NTIRE 2023 Challenge on Image Super-Resolution ($\times$4): Methods and Results

Yulun Zhang[†] Kai Zhang[†,*] Zheng Chen[†] Yawei Li[†] Radu Timofte[†]
Junpei Zhang Kexin Zhang Rui Peng Yanbiao Ma Licheng Jia Huaibo Huang
Xiaoqiang Zhou Yuang Ai Ran He Yajun Qiu Qiang Zhu Pengfei Li
Qianhui Li Shuyuan Zhu Dafeng Zhang Jia Li Fan Wang Chunmiao Li
TaeHyung Kim Jungkeong Kil Eon Kim Yeonseung Yu Beomyeol Lee
Subin Lee Seokjae Lim Somi Chae Heungjun Choi ZhiKai Huang
YiChung Chen YuanChun Chiang HaoHsiang Yang WeiTing Chen HuaEn Chang
I-Hsiang Chen ChiaHsuan Hsieh SyYen Kuo Ui-Jin Choi Marcos V. Conde
Sunder Ali Khowaja Jiseok Yoon Ik Hyun Lee Garas Gendy Nabil Sabor
Jingchao Hou Guanghui He Zhao Zhang Baiang Li Huan Zheng Suiyi Zhao
Yangcheng Gao Yanyan Wei Jiahuan Ren Jiayu Wei Yanfeng Li Jia Sun
Zhanyi Cheng Zhiyuan Li Xu Yao Xinyi Wang Danxu Li Xuan Cui
Jun Cao Cheng Li Jianbin Zheng Anjali Sarvaiya Kalpesh Prajapati
Ratnadeep Patra Pragnesh Barik Chaitanya Rathod Kishor Upla Kiran Raja
Raghavendra Ramachandra Christoph Busch

Abstract

This paper reviews the NTIRE 2023 challenge on image super-resolution ($\times$4), focusing on the proposed solutions and results. The task of image super-resolution (SR) is to generate a high-resolution (HR) output from a corresponding low-resolution (LR) input by leveraging prior information from paired LR-HR images. The aim of the challenge is to obtain a network design/solution capable to produce high-quality results with the best performance (e.g., PSNR). We want to explore how high performance we can achieve regardless of computational cost (e.g., model size and FLOPs) and data. The track of the challenge was to measure the restored HR images with the ground truth HR images on DIV2K testing dataset. The ranking of the teams is determined directly by the PSNR value. The challenge has attracted 192 registered participants, where 15 teams made valid submissions. They achieve state-of-the-art performance in single image super-resolution.

1. Introduction

Single image super-resolution (SR) is a field of research that focuses on the recovery of high-resolution (HR) images from their low-resolution (LR) counterparts that have undergone a certain degradation process. The topic has recently garnered significant attention from both the vision and graphics communities, with a surge of interest in the field. There has been a constant growth of related papers, and substantial progress has been made with employing CNN and Transformer [5, 7, 13, 23, 30, 33, 37, 41, 59, 63, 65].

Advancements in image SR techniques have facilitated the utilization of images for various tasks. Consequently, the range of applications for these techniques has constantly expanded to various fields, including surveillance, remote sensing, automotive industry, medical image analysis, and electronics. Furthermore, the widespread adoption and usage of mobile and wearable devices provide an excellent opportunity for developing new applications and faster methods. In this challenge, we mainly focus on higher performance with larger models usually.

For image SR, the LR image is obtained by applying a specific degradation process to its HR counterpart. Variations in this process can introduce different types of noise, blurring, or other artifacts, ultimately resulting in the loss of high-frequency information. The primary objective of image SR methods is to recover as much high-frequency information as possible. There are various standard prob-

[†] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, and Radu Timofte are the challenge organizers, while the other authors participated in the challenge. * Corresponding author: Kai Zhang. Appendix A contains the authors' teams and affiliations. NTIRE 2023 webpage: https://cvlai.net/ntire/2023/. Code: https://github.com/zhengchen1999/NTIRE2023_ImageSR_x4.

lems for image SR that depend on the degradation process. The most commonly used degradation model is bicubic down-sampling, which involves different downscaling factors. This classical degradation model facilitates the comparison of different image SR methods directly. In practice, a model performing pretty well under Bicubic degradation can also obtain consistent performance in other degradations or even other related applications. So, Bicubic degradation can serve as a testing ground for validating the superiority of newly proposed image SR methods.

In recent times, methods based on neural networks (*e.g.*, CNN and Transformer) have exhibited impressive performance for image restoration, especially image super-resolution (SR). SRCNN [12] first utilizes a three-layer convolutional neural network (CNN), achieving noteworthy enhancements over conventional SR methods (e.g., sparse coding based methods). Due to the fast-paced evolution of hardware technologies, numerous larger and deeper neural networks are being trained for image super-resolution. VDSR [20] builds a 20-layer network based on residual learning. RCAN [62] proposes a residual-in-residual structure to train a model over 400 layers. Moreover, numerous spatial and channel attention mechanisms [33,37,63,65] are proposed to improve the reconstruction quality. Apart from the development of large CNN models with high performance, Transformer proposed in the natural language processing (NLP) field is introduced to alternate CNN. The core component of the Transformer is the self-attention (SA) mechanism, which can directly model long-range dependencies for an accurate restoration. Several methods have successfully applied Transformer to image SR [5,7,30, 59, 61]. Those methods further achieve performance gains and show promising potential for future research.

Collaborating with the NTIRE workshop, we organized a challenge specifically focused on example-based single-image super-resolution ($\times$4). This task requires restoring high-frequency information in a high-resolution image using a single low-resolution input image and a set of prior examples that include low and corresponding high-resolution images. The aim of the challenge is to obtain a network design/solution capable to produce high-quality results with the best performance (*e.g.*, PSNR).

This challenge is one of the NTIRE 2023 Workshop series of challenges on: night photography rendering [43], HR depth from images of specular and transparent surfaces [57], image denoising [29], video colorization [19], shadow removal [47], quality assessment of video enhancement [35], stereo super-resolution [48], light field image super-resolution [50], image super-resolution ($\times$4) [64], 360° omnidirectional image and video super-resolution [4], lens-to-lens bokeh effect transformation [10], real-time 4K super-resolution [11], HR nonhomogenous dehazing [3], efficient super-resolution [28].

https://cvlai.net/ntire/2023/

2. NTIRE 2023 Image Super-Resolution ($\times$4) Challenge

This challenge is part of the NTIRE 2023 associated challenges, which aims are: (1) provide an overview of the latest trends and advances in image SR; (2) offer a platform for academic and industrial attendees to interact and explore collaborations. This section will elaborate on the specifics of the challenge.

2.1. DIV2K Dataset [2,45]

The DIV2K dataset comprises 1,000 RGB images with a 2K resolution, exhibiting diverse content. It is split into three sets: 800 images for training, 100 images for validation, and 100 images for testing. In this challenge, the low-resolution (LR) version of the DIV2K dataset is created by down-sampling the high-resolution images with a bicubic interpolation using a downscaling factor of 4. The validation set has already been provided to the participants, while the high-resolution images in the testing set are kept hidden throughout the challenge.

2.2. Flickr2K Dataset [31,45]

The Flickr2K dataset is a large-scale dataset of high-resolution images and is commonly used in image SR. The dataset contains 2,650 Flickr 2K images from the online photo-sharing platform Flickr. The images in the Flickr2K dataset are diverse in content, ranging from landscapes and nature scenes to portraits and still life photography. They are also diverse in quality, with some images being sharp and clear while others are blurry or contain noise.

2.3. LSDIR Dataset [27]

The LSDISR is a large-scale dataset where all high-resolution images are collected from Flickr. To ensure the pixel-level quality of the collected dataset, annotators were invited to manually inspect each of the collected images and remove the low-quality ones. The LSDIR contains 86,991 high-resolution images divided into a training set with 84,991 images, a validation set with 1,000 images, and a testing set with 1,000 images.

2.4. Track and Competition

The aim is to obtain a network design/solution capable to produce high-quality results with the best performance (*e.g.*, PSNR).

Track: Restoration Track. The ranking of participating teams was determined based on the PSNR value of their restored high-resolution images compared to the ground truth high-resolution images on the DIV2K testing dataset.

Challenge phases. *(1) Development and validation phase:* The participants were given access to the training and validation image pairs of the DIV2K dataset. Additionally, participants were allowed to utilize supplementary data, such as the Flickr2K dataset and LSDIR Dataset, during training. To obtain prompt feedback, participants could submit their restored high-resolution images to the evaluation server, which would calculate the PSNR of the super-resolved image generated by their models. *(2) Testing phase:* Participants were given access to 100 LR testing images during the final testing phase, while the corresponding high-resolution ground-truth images were kept hidden from them. Subsequently, participants submitted their restored high-resolution images to the Codalab server and provided their factsheet and code to the organizers via e-mail. The organizers then verified and executed the submitted code to obtain the final results, which were later communicated to the participants at the end of the challenge.

Evaluation protocol. The evaluation process involves comparing the super-resolved images with their corresponding ground truth high-resolution images using standard metrics such as Peak Signal-to-Nnoise Ratio (PSNR) and Structural Similarity (SSIM) [51] index, which are widely used in the literature. We calculate the PSNR and SSIM values on the Y channel of the YCbCr color space after discarding a 4-pixel boundary around each image. The average results over all the processed images of the DIV2K testing dataset are reported. We allow for a slight drop in accuracy to determine the final ranking. A code example for computing these metrics is provided at https://github.com/zhengchen1999/NTIRE2023_ImageSR_x4. The repository also includes the code and pre-trained weights of the submitted solutions.

3. Challenge Results

The final test results and rankings of the participating teams are presented in Tab. 1. The evaluated methods are briefly described in Sec.4, and the team members are listed in Appendix A. As can be observed from Tab. 1, the ZZPM team achieved the highest overall ranking in this Image Super-Resolution Challenge. Moreover, the average PSNR value of the top five teams is above 31 dB.

Team	Rank	PSNR (primary)	SSIM
ZZPM	1	31.23	0.8750
Graphene	2	31.21	0.8665
IPLAB	3	31.18	0.8660
SRC-B	4	31.16	0.8656
LDCC	5	31.16	0.8655
NTU607-SR	6	30.97	0.8617
Swin2SR	7	30.86	0.8603
TUK-IKLAB	8	30.80	0.8595
GarasSjtu	9	30.78	0.8582
LVGroup_HFUT	10	30.68	0.8563
AhRightRightRight	11	30.65	0.8555
helloooo	12	30.57	0.8551
chaobaer	13	30.23	0.8460
Alpha	14	30.02	0.8408
SVNIT_NTNU	15	29.92	0.8436

Table 1. Results of NTIRE 2023 Image Super-Resolution Challenge. PSNR/SSIM results are measured on the DIV2K testing dataset. The ranking of the teams is determined directly by the PSNR (primary) and SSIM (secondary).

3.1. Architectures and main ideas

Several techniques were proposed to improve the performance of image SR models during this challenge. Below, we list some of the typical ideas proposed by the participants.

