iterations with a batch size of 8. The attention head number and depth are set to 4 each. The code is implemented using the PyTorch library. The loss function is Charbonnier loss with a learning rate of 1×10^{-4}, which is decayed by 0.5 every 1×10^4 iterations. The model is optimized using the Adam optimizer.

A.14. SAK_DCU

Description. The network architecture of the proposed method is shown in Figure 15a and Figure 15b. The proposed image super-resolution framework, SAKSRNet, as shown in Figure 14,, integrates multiple components that are designed to enhance feature extraction, preserve fine details, and improve upsampling quality. The architecture consists of a multi-scale convolution block (MSCB) [42], gated convolution feature enhancement (GCFE) module [4], Swin Transformer block (STB) [28], and a lightweight recurrent mechanism [53], which includes gated convolution (GConv) and progressive pixel-shuffle upsampling (PPSU) [61].

The network takes a low-resolution (LR) image as input, which undergoes hierarchical feature extraction through the MSCB. This block captures multi-scale spatial and contextual information using varied strides and kernel sizes. It also employs residual connections to maintain training stability.

Next, the GCFE module replaces traditional convolutions with gated convolutions, introducing dynamic feature selection to suppress noise while preserving edges and texture details. This is followed by the Swin Transformer block, which enhances long-range feature modeling via hierarchical window-based attention and LayerNorm, effec-

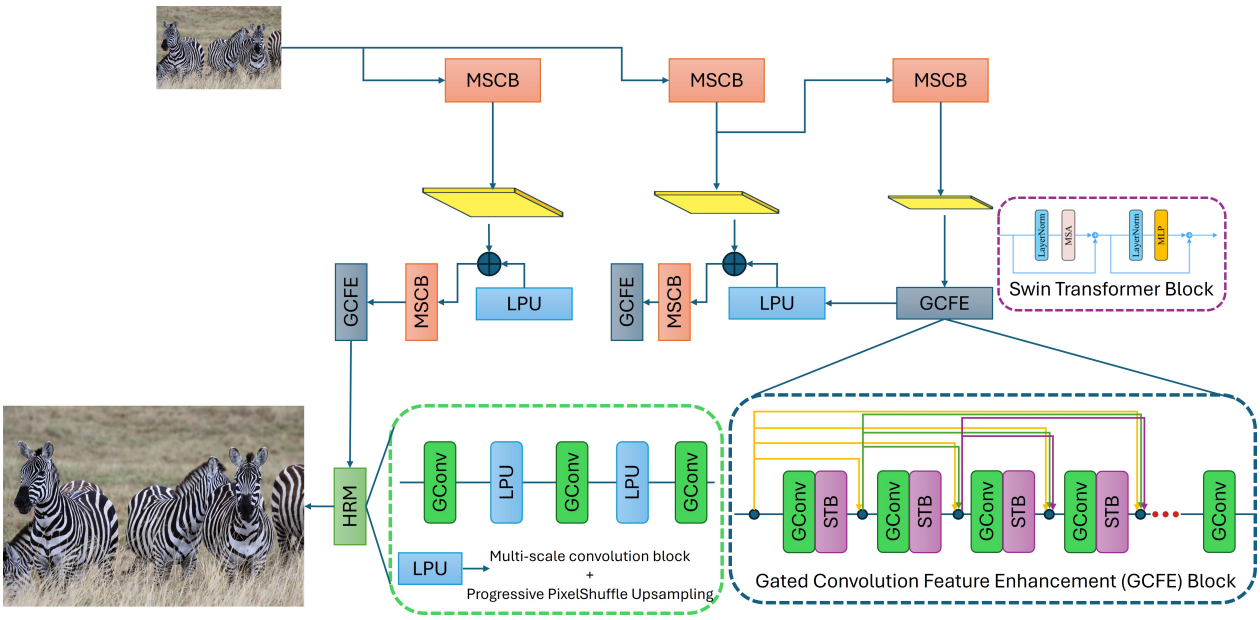
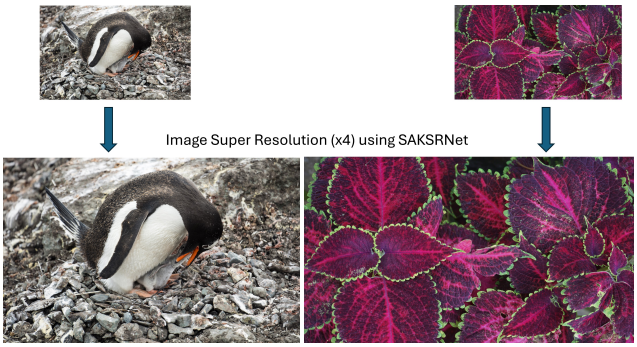
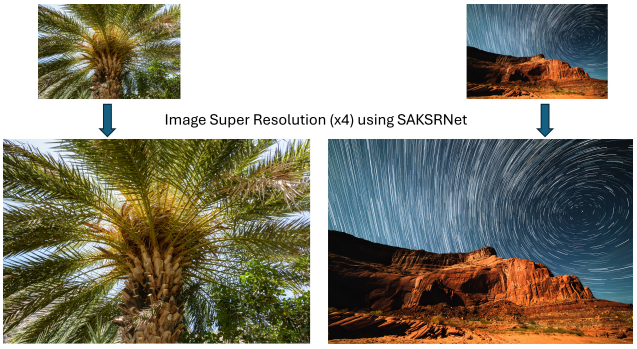


Figure 14. **Team SAK_DCU**



(a) **Team SAK_DCU Result**



(b) **Team SAK_DCU Result**

Figure 15. **Team SAK_DCU**

tively refining global structures.

The recurrent module, named the lightweight processing unit, is integrated to optimize spatial feature aggregation across varied receptive fields in a memory-efficient manner. Finally, the PPSU module performs resolution enhancement via sub-pixel convolutions. Its progressive design enables fine-grained sharpening of textures and reduction of artifacts in the HR output.

Implementation Details. The network is trained using the Adam optimizer [26] with a composite loss function composed of L1 loss, perceptual loss, and adversarial loss [66], aiming to balance pixel accuracy, perceptual similarity, and realism. The initial learning rate is set to 2×10^{-4} and decayed to 10^{-6} via a cosine annealing schedule. A warm-up strategy is applied in the first 5000 iterations to stabilize training.

Training is conducted on an A6000 GPU with 48GB memory. The batch size is set to 16, and patch size is 64×64. Data augmentation strategies include random rotations, horizontal and vertical flips, and Gaussian noise injection. No external datasets are used beyond the provided training data.

The team implements the method based on the RTSR codebase (NTIRE 2023), and intends to submit an extended version of this work to a journal. Visual results demonstrating the model's performance are presented in Figures 15a and 15b.

