arXiv:2409.16947v1 [cs.CV] 25 Sep 2024

NTIRE 2024 Challenge on Stereo Image Super-Resolution: Methods and Results

Longguang Wang, Yulan Guo†, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, Radu Timofte, Davinci, Saining Zhang, Rongxin Liao, Ronghui Sheng, Feng Li, Huihui Bai, Wei Zhang, Runmin Cong, Yuqiang Yang, Zhiming Zhang, Jingjing Yang, Long Bao, Heng Sun, Kanglun Zhao, Enyuan Zhang, Huiyuan Fu, Huadong Ma, Yuanbo Zhou, Wei Deng, Xintao Qiu, Tao Wang, Qinquan Gao, Tong Tong, Yinghao Zhu, Yongpeng Li, Zhitao Chen, Xiujuan Lang, Kanghui Zhao, Bolin Zhu, Wenbin Zou, Yunxiang Li, Qiaomu Wei, Tian Ye, Sixiang Chen, Weijun Yuan, Zhan Li, Wenqin Kuang, Ruijin Guan, Jian Wang, Yuqi Miao, Baiang Li, Kejie Zhao, Wenwu Luo, Jing Wu, Yunkai Zhang, Songyan Zhang, Jingyi Zhang, Junyao Gao, Xueqiang You, Yanhui Guo, Hao Xu, Sahaj K. Mistry, Aryan Shukla, Sourav Saini, Aashray Gupta, Vinit Jakhetiya, Sunil Jaiswal, Zhao Zhang, Bo Wang, Suiyi Zhao, Yan Luo, Yanyan Wei, Yihang Chen, Ruting Deng, Yifan Deng, Jingchao Wang, Zhijian Wu, Dingjiang Huang, Yun Ye

Abstract

This paper summarizes the 3rd NTIRE challenge on stereo image super-resolution (SR) with a focus on new solutions and results. The task of this challenge is to super-resolve a low-resolution stereo image pair to a high-resolution one with a magnification factor of $\times 4$ under a limited computational budget. Compared with single image SR, the major challenge of this challenge lies in how to exploit additional information in another viewpoint and how to maintain stereo consistency in the results. This challenge has 2 tracks, including one track on bicubic degradation and one track on real degradations. In total, 108 and 70 participants were successfully registered for each track, respectively. In the test phase, 14 and 13 teams successfully submitted valid results with PSNR (RGB) scores better than the baseline. This challenge establishes a new benchmark for stereo image SR.

1. Introduction

Recently, dual cameras have been increasingly popular in AR/VR, mobile phones, autonomous vehicles and robots to record and perceive the 3D environment. For higher perceptual quality and finer-grained parsing of the real world, increasing the resolution of stereo images is highly demanded. To this end, stereo image super-resolution (SR) has been introduced to reconstruct a high-resolution (HR) stereo image pair with finer details from a low-resolution (LR) one.

Compared with a single image, stereo images can provide additional cues from a second viewpoint to better recover image details. However, since an object is projected onto different locations in left and right views, how to make full use of these cross-view information still remains challenging. On the one hand, stereo correspondence for objects at different depths can vary significantly. On the other hand, the occlusion between left and right views hinders correspondences to be incorporated.

To develop and benchmark stereo SR methods, stereo image SR challenge was hosted in the NTIRE 2022 and NTIRE 2023 workshops [1, 2]. The objective of previous challenges is to minimize the distortion between super-resolved stereo images and the groundtruth under both bicubic and realistic degradations. However, the computational cost of the stereo SR methods is not fully considered, which hinders these methods to be deployed on resource-limited devices.

Succeeding the previous years, NTIRE 2024 Stereo Image SR Challenge presents two competition tracks. These two tracks are inherited from the NTIRE 2023 challenge with an additional constraint of computational complexity. Specifically, both the memory and computational cost are considered for real-world applications.

†Corresponding author: Yulan Guo.
Section 7 provides the authors and affiliations of each team.
NTIRE 2024 webpage: https://cvlai.net/ntire/2024/
Challenge webpage (Track 1): https://codalab.lisn.upsaclay.fr/competitions/17245
Challenge webpage (Track 2): https://codalab.lisn.upsaclay.fr/competitions/17246
Github: https://github.com/The-Learning-And-Vision-Atelier-LAVA/Stereo-Image-SR/tree/NTIRE2024

This challenge is one of the NTIRE 2024 Workshop associated challenges on: dense and non-homogeneous dehazing [3], night photography rendering [4], blind compressed image enhancement [5], shadow removal [6], efficient super resolution [7], image super resolution ($\times 4$) [8], light field image super-resolution [9], stereo image super-resolution [10], HR depth from images of specular and transparent surfaces [11], bracketing image restoration and enhancement [12], portrait quality assessment [13], quality assessment for AI-generated content [14], restore any image model (RAIM) in the wild [15], RAW image super-resolution [16], short-form UGC video quality assessment [17], low light enhancement [18], and RAW burst alignment and ISP challenge.

2. Related Work

In this section, we briefly review recent advances in single image SR and stereo image SR.

2.1. Single Image SR

In the last decade, learning-based approaches have dominated the area of single image SR [19–23]. Dong *et al.* [24] proposed the first CNN-based SR model (*i.e.*, SRCNN) to learn an LR-to-HR mapping. Following SRCNN, early methods focus on developing larger and more effective network architectures to achieve higher SR performance [25–27]. Specifically, Zhang *et al.* [28] proposed a residual dense network (*i.e.*, RDN) to fully use hierarchical features by combining residual connection [29] with dense connection [30]. Subsequently, Li *et al.* [31] suggested to use image features at different scales for single image SR, and proposed a multi-scale residual network (*i.e.*, MSRN). Dai *et al.* [32] proposed a second-order attention network (*i.e.*, SAN) for more powerful feature correlation learning, which achieves superior performance. Recently, the efficiency of SR models has drawn increasing interests, with numerous lightweight network architectures being developed [33–35]. For example, distillation blocks were proposed for feature learning in IDN [36]. Then, a cascading mechanism was introduced to encourage efficient feature reuse in CARN [37]. Different from these manually designed networks, Chu *et al.* [38] developed a compact architecture using neural architecture search (NAS).

Inspired by the great success of Transformer in computer vision, Transformer has been widely studied to promote single image SR. Liang *et al.* [39] designed a SwinIR model for image restoration by applying Swin Transformer [40]. Lu *et al.* [41] proposed an effective super-resolution Transformer (*i.e.*, ESRT) for single image SR, which introduced a lightweight Transformer and feature separation strategy to reduce GPU memory consumption. Zamir *et al.* [42] proposed an encoder-decoder Transformer (*i.e.*, Restormer) for image restoration with multi-scale local-global representation learning.

2.2. Stereo Image SR

Jeon *et al.* [43] developed the first learning-based stereo image SR method termed StereoSR. This method incorporates cross-view information by concatenating the left image and a stack of right images with different pre-defined shifts. Later, Wang *et al.* [44, 45] developed PASSRnet by introducing a parallax attention module (PAM) to capture stereo correspondence along the epipolar line. Inspired by PASSRnet, Song *et al.* [46] further combined self-attention with parallax attention to better model global correspondence. Wang *et al.* [47] introduced a Siamese network with a bi-directional parallax attention module to simultaneously super-resolve left and right images in a symmetric manner. Guo *et al.* [48] proposed a new Transformer-based parallax fusion model called Parallax Fusion Transformer.

Instead of using parallax-attention mechanism, several efforts have also been made to employ stereo matching approach to capture stereo correspondence. Yan *et al.* [49] proposed a domain adaptive stereo SR network (DASSR) to incorporate cross-view information through explicit disparity estimation using a pre-trained stereo matching network. Dai *et al.* [50] proposed a feedback network to alternately solve disparity estimation and stereo image SR in a recurrent manner. Wan *et al.* [51] proposed a multi-stage network to progressively obtain cross-view features and an edge-guided supplementary branch to refine the cross-view features.

In the NTIRE 2022 Stereo Image SR Challenge, the champion team developed NAFSSR network [52] by using nonlinear activation-free network (NAFNet) for feature extraction and PAM for cross-view information interaction. In the NTIRE 2023 Stereo Image SR Challenge, the champion team of track 1 developed a Hybrid Transformer and CNN Attention Network (HTCAN), which employs a Transformer-based network for single image enhancement and a CNN-based network for stereo information aggregation. The champion team of track 2 proposed a SwinFIRSSR by using Swin Transformer [40] and fast Fourier convolution [53]. The champion team of track 3 combined NAFSSR [54] with LTE [55] and proposed LTFSSR.