1. **Adopting and modifying the Transformer architecture is the mainstream technology.** Recently, some methods that adopt Transformer architecture, such as SwinIR [30], ART [59], and HAT [5], have achieved better performance than CNNs in image SR. Therefore, several teams participating in this challenge have investigated modifying existing architectures to enhance the model performance. For example, the Graphene team proposed a cross-scale attention and wavelet hallucination on the architecture of SwinIR. The IPLAB team added a convolution module FEB to the ART model.

2. **Enhance model performance with global information.** Global information plays a crucial role in image SR as it can activate more pixels and significantly enhance the performance of the reconstruction process. Some teams try to utilize global information. For example, the SRC-B team proposed a spatial frequency block to extract the global information based on fast fourier convolution. The LVGroup_HFUT team introduced an additional global feature branch based on SwinIR.

3. **Image augmentation methods could effectively improve performance.** The self-ensemble strategy [46] has been shown to be effective in improving the performance and is widely used. Some teams tried more image augment methods. The ZZPM team adopted six data augmentation methods to train five models, and proposed a new fusion method to generate the results.

4. **More training data is an important factor.** In addition to DIV2K, the Flickr2K and LSDIR datasets are

allowed as training data. Models trained on large-scale datasets have better performance. Therefore, some teams try to use more extra data, *e.g.*, OST and FFHQ. The LDCC team also selected the training data with the proposed latent discriminative cosine criterion.

5. **Advanced training strategy also boosts performance.** It is observed that some teams employed well-designed training strategies in their methods. For instance, the progressive training strategy is used, gradually increasing the patch size during the training phase.

6. **Various loss functions are considered.** The loss function is also a crucial element in neural networks. Apart from the commonly used L_1 and L_2 loss, several teams used more loss functions, like the fast fourier transform (FFT) loss and MSE loss, to achieve PSNR gain. For instance, the chaobaer team uses different losses at different stages of training.

7. **Some other techniques are also attempted.** Some teams also proposed solutions based on the U-Net architecture and Hadamard product.

3.2. Participants

The challenge attracted 33 registered participants, out of which 15 teams submitted valid entries. These entries set the state-of-the-art in image SR ($\times 4$).

3.3. Fairness

To ensure the fairness of the challenge, several rules were established regarding fair and unfair practices. Firstly, training with the DIV2K test LR images is prohibited. Secondly, training with extra datasets, such as the Flickr2K and LSDIR datasets, is allowed. Thirdly, using advanced data augmentation strategies during training is considered a fair practice.

3.4. Conclusions

The following conclusions can be drawn based on the above analysis of the image SR challenge results.

1. The proposed methods submitted during this challenge have contributed to the advancement of research and implementation in the field of image SR.

2. The methods based on the Transformer architecture demonstrate impressive performance and set new records regarding restoration quality.

3. High-quality, large-scale datasets play a critical role in image SR, especially for large neural networks.

4. Challenge Methods and Teams

4.1. ZZPM

Network Architecture. The overall workflow of the solution proposed by the ZZPM team is illustrated in Fig. 1. The method first performs data augmentation locally on the training set on the DIV2K and LSDIR datasets. Multiple models are then used to train on the augmented data. The final result consists of the fusion of multiple models. A total of six data augmentation methods are used, including CutBlur, Blend, RGB permute, Mixup, CutMix, and CutMixup. These methods greatly increase the diversity of the data and improve the robustness of the model.

In the model selection phase, the team selects a total of five models, including Liif-EDSR, Liif-RDN, RDN-LTE, SwinIR-LTE, and SwinIR, which are widely used and have shown excellent performance in super-resolution tasks.

I. The Liif-EDSR is a deep learning model used for image super-resolution. It combines the Liif (Learned Intrinsic Image Filtering) model with the EDSR (Enhanced Deep Residual Networks) architecture to produce high-quality, high-resolution images. The Liif component is responsible for image filtering and refinement, and the EDER part is for image super-resolution. EDSR consists of multiple residual blocks, each containing convolutional layers and skip connections.

II. The Liif-RDN is similar to Liif-EDSR. The RDN component is used for super-resolution, and it consists of multiple dense blocks, each containing convolutional layers with densely connected feature maps.

III. The RDN-LTE is a variant of the Residual Dense Networks (RDN) architecture that has been optimized for low-latency applications. It uses densely connected convolutional layers to enable the model to learn and refine image features at multiple scales.

IV. The SwinIR architecture consists of three main components: the Swin Transformer encoder, the Swin Transformer decoder, and the image restoration module. It combines the Swin Transformer architecture with an image restoration module for super-resolution.

V. The SwinIR-LTE is a variant of the SwinIR model optimized for low-latency applications. It uses a smaller model architecture and a more lightweight image restoration module to reduce the computational cost and latency of the model while maintaining high-quality super-resolution results.

The main purpose of the post-processing part is to fuse the outputs of the models. Here a new fusion method is proposed. In order to fuse the outputs of the five models, first calculate the mean of the five images to be fused. And then, the MSE is calculated between each image and the mean. The smaller the MSE value, the greater the weight

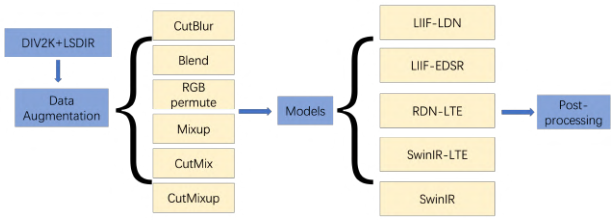


Figure 1. *ZZPM Team:* Overall architecture of the method.

assigned to the model during fusion. Finally, the final output result is obtained by performing a weighted fusion of the outputs of the five models using the calculated weights. The significance of this method is that it can avoid the final result being skewed by special bad points.

Training strategy. During the training phase, an L_1 loss is used for SR, and an MSEloss is used for enhancement. The training of the aforementioned models is conducted on 8 Nvidia V100 GPUs, using the Adam optimizer and multi-step learning rate decay method. The initial learning rate is set to 1×10^4. In the testing phase, five models are used for the ensemble, and a series of test-time augmentation approaches, including RGB channel shuffling and horizontal and vertical flipping.

4.2. Graphene

Network Architecture. The Graphene team designed an Image Super-Resolution Transformer with Cross-Scale Attention (CSA) and Wavelet Hallucination (WH) to address the image super-resolution task. The framework is shown in Fig. 2.

In cross-scale attention, the method adds a 3×3 depth-wise convolution in the query transformation and adds multi-scale depth-wise convolutions in the key and value transformation. Such a depth-wise convolution can extract the local features and model the relationship between features across multiple scales. The cross-scale attention is used to improve the vanilla and shifted window attention in SwinIR [30].

For the wavelet hallucination, the method first hallucinates four frequency sub-bands using a linear layer, then perform depth-wise convolution followed by a linear projection separately for each sub-band. A Haar wavelet reconstruction function reconstructs the higher-resolution feature maps. Finally, a 3×3 convolution with stride 2 is applied to reduce the size of feature maps into the original one. A higher-resolution hidden feature is predicted by hallucinating features in different frequency sub-bands. Such hidden high-resolution features help to extract finer details. The wavelet hallucination is used to replace the 3×3 convolution at the end of each Transformer layer in SwinIR.

Training strategy. The model is trained on the combination of the trainsets of DIV2K, Flickr2K, and LSDIR. The training process is divided into two stages. In the first stage, the model is trained on 64×64 randomly cropped images with The Adam optimizer. The batch size, initial learning rate, and total iterations are 32, 2×10^4 and 800K, respectively. The learning rate is halved at 300K, 500K, 650K, 700K, and 750K iterations. In the second stage, the model is fine-tuned on 128×128 images with 200K iterations.

4.3. IPLAB

Network Architecture. The NEESR team proposed Attention Retractable Frequency Transformer (ARFT) for image super-resolution [66]. The overall architecture of the ARFT is shown in Fig. 3. Following ART [59], ARFT employs residual in residual structure to construct a deep feature extraction module. Given a low-resolution image $I_{LR} \in \mathbb{R}^{H\times D\times C_{in}}$ (H, D, and C_{in} are the height, width, and input channels of the input), ARFT firstly applies a 3×3 convolution layer to obtain shallow feature $F_0 \in \mathbb{R}^{H\times D\times C}$, where C is the dimension size of the new feature embedding. Next, the shallow feature is normalized and fed into the residual groups, which consist of core Transformer blocks. The deep feature is extracted and then passes through another 3×3 convolution layer to get further feature embeddings F_1. Then the element-wise sum is used to obtain the final feature map $F_R{=}F_0{+}F_1$. Finally, ARFT employs the pixel shuffle layer to generate the high-resolution image I_{SR} from the feature F_R.

As shown in Fig. 3(b), two attention strategies, *i.e.*, D-MSA and S-MSA, are applied to design two types of self-attention blocks named dense attention block (DAB) and sparse attention block (SAB). In the dense attention block (DAB), the dense multi-head self-attention module (D-MSA) allows each token to interact with a smaller number of tokens from the neighborhood position of a non-overlapping $W\times W$ window. We apply these groups to compute self-attention for W times. Meanwhile, in sparse attention block (SAB), the dense multi-head self-attention module (S-MSA) is proposed to allow each token to interact with a smaller number of tokens from sparse positions with interval size I. After that, the updates of all tokens are split into several groups, each with tokens.

As shown in Fig. 3(c), the FEB network architecture is composed of two primary components: a frequency branch on the up and a spatial branch on the down. Two distinct domains process the input feature to generate frequency feature $X_{\text{frequency}}$and spatial feature X_{spatial}. The outputs of two branches are concatenated and operated by a convolution layer to obtain the final result. Specifically, $X_{\text{frequency}}$ is intended to capture the long-range context in the frequency domain, and X_{spatial} is utilized in the spatial domain.

Progressive Model Training Strategy The team proposed a novel progressive model training strategy to improve SR performance. Specifically, the progressive model

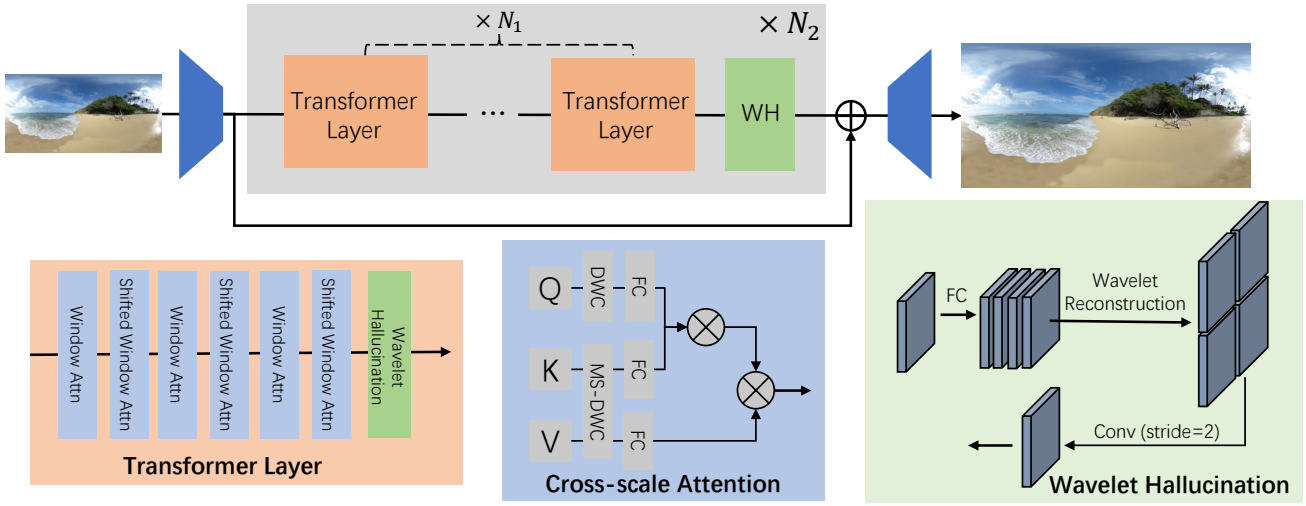


Figure 2. *Graphene Team:* The pipeline of the proposed method. A Transformer-based architecture is adopted. The cross-scale attention and wavelet hallucination are proposed to enhance the feature extraction.