A.15. VAI-GM

Description. The VAI-GM team builds upon the DRCT methodology [8] to further improve its performance on $4\times$ super-resolution (SR). As illustrated in Figure 16, the base network is DRCT, an established state-of-the-art method for $4\times$ SR. The team enhances the original architecture by introducing additional skip connections between every two

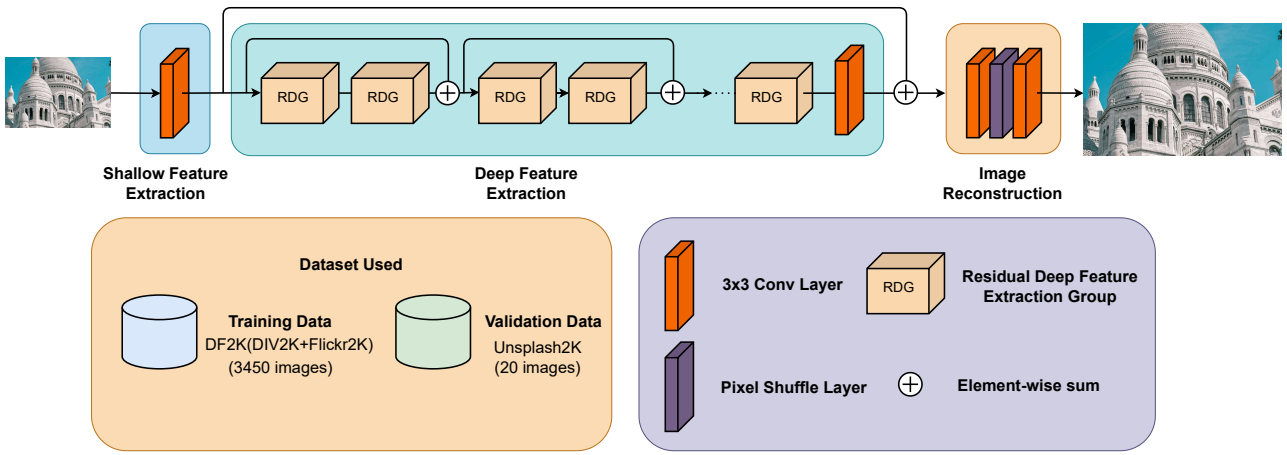


Figure 16. **Team VAI-GM**

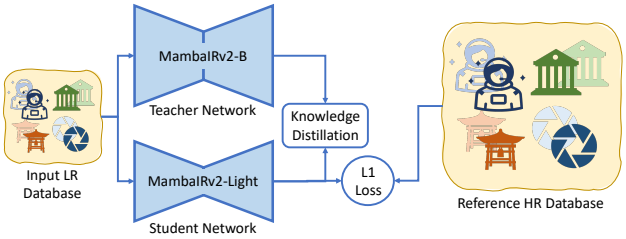


Figure 17. **Team Quantum Res**

RDG (Residual Dense Group) blocks and employing 12 RDG blocks in total. These *multi-step* residual connections allow features from shallower layers to be preserved and passed to deeper layers, mitigating gradient vanishing and helping retain low-level details critical for high-quality image reconstruction.

This architectural enhancement draws inspiration from traditional residual learning strategies such as ResNet [17], as well as more recent transformer-based designs like HAT [6], which combine hierarchical attention with skip connections to maintain robust feature flow. By integrating these principles into the DRCT framework, the team improves information propagation and training stability, resulting in better image quality.

Implementation Details. Training is conducted on 800 images from DIV2K [48] and 2,650 images from Flickr2K [29], with validation on 20 images from Unsplash2K [25]. Pretraining is used by initializing from a DRCT-L model trained on ImageNet and fine-tuning it on the SR task. The model is optimized using Adam with a learning rate of 5×10^{-4}, weight decay set to 0, and β values of 0.9 and 0.99. The learning rate is scheduled using MultiStepLR, with milestones at 125000, 200000, 225000, and 240000 iterations, and a decay factor of 0.5.

Training is run for 250,000 iterations. Data augmentations include random horizontal flips and rotations. The batch size is set to 4 per GPU, with a ground-truth patch size of 256. The model is trained using the L1 loss, which minimizes the average absolute difference between the predicted SR image $\hat{I}$ and the HR ground truth image I:

$$\mathcal{L}_{L1} = \frac{1}{N} \sum_{i=1}^{N} \left| \hat{I}_i - I_i \right| \tag{4}$$

where N is the total number of pixels. This loss provides robustness to various noise levels in super-resolved outputs.

A.16. Quantum Res

Description. In this work, the team members propose a novel student-teacher framework for super-resolution as shown in Figure 17 that enables a lightweight student model to achieve better performance comparable to heavier models. Specifically, to adopt this architecture the team members used MambaIRv2-Light [14] as the student model, while MambaIRv2-base [14] serves as the teacher. While the team members use MambaIRv2-light as an efficiency, the key contribution is demonstrating that a guided student-teacher learning strategy can significantly improve SR performance while keeping model complexity low [50].

The student model extracts the initial low-level features from the input low-resolution image using the 3×3 convolutional layer. The core of the network comprises a series of Attentive State-Space Blocks (ASSBs) [14]to capture long-range dependencies efficiently. For each block, residual connections are used to facilitate stable gradient propagation. Finally, a pixel-shuffle-based upsampling module reconstructs the final high-resolution image. [14]

Mathematically, the feature extraction and transformation process in a single ASSB can be formulated as:

$$F_{\text{out}} = \text{ASSB}(F_{\text{in}}) + F_{\text{in}} \tag{5}$$

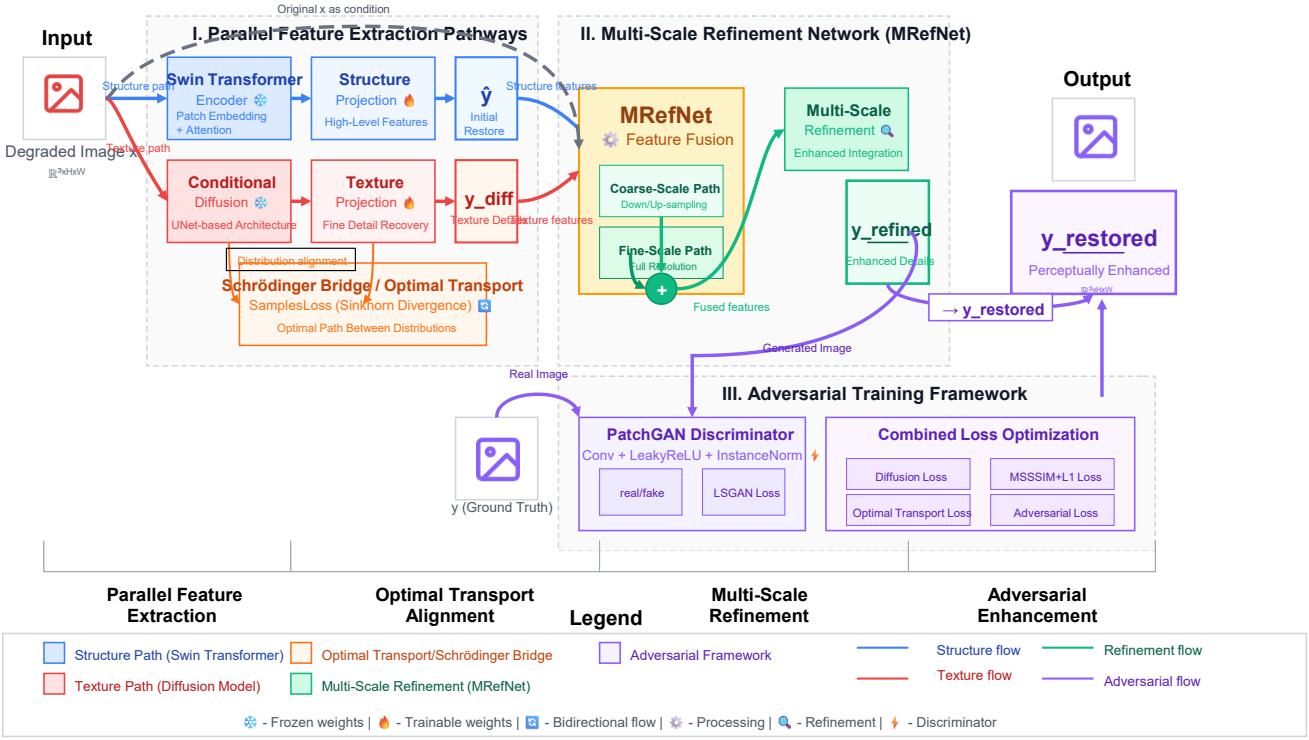


Figure 18. **Team PSU**

where F_{in} and F_{out} are the input and output feature maps, respectively. [14]

The teacher model, MambaIRv2, follows the same architectural design but with increased depth and wider feature dimensions. This model has significantly more parameters and serves as an upper-bound reference for the student.

Teacher-Guided Inference: The teacher model remains frozen throughout training and is only used as a qualitative reference to validate architectural choices and improvements. The student model inherits refined architectural principles from the teacher rather than weight transfer or feature alignment. This allows the student to retain its original lightweight nature while benefiting from structural knowledge obtained from a larger-capacity model. [50]

Inference Strategy: During inference, an efficient patch-based processing method is applied to handle high-resolution images. Given an input image, it is divided into overlapping patches. Each patch is processed independently by the student network, and final predictions are blended using a weighted averaging scheme to ensure seamless reconstruction. [14]

$$I^{SR} = \sum_k W_k S(P_k) \tag{6}$$

where P_k are the patches processed independently by the student network, W_k are blending weights ensuring smooth transitions between adjacent patches.

Implementation Details. The student model is initialized using pre-trained weights of MambaIRv2-light. The teacher model is loaded with pre-trained weights from a high-performing MambaIRv2-base variant.

Fine-tuning was performed on DIV2K and LSDIR, with the number of feature channels set to 48. The training was conducted on patches of size 192×192 extracted from high-resolution images, using a batch size of 8. The model is finetunned by minimizing the L1 loss function using the Adam optimizer.

The initial learning rate is set to 1×10^{-5} and is reduced when training iterations reach specific milestones, following a MultiStepLR decay strategy with a factor of 0.5. The total number of iterations is 150k. The teacher model is only used as a reference for guiding architectural refinement and remains frozen throughout the training.

A.17. PSU

Description. The PSU team proposes a novel deep learning framework named OptiMalDiff, which formulates image restoration as an optimal transport problem based on Schrödinger Bridge theory. The method aims to learn the most efficient stochastic path connecting the distributions of degraded and clean images. As shown in Figure 18, the architecture integrates multiple components:

- **Hierarchical Swin Transformer Backbone:** Utilizes an encoder-decoder structure with window-based multi-head

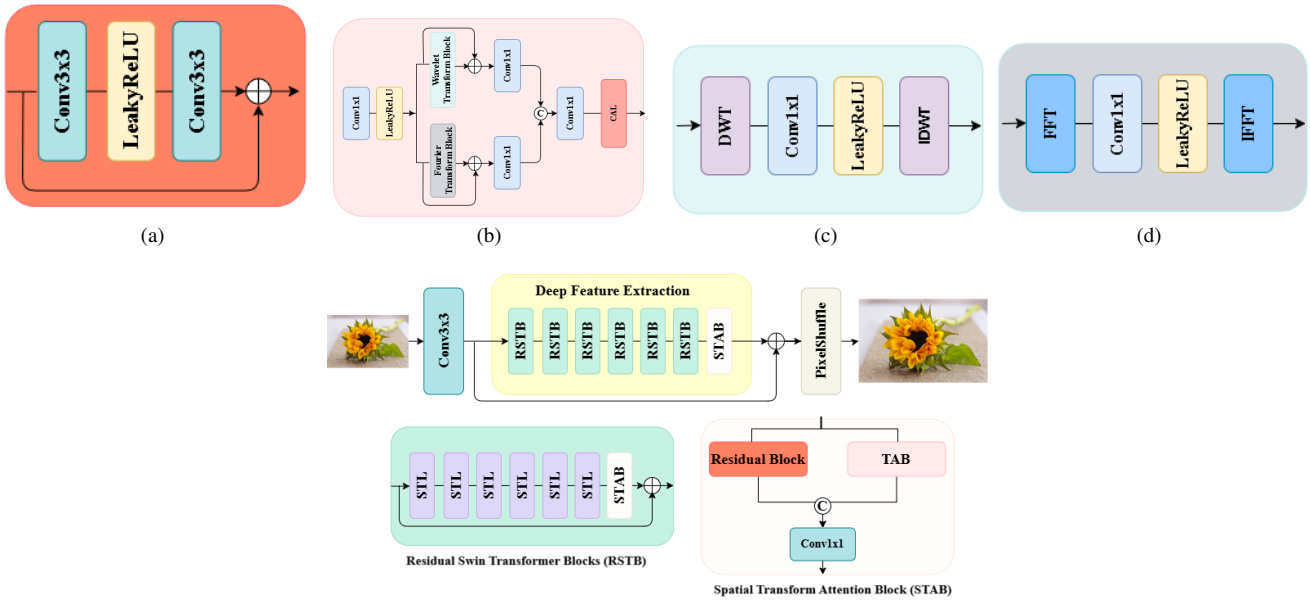


Figure 19. **Team IVPLAB-sbu**

self-attention to efficiently extract both local and global features.

- **Schrödinger Bridge Diffusion Module:** Learns an optimal transport plan via a forward diffusion process and a conditional reverse denoising UNet.
- **Multi-Scale Refinement Network (MRefNet):** Comprises a coarse-scale UNet branch and a fine-scale refinement stage guided by the coarse output.
- **Adversarial Training Module:** Uses a PatchGAN discriminator to ensure texture realism by classifying local image patches.

Implementation Details. The model is trained from scratch without any pre-trained weights, using the DIV2K dataset provided for the NTIRE 2025 Image Super-Resolution ($\times$4) Challenge. The team employs a composite loss function that combines diffusion loss for accurate noise prediction, optimal transport loss based on Sinkhorn divergence, multi-scale SSIM and L1 loss for perceptual quality enhancement, and adversarial loss to improve texture realism and natural appearance.

The model is trained for 300 epochs with a batch size of 8, totaling 35,500 iterations per epoch. All components are optimized jointly in an end-to-end manner using PyTorch. The final model contains approximately **41 million** parameters. Efficient inference is achieved through the use of windowed attention in the backbone. According to the team, this is the first application of Schrödinger Bridge theory combined with diffusion modeling for image restoration in a multi-scale adversarial training framework. The method has not been previously published.