3. NTIRE 2024 Challenge

The objectives of the NTIRE 2024 challenge on example-based stereo image SR are: (i) to gauge and push the state-of-the-art in SR under given computational constraints; and (ii) to compare different solutions.

3.1. Dataset

Training Set. The training set of the Flickr1024 dataset [56] (with 800 images) is used as the training set of this challenge. Both original HR images and their LR versions will be released. The participants can use these HR images as ground-truth to train their models.

Validation Set. The validation set of the Flickr1024 dataset (with 112 images) is used as the validation set of this challenge. Similar to the training set, both HR and LR images in the validation set are provided. The participants can download the validation set to evaluate the performance of their developed models by comparing their super-resolved images with the HR ground-truth images. Note that the validation set should be used for validation purposes only but cannot be used as additional training data.

Test Set. To rank the submitted models, a test set consisting of 100 stereo images is provided. Unlike the training and validation sets, only LR images will be released for the test set. The participants must apply their models to the released LR stereo images and submit their super-resolved images to the server. It should be noted that the images in the test set (even the LR versions) cannot be used for training.

3.2. Tracks

• Track 1: Constrained SR & Bicubic Degradation

In this track, bicubic degradation (Matlab *imresize* function in bicubic mode) is used to generate LR images:

$$I^{LR} = I^{HR} \downarrow_s, \tag{1}$$

where I^{LR} and I^{HR} are LR and HR images, $\downarrow_s$ represents bicubic downsampling with scale factor s.

• Track 2: Constrained SR & Realistic Degradation

In this track, a realistic degradation model consisting of blur, downsampling, noise, and compression is adopted to synthesize LR images:

$$I^{LR} = \mathcal{C}\left(\left(I^{HR} \otimes k\right) \downarrow_s + n\right), \tag{2}$$

where k is the blur kernel, n is additive Gaussian noise, and $\mathcal{C}$ represents JPEG compression.

In these two tracks, the model size (*i.e.*, number of parameters) is restricted to 1 MB, and the computational complexity (*i.e.*, number of MACs) is restricted to 400 G (a stereo image pair of size 320×180). Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used as metrics for performance evaluation. The average results of left and right views over all of the test scenes are reported. Note that only PSNR (RGB) is used for ranking.

3.3. Challenge Phases

Development Phase. The participants were provided with pairs of LR and HR training images and LR validation images of the Flickr1024 dataset. The participants had the opportunity to test their solutions on the LR validation images and to receive immediate feedback by uploading their results to the server. A validation leaderboard is available online.

Testing Phase. The participants were provided with the LR test images and were asked to submit their super-resolved images, codes, and a fact sheet for their methods before the challenge deadline. After the end of the challenge, the final results were released to the participants.

4. Challenge Results

4.1. Track 1: Fidelity & Bicubic Degradation

Among the 108 registered participants, 14 teams successfully participated the final phase and submitted their results, codes, and fact sheets. Table 1 reports the final test results, rankings of the challenge, and major details from the fact sheets of 14 teams. These methods are briefly described in Section 5 and the team members are listed in Section 7.

4.2. Track 2: Fidelity & Realistic Degradation

Among the 70 registered participants, 13 teams successfully participated the final phase and submitted their results, codes, and fact sheets. Table 2 reports the final test results, rankings of the challenge, and major details from the fact sheets of 13 teams. These methods are briefly described in Section 5 and the team members are listed in Section 7.

4.3. Summary

Architectures and main ideas. All the proposed methods are based on deep learning techniques. Transformers and the winner method in the NTIRE 2022 challenge (*i.e.*, NAFSSR) are widely used as the basic architecture. To exploit cross-view information, the idea of parallax-attention mechanism (PAM) are widely adopted in most solutions to capture stereo correspondence.

Data Augmentation. Widely applied data augmentation approaches such as random flipping and RGB channel shuffling are used for most solutions. In addition, random horizontal shifting, Mixup, CutMix, and CutMixup are also used in several solutions and help to achieve superior performance.

Ensembles and Fusion. Due to the constraints in computational complexity, ensemble strategy (data ensemble and model ensemble) is only adopted in a few solutions. Several solutions employ a limited number of transformed inputs for enhanced prediction. In addition, the champion solution in track 1 employs model exponential moving average for model ensemble without additional overhead during the inference phase.

Conclusions. By analyzing the settings, the proposed methods and their results, it can be concluded that: 1) The proposed methods strike better balance between accuracy and efficiency. 2) With recent renaissance of CNNs (*e.g.*,

Table 1. NTIRE 2024 Stereo Image SR Challenge (Track 1) results, rankings, and details from the fact sheets. Note that, PSNR (RGB) is used for the ranking. "Transf" denotes Transformer and "PAM" denotes parallax attention mechanism.

Rank	Team	Authors	PSNR (RGB)	Architecture	Disparity	Ensemble
1	Davinci	Davinci and S. Zhang	23.6503	CNN+Transf	PAM	Data+Feature
2	HiSSR	R. Liao, R. Sheng, F. Li, H. Bai, R. Cong, and W. Zhang	23.6105	CNN	PAM	N.A.
3	MiVideoSR	Y. Yang, Z. Zhang, J. Yang, L. Bao, and H. Sun	23.6070	CNN+Transf	PAM	Data
4	webbzhou	Y. Zhou, W. Deng, X. Qiu, T. Wang, Q. Gao, and T. Tong	23.5941	CNN	PAM	N.A.
5	Qi5	Y. Zhu and Y. Li	23.5896	CNN+Transf	PAM	N.A.
6	WITAILab	Z. Chen, X. Lang, K. Zhao, and B. Zhu	23.5725	CNN+Transf	PAM	N.A.
7	Giantpandacv	W. Zou, Y. Li, Q. Wei, T. Ye, and S. Chen	23.5271	CNN	PAM	N.A.
8	JNU_620	W. Yuan, Z. Li, W. Kuang, and R. Guan	23.4851	CNN	PAM	N.A.
9	GoodGame	J. Wang, Y. Miao, B. Li, and K. Zhao	23.4598	CNN	PAM	N.A.
10	Fly_Flag	W. Luo, and J. Wu	23.4510	CNN+Transf	PAM	Data
11	Mishka	Y. Zhang, B. Li, S. Zhang, J. Zhang, J. Gao, and X. You	23.4270	CNN+Mamba	PAM	N.A.
12	LightSSR	Y. Guo and H. Xu	23.3888	CNN	N.A.	N.A.
13	DVision	S. Mistry, A. Shukla, S. Saini, A. Gupta, V. Jakhetiya, and S. Jaiswal	23.1895	CNN	PAM	N.A.
14	LVGroup_HFUT	Z. Zhang, B. Wang, S. Zhao, Y. Luo, and Y. Wei	23.0977	CNN	N.A.	Data

Table 2. NTIRE 2024 Stereo Image SR Challenge (Track 2) results, rankings, and details from the fact sheets. Note that, PSNR (RGB) is used for the ranking. "Transf" denotes Transformer and "PAM" denotes parallax attention mechanism.

Rank	Team	Authors	PSNR (RGB)	Architecture	Disparity	Ensemble
1	Davinci	Davinci and S. Zhang	21.8724	CNN+Transf.	PAM	Data+Feature
2	MiVideoSR	Y. Yang, Z. Zhang, J. Yang, L. Bao, and H. Sun	21.6983	CNN+Transf	PAM	Data
3	BUPTMM	K. Zhao, E. Zhang, H. Fu, and H. Ma	21.6702	CNN	PAM	N.A.
4	webbzhou	Y. Zhou, W. Deng, X. Qiu, T. Wang, Q. Gao, and T. Tong	21.6691	CNN	PAM	N.A.
5	JNU_620	W. Yuan, Z. Li, W. Kuang, and R. Guan	21.5935	CNN	PAM	N.A.
6	Liz620	Y. Chen, R. Deng, and Y. Deng	21.5655	CNN	PAM	N.A.
7	Mishka	Y. Zhang, B. Li, S. Zhang, J. Zhang, J. Gao, and X. You	21.5313	CNN+Mamba	PAM	N.A.
8	ECNU-IDEALab	J. Wang, Z. Wu, and D. Huang	21.5238	CNN+Transf	PAM	N.A.
9	Giantpandacv	W. Zou, Y. Li, Q. Wei, T. Ye, and S. Chen	21.4970	CNN	PAM	N.A.
10	HiYun	Y. Ye	21.1994	CNN	PAM	Data
11	GoodGame	J. Wang, Y. Miao, B. Li, and K. Zhao	20.7642	CNN	PAM	N.A.
12	Fly_Flag	W. Luo, and J. Wu	20.7518	CNN+Transf	PAM	Data
13	LVGroup_HFUT	Z. Zhang, B. Wang, S. Zhao, Y. Luo, and Y. Wei	20.6167	CNN	N.A.	Data

NAFNet), Transformers and CNNs are comparably popular in this challenge and produce competitive performance. 3) Cross-view information lying at varying disparities is critical to the stereo image SR task and helps to achieve higher performance. 4) To meet the efficiency requirements of the challenge, techniques like depth-wise convolution are widely applied and produce promising results. 5) One recent remarkable technique (*i.e.*, Mamba) has been introduced to achieve efficient image SR and produces promising results.