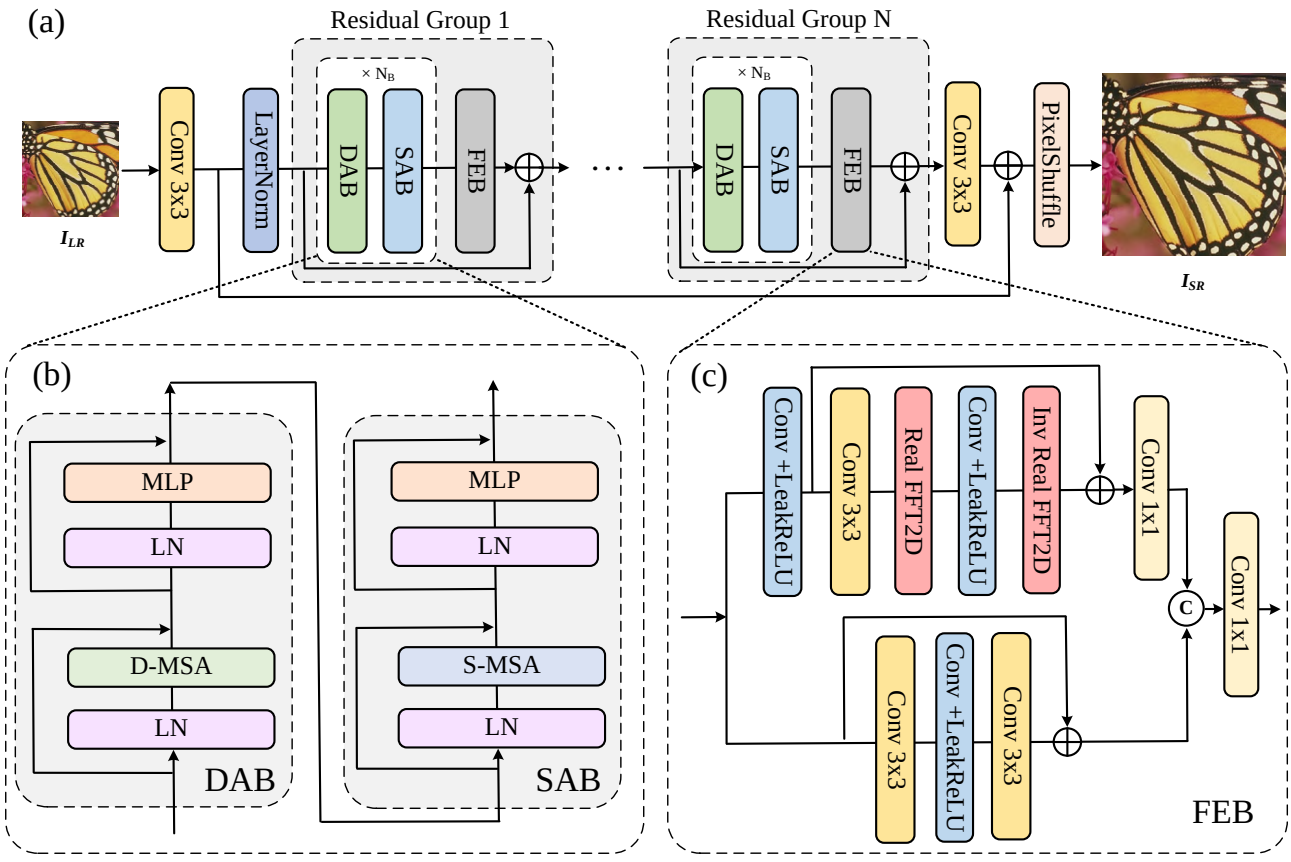


Figure 3. *IPLAB Team:* (a) The architecture of the proposed ARFT for image SR. (b) The structure of two successive attention blocks, DAB and SAB, with two attention modules, D-MSA and S-MSA. (c) The structure of frequency enhancement block.

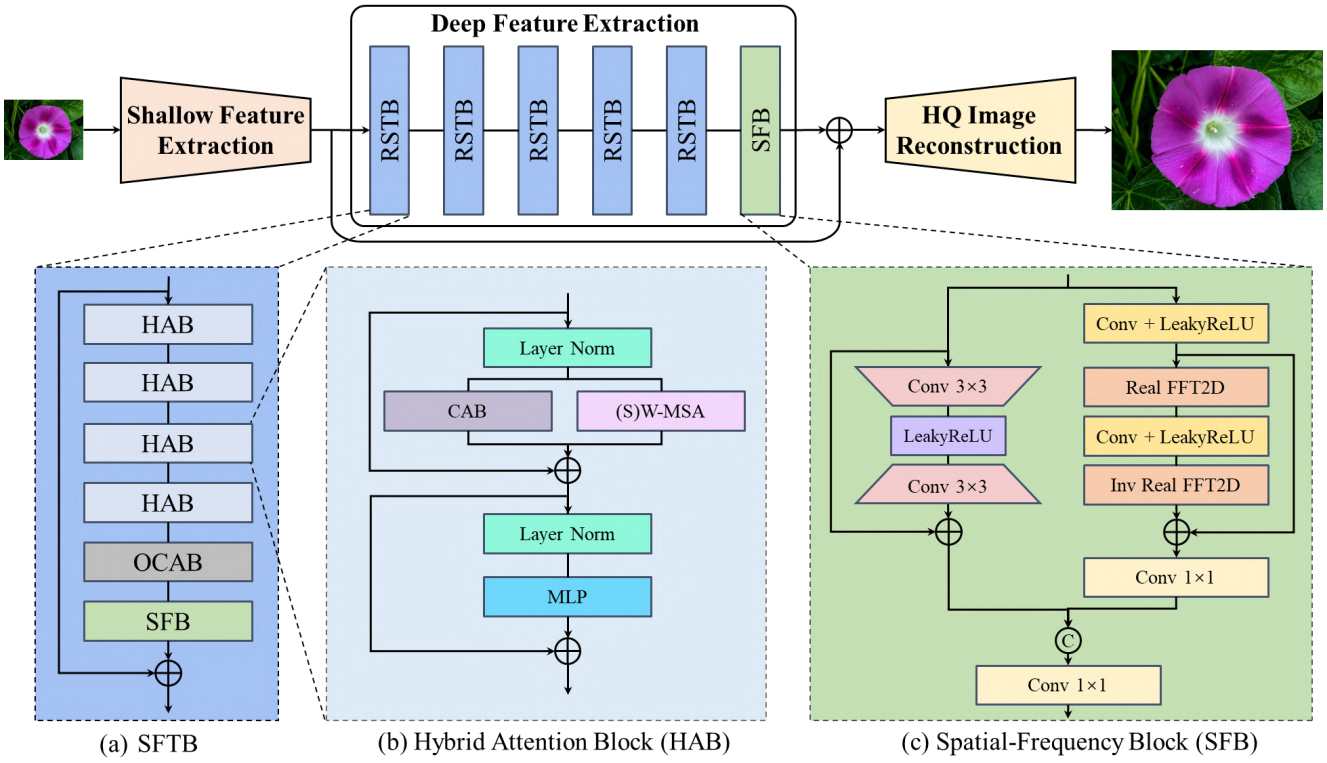


Figure 4. *SRC-B Team:* The network architecture of the proposed SwinFIR for image super-resolution.

training strategy combines the inference results from various models. The model is trained with different patch sizes of training datasets in multi-progressive stages. Specifically, three training stages are employed, and the patch size of each stage is set to 48, 64, and 84, respectively.

Loss Function. Recently, the Fast Fourier Transform loss (FFTLoss) [16] is proposed to constrain the frequency information to get better performance in SR tasks. The team uses L_1 loss, L_2 loss, and the FFTLoss [16] to optimize their network for generating promising SR results.

In each training stage, the basic loss function composed of L_1 loss and the FFTLoss are used to obtain the basic SR performance,

$$Loss_1 = \|I_{\mathrm{HR}} - I_{\mathrm{SR}}\|_1 + \alpha FFTLoss(I_{\mathrm{HR}}, I_{\mathrm{SR}}), \quad (1)$$

where I^{HR} is the corresponding HR image and α is the penalty factor with a value of 0.1. Then, the L_2 loss is applied to continuously train the model to improve the SR performance,

$$Loss_2 = \|I_{\mathrm{HR}} - I_{\mathrm{SR}}\|_2. \quad (2)$$

Implementation details. The ARFT is trained on a large combination training dataset, which is composed of DIV2K, Flicker2K, and LSDIR. Data augmentation is performed on the training data through the horizontal flip and random rotation of 90°, 180°, and 270°. Besides, the original images are cropped into 64×64 patches as the basic training inputs for image SR. Due to using the progressive model fusion strategy, different batch sizes and patch sizes are used in each stage. Specifically, in three stages, the training batch and patch size are (32, 48), (16, 64), and (8, 84), respectively. The ARFT is optimized by the Adam optimizer with β_1=0.9, β_2=0.999, and zero weight decay. The initial learning rate is set as 2×10^{-4} and is reduced by half as the training iteration reaches a certain number. Taking image SR as an example, the total iterations are 500K, and the learning rate is halved when training iterations reach 250K, 400K, 450K, and 475K. The ARFT is implemented on PyTorch with 4 NVIDIA RTX 3090 GPUs.