A.18. IVPLAB-sbu

Description. Inspired by SwinFIR [62] and using the SwinIR [28] architecture, the model is slightly similar to the latter by incorporating the Spatial Transform Attention Block into it. As shown in Figure 19e, it consists of a conv3×3 layer to capture shallow features, as it showed its strength and stability to start with transformers, a Deep Feature Extraction Module, and a Reconstruction phase.

The deep feature extraction module, including several Residual Swin Transformer Blocks (RSTB), is structured as follows: Each RSTB consists of several Swin Transformer Layers (STL), and the Spatial Transform Attention Block (STAB) is the last layer instead of the convolution layer in SwinIR [28]. The Swin Transformer Layers (STL) use a Multi-head Self-Attention and a shifted window mechanism to calculate self-attention for windows separately. The PixelShuffle upsamples features in the Reconstruction part.

Spatial Transform Attention Block. Convolution layers bring the inductive bias to the transformer model, but the size of the convolution kernels is an issue. Small kernel sizes exert minimal influence on receptive fields, whereas excessively large kernels may lead to saturation, resulting in degraded performance.

Spatial Transform Attention Block (STAB), as shown in Figure 19e, consists of a Transform Attention Block (TAB) to capture the feature maps in the transform domain, in parallel with a Residual Block, in order to take feature maps in the spatial domain. Each part is explained in detail as follows.

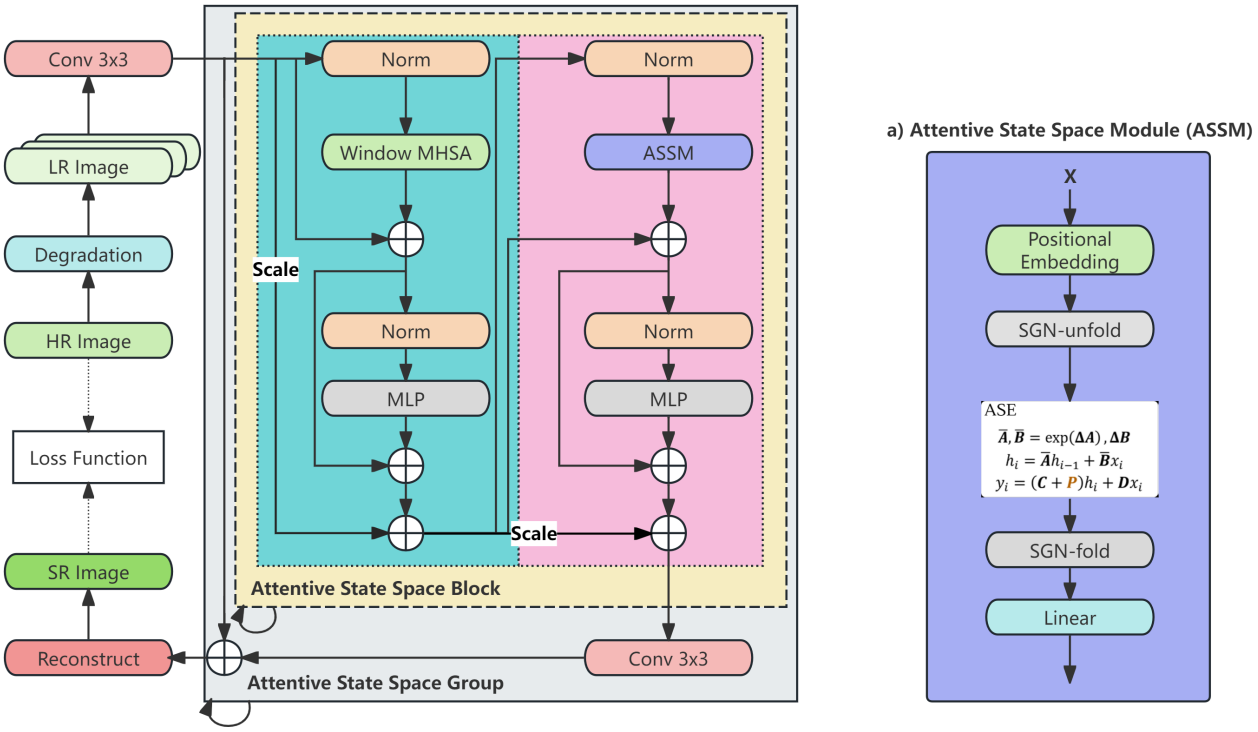


Figure 20. **Team MCMIR**

Transform Attention Block. The performance of the model increases by using a Transform Attention Block instead of a convolution layer, as convolution layers focus on low resolutions. Transform methods sharpen the edges of input features effectively and thus could be a good choice to focus more on edges.

The Transform Attention Block, depicted in Figure 19b, includes a Wavelet Transform Attention Block, a Fourier Transform Attention Block, and a Channel Attention Layer. A Wavelet Transform Attention Block, alongside the Fourier Transform Attention Block, is responsible for capturing the information of transform domains. Then, all information is concatenated, and a conv1×1 reduces the dimension of channels. Output of this convolution is then fed to a Channel Attention Layer to select the best channels.

Wavelet Transform Block. Wavelet Transform is a technique to capture feature maps and expand the receptive fields, as it extracts frequencies in three different directions: horizontally, vertically, and diagonally. It also maintains low-resolution spatial input decomposition. Therefore, extracting feature maps from it using a convolution layer is like extracting features from a low-resolution input.

The Wavelet Transform Block, depicted in Figure 19c, uses a Daubechies Wavelet Transform in four levels. Each decomposition of the input is fed into a conv1×1, followed by a LeakyReLU, before applying the inverse wavelet transform to extract more accurate features.

Fourier Transform Block. The Fourier Transform Block, depicted in Figure 19d is the same as the Wavelet Transform Block, but instead of the wavelet, the team members used a Fast Fourier Transform (FFT). As it extracts the global features of the input image in the frequency domain, it can leverage the impact of global information and expand the receptive fields.

Residual Block. The Residual Block consists of two conv3×3 layers, a LeakyReLU activation function in between, and a skip connection to extract more local features in the spatial domain, as shown in Figure 19a.

Implementation Details. The model was trained on the DIV2K dataset, and Set5, Set14, and DIV2K validation datasets were used for the validation phase. PSNR, SSIM, and LPIPS are metrics to measure model results. The batch size, window size, and patch size were set to 4, 12, and 60, respectively. The number of RSTB blocks, STL blocks, and heads in multi-head self-attention was all set to 6, while the channel dimensions were set to 180.

The model underwent training for 500k iterations using the Charbonier loss function and Adam optimizer (with β_1 = 0.9 and β_2 = 0.99), without weight decay. The initial learning rate was set to 2e-4 and halved at 250000, 400000, 450000, and 475000 iterations. Implementation was carried out using the Pytorch framework on a single RTX 3090 GPU. Additionally,

- Total method complexity (parameters: 15,479,531).
- The model is trained on the DIV2K train dataset.
- The team uses BasicSR API and Pytorch-wavelets. Additionally, the team members used the SwinFIR code to implement the model.