5. Challenge Methods and Teams

5.1. Davinci - Track 1★, Track 2★

Inspired by SwinFIR [57], HAT [58], and NAFSSR [52], they proposed SwinFIRSSR using Swin transformer [40] and fast fourier convolution [53], as shown in Fig. 2. HAT employs Residual Hybrid Attention Group (RHAG) to activate more pixel in image SR transformer to improve the performance. Each RHAG contains N hybrid attention blocks (HAB), an overlapping cross-attention block (OCAB) and a 3×3 convolutional layer. They replaced the convolution (3×3) with fast fourier convolution and a residual module to fuse global and local features, namely Spatial-frequency Block (SFB), to improve the representation ability of the model. They also followed NAFSSR to fuse left/right features using stereo cross-attention module (SCAM).

During the training phase, HR images were cropped to 128×384 sub-images. The Adam [59] optimizer with default parameters and the Charbonnier L1 loss [60] were employed for training. The initial learning rate was set to 2×10^{-4} and decayed at 600,000, 650,000, 700,000, 750,000 iterations. The batch size was 4 and patch size was 32×96. The models were implemented usingPyTorch 1.8.1, NVIDIA A6000 GPU with CUDA11.1. Random horizontal flip, vertical flip, rotation, RGB perm and mixup [61] were adopted for data augmentation.

Inspired by [62], model ensemble was employed to improve the performance. Widely-applied multi-model and data ensemble strategy inevitably introduces additional computational overhead. To remedy this, they proposed to conduct model ensemble from the perspective of network parameters, as illustrated in Fig. 1. Specifically, the parameters of several models are aggregated as:

$$SwinFIRSSR(\theta) = \sum_{i=1}^{n} SwinFIRSSR(\theta)^i * \alpha^i, \quad (3)$$

where θ denotes the parameter sets of SwinFIRSSR, n is the numbers of models. α is the weight of each model and the $\alpha = \frac{1}{n}$ in our solution.

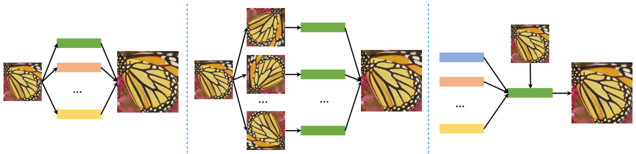


Figure 1. Comparison between different ensemble strategies. (a) Multi-model ensemble. (b) Data ensemble. (c) The proposed model ensemble. Rectangles with different colors represent different model parameters.

5.2. HiSSR - Track 1★

Figure 3 depicts the network structure of the proposed method. Specifically, CVHSSR [63] is used as the backbone, which consists of a cross-hierarchy information mining block (CVIM) and a cross-hierarchy information mining block (CHIMB). CHIMB leverages both spatial and channel-wise attention mechanisms to model information at different levels within a single view. Meanwhile, CVIM integrates similar information across different views through cross-view attention. They employed CHVIMB to extract both global and local information from a single view, while CVIM is utilized to integrate analogous information across each views. Following CFSR [64], they further reparameterized the depth-wise convolution in the Information Refinement Feedforward Block in CHIMB. To further enhance the model's performance, they introduced an iterative interaction mechanism, leveraging the restoration results as reference images to enhance the performance.

Training Settings. The number of CHIMB in RIISSR is set to 20, while the channels of all the convolutional layers are set to 48. Residual connections are inserted between every four blocks. The number of parameters in the model is 0.918M, and the MACs are 235.28G for a 320×180 stereo image pair. During training, HR images were cropped into (30 $\times$90) patches with a stride of 10. Random horizontal, flips, rotations, mixup and RGB channel shuffle were adopted for data augmentation. The AdamW with β_1 = 0.9 and β_2 = 0.99 was used for optimization. The learning rate was initialized to (1×10^{-3}) and decreased using the multistep strategy with γ set to 0.5. The proposed model was trained for (2.4×10^5) iterations with a batch size of 24. A $2\times$ SR model was first trained from scratch, which was then used to initialize the $4\times$ SR model.

5.3. MiVideoSR - Track 1★, Track 2★

Transformer-like methods have achieved advanced performance on low-level tasks. Motivated by NAFSSR [52], HAT [58] and SRFormer [65], this team proposed a HCASSR by plugging the SCAM of NAFSSR into the HAT to aggregate features from two views, as illustrated in Fig. 4. In addition, to improve the performance and efficiency of the model, they replaced the self-attention module with permuted self-attention (PSA) [65] to transfer the spatial information to the channel dimension.

Training Settings. The proposed model was first trained with a Charbonnier loss using an Adam optimizer and stopped after 400k iterations. Then, the resultant model was fine-tuned with the MSE loss. The batch size was set to 16 and the patch size was first set to 96×96 and then enlarged to 192×192 for fine-tuning. The learning rate was initialized as 5×10^{-4} and updated using a cosine annealing strategy. Data augmentation was performed through horizontal/vertical flipping, RGB channel random shuffing, and Mixup operations.

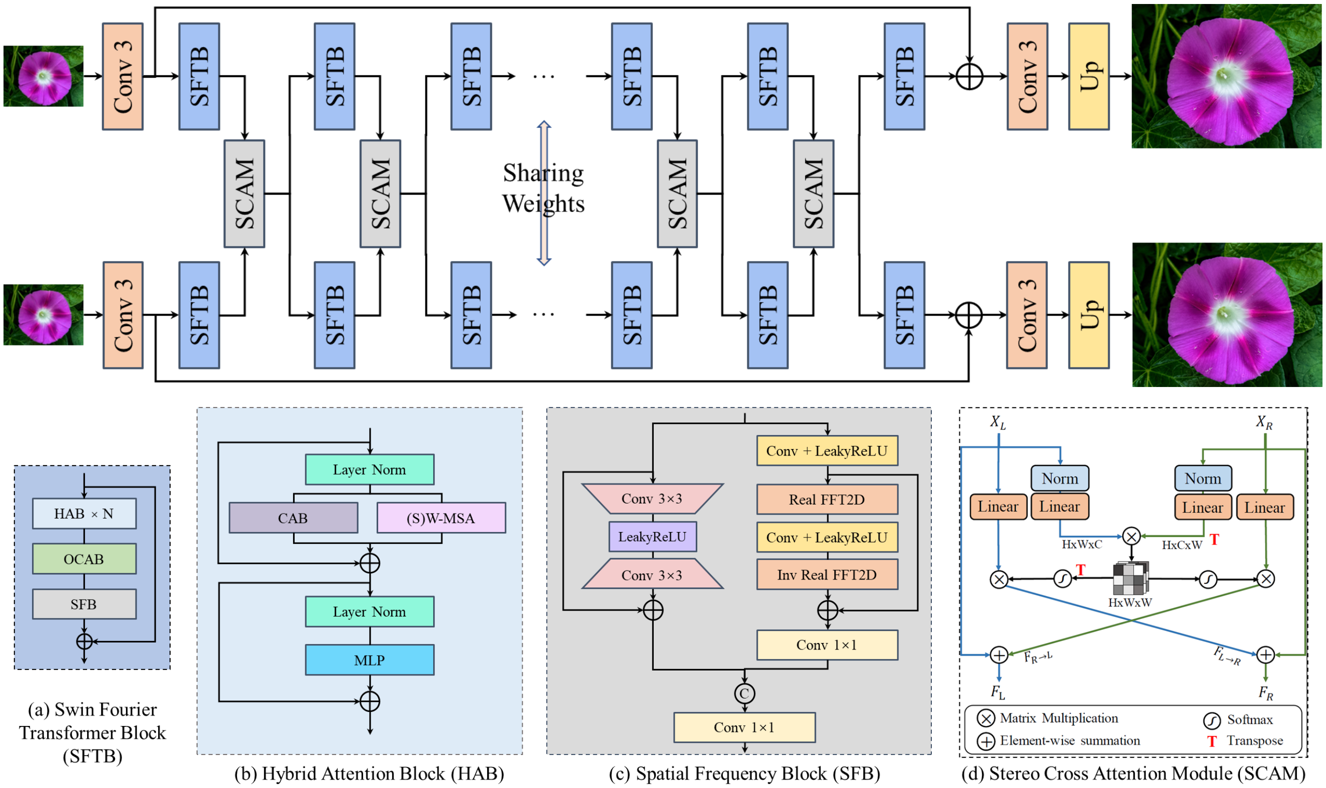
Data Ensemble. Due to the limitation of the number of parameters, model ensemble strategy was abandoned. Besides, due to the limitation of the computational complexity, only horizontal flipping and vertical flipping were used to generate three images for data ensemble.