4.4. SRC-B

Network Architecture. Inspired by SwinIR [30] and HAT [5], the SRC-B team proposed SwinFIR [58] using Swin Transformer [36] and fast fourier convolution [8], as shown in Fig. 4. HAT proposed the Residual Hybrid Attention Group (RHAG) to activate more pixels in the image super-resolution transformer to improve performance. RHAG contains N hybrid attention blocks (HAB), an overlapping cross-attention block (OCAB), and a 3×3 convolutional layer. SwinFIR replaces the convolution (3×3) with Fast Fourier Convolution and a residual module to fuse global and local features, Spatial-frequency Block (SFB), to

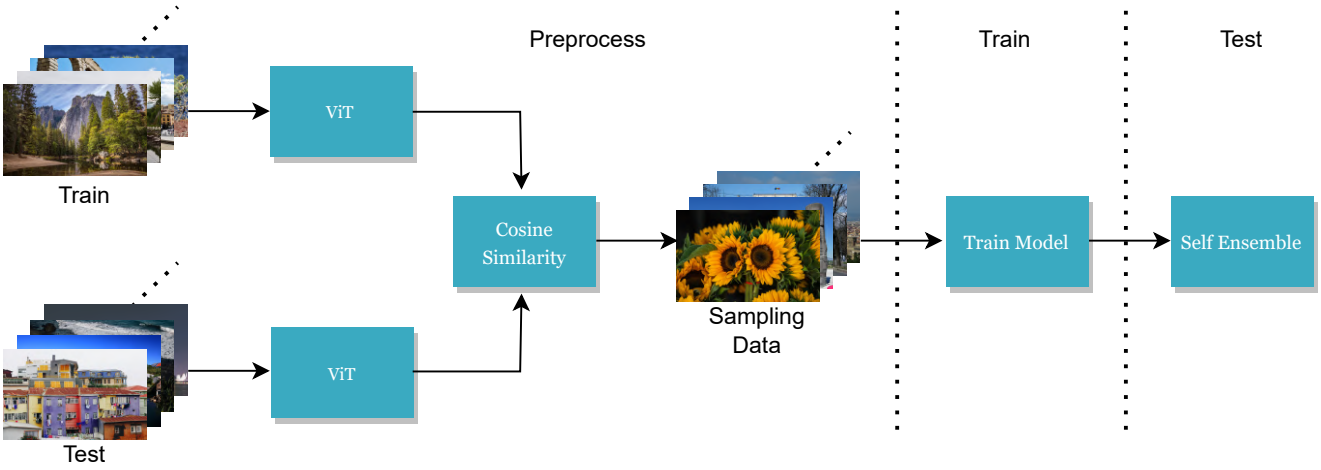


Figure 5. *LDCC Team:* The flow chart of the proposed method

improve the representation ability of the model.

Training strategy. The high-resolution images are cropped to sub-images (384×384). The Adam [21] optimizer with default parameters and the Charbonnier L_1 loss [24] are used to train the model. The initial learning rate is 2×10^{-4}, and the MultiStepLR learning rate scheduler with about 800K iterations and milestones is [600K, 650K, 700K, 750K]. The batch size is 4, and the patch size is 64. The model is implemented by PyTorch 1.8.1 and trained on NVIDIA A6000 GPU with CUDA 11.1. The horizontal flip, vertical flip, rotation, RGB perm, and mixup [55] are used for data augmentation. Inspired by previous works [46, 52], the self-ensemble and multi-model ensemble is applied to improve the performance.

4.5. LDCC

Network Architecture. The LDCC team proposed the latent discriminative cosine criterion (LDCC) to select train data, which helps improve performance by computing the cosine similarity with test data. For this, the latent features are extracted by using the pre-trained VIT [14] model as an encoder and measure the cosine similarity [44] between the train image and the test image as follows:

$$cos(\Theta) = \frac{1}{H \times W} \sum_{i \in (H,W)} \frac{I_{train}(i) \cdot I_{test}(i)}{||I_{train}(i)||^2 ||I_{test}(i)||^2}, \quad (3)$$

where I_{train} and I_{test} denote low-resolution images in train and test datasets, respectively. H and W are the height and width of the low-resolution image. i means the pixel position of I_{train}. $(\cdot)$ indicates the inner product. After that, the top 200 samples are chosen as training data according to the similarity values for each test image. By doing so, about 18K image samples are obtained from DIV2K [45], Flickr2K [31], LSDIR [27], OST [49], and AI-HUB [1] datasets.

Training strategy. The training process contains two stages according to competition phases, *i.e.*, development and testing. At the development phase, the SR model is pre-trained on DIV2K, Flickr2K, and LSDIR datasets with initial model parameters, which are publicly available from HAT [5]. After that, the latent discriminative cosine criterion (LDCC) is applied to DIV2K, FLickr2K, LSDIR, OST, and the additional dataset obtained from AI-HUB in the testing phase. Finally, the SR model is fine-tuned on selected samples.

To train the SR model, the random rotation, horizontal flip, and random crop (i.e., 64×64 patches) are used as data augmentation and mean absolute error loss. The Adam [21] optimizer is used with a batch size of 4 per GPU, where the power and momentum are set to 0.9 and 0.99, respectively. The initial learning rate was initialized as 1×10^{-5} in the development phase and 6.25×10^{-6} in the testing phase and reduced by half at 100K, 200K, 225K, and 240K iterations. The model is implemented with 8 NVIDIA V100 GPUs.

4.6. NTU607-SR

Training strategy. The approach of the NTU607-SR is based on ART [59] and involves a 3-stage training process for fine-tuning the model.

The model is trained on DIV2K and LSDIR datasets. In the entire training process, the batch size is set to 8, and the Adam optimizer with β_1=0.9, β_2=0.99, weight decay=0 is used. The train iterations are 250K in each stage. All training images are cropped to a size of 256×256. Different loss functions are used for the three stages of the training, including L_1 loss, MSE loss, and PSNR loss, respectively. The PSNR loss is calculated only on the Y channel, and for the learning rate, a lower rate is used when optimizing the model later. Regarding data augmentation, more augmentation tricks are applied in the early stages of the training,

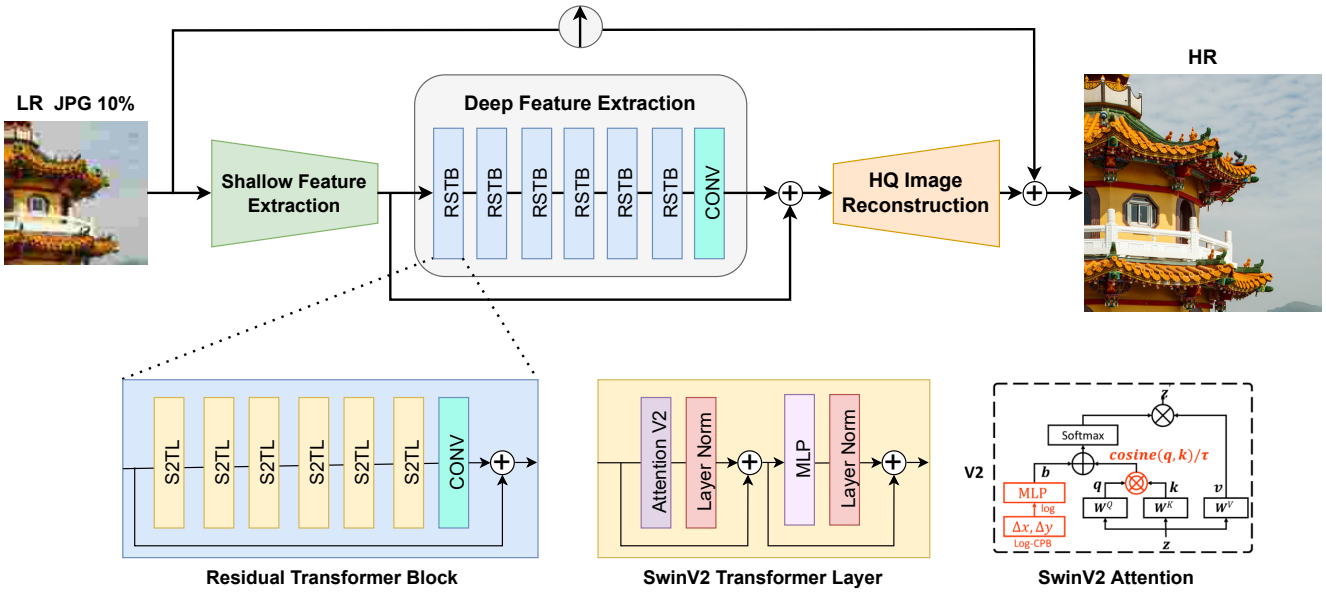


Figure 6. *Swin2SR Team:* The diagram of Swin2SR [9].

which allows the model to converge quickly and steadily. As the training progresses, the usage of augmentation is reduced.

Specifically, in the first stage, the L_1 loss is used with the learning rate of 1×10^{-4}. Data augmentation includes the horizontal flip, the vertical flip, and the rotation with 90 degrees. In the second stage, the MSE loss with a learning rate of 5×10^{-5} is applied. Finally, the PSNR loss on the Y channel is adopted, and the learning rate is set to 1×10^{-5}.

Testing description. The test time augmented images are used to compute the predictions and ensembled the results to obtain the final prediction. Specifically, the average of the predictions made on the original image and its three flips (horizontal, vertical, and both) are taken to generate the final prediction. This ensemble approach [46] helped reduce the impact of the noise in individual predictions and improved the overall accuracy of the model.

4.7. Swin2SR

Network Architecture. The solution of Swin2SR team studies Swin2SR [9] and HAT [5]. The Swin2SR is illustrated in Fig. 6. Note that Swin2SR is **a previously published work by the team**. They aim to test these methods on the novel LSDIR dataset.

In Swin2SR, the novel Swin Transformer V2 is explored as a possible update and improvement of SwinIR [30] for image super-resolution. Through this method, the major issues in training transformer vision models, such as training instability, resolution gaps between pre-training and fine-tuning, and hunger for data, can be tackled. Swin2SR is also tested on: JPEG compression artifacts removal, real-world image super-resolution, and compressed image super-resolution [9]. Experimental results also demonstrate that Swin2SR can improve the training convergence and performance of SwinIR [30].

Compare HAT [5] and Swin2SR [9]: **(1)** HAT has 40.84M parameters, 1998 GFLOPs, while Swin2SR has 12.23M parameters, 515 GFLOPs, and max memory allocation of 2677M. Swin2SR, therefore, is more efficient than HAT (3× fewer params and GFLOPs) and achieves similar results.

Training strategy. Training and testing description can be consulted in Swin2SR [9]. The model is trained with the Adam optimizer, L_1 loss, flips, and rotation augmentations. The datasets: DIV2K, Flickr2K, OST, WED, FFHQ, Manga109, and LSDIR are used. The training iterations are 500K, and the progressive lr decay is the same as SwinIR [30].

4.8. TUK-IKLAB

Network Architecture. The proposed method of team TUK-IKLAB is named Dense Residual Swin Transformers (DRSTNet) for image super-resolution. The proposed method comprises four modules, *i.e.*, Hierarchical Feature Extraction, Dense Residual Feature Enhancement, Fusion, and HR Reconstruction modules, as shown in Fig. 7.