A.19. MCMIR

Description. In recent years, Transformer-based methods [49] have emerged as the dominant approach in the field of image restoration [1, 5, 16, 21, 28, 33], particularly excelling in image super-resolution tasks and surpassing the performance of CNN-based approaches [11, 31]. However, despite their impressive performance, the attention mechanism in Transformers is constrained by the quadratic computational cost of vanilla self-attention. While some works address this issue through model compression [36–39], such approaches often lead to performance degradation. To mitigate this, various linear attention mechanisms, such as RWKV [41] and Mamba [13], have been proposed, achieving notable success in large language model applications. Among these, Mamba has recently been widely adopted in computer vision tasks [13–15, 54, 67], demonstrating promising results.

Traditional Mamba-based methods [2, 12, 30, 43, 44, 55, 56, 64, 68] for image restoration sequentially unfold 2D images into 1D token sequences using specific scanning strategies, which limit each pixel's ability to access global context and require multi-directional scans to expand the receptive field, leading to increased computational complexity and redundancy. Due to their rigid sequential nature, these methods also fail to fully utilize semantic relationships between pixels. FreqMamba [67] perceive global degradation using the state space model in the Fourier domain. MambaLLIE [54] locality enhancement for low-light image enhancement tasks. MambaIRV2 [14] employs the routing matrix from Adaptive Semantic Encoding (ASE) to group semantically similar pixels together in the unfolded sequence, thus enhancing global context utilization while reducing computational complexity and redundancy.

The method introduces slight modifications to the MambaIRV2 framework and trains the model on a dataset processed through the custom-designed degradation pipeline. The detailed pipeline of the approach is illustrated in Figure 20. More specifically, the version of $network\_g$ simplifies the architecture by reducing the `depths` from 9 layers to 3 and aligning the num_heads accordingly, significantly lowering computational complexity while maintaining sufficient depth for feature extraction. Additionally, the team members increase the $window\_size$ from 16 to 64, enabling the model to capture broader spatial dependencies, which is crucial for high-resolution super-resolution tasks. Other key parameters remain unchanged, preserving the strong baseline architecture. These adjustments strike a balance between efficiency and performance while enhancing the model's spatial awareness and practicality for training and deployment.

Implementation Details. The team members pretrain the MambaIRv2 model using the Flickr2K [29] and LSDIR [27] datasets, incorporating a custom-designed degradation pipeline to generate diverse low-resolution (LR) variants from high-resolution (HR) images. This addresses the limited availability of degraded images by simulating realistic LR-HR pairs for robust training. Subsequently, the team members fine-tune the model on the DIV2K [48] training dataset, utilizing the provided HR-LR pairs to further enhance super-resolution performance. For validation, the team members evaluate the model on the DIV2K validation dataset, using PSNR and other metrics with a focus on the Y-channel for precise assessment.

The degradation pipeline assumes that a degraded image y is generated by the linear model:

$$y = kx + n$$

where x is the original (clean) image, k represents the degradation operator (e.g., a Gaussian blur kernel), and n is additive white Gaussian noise (AWGN). This model is commonly used in image processing to simulate how real-world distortions affect image quality. The convolution of the Gaussian kernel with the original image (kx) reduces the clarity of the image. The pipeline uses a 25×25 blur kernel with 9 modes. For blur_mode $= 0$, a Gaussian blur is applied using a 2D Gaussian kernel, with the blur strength controlled by σ_{blur}. For blur_mode $= 1$ to 8, precomputed 25×25 motion blur kernels from PnP_GS [19] are used to simulate various motion patterns and directions. These kernels are applied via convolution, with proper alignment ensured using a `torch.roll` operation. Additionally, Gaussian noise with a standard deviation σ_{noise} (scaled to the $[0, 1]$ range) is added to simulate sensor noise. By combining blurring and noise, the degradation pipeline effectively simulates realistic distortions, enriching the training data with diverse low-resolution images. This enables super-resolution models to generalize effectively, improving their performance in handling real-world degradations.

A.20. Aimanga

Description. This model is trained based on the RealESRGAN [52]. They have done in-depth thinking in cleaning data and image degradation models to achieve higher image reconstruction effects and better generation quality.

Implementation Details. As shown in Figure 21, the generator is based on an RRDBNet architecture, which primarily consists of multiple stacked Residual-in-Residual Dense Blocks (RRDB). Each RRDB integrates dense connections and residual skip connections, with residual scaling applied to stabilize training. The network removes batch normalization (BN) layers to avoid artifacts and incorporates pixel

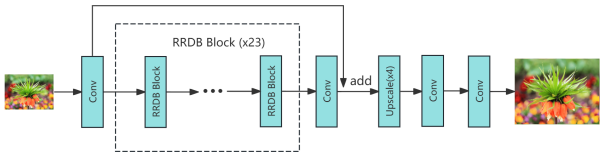


Figure 21. **Team Aimanga.** Architecture of Generator.

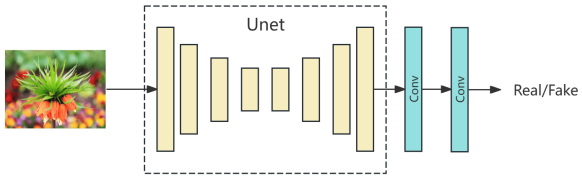


Figure 22. **Team Aimanga.** Architecture of Discriminator.

attention or channel attention mechanisms to enhance detail recovery. The input image first undergoes shallow feature extraction via convolutional layers, then passes through multiple RRDB blocks for deep feature refinement. Finally, PixelShuffle [46] upsampling reconstructs the high-resolution output. This design prioritizes preserving fine details and natural textures, making it effective for super-resolution tasks.

As shown in Figure 22, the discriminator adopts a U-Net-based [45] architecture to better capture both global and local details. It consists of an encoder with strided convolutions for downsampling and a decoder with transposed convolutions for upsampling, connected by skip connections to preserve spatial information. The network employs spectral normalization for stable training and uses PatchGAN-style [20] prediction to distinguish real vs. fake patches at multiple scales. Additionally, LeakyReLU activation and instance normalization help improve discrimination ability while maintaining gradient flow. This design enables effective adversarial training by providing detailed feedback to the generator on both high-level structures and fine textures.

They used nomos2k as the training set. To extract the detailed areas in the image, they designed an effective data screening process. Finally, they selected 1342 images from nomos2k dataset, which contains 2436 images in total. They carefully analyzed and tuned the image degradation model. Through continuous experiments, they found very effective image degradation models and parameters. The training process takes over 100k iterations with a batch size of 16 and a fixed learning rate of 1e-4.

To evaluate the performance of our trained model, they created a synthetic evaluation set of 125 images. The evaluation set contains 10 categories, such as human, animals, animation, and AI. This can more accurately evaluate the model's capabilities for various scenes. They further conducted experiments on the real-world dataset RealSR [3], which contains 100 images captured by Canon 5D3 and Nikon D810 cameras.

They compared our model with RealESRGAN and the latest work InvSR [59] in CVPR2025. They used niqe, brisque, nrqm, pi, clipiqa and musiq as evaluation indicators. The evaluation results are shown in the following table, from which they can see that our model is better than these two models in most indicators.