5.4. BUPTMM - Track 2★

This team developed a Cross-View Hierarchy Network for Stereo Image Super-Resolution (CVHSSR) [63] by leveraging the complementary information between different viewpoints (Fig. 5). CVHSSR consits of two modules: the Cross-Hierarchy Information Mining Block (CHIMB) and the Cross-View Interaction Module (CVIM). CHIMB is developed to simulate and recover intra-view information across different levels, employing large-kernel convolutional attention and channel attention mechanisms. Meanwhile, CVIM utilizes a cross-view attention mechanism to effectively consolidate similar information from different views. These two modules facilitate CVHSSR to better aggregate cross-view information for higher performance.

In order to improve the SR performance of the model more efficiently with a limited number of parameters and MACs, the number of channels and the number of the CHIMB and CVIM modules are tuned. The final number of parameters is 999.54 K.

Training Settings. The training of the proposed model consists three phases. In the first phase, only 700 stereo image pairs were used as the training set. The proposed model was trained for 200K iterations with an MSE loss and a frequency Charbonnier loss. The Lion optimizer was employed in this phase. The batch size was set to 18 and the patch size was set to 64×64. The learning rate was initialized as 1×10^{-4} and updated using a cosine annealing strategy. The minimum learning rate was set to 1×10^{-8}. Data augmentation was performed through RGB channel shuffing and horizontal/vertical flipping. In the second phase, the resultant model was used for initialization and all 800 stereo image pairs were included for training. The model was further trained for 100K iterations, with the same settings as the first phase. In the third phase, the resultant model is fur-



(a) Swin Fourier Transformer Block (SFTB)

(b) Hybrid Attention Block (HAB)

(c) Spatial Frequency Block (SFB)

(d) Stereo Cross Attention Module (SCAM)

Figure 2. Davinci: The structure of the proposed SwinFIRSSR.

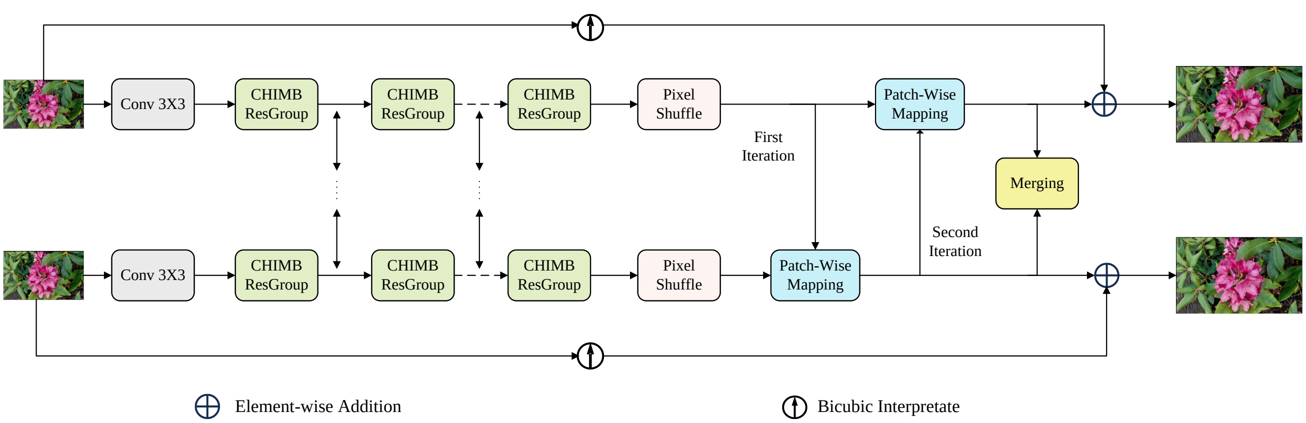


Figure 3. HiSSR: The structure of the proposed RISSR.

ther trained for 100K iterations. The batch size was set to 16 and the learning rate was updated to 2×10^{-6} for further training.

5.5. CV_IITRPR - Track 1, 2

This team proposed a stereo image SR network with a two-branch structure. As shown in Fig. 6, the proposed network consists of three stages, including initial feature extraction, cross-view feature merging, and reconstruction. Within the initial feature extraction module, attentive transition blocks are developed for better exploitation of features at different channels. After that, a bidirectional parallax attention module [47] is employed to aggregate features from both views. Finally, intra-view and inter-view features are collected to produce the super-resolved results.

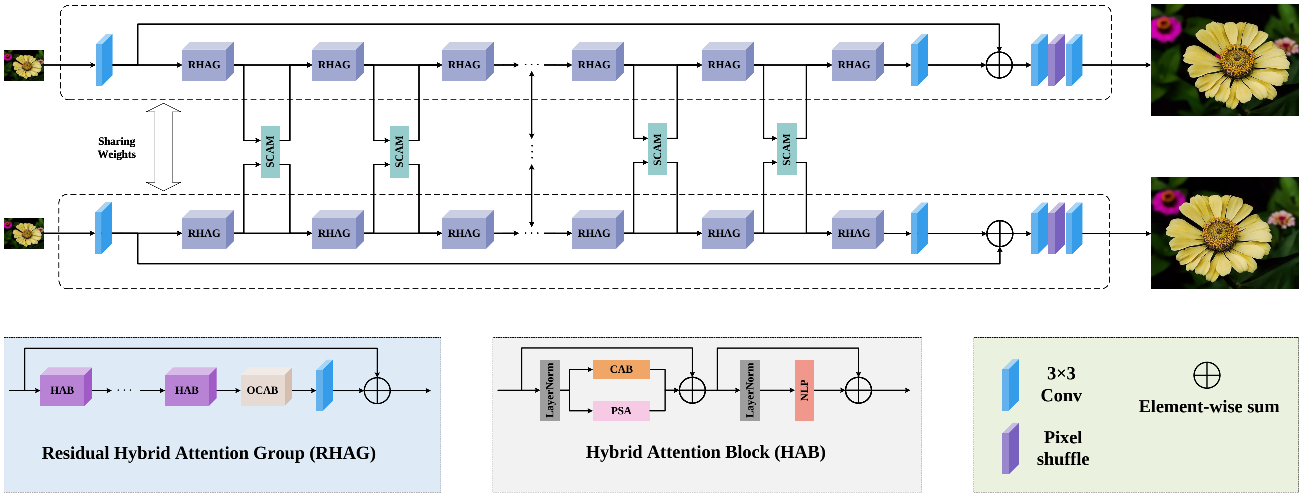


Figure 4. MiVideoSR: The structure of the proposed HCASSR.

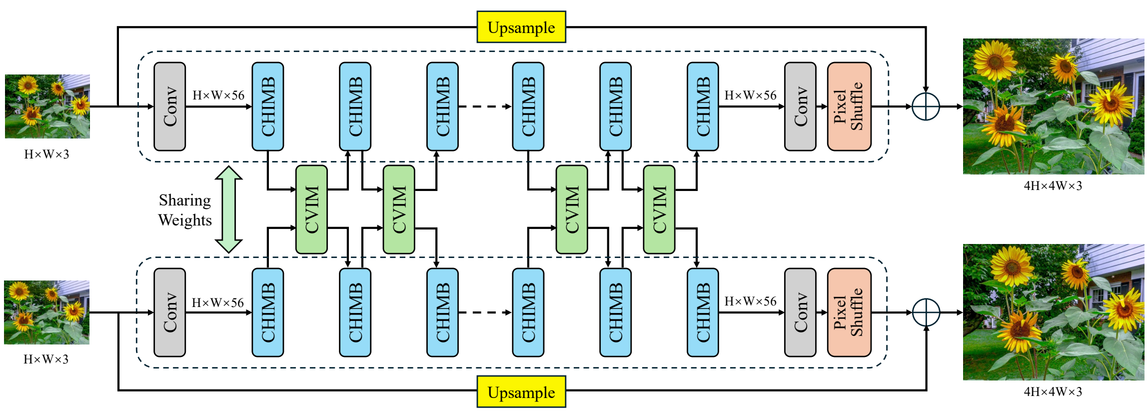


Figure 5. BUPTMM: The structure of the proposed Efficient CVHSSR.