Existing studies have revealed that using a hierarchical feature extraction module allows the network to extract meaningful representations from images at different scales in a divide-and-conquer manner [6, 39]. Furthermore, it helps the network deal with complex and severe degradation in an efficient manner. The term hierarchical is used for this

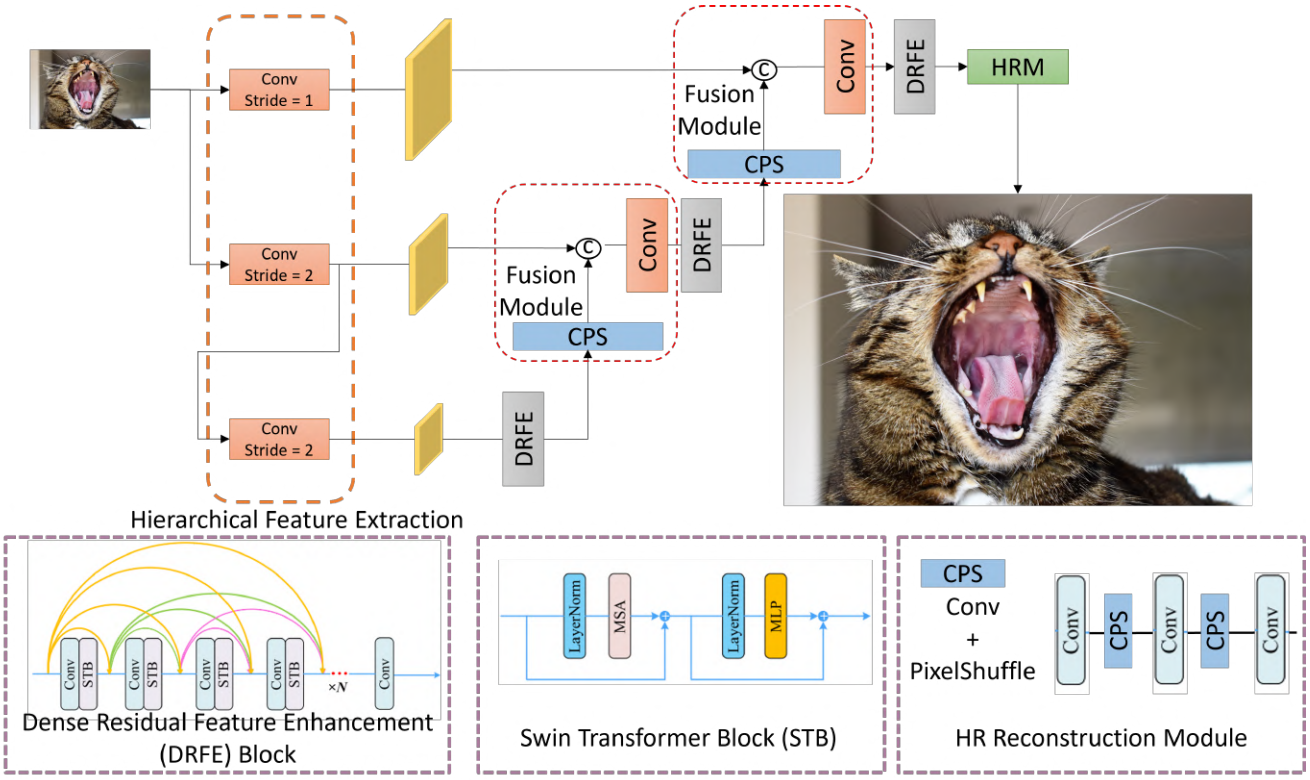


Figure 7. *TUK-IKLAB Team:* Architecture for the proposed dense residual Swin Transformer network (DRSTNet) for Image Super Resolution.

module as it extracts the representation from low-resolution (LR) images with three step architecture that applies convolution operation with varying strides and kernels using three different scales. For the implementation of hierarchical feature extraction modules, DRSTNet refers to the work [39]. The first step comprises padding, stride, number of channels, and kernel size, which are set to be 3, 1, 60, and 7, respectively. For the second step in this hierarchy, DRSTNet follows the same convention for feature extraction but with the values 2, 2, 60, and 5, followed by the third step that takes the values 1, 2, 60, and 3, respectively. Existing works for image restoration and super-resolution consider convolutional neural networks as feature enhancers [26, 39, 60].

Recently, some of the studies considered Swin Transformer block to enhance the features as well as model long-range dependencies [25, 30]. Shifted Windows (Swin) transformers have proven to be effective for such degradation tasks, including image restoration and super-resolution, while yielding less number of parameters. Different from the study [25, 30], the team proposes a dense residual feature enhancement (DRFE) block. As shown in Fig. 7, the DRFE block combines the Swin Transformer layers with dense residual convolutional blocks. Four layers are used for DRFE. The residual convolutional and Swin Transformer blocks are then connected in a dense scheme. This dense connection scheme was inspired by DenseNet, which helps to deal with such a complex degradation task of image super-resolution. The Swin Transformer block is further divided into Swin Transformer layers, multi-head self-attention, and layer normalization, accordingly.

The feature fusion module undertakes the enhanced features from DRFE and performs the feature level fusion with the ascending hierarchical step module. Such a fusion strategy leverages contextual information while fusing features extracted and enhanced from middle and lower branches. The features are upscaled and concatenated within the fusion module with the middle branch features. Similarly, the features from the middle branch are extracted and enhanced using DRFE, followed by the upscaling and concatenation with the feature maps from the first branch, respectively. The upsampling operation is performed using convolutional layers and pixel shuffle layer [42].

Finally, the last module is the HR reconstruction module that undertakes the enhanced features and outputs 4× higher resolution RGB image. As shown in Fig. 7, the module has a two-pixel shuffle, two convolutional layers, and two sub-pixel convolutional layers. Lastly, a convolutional layer complies with the existing works to generate the high-

resolution image [25, 30].

Training strategy. The model is trained on Flickr2K, and the training dataset available from the competition website. For the testing stage, the DIV2K validation images and testing images from the competition website are used. The number of channels, stride, padding, and kernel size for each hierarchical branch is mentioned in the previous section. For the DRFE module, the number of channels is set to 60 for each hierarchical branch, and the number of STL layers is 6. The window size is set to 8, and the number of STBs is 2, 6, and 8 for the third, second, and first steps of the hierarchy. The number of connections in DRFE is set to 6. The network is trained using 2 NVIDIA 3060Ti GPUs with a batch size of 8. The pre-trained network is used and fine-tuned with the learning rate of 5×10^{-5} and trained the network for 15K iterations.

4.9. GarasSjtu

Network Architecture. The GarasSjtu team designed a simple Transformer-style network for image super-resolution (STSN) [17], as shown in Fig. 8. The low-level vision model is developed based on Conv2former [18] that proposed a simple Transformer-style ConvNet model for high-level vision. In order to improve the model further, the local features can be extracted based on 3×3 convolutions in the main block of the model. In addition, the ESA [33] block is used to further improve the model.

The model is built based on three stages of shallow feature extraction, deep feature extraction, and high-resolution image reconstruction, as shown in Fig. 8. Shallow feature extraction is performed using 3×3 convolutions, which change the domain from image to feature. After that, there are five blocks of Conv2FormerGroup included in the deep feature extraction modules. In each Conv2FormerGroup contains 4 Conv2FormerB with 3×3 convolution and ESA [33]. The Conv2FormerB block is designed based on the Conv2Former block and multi-layer perception (MLP) using Layernorm before each one and using residual learning, as shown in Fig. 8. The MLP is composed of two linear layer intermediate with one point-wise convolution. The Conv2Former block is further improved by changing 1×1 convolution to 3×3 convolution for extracting local features. The Conv2Former includes two branches, the first is only 1×1 convolution, and the second is 1×1 convolution followed by depth-wise convolution with a kernel size of $k\times k$.

The Hadamard product is used to multiply the outputs of the first and second branches. Then, this Conv2Former block ends by using 3×3 convolution instead 1×1 in the original block [18]. At the end of the deep feature extraction stage, 1×1 convolution and 3×3 convolution are used. Then, residual learning is used between the input and the output of this STSN network. The final stage is the image reconstruction made by utilizing one 3×3 convolution. Then, the pixel shuffle layer is utilized for mapping features to HR image space.

Training strategy. The STSN model contains five Conv2FormerGroup blocks containing 4 Conv2FormerB, in which the number of feature maps is set to 150. Also, the channel number of the ESA is set to 32, similar to previous work [18]. DIV2K and LSDIR are used to train the model. At the starting stage, the model is trained from scratch using the DIV2K and LSDIR datasets, with a patch size of 192×192. The bach siz is 16 for 70 epochs. Then, the pre-trained weights are used to train it again for 450 epochs with the same setting based on using Warm-Start Strategy [22]. The L_1 loss function is used with the Adam optimizer. After the previous stage, the model is trained starting from the pre-trained weights using the DIV2K and Flickr2K datasets with an initial learning rate of 5×10^{-5} for 200 epochs using L_1 loss.

4.10. LVGroup_HFUT

Network Architecture. The LVGroup_HFUT team proposed a GlobalSwinIR based on the SwinIR model since existing representative methods pay less attention to the global features of images and mainly focus on the local features of images. The diagram of the GlobalSwinIR is shown in Fig. 9. Specifically, the GlobalSwinIR follows the basic architecture of SwinIR [30]. While increasing the depth of SwinIR to obtain better detail recovery quality, another global feature branch [53] is introduced to better capture the global features of the image, thereby guiding image recovery.

Training strategy. The GlobalSwinIR is trained a total of 500K iterations. The patch size(random crop) is 192 for a high-resolution image. The Adam optimizer is used. The learning rate is initially set to 2×10^{-4} and is halved every 100K iterations. Meanwhile, in the first 400K iterations, L_1 loss is used as our loss function, and in the last 100,000 iterations, L_2 loss is adopted.

4.11. AhRightRightRight

Network Architecture. Inspired by HNCT [15], NeWCRFs [56], the AhRightRightRight team proposed the Hybrid Attention Single Image Super Resolution Network with Conditional Random Field (HANCRF+). The HANCRF+ combines spatial attention, channel attention, and self-attention. While retaining the advantage of performing local feature extraction quickly, it can explore feature correlations along spatial and channel dimensions and activate more input pixels. Feature fusion using NeW FC-CRFs [56] uses pixel-to-pixel relations to constrain features and fuse the different layers of features.

As shown in Fig. 11, HANCRF+ consists of four parts: shallow feature extraction, deep feature extraction, feature

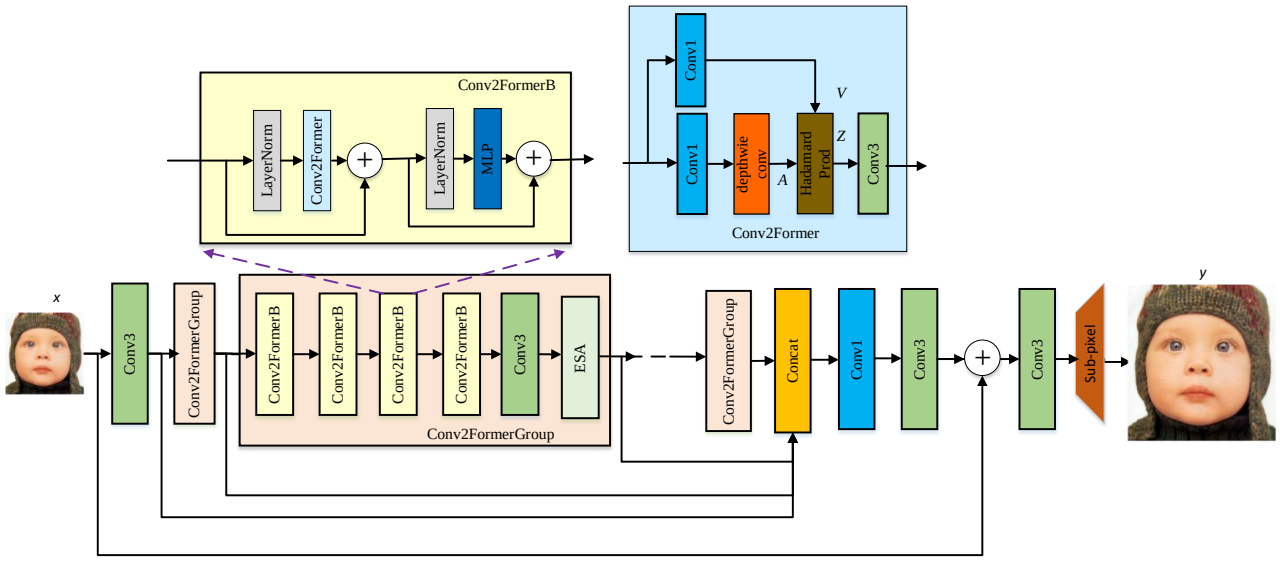


Figure 8. *GarasSjtu Team:* The structure of the proposed STSN.