A.21. IPCV

Description. HMANet (Hybrid Multi-Axis Aggregation Network) [10] is a novel deep learning-based approach for Single Image Super-Resolution (SISR) that enhances both local and global feature learning through a hybrid aggregation strategy. The network consists of four key components: Shallow Feature Extraction Module (SFEM), Hybrid Multi-Axis Aggregation Module (HMAA), Residual Feature Aggregation Module (RFAM), and an Upsampling & Reconstruction module. HMAA integrates multi-axis attention mechanisms, capturing spatial and channel-wise dependencies by leveraging axial attention, dilated convolutions, and feature fusion from multiple orientations. This allows the model to efficiently extract long-range dependencies while preserving fine details.

To further refine the feature representation, RFAM employs hybrid convolutions (standard, dilated, and grouped) along with residual skip connections and squeeze-and-excitation attention, enabling hierarchical residual learning. The final upsampling is performed using a sub-pixel convolution layer (PixelShuffle) to reconstruct the high-resolution image. By integrating multi-axis feature learning and efficient upsampling, HMANet achieves superior super-resolution performance while maintaining computational efficiency.

Implementation Details. The model is trained using a combination of L1 loss, perceptual loss (VGG-19), and optionally adversarial loss, ensuring both pixel accuracy and perceptual quality.

B. Teams and Affiliations

NTIRE 2025 team

Title: NTIRE 2025 Image Super-Resolution (×4) Challenge

Members:

Zheng Chen[1] (zhengchen.cse@gmail.com),
Kai Liu[1] (normal.kliu@gmail.com),
Jue Gong[1] (g1017325431@gmail.com),
Jingkai Wang[1] (jingkaiwang100@gmail.com),
Lei Sun[2] (leosun0331@gmail.com),
Zongwei Wu[3] (zongwei.wu@uni-wuerzburg.de),
Radu Timofte[3] (radu.timofte@uni-wuerzburg.de),

Yulun Zhang[1*] (yulun100@gmail.com)
Affiliations:
[1] Shanghai Jiao Tong University, China
[2] INSAIT, Sofia University St. Kliment Ohridski, Bulgaria
[3] Computer Vision Lab, University of Würzburg, Germany

SamsungAICamera

Title: Detail Enhanced Image Super-Resolution Network with Hybrid Transformer-CNN Architecture
Members:
Xiangyu Kong[1](xiangyu.kong@samsung.com), Xiaoxuan Yu[1], Hyunhee Park[2], Suejin Han[2], Hakjae Jeon[2], Dafeng Zhang[1], Hyung-Ju Chun[2]
Affiliations:
[1]Samsung Research China, Beijing
[2]Department of Camera Innovation Group, Samsung Electronics

SNUCV

Title: Learning to Upsample for One-Step Diffusion Model
Members:
Donghun Ryou[1](dhryou@snu.ac.kr), Inju Ha[1], Bohyung Han[1]
Affiliations:
[1]Seoul National University, Republic of Korea

BBox

Title: SMT: Scale-Aware Mamba-Transformer for Image Super-Resolution
Members:
Lu Zhao[1] (zlcossiel@gmail.com), Yuyi Zhang[1,2], Pengyu Yan[1,2], Jiawei Hu[1], Pengwei Liu[1], Fengjun Guo[1]
Affiliations:
[1]Intsig Information Co., Ltd.
[2]South China University of Technology.

XiaomiMM

Title: Enhancing HAT with Multi-Branch Statistical Features and Mamba Integration
Members:
Hongyuan Yu[1](yuhyuan1995@gmail.com), Pufan Xu[2], Zhijuan Huang[1], Shuyuan Cui[3], Peng Guo[4], Jiahui Liu[1], Dongkai Zhang[1], Heng Zhang[1] , Huiyuan Fu[5] , Huadong Ma[5]
Affiliations:
[1]Multimedia Department, Xiaomi Inc.
[2]School of Integrated Circuits, Tsinghua University
[3]Huatai Insurance Group Co., Ltd.
[4]Hanhai Information Technology (Shanghai) Co., Ltd.
[5]Beijing University of Posts and Telecommunications

MicroSR

Title: Perceptual-aware MicroSR
Members:
Yanhui Guo[1](guoy143@mcmaster.ca), Sisi Tian[2]
Affiliations:
[1]McMaster University, Hamilton, Ontario, Canada
[2]Northeastern University, Seattle, Washington, United States

NJU_MCG

Title: Adaptive Token Mamba
Members:
Xin Liu[1](xinliu2023@smail.nju.edu.cn), Jinwen Liang[1], Jie Liu[1], Jie Tang[1], Gangshan Wu[1]
Affiliations:
[1]Nanjing University, China

X-L

Title: Methods based on Self Ensemble and Model Ensemble
Members:
Zeyu Xiao[1](zeyuxiao@mail.ustc.edu.cn), Zhuoyuan Li[2]
Affiliations:
[1]National University of Singapore
[2]University of Science and Technology of China

Endeavour

Title: Endeavour
Members:
Yinxiang Zhang[1](zhangyinxiang@mail.nwpu.edu.cn), Wenxuan Cai[1]
Affiliations:
[1]Northwestern Polytechnical University, Xi'an, China

KLETech-CEVI

Title: WHAT: Wavelet Hybrid Attention Transformer for Image Super-Resolution $\times 4$
Members:
Vijayalaxmi Ashok Aralikatti[1,3](01fe21bcs181@kletech.ac.in), Nikhil Akalwadi [1,3], G Gyaneshwar Rao[2,3] , Chaitra Desai[1,3], Ramesh Ashok Tabib[2,3], Uma Mudenagudi[2,3]
Affiliations:
[1] School of Computer Science and Engineering, KLE Technological University
[2] School of Electronics and Communication Engineering, KLE Technological University
[3] Center of Excellence in Visual Intelligence (CEVI) KLE Technological University

CidautAi

Title: SuPEm Super Resolution Perceptual Emsambler
Members:
Marcos V. Conde[1](marcos.conde@uni-wuerzburg.de), Alejandro Merino[1], Bruno Longarela[1], Javier Abad[1]
Affiliations:
[1]CidautAI

JNU620

Title: PMELSR: Pre-trained Model Ensemble for Image Super-Resolution
Members:
Weijun Yuan[1](yweijun@stu2022.jnu.edu.cn), Zhan Li[1], Zhanglu Chen[1], Boyang Yao[1]
Affiliations:
[1]Jinan University, Guangzhou, China

CV_SVNIT

Title: Mix Attention based Transformer for Single Image Super-Resolution
Members:
Aagam Jain[1](aagamjainaj1805@gmail.com), Milan Kumar Singh[1], Ankit Kumar[1], Shubh Kawa[1], Divyavardhan Singh[1], Anjali Sarvaiya[1], Kishor Upla[1], Raghavendra Ramachandra[2]
Affiliations:
[1]Sardar Vallabhbhai National Institute of Technology, India
[2]Norwegian University of Science and Technology, Norway