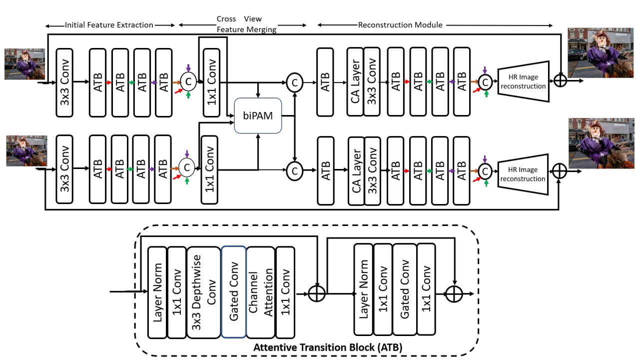


Figure 6. CV_IITRPR: The structure of the proposed network.

5.6. webbzhou - Track 1, 2

This team proposed a Stereo Omnidirectional Aggregation Network (SOAN) to efficiently super-resolve the stereo images, as illustrated in Fig. 7. SOAN consists of three parts: shallow feature extraction, deep feature extraction, and image reconstruction modules. The shallow feature extraction module comprises a 3×3 convolution, while the image reconstruction module consists of convolutional layers and pixel shuffle layers. The deep feature extraction part consists of MOmni-Scale Stereo Aggregation Groups (OSSAG), with each OSSAG comprising LCB [66], Meso-OSA [66], Global-OSA [66], SCATM [67], DWSCGLAM, and ESA [68]. Specifically, LCB, Meso-OSA, and Global-OSA are used to aggregate local information at different scales. In addition, SCATM [67] is adopted to aggregate features from two views.

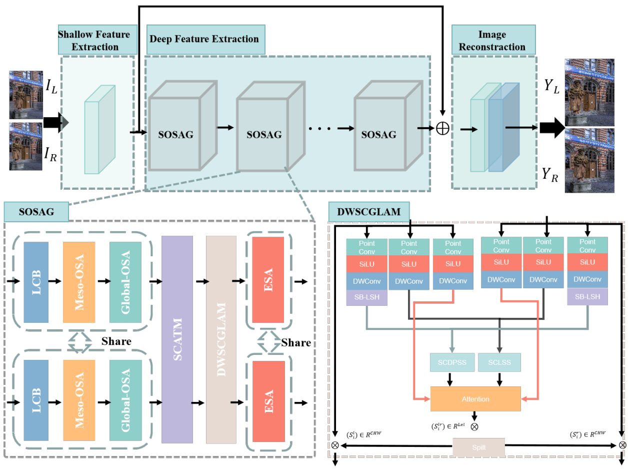


Figure 7. webbzhou: The structure of the proposed SOAN model.

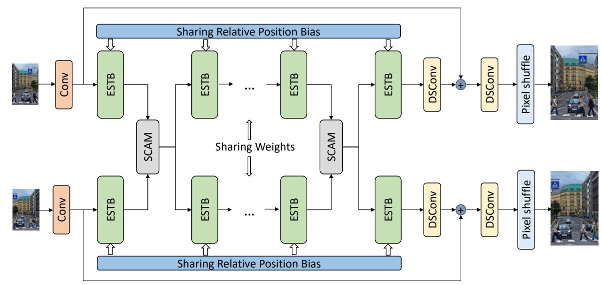


Figure 8. Qi5: The structure of the proposed efficient Swin Transformer network.

Track 1 Training Settings. Considering that rotational augmentation would disrupt the epipolar constraints, it cannot be employed for data augmentation. To address this issue and enhance the robustness of the model, this team divided the training into two stages:

- At the first stage, the proposed model was trained on the Flickr1024 dataset for single-image SR. The SCATM and DWSCGLAM modules are excluded at this stage. Random rotation and flipping are included for data augmentation. L1 loss was utilized for training with a batch size of 32. Cosine annealing schedule with 500,000 iterations was employed to update the learning rate.

- At the second stage, the resultant model was further fine-tuned for stereo image SR. The SCATM and DWSCGLAM modules were recovered and the batch size was set to 64. RGB channel shuffling was introduced for data augmentation. Cosine annealing schedule with 500,000 iterations was employed to update the learning rate. Mean squared error (MSE) loss is used for training.

Track 2 Training Settings. For track 2, the model trained in track 1 was used for initialization and then further fined-tuned on the training set of track 2.

5.7. Qi5 - Track 1

This team proposed an efficient Swin Transformer network for stereo image SR. First, they replaced all the vanilla convolutions in the deep feature extraction and image reconstruction stages with depth-wise convolutions. Second, they shared the relative position encoding parameters across all window attention modules, which allows the network to adopt larger window sizes while reducing parameters. Third, they removed the masking mechanism in shifted window multi-head self-attention (SW-MSA). In addition, the stereo cross-attention modules (SCAM) were employed to aggregate information from both views. The architecture is shown as Fig. 8

Training Settings. During the training phase, the model was first trained with an L1 loss and then fine-tuned with an L2 loss. The Adam optimizer was utilized for optimization. The batch size was 16 and the patch size was 64×64. The learning rate was initiated as 2×10^{-4} and halved at 150K, 250K, 350K, 450K, and 500K iterations.

5.8. WITAILab - Track 1

This team proposed a novel method named cross-view aggregation network for stereo image super-resolution(CANSSR) which exploits multi-directional parallax attention to capture both horizontal and vertical stereo correspondences while enhancing long-range dependence. Specifically, they proposed a multi-directional cross-view aggregation module(MCAM, as shown in Fig. 11) to aggregate the horizontal and vertical stereo correspondences to obtain more reliable complementary information from cross-view. Furthermore, to effectively aggregate the high-frequency detailed information for intra-view, we propose a channel-spatial aggregation module(CSAM, as shown in Fig. 10(a)) to capture long-range dependencies. Finally, we introduce a large-kernel convolution feed-forward network(LGFN, as shown in Fig. 10(b)) to aggregate richer spatial texture information, and a non-linear free activation function is introduced to enhance the non-linear representation. The overall architecture is shown as Fig.9.

Training Settings. All the models were optimized using AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was initialized as 1×10^{-3} and decayed to 1×10^{-7} using the cosine annealing strategy. Batch size was set to 16. The proposed model was trained for 400,000 iterations using an MSE loss on two NVIDIA RTX 4090 GPUs.

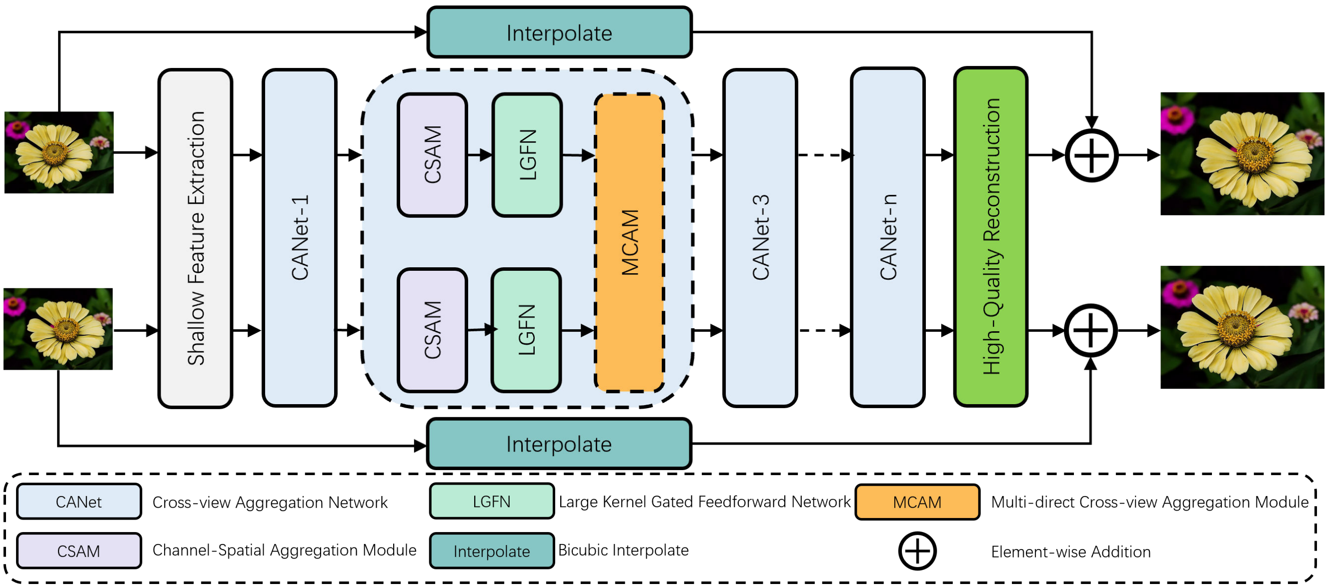


Figure 9. WITAILab: The structure of the proposed CANSSR.