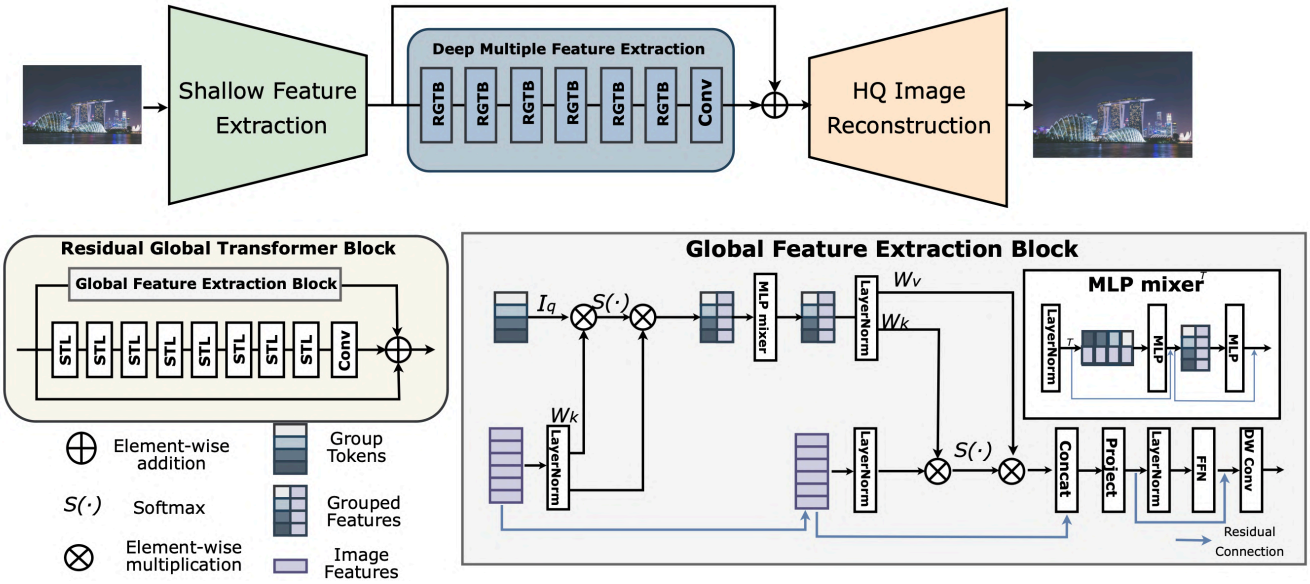


Figure 9. *LVGroup_HFUT Team:* The structure of the proposed GlobalSwinIR. The Swin Transformer Layer(STL) is the same as SwinIR [30].

fusion, and up-sampling module. Specifically, a 3×3 convolution is used for extracting shallow features f_0 from the input LR image I_{LR}:

$$f_0 = H_s(I_{LR}), \tag{4}$$

where H_s represents a 3×3 convolutional layer. Subsequently, the shallow feature is used for the deep feature extraction by a stack of Enhanced Hybrid Blocks of CNN and Transformer (EHBCT+). This process can be formulated as:

$$f_k = H^k_{EHBCT+}(f_{k-1}), \quad k = 1, ..., 6, \tag{5}$$

where H^k_{EHBCT+} denotes the n-th EHBCT+. The outputs of all EHBCTs will concatenate together, and the number of channels is adjusted by a 1×1 convolution to obtain the residual features f_c.

Subsequently, f_c and f_k will be input to Neural Window Fully-connected CRFs (NeW FC-CRFs) to aggregate

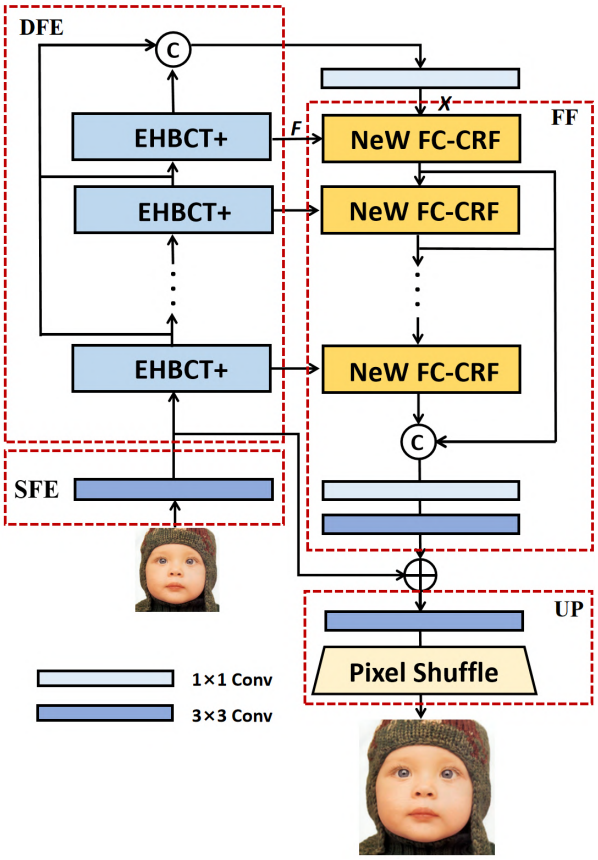


Figure 10. *AhRightRightRight Team:* The architecture of the proposed HANCRF+ for image super-resolution.

the observed information from each module of the feature extraction section, and the output of each layer can be denoted as f_m. The outputs of the NeW FC-CRFs are concatenated together, and the number of channels is adjusted by a 1×1 convolution to obtain the features $f_{c'}$, followed by a final adjustment by a 3×3 convolution to obtain f_{fin}.

$$f'_k = H^k_{NeWFC-CRF}(f_k, f'_{k+1}), \quad k = 1, ..., 6, \tag{6}$$

where $H^k_{NeWFC-CRF}$ denotes the k-th NeW FC-CRFs. Finally, the image is reconstructed by using 3×3 convolutional layers and pixel shuffle. The results of the reconstructed super-resolution image are as follows:

$$I_{SR} = H_{up}(H_{rec}(f_{fin} + f_0)). \tag{7}$$

The L_1 pixel loss is minimized to optimize the parameters of HANCRF+.

$$L_1 = ||I_{SR} - I_{HR}||_1. \tag{8}$$

Enhanced hybrid block of C'N'N and transformer As shown in Fig. 11, the EHBCT+ be composed of two Swin Transformer Blocks(STB) [30], two Enhanced Spatial Attention(ESA) modules [34], Multi-Spectral Channel Attention Module [40] and one convolutional layer.

Neural Window Fully-connected CRFs CRFs mainly

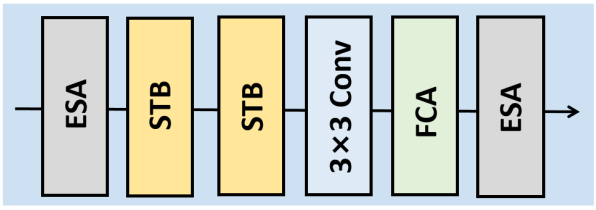


Figure 11. *AhRightRightRight Team:* The module of Enhanced Hybrid Blocks of CNN and Transformer (EHBCT+).

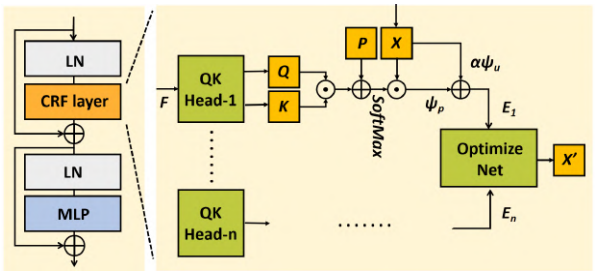


Figure 12. *AhRightRightRight Team:* The architecture of the NeW FC-CRFs.

consist of unary potentials on individual pixels or image patches and pairwise potentials on neighboring pixels or patches. The energy function of the fully connected CRFs is usually defined as

$$E(x) = \sum_i \psi_u(x_i) + \sum_{ij} \psi_p(x_i, x_j), \tag{9}$$

where x_i is the value of a point in the image, j is the other points in the image, ψ_u represents the unary potential of the pixel, and ψ_p represents the pairwise potential calculated by the pixel and its surrounding pixels.

Using neural networks instead of hand-designed potential functions can discover more potential information while performing end-to-end training. NeW FC-CRFs [56] unary potential is obtained directly from the network based on image features.

$$\psi_u(x_i) = \theta_u(I, x_i), \tag{10}$$

where θ is a parameter of the unary network, I is the input color image. NeW FC-CRF divides the image into block-based windows and uses the Swin Transformer strategy [30] to calculate the pairwise potential energy within the window.

$$\psi_p = SoftMax(qK^T + b)X, \tag{11}$$

In HANCRF+, the New FC-CRFs calculate the unary potential and pairwise potential through the feature map F output by EHBCT+ and the feature map X output by the previous layer network, as shown in Fig. 12. A learnable weight α is added in front of the unary potential, allowing the network to learn adaptively from the pixels themselves and from pairs of pixels.