ACVLAB

Title: Self-ensemble Fusion Solution for Image Super-resolution
Members:
Chia-Ming Lee[1](zuw408421476@gmail.com), Yu-Fan Lin[1], Chih-Chung Hsu[1,2]
Affiliations:
[1]Institute of Data Science, National Cheng Kung University
[2]Institute of Intelligent Systems, National Yang Ming Chiao Tung University

HyperPix

Title: ETSRN: Enhancing Super-Resolution with Residual Hybrid Attention and Overlapping Cross-Attention
Members:
Risheek V Hiremath[1](hiremathrisheek745@gmail.com), Yashaswini Palani[1]
Affiliations:
[1]KLE Technological University

BVIVSR

Title: Mamba-Driven Implicit Image Function for Image Super-Resolution
Members:
Yuxuan Jiang[1](dd22654@bristol.ac.uk), Qiang Zhu[2,1], Siyue Teng[1], Fan Zhang[1], Shuyuan Zhu[2], Bing Zeng[2], David Bull[1]
Affiliations:
[1]University of Bristol, United Kingdom
[2]University of Electronic Science and Technology of China, China

AdaDAT

Title: AdaDAT: Enhancing Dataset Adaptability in Pre-trained SR Models
Members:
Jingwei Liao[1](jliao2@gmu.edu), Yuqing Yang[1], Wenda Shao[1]
Affiliations:
[1]George Mason University

Junyi

Title:
Members:
Junyi Zhao[1](z15236936309@gmail.com), Qisheng Xu[1], Kele Xu[1]
Affiliations:
[1]Key Laboratory for Parallel and Distributed Processing, Changsha, China

ML_SVNIT

Title: Pyramid Attention based Transformer Single Image Super-Resolution
Members:
Ankit Kumar[1](ankitkumar735226@gmail.com), Milan Kumar Singh[1], Aagam Jain[1], Divyavardhan Singh[1], Shubh Kawa[1], Anjali Sarvaiya[1], Kishor Upla[1], Raghavendra Ramachandra[2]
Affiliations:
[1] Sardar Vallabhbhai National Institute of Technology, India
[2] Norwegian University of Science and Technology, Norway

SAK_DCU

Title: Gated Convolution and Swin Transformer-based Single Image Super Resolution
Members:
Sunder Ali Khowaja[1](sunderali.khowaja@dcu.ie), Ik Hyun Lee[2]
Affiliations:
[1]Dublin City University, Dublin, Ireland

[2]IKLab and Tu Korea, u Korea, Siheung-Si, Republic of Korea

VAI-GM

Title: MDRCT: Multi-step Dense Residual Connected Transformer
Members:
Snehal Singh Tomar[1](stomar@cs.stonybrook.edu), Rajarshi Ray[1], Klaus Mueller[1]
Affiliations:
[1]Stony Brook University, USA

Quantum Res

Title: Mamba-Based Image Super-Resolution via Knowledge Distillation
Members:
Sachin Chaudhary[1](sachin.chaudhary@ddn.upes.ac.in), Surya Vashisth[2], Akshay Dudhane[3], Praful Hambarde[4], Satya Naryan Tazi[5], Prashant Patil[6], Santosh Kumar Vipparthi[7], Subrahmanyam Murala[8]
Affiliations:
[1]UPES Dehradun, India
[2]Amity University, Punjab, India
[3]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi
[4]Indian Institute of Technology Mandi, India
[5]Government Engineering College Ajmer, India
[6]Indian Institute of Technology Guwahati, India
[7]Indian Institute of Technology Ropar, India
[8]Trinity College Dublin, Ireland

PSU

Title: OptimalDiff: High-Fidelity Image Enhancement Using Schrödinger Bridge Diffusion and Multi-Scale Adversarial Refinement
Members:
Bilel Benjdira[1](bbenjdira@psu.edu.sa), Anas M. Ali[1], Wadii Boulila[1]
Affiliations:
[1]Robotics and Internet-of-Things Laboratory, Prince Sultan University, Riyadh 12435, Saudi Arabia

IVPLAB-sbu

Title: Swin Spatial Transform Attention Super-Resolution
Members:
Zahra Moammeri[1](zahramoammeri1@gmail.com), Ahmad Mahmoudi-Aznaveh[1], Ali Karbasi[1], Hossein Motamednia[2]
Affiliations:
[1]Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran
[2]Institute for Research in Fundamental Sciences, Tehran, Iran

MCMIR

Title: MCMIR
Members:
Liangyan Li[1](lil61@mcmaster.ca), Guanhua Zhao[1], Kevin Le[1], Yimo Ning[1], Haoxuan Huang[1], Jun Chen[1]
Affiliations:
[1]McMaster University

Aimanga

Title: Aimanga
Members:
Zonghao Chen[1](chenzonghao@k-fashionshop.com), Yang Ji[1], Xi Wang[1]
Affiliations:
[1]KUNBYTE, Hangzhou, Zhejiang Province, China

IPCV

Title: HMAx4
Members:
Jameer Babu Pinjari[1](jameer.jb@gmail.com), Kuldeep Purohit[1]
Affiliations:
[1]Independent Researchers

References

[1] Anas M Ali, Bilel Benjdira, Anis Koubaa, Walid El-Shafai, Zahid Khan, and Wadii Boulila. Vision transformers in image restoration: A survey. *Sensors*, 23(5):2385, 2023. 16

[2] Jiesong Bai, Yuhao Yin, Qiyuan He, Yuanxian Li, and Xiaofeng Zhang. RetinexMamba: Retinex-based Mamba for low-light image enhancement. *arXiv preprint arXiv:2405.03349*, 2024. 16

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *CVPR*, 2019. 17

[4] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1375–1383, 2019. 10

[5] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1764–1774, 2023. 16

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023. 2, 4, 5, 6, 10, 12

[7] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023. 5, 8, 9, 10

[8] Zheng Chen, Zongwei WU, Eduard Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. NTIRE 2024 challenge on image super-resolution (×4): Methods and results. In *CVPRW*, 2024. 11

[9] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. In *ICLR*, 2024. 2

[10] Shu-Chuan Chu, Zhi-Chao Dou, Jeng-Shyang Pan, Shaowei Weng, and Junbao Li. Hmanet: Hybrid multi-axis aggregation network for image super-resolution, 2024. 17

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 16

[12] Hu Gao, Bowen Ma, Ying Zhang, Jingfan Yang, Jing Yang, and Depeng Dang. Learning enriched features via selective state spaces model for efficient image deblurring. In *ACM MM*, pages 710–718, 2024. 16

[13] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 16

[14] Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. Mambairv2: Attentive state space restoration. *arXiv preprint arXiv:2411.15269*, 2024. 5, 7, 12, 13, 16

[15] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *ECCV*, pages 222–241. Springer, 2025. 16

[16] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE TPAMI*, 2025. 16

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 12

[18] Chih-Chung Hsu, Chia-Ming Lee, and Yi-Shiuan Chou. Drct: Saving image super-resolution away from information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6133–6142, 2024. 3

[19] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022. 16

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 17

[21] Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *arXiv preprint arXiv:2410.15067*, 2024. 16