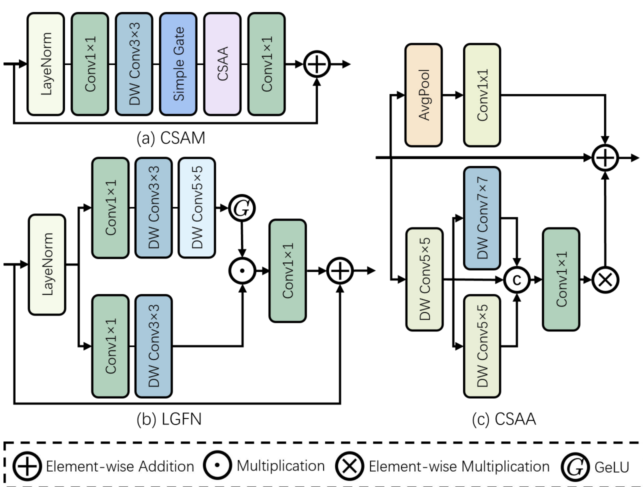


Figure 10. WITAILab: The structure of the proposed modules. (a) Channel-Spatial Aggregation Module (b) Large Kernel Gated Feedforward Network (c) Channel-Spatial Aggregation Attention

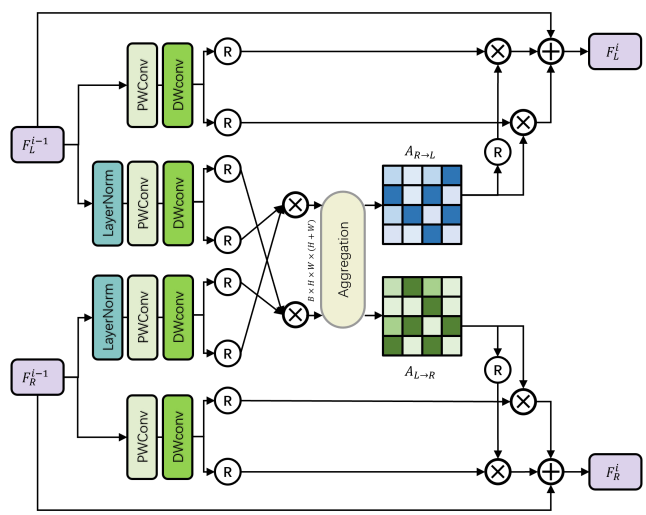


Figure 11. WITAILab: The structure of the proposed multi-direction cross-view aggregation module.

5.9. Giantpandacv - Track 1, 2

This team proposed an efficient multi-level information extraction network for stereo image SR (MIESSR), as shown in Fig. 12. Specifically, MIESSR consists of mixed attention feature extraction blocks (MAFEB, as shown in Fig. 13) and correlation matching information modules (CMIM, as shown in Fig. 14). MAFEBlocks aims to extract multi-scale view features and efficiently interact with CMIMs. CMIMs are mainly used to extract cross-view differential features and fully utilize the complementary information of stereo images.

Mixed Attention Feature Extraction Block. The MAFEB consists of two parts: (1) The multi-level information extractor (MIE), and (2) The simplified information refinement feedforward network (SIRFFN). In addition to introducing channel attention and large kernel convolution attention, MIE employs reparameterized convolution to further improve the model capacity and flexibility. Besides, SIRFFN is adopted to reduce structural redundancy.

Cross-View Matching Module. Parallax attention has been widely used in previous stereo image SR works to aggregate information from both left and right views. Nevertheless, parallax calculation introduces considerable com-

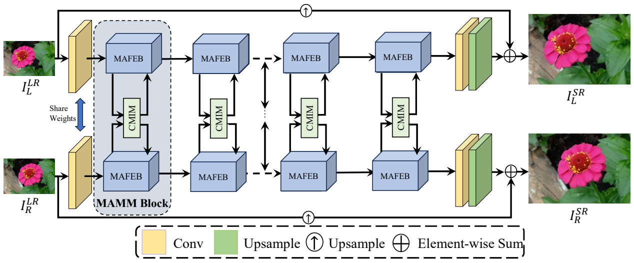


Figure 12. Giantpandacv: An overview of the efficient multi-level information extraction network for stereo image super-resolution (MIESSR).

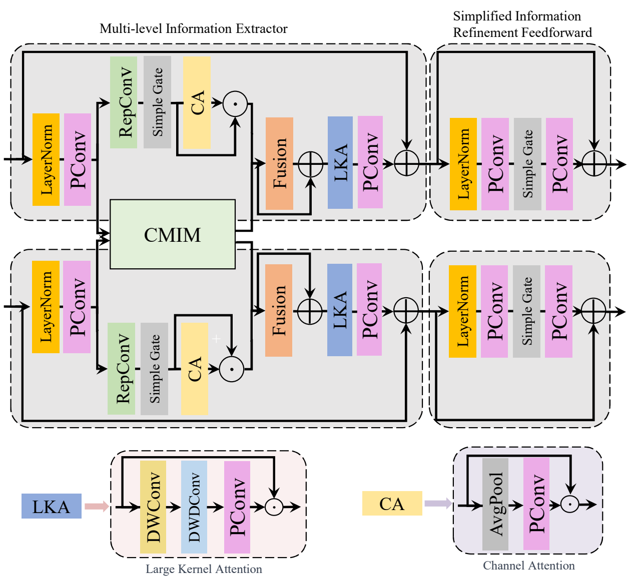


Figure 13. Giantpandacv: The structure of the proposed mixed attention feature extraction block (MAFEB).

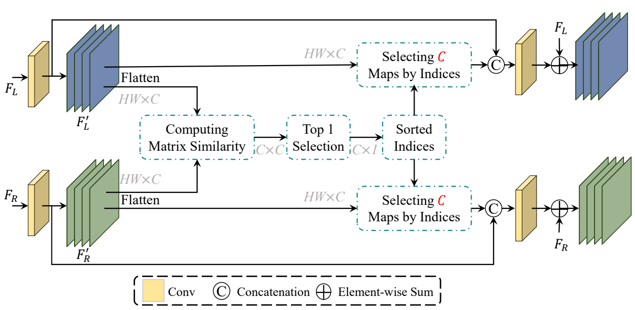


Figure 14. Giantpandacv: The structure of the proposed cross-view matching information modules (CMIM).

putational cost. In addition, erroneous textures and artifacts can be introduced due to inaccurate correspondence. To remedy this, this team developed a cross-view matching information module (CMIM) for cross-view information interaction, as shown in Fig. 14. Specifically, CMIM first computes the similarity along the channel dimension between left and right view feature F_L, F_R to generate similarity matrix $M \in \mathcal{R}^{C\times C}$. Then, features with the highest similarities $F_L^{Select}, F_R^{Select}$ were selected from F_L, F_R and concatenated with F_L, F_R for subsequent feature aggregation. Note that, the channel and block number of the NAFBlock were set to 56 and 24 in this method, which meets the requirements of the model size and computational complexity.

Training Settings. The numbers of MAFEB blocks and feature channels are set to 64 and 20, respectively. The whole network was trained using an MSE loss and a Frequency Charbonnier loss, and optimized using the Lion method [69] with β_1=0.9, β_2=0.999, and a batch size of 64. The proposed CVHSSR was implemented using PyTorch on a PC with four NVidia A100 GPUs. The learning rate was initially set to 5×10^{-4} and decayed using a cosine annealing strategy. The proposed model was trained for 200,000 iterations.

5.10. JNU_620 - Track 1, 2

This team contructed their solution based on NAFSSR [52], as shown in Fig. 15. For different tracks, different losses were proposed for training.