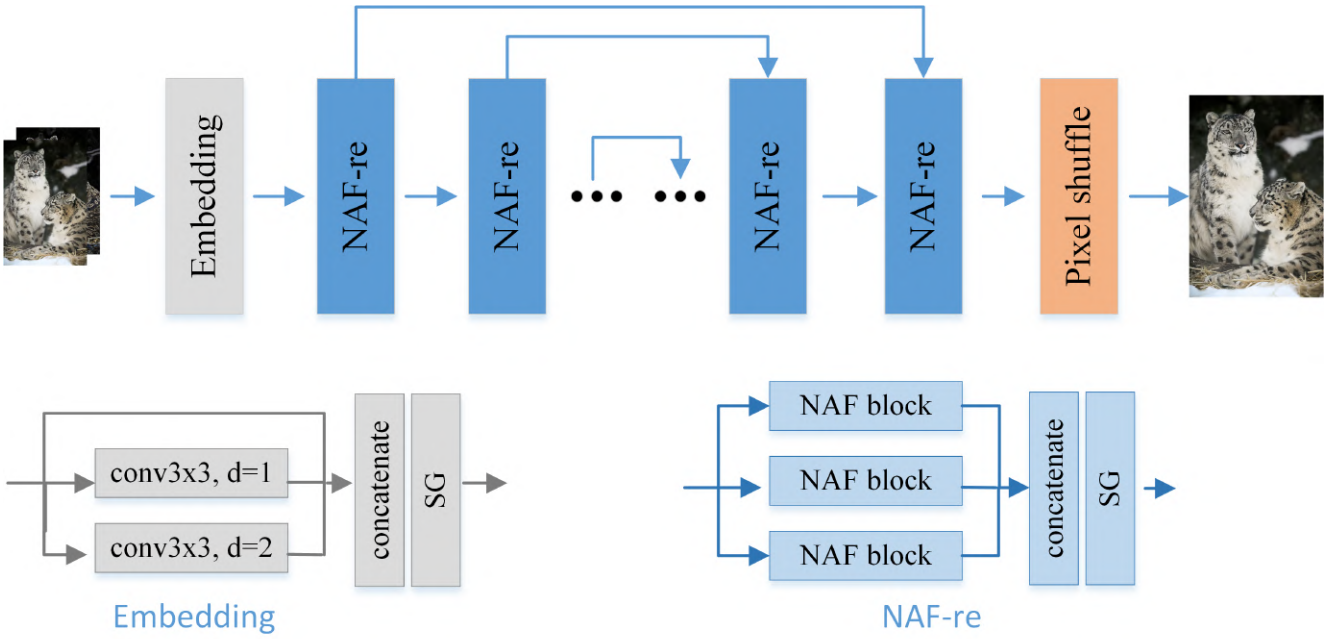


Figure 13. *chaobaer Team:* The proposed method.

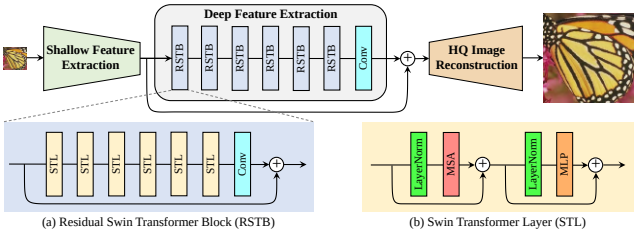


Figure 14. *helloooo Team:* The network architecture.

Training strategy. The LSDIR dataset is adopted to perform pre-training, and the DIV2K dataset is used for fine-tuning. The patch size is 64×64 for input images, and the batch size is set to 16. Data augmentation methods, including random rotation and horizontal and vertical flipping, are applied during our training. All models are trained using the Adam with L_1 loss. For pre-training, the learning rate is initialized to $5\times\times10^{-4}$, halved per 50 epochs. The total number of epochs is 300. For fine-tuning, the learning rate is initialized to 6.25×10^{-5}, halved per 100 epochs. The total number of epochs is 300. The model is implemented by Pytorch 1.11.0 and trained with NVIDIA GeForce RTX 3090 GPUs.

4.12. helloooo

Network Architecture. The helloooo team adopted SwinIR [30] as the base network model and combined the advantages of CNN and Transformer to further improve the model effect. As shown in Fig. 14, the network mainly comprises three modules: shallow feature extraction, deep feature extraction, and high-quality image reconstruction module.

The shallow feature extraction module adopts 3×3 convolutional layer to extract shallow features. The network adopts a long-distance connection to directly transmit low-frequency information to the reconstruction module, which can help the deep feature extraction module focus on high-frequency information and stabilize training.

The extraction of deep features is different from that of shallow features. Deep features focus on recovering lost high-frequency information. The extraction module is mainly composed of residual Swin Transformer blocks(RSTB), and each Transformer block utilizes multiple Swin Transformers for local attention and cross-window interaction. In addition, a convolution layer is added at the end of the block for feature enhancement and uses a residual join to provide a shortcut for feature aggregation.

The reconstruction module integrates shallow layer and depth features for high-quality image reconstruction. In the image reconstruction module, pixel shuffle replaces the convolution operation, and the features are up-sampled to realize reconstruction and save calculation.

Training strategy. The model is trained on DIV2K, and data augmentation, including vertical flipping, is used. The input patch size is set to 48×48. The model is trained by Adam optimizer, β_1=0.9, β_2=0.99. The initial learning rate is set as 0.0002, and the number of iterations is 500K. The warmup is performed 500 times per iteration. Moreover, the

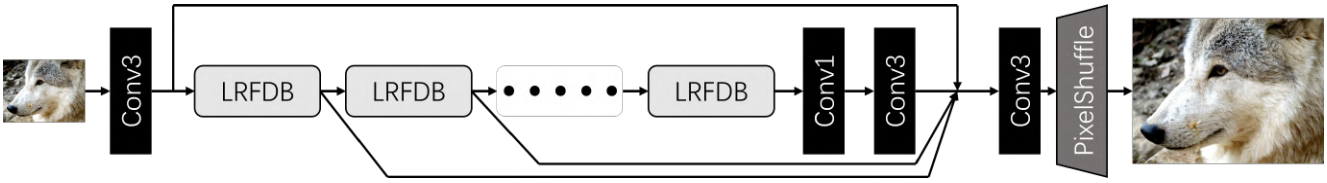


Figure 15. *Alpha Team:* Framework of LRFDN.

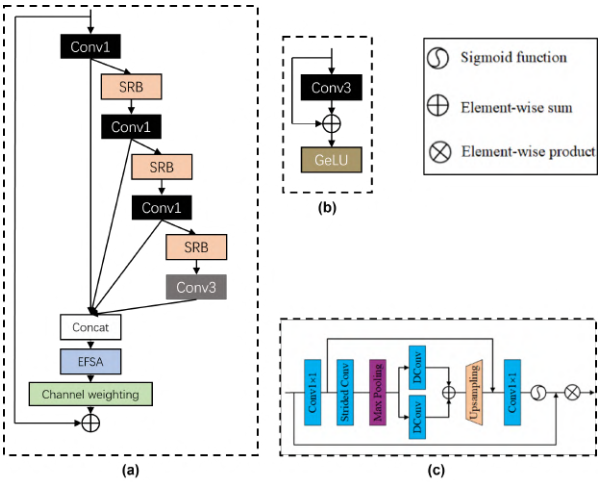


Figure 16. *Alpha Team:* (a) Structure of IRFDN. (b) Structure of SRB. (c) Structure of EFSA.

training strategy is divided into two stages. In the first stage, the model is trained from scratch with 255K iterations. In the second stage, the pre-trained weights of the first stage are used to accelerate the convergence rate of the model. The model is implemented using Pytorch 1.13.1 and trained using two GeForce RTX 2080ti GPUs.

4.13. chaobaer

Network Architecture. The chaobaer team developed an improved version of NAFNet to super-resolve the single image. The network architecture of the proposed method is shown in Fig. 13. NAFBlk has been proven to be efficient and easy to train in image super-resolution, so it is used as the basic stem. Most current DL-based super-resolution methods aim to learn the residual between LR and HR images. Since the residuals are mainly concentrated in edge regions, the boundary maps generated by Sobel are introduced as prior information and combine the features of a larger receptive field as the input embedding. In addition, skip connections are used for residual in residual learning.

Training strategy. During the training phase, only the DVI2K is used as the training dataset. The generated LR images are cropped into patches of size 128×128, and data is augmented with random flipping. CosineAnnealingLR is used to decay the learning rate. The L_1 loss to train for 200 epochs and then switched to L_2 loss for fine-tuning. The learning rate for fine-tuning is 5×10^{-5}.

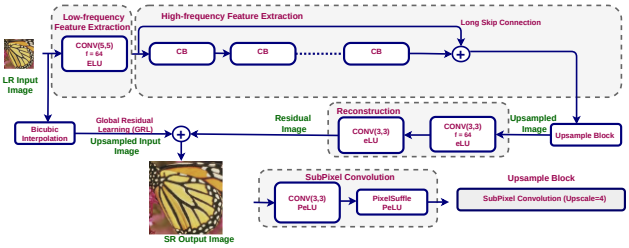
4.14. Alpha

Network Architecture. The Alpha team proposed a Lightweight Residual Feature Distillation Network (LRFDN) for efficient image super-resolution, as shown in Fig. 15. The LRFDN is mainly inspired by RFDN [32] and MAFFSRN [38]. Following the overall architecture of RFDN, LRFDN consists of four stages: shallow feature extraction, deep feature extraction, and reconstruction. To further reduce the parameters and computational complexity of the original RFDB module, the number of channels of layered distillation is effectively compressed. These distillation features are extracted by three shared 1×1 and one 3×3 convolutional filter. The design of the LRFDB block is shown in Fig. 16(a).

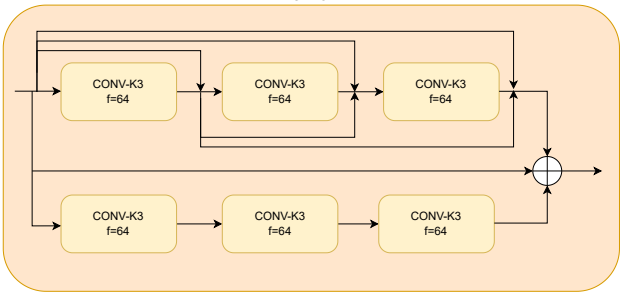
Inspired by the residual block proposed in the RFANet [33] and MAFFSRN [38], The team proposed an enhanced fast spatial attention module (EFSA). It aims to realize spatial attention weighting to make the features more concentrated in some desired regions to obtain more representative features. The design of the EFSA module is shown in Fig. 16(c). Using the blocks above, the proposed model can better extract and integrate compact contextual information with fewer parameters.

Furthermore, it has been found that channel-wise feature rescaling is effective for shallow SR models to boost reconstruction accuracy. Therefore, a channel weighting layer is involved in each LRFDB for modeling channel-wise relationships to utilize inter-dependencies among channels with slightly additional cost. Additionally, the GeLU activation function is adopted to replace LeakyReLU in RFDN.

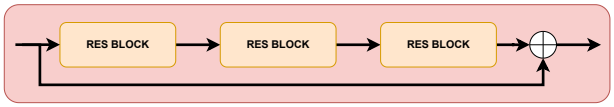
Training strategy. The proposed LRFDN has four LRFDBs, in which the number of feature channels is 64. During training, DIV2K and Flickr2K datasets are used for the whole process. The LRFDN model is trained from scratch with only one stage. HR patches of size 640×640 are randomly cropped from HR images, random horizontal flip, vertical flip, and rotation are introduced as the data augmentation, and the mini-batch size is set to 64. The Adam optimizer with β_1=0.9, β_2=0.999 is used to train our LRFDN model by minimizing the L_1 loss function. The base learning rate is set to 5×10^{-4} equipped with co-



(a) The block schematic of the proposed architecture.