[22] Yuxuan Jiang, Ho Man Kwan, Tianhao Peng, Ge Gao, Fan Zhang, Xiaoqing Zhu, Joel Sole, and David Bull. HIIF: Hierarchical encoding based implicit image function for continuous super-resolution. *arXiv preprint arXiv:2412.03748*, 2024. 7, 8

[23] Yuxuan Jiang, Chengxi Zeng, Siyue Teng, Fan Zhang, Xiaoqing Zhu, Joel Sole, and David Bull. C2D-ISR: Optimizing attention-based image super-resolution from continuous to discrete scales. *arXiv preprint arXiv:2503.13740*, 2025. 8

[24] Amogh Joshi, Nikhil Akalwadi, Chinmayee Mandi, Chaitra Desai, Ramesh Ashok Tabib, Ujwala Patil, and Uma Mudenagudi. Hnn: Hierarchical noise-deinterlace net towards image denoising. In *CVPRW*, pages 3007–3016, 2024. 4

[25] Younggeun Kim and Donghee Son. Noise conditional flow model for learning the super-resolution. In *CVPRW*, 2021. 12

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2, 8, 11

[27] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *CVPRW*, 2023. 3, 8, 16

[28] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 9, 10, 14, 16

[29] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 3, 8, 12, 16

[30] Wei-Tung Lin, Yong-Xiang Lin, Jyun-Wei Chen, and Kai-Lung Hua. PixMamba: Leveraging state space models in a dual-level architecture for underwater image enhancement. *arXiv preprint arXiv:2406.08444*, 2024. 16

[31] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5831–5840, 2022. 16

[32] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 9

[33] Yangyi Liu, Huan Liu, Liangyan Li, Zijun Wu, and Jun Chen. A data-centric solution to nonhomogeneous dehazing via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1406–1415, 2023. 16

[34] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. 1

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[36] Yu Mao, Yufei Cui, Tei-Wei Kuo, and Chun Jason Xue. Accelerating general-purpose lossless compression via simple and scalable parameterization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3205–3213, 2022. 16

[37] Yu Mao, Yufei Cui, Tei-Wei Kuo, and Chun Jason Xue. Trace: A fast transformer-based general-purpose lossless compressor. In *Proceedings of the ACM Web Conference 2022*, pages 1829–1838, 2022.

[38] Yu Mao, Jingzong Li, Yufei Cui, and Jason Chun Xue. Faster and stronger lossless compression with optimized autoregressive framework. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2023.

[39] Yu Mao, Weilan Wang, Hongchao Du, Nan Guan, and Chun Jason Xue. On the compressibility of quantized large language models. *arXiv preprint arXiv:2403.01384*, 2024. 16

[40] Jakub Nawała, Yuxuan Jiang, Fan Zhang, Xiaoqing Zhu, Joel Sole, and David Bull. Bvi-aom: A new training dataset for deep video compression optimization. In *VCIP*, 2024. 8

[41] Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Teddy Ferdinan, Haowen Hou, Przemysław Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 3, 2024. 16

[42] Qi Qi, Kunqian Li, Haiyong Zheng, Xiang Gao, Guojia Hou, and Kun Sun. Sguie-net: Semantic attention guided underwater image enhancement with multi-scale perception. *IEEE Transactions on Image Processing*, 31:6816–6830, 2022. 10

[43] Junbo Qiao, Jincheng Liao, Wei Li, Yulun Zhang, Yong Guo, Yi Wen, Zhangxizi Qiu, Jiao Xie, Jie Hu, and Shaohui Lin. Hi-Mamba: Hierarchical Mamba for efficient image super-resolution. *arXiv preprint arXiv:2410.10140*, 2024. 16

[44] Shiyu Qin, Jinpeng Wang, Yimin Zhou, Bin Chen, Tianci Luo, Baoyi An, Tao Dai, Shutao Xia, and Yaowei Wang. MambaVC: Learned visual compression with selective state spaces. *arXiv preprint arXiv:2405.15413*, 2024. 16

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 17

[46] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 17

[47] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 2001. 3

[48] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 8, 12, 16

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 16

[50] Cheng Wan, Hongyuan Yu, Zhiqi Li, Yihang Chen, Yajun Zou, Yuqing Liu, Xuanwu Yin, and Kunlong Zuo. Swift parameter-free attention network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 12, 13

[51] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 5, 9

[52] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021. 9, 16

[53] Zhengxue Wang, Guangwei Gao, Juncheng Li, Yi Yu, and Huimin Lu. Lightweight image super-resolution with multi-scale feature interaction network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021. 10

[54] Jiangwei Weng, Zhiqiang Yan, Ying Tai, Jianjun Qian, Jian Yang, and Jun Li. MambaLLIE: Implicit retinex-aware low light enhancement with global-then-local state space. *arXiv preprint arXiv:2405.16105*, 2024. 16

[55] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. RainMamba: Enhanced locality learning with state space models for video deraining. In *ACM MM*, pages 7881–7890, 2024. 16

[56] Xinyu Xie, Yawen Cui, Chio-In Ieong, Tao Tan, Xiaozhi Zhang, Xubin Zheng, and Zitong Yu. FusionMamba: Dynamic feature enhancement for multimodal image fusion with Mamba. *arXiv preprint arXiv:2404.09498*, 2024. 16

[57] Zhiwen Yang, Jiayin Li, Hui Zhang, Dan Zhao, Bingzheng Wei, and Yan Xu. Restore-rwkv: Efficient and effective medical image restoration with rwkv. *arXiv preprint arXiv:2407.11087*, 2024. 1

[58] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 3

[59] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. *arXiv preprint arXiv:2412.09013*, 2024. 17

[60] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 1

[61] Dongyang Zhang, Changyu Li, Ning Xie, Guoqing Wang, and Jie Shao. Pffn: Progressive feature fusion network for lightweight image super resolution. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3682–3690, 2021. 10

[62] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022. 14

[63] Leheng Zhang, Yawei Li, Xingyu Zhou, Xiaorui Zhao, and Shuhang Gu. Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In *CVPR*, 2024. 1

[64] Mingjin Zhang, Longyi Li, Wenxuan Shi, Jie Guo, Yunsong Li, and Xinbo Gao. VmambaSCI: Dynamic deep unfolding

network with mamba for compressive spectral imaging. In *ACM MM*, pages 6549–6558, 2024. 16

[65] Xiang Zhang, Yulun Zhang, and Fisher Yu. Hit-sr: Hierarchical transformer for efficient image super-resolution. In *European Conference on Computer Vision*, pages 483–500. Springer, 2024. 3

[66] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, Junpei Zhang, Kexin Zhang, Rui Peng, Yanbiao Ma, Licheng Jia, et al. Ntire 2023 challenge on image super-resolution (x4): Methods and results. In *CVPRW*, 2023. 11

[67] Zou Zhen, Yu Hu, and Zhao Feng. FreqMamba: Viewing Mamba from a frequency perspective for image deraining. *arXiv preprint arXiv:2404.09476*, 2024. 16

[68] Zhuoran Zheng and Chen Wu. U-shaped vision Mamba for single image dehazing. *arXiv preprint arXiv:2402.04139*, 2024. 16