Track 1 Training Settings. A back-projection (BP) loss [70] was introduced for optimization, which enforces that the downscaled super-resolved images match the original LR observations:

$$L_{BP} = \left|\left|S(I_L^{SR}, s) - I_L^{LR}\right|\right|_1 + \left|\left|S(I_R^{SR}, s) - I_R^{LR}\right|\right|_1, \tag{4}$$

where S denotes the bicubic downsampling operation and s represents the downscale factor (*i,e.*, 4). The overall loss function is formulated as:

$$L_{total} = L_1 + \lambda L_{BP}, \tag{5}$$

where λ is set to 0.1.

Track 2 Training Settings. Since the realistic degradation is more complex than the bicubic degradation, only an L1 loss was used for training. During the training phase, the network was trained for 4× SR with a batch size of 24 and a patch size of 30 × 90. The AdamW method was employed for optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The initial learning rate was set to 3×10^{-3}, and decreased to 1×10^{-7} with a cosine annealing strategy. Random horizontal/vertical flipping and RGB channel shuffling were adopted for data augmentation.

5.11. GoodGame - Track 1, 2

This team used the CVHSSR [63] model (Figure 16) as the baseline. The proposed method pays special attention to the complementary information between left and

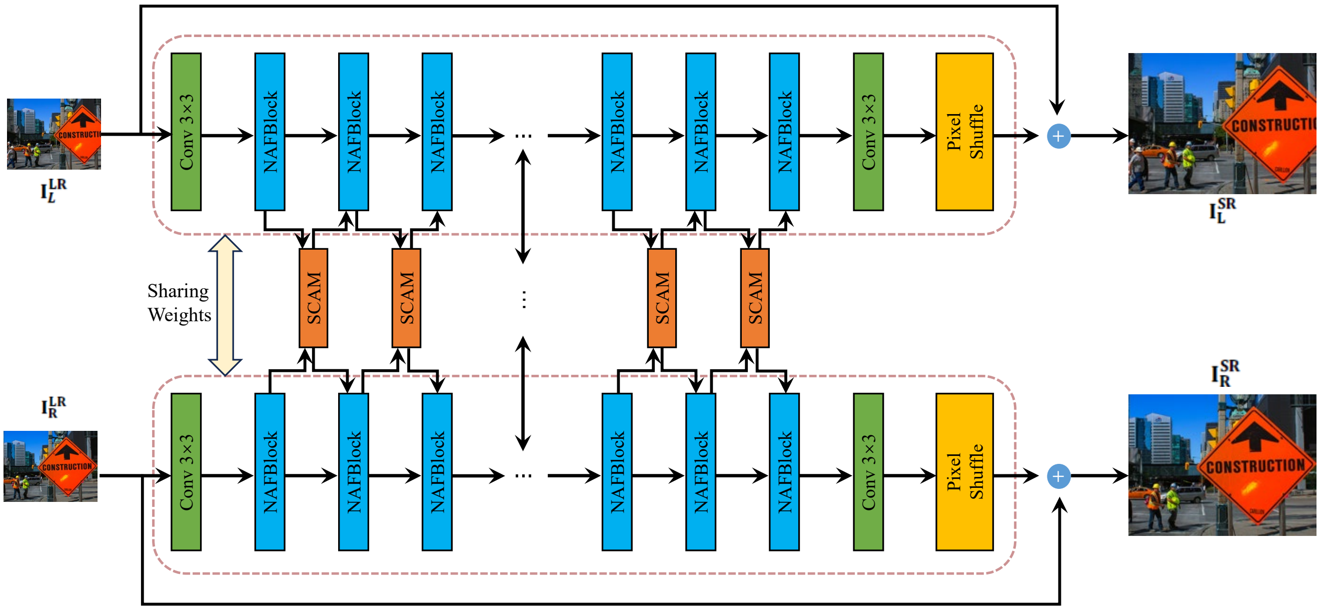


Figure 15. JNU_620: The structure of the proposed modified NAFSSR.

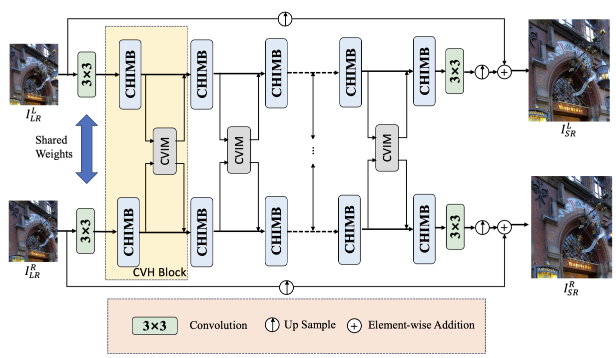


Figure 16. GoodGame: The structure of the proposed CVHSSRPlus model. CVIM and CHIMB in addition to the Information Refinement Feedforward module is the same as CVHSSR [63]

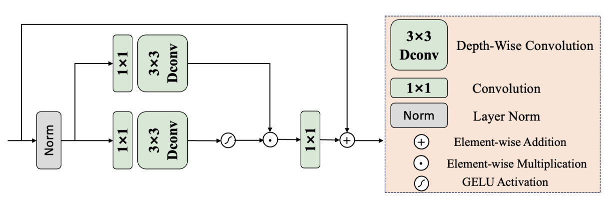


Figure 17. GoodGame: The structure of the proposed module that is used to replace Information Refinement Feedforward for feature extraction.

right views in stereo image pair, as well as the importance of the information within the respective views. In CVHSSR, CHIMB uses channel attention and large-kernel convolutions to extract global and local features within each view. CVIM aims to fuse features from different views through a cross-view attention mechanism. To further boost CVHSSR, the Information Refinement Feedforward module in CHIMB was replaced with the structure shown in Fig. 17.

Training Settings. The number of CVH Blocks was set to 17, and the total parameter amount of the model was approximately 985K. During the training phase, images were cropped to 128 $\times$ 128 patches with horizontal/vertical flipping being employed for data augmentation.

5.12. Fly_Flag - Track 1, 2

This team developed a model termed combined substitution self-attention and fast fourier convolution for stereo image super-resolution (CPFSSR). The framework of the proposed CPFSSR is shown in Figure 18. Specifically, they constructed permuted Swin Fourier Transformer block (PSFTB, as shown in Figure 18 (b)) based on permutation self-attention (PSA, as shown in Figure 18 (d)) layer, better captured global features and local features with limited window size size by cascading Permuted Self-Attention Blocks (PSABs, as shown in Figure 18 (c)), and built a spatial frequency reinforcement block (SFRB, as shown in Figure 18 (e)) based on the fast Fourier convolution block to improve the extraction of frequency domain information. In addition, they designed a deep cross-attention module (DCAM) to model the depth of hierarchical information before interaction, which facilitates the realization of information interaction between different views.

Training Settings. A pixel loss and a frequency loss are

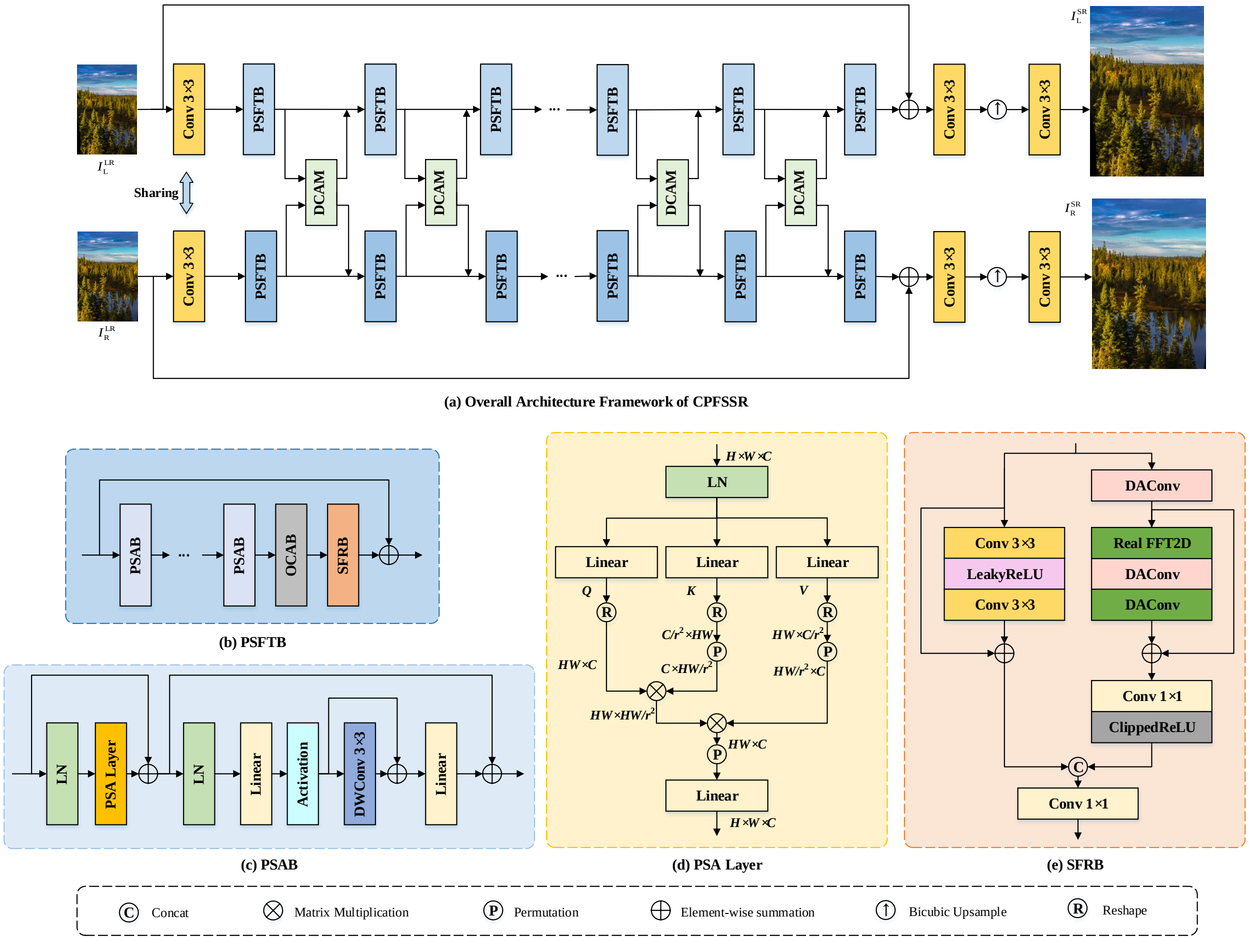


Figure 18. Fly_Flag: The structure of the proposed CPFSSR.

used for training:

$$\mathcal{L} = \mathcal{L}_{\text{Charbonnier}} + \lambda \mathcal{L}_{\text{FFT}}, \tag{6}$$

$$\mathcal{L}_{\text{Charbonnier}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left\| I_{\text{L,R}}^{\text{HR}} - I_{\text{L,R}}^{\text{SR}} \right\|^2 + \varepsilon_1^2}, \tag{7}$$

$$\mathcal{L}_{\text{FFT}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left\| \text{FFT}\left(I_{\text{L,R}}^{\text{HR}}\right) - \text{FFT}\left(I_{\text{L,R}}^{\text{SR}}\right) \right\|^2 + \varepsilon_2^2}, \tag{8}$$

where ε_1 is set to 10^{-3}, $I_{\text{L,R}}^{\text{SR}}$ represents the left and right SR image, and $I_{\text{L,R}}^{\text{HR}}$ represents the corresponding HR image. ε_2 is set to 10^{-3} and $\text{FFT}(\cdot)$ represents the Fast Fourier Transform. λ is a hyperparameter that controls the frequency Charbonnier loss function and set to 0.1.

During training, the Adam method with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ was employed for optimization. The learning rate was initialized to 2×10^{-4} and reduced to 1×10^{-7} using the cosine annealing strategy. Random cropping, random vertical and horizontal flipping, random horizontal shifting, and random RGB channel shuffling are adopted for data augmentation.

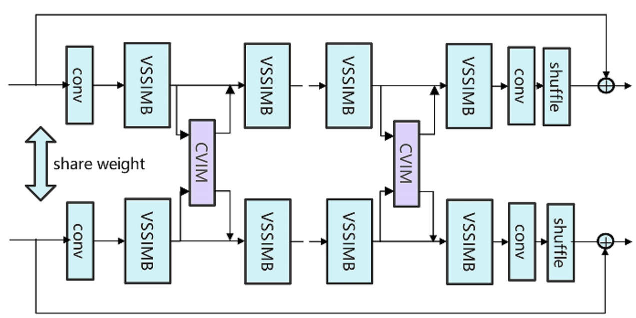


Figure 19. Mishka: The structure of the proposed VSSSR.

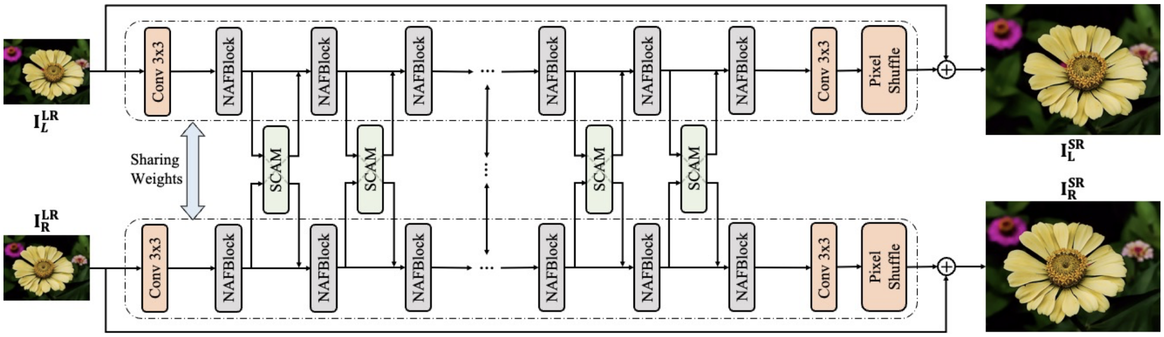


Figure 22. DVision: The structure of the proposed modified NAFSSR. This team modifies NAFSSR by using ODConv layers and knowledge distillation.

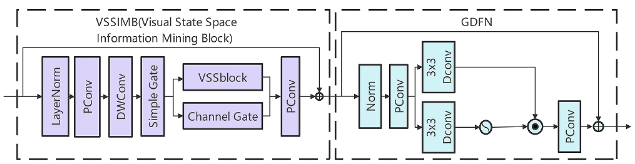


Figure 20. Mishka: The structure of the visual state Space Information Mining Block (VSSIMB).

5.13. Mishka - Track 1, 2

This team developed a VSSR network based on CVHSSR [63], as illustrated in Fig. 19). Specifically, the VSSBlock from VMamba [71] was adapted to extract image features in the spatial dimension. Concurrently, the ChannelGate [72] module is utilized to obtain image feature in the channel dimension, forming the VSSIMB (Fig. 20). Additionally, we incorporated the GDFN (Gated-Dconv Feed-Forward Network) module from Restormer [42] to refine the features extracted by VSSIMB.

Training Settings. For track 1, the proposed network was trained using an MSE loss function and a frequency Charbonnier loss. The model was optimized using the Lion method with $\beta_1 = 0.9$, $\beta_2 = 0.9$, and a batch size of 4 per GPU. The initial learning rate was set to 1×10^{-4} and updated using a cosine decay strategy. The model was first trained for 200,000 iterations and then fine-tuned for an additional 200,000 iterations using an MSE loss with the learning rate being set to 1×10^{-5}. Additionally, stochastic depth strategy was employed for network regularization, with the drop path rate being set to 0.1. For track 2, the model obtained from track 1 was used for initialization. Then, the model was further fine-tuned using an MSE loss with the same settings. During the training phase, horizontal flipping, vertical flipping, and RGB flipping were adopted for data augmentation.

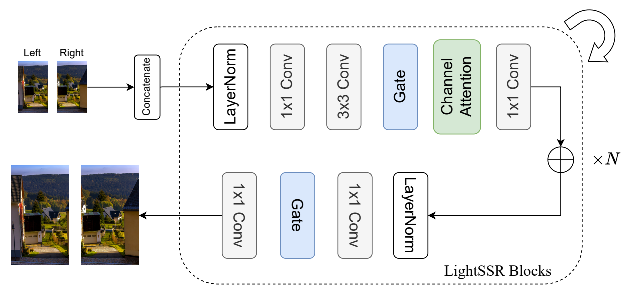


Figure 21. LightSSR: The structure of the proposed LightSSR block.

5.14. LightSSR - Track 1

Drawing inspiration from [54], this team introduced an efficient and lightweight stereo super-resolution method named LightSSR. The complexity of a stereo image SR network can be decomposed into inter-block computational complexity and intra-