(b) The design of the ResBlock.



(c) The design of the Concat Block.

Figure 17. *SVNIT_NTNU Team:* The architecture of the proposed model for image super-resolution.

sine learning rate decay and 3000 warm up steps. The total number of epochs is 1000. An exponential moving average (EMA) of LRFDN weights is also maintained over training with a decay of 0.9999, which improves the performance of the model.

4.15. SVNIT_NTNU

Network Architecture. The SVNIT_NTNU used the dense convolutional neural network approach in the proposed method. Figure. 17(a) depicts the proposed architecture for single image SR for scaling factors of $\times 4$. The LR image is applied as input to the network, and it is passed to extract the salient features from it. The low-frequency features are extracted with first layers that employ a convolutional layer, while high-frequency features are extracted with residual blocks. The architecture uses the Exponential Linear Unit (ELU) activation function to improve learning performance at each layer in an efficient manner.

A new and core element of the proposed architecture is the partially densely connected design of ResBlock that preserves the high-frequency details of the SR image by retaining salient features, which is displayed in Fig. 17(b). The kernel sizes (*i.e.*, 3×3) adopted in ResBlock recover details distributed at local and global regions. The channel attention modules are further used in concat block, shown to perform adaptive re-scaling of features on a per-channel basis. The pixel shuffle is used to upscale the feature maps to the desired scaling factor (*i.e.*, $\times 4$) [54]. Additionally, low-frequency features are upscaled and added to the reconstructed output to retain more versatile information.

Training strategy. The proposed network is trained using a weighted combination of L_1, SSIM functions with a learning rate of 1×10^{-4}, which is decayed by 1×10^4 iterations, and the same is optimized using Adam optimizer. The model has trained up to 1×10^5 iterations with a batch size of 8. The model is implemented using Pytorch.

5. Acknowledgements

This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2023 sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).

A. Teams and Affiliations

NTIRE 2023 team

Title: NTIRE 2023 Image Super-Resolution ($\times4$) Challenge

Members:
Yulun Zhang[1] (yulun100@gmail.com),
Kai Zhang[1] (cskaizhang@gmail.com),
Zheng Chen[2] (463465810cz@gmail.com),
Yawei Li[1] (yawei.li@vision.ee.ethz.ch),
Radu Timofte[3] (radu.timofte@uni-wuerzburg.de)
Affiliations:
[1] Computer Vision Lab, ETH Zurich, Switzerland
[2] Shanghai Jiao Tong University, China
[3] University of Würzburg, Germany

ZZPM

Title: SwinIR-LTE Fusion Net

Members: Junpei Zhang (22171214671@stu.xidian.edu.cn), Kexin Zhang, Rui Peng, Yanbiao Ma, Licheng Jiao
Affiliation:
Xidian University

Graphene

Title: Image Super-Resolution Transformer with Cross-Scale Attention and Wavelet Hallucination

Members: Huaibo Huang[1,2] (huaibo.huang@cripac.ia.ac.cn), Xiaoqiang Zhou[1,3], Yuang Ai[1,4], Ran He[1,2,5]
Affiliation:

[1] MAIS&CRIPAC, Institute of Automation, Chinese Academy of Sciences, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3] University of Science and Technology of China
[4] Beijing Institute of Technology
[5] School of Information Science and Technology, ShanghaiTech University, China

IPLAB

Title: Attention Retractable Frequency Transformer for Image Super-Resolution
Members: Yajun Qiu (qyjyun@gmail.com), Qiang Zhu, Pengfei Li, Qianhui Li, Shuyuan Zhu
Affiliation:
School of Information and Communication Engineering, University of Electronic Science and Technology of China

SRC-B

Title: SwinFIR: Revisiting the SwinIR with Fast Fourier Convolution and Improved Training for Image Super-Resolution
Members: Dafeng Zhang (dfeng.zhang@samsung.com), Jia Li, Fan Wang, Chunmiao Li
Affiliation:
Samsung Research China - Beijing (SRC-B)

LDCC

Title: LDCC: Latent Discriminative Cosine Criterion
Members: TaeHyung Kim (thkim07@lotte.net), Jungkeong Kil, Eon Kim, Yeonseung Yu, Beomyeol Lee, Subin Lee, Seokjae Lim, Somi Chae, Heungjun Choi
Affiliation:
LOTTE DATA COMMUNICATION COMPANY, Seoul, Korea

NTU607-SR

Title: Transfer learning with ART for image super-resolution
Members: *ZhiKai Huang[2] (brent5481@gmail.com)*, YiChung Chen[3], YuanChun Chiang[2], HaoHsiang Yang[2], WeiTing Chen[1], HuaEn Chang[2], I-Hsiang Chen[2], ChiaHsuan Hsieh[4], SyYen Kuo[2]
Affiliation:
[1] Graduate Institute of Electronics Engineering, National Taiwan University, Taiwan
[2] Department of Electrical Engineering, National Taiwan University, Taiwan
[3] Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[4] ServiceNow, USA

Swin2SR

Title: Swin2SR
Members: Ui-Jin Choi[1] (choiuijin1125@gmail.com), Marcos V. Conde[2]
Affiliation:
[1] MegaStudyEdu, South Korea
[2] Computer Vision Lab, CAIDAS, University of Würzburg, Germany

IKLAB-TUK

Title: Dense Residual Swin Transformers for Image Super Resolution
Members: Sunder Ali Khowaja[1] (sandar.ali@usindh.edu.pk), Jiseok Yoon[2], Ik Hyun Lee[2]
Affiliation:
[1] University of Sindh, Pakistan
[2] IKLAB Inc., Tech University of Korea, Siheung-Si, South Korea

GarasSjtu

Title: A Simple Transformer-style Network for Image Super-resolution
Members: Garas Gendy[1] (garasgaras@yahoo.com), Nabil Sabor[2], Jingchao Hou[1], Guanghui He[1]
Affiliation:
[1] Micro-Nano Electronics Department, Shanghai Jiao Tong University
[2] Electrical Engineering Department, Faculty of Engineering, Assiut University

LVGroup_HFUT

Title: Global Swin Transformer for Image Super-Resolution
Members: Zhao Zhang (cszzhang@gmail.com), Baiang Li, Huan Zheng, Suiyi Zhao, Yangcheng Gao, Yanyan Wei, Jiahuan Ren
Affiliation:
Hefei University of Technology

AhRightRightRight

Title: The Hybrid Attention Single Image Super Resolution Network with Conditional Random Field
Members: Jiayu Wei (jy.wei@bjtu.edu.cn), Yanfeng Li, Jia Sun, Zhanyi Cheng, Zhiyuan Li, Xu Yao
Affiliation:

Beijing Jiaotong University

helloooo

Title: Image super-resolution reconstruction algorithm SwinIR based on warm-up learning rate strategy.
Members: Xinyi Wang (493983152@qq.com), Danxu Li, Xuan Cui
Affiliation:

chaobaer

Title: NAF-Reload: Multilayer Residuals for Image Super Resolution
Members: Jun Cao (jonahcao0109@gmail.com), Cheng Li
Affiliation:

Alpha

Title: Lightweight Residual Feature Distillation Network for Efficient Image Super-Resolution
Members: Jianbin Zheng (jabir.zheng@outlook.com)
Affiliation:
South China University of Technology, Guangzhou, Guangdong, China

SVNIT_NTNU

Title: Convolution Neural Network based Single Image Super-Resolution
Members: Anjali Sarvaiya[1] (anjali.sarvaiya.as@gmail.com), Kalpesh Prajapati[1], Ratnadeep Patra[1], Pragnesh Barik[1], Chaitanya Rathod[1], Kishor Upla[1], Kiran Raja[2], Raghavendra Ramachandra[2], Christoph Busch[2]
Affiliation:
[1] Sardar Vallabhbhai National Institute of Technology
[2] Norwegian University of Science and Technology, Norway

References

[1] Aihub super resolution dataset. https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=77, 2020. "Accessed: 2023-03-21". 8

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 2

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[4] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[5] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023. 1, 2, 3, 7, 8, 9

[6] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *ICCV*, 2019. 9

[7] Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022. 1, 2

[8] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 2020. 7

[9] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *ECCVW*, 2023. 9

[10] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[11] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2016. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 8

[15] Jinsheng Fang, Hanjiang Lin, Xinyu Chen, and Kun Zeng. A hybrid network of cnn and transformer for lightweight image super-resolution. In *CVPRW*, 2022. 11

[16] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *ICCV*, 2021. 7

[17] Garas Gendy, nabil sabor, Jingchao Hou, and Guanghui He. A simple transformer-style network for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 11

[18] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *arXiv preprint arXiv:2211.11943*, 2022. 11

[19] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[21] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 8

[22] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPR*, 2022. 11

[23] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 1

[24] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 8

[25] Bingchen Li, Xin Li, Yiting Lu, Sen Liu, Ruoyu Feng, and Zhibo Chen. Hst: Hierarchical swin transformer for compressed image super-resolution. In *ECCVW*, 2023. 10, 11

[26] Xin Li, Simeng Sun, Zhizheng Zhang, and Zhibo Chen. Multi-scale grouped dense network for vvc intra coding. In *CVPRW*, 2020. 10

[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. 2, 8

[28] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[29] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[30] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCVW*, 2021. 1, 2, 3, 5, 7, 9, 10, 11, 12, 13, 14

[31] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2, 8

[32] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCVW*, 2020. 15

[33] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 1, 2, 11, 15

[34] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *CVPR*, 2020. 13

[35] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[36] 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. 7

[37] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *ICCV*, 2021. 1, 2

[38] Abdul Muqeet, Jiwon Hwang, Subin Yang, JungHeum Kang, Yongwoo Kim, and Sung-Ho Bae. Multi-attention based ultra lightweight image super-resolution. In *ECCVW*, 2020. 15

[39] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, Jianzhao Liu, Sen Liu, and Zhibo Chen. Fan: Frequency aggregation network for real image super-resolution. In *ECCVW*, 2020. 9, 10

[40] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *ICCV*, 2021. 13

[41] Mehdi SM Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 1

[42] 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. 10

[43] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[44] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 2001. 8

[45] 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. 2, 8

[46] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, 2016. 3, 8, 9

[47] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[48] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[49] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 8

[50] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 3

[52] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint arXiv:2203.05482*, 2022. 8

[53] Chenhongyi Yang, Jiarui Xu, Shalini De Mello, Elliot J Crowley, and Xiaolong Wang. Gpvit: A high resolution non-hierarchical vision transformer with group propagation. In *ICLR*, 2023. 11

[54] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019. 16

[55] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *CVPR*, 2020. 8

[56] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *CVPR*, 2022. 11, 13

[57] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, et al. NTIRE 2023 challenge on hr depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[58] 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. 7

[59] Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. In *ICLR*, 2023. 1, 2, 3, 5, 8

[60] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In *ICIP*, 2018. 10

[61] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, 2022. 2

[62] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2

[63] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *ICLR*, 2019. 1, 2

[64] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[65] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, 2020. 1, 2

[66] Qiang Zhu, Li Peng Fei, and Qianhui Li. Attention retractable frequency fusion transformer for image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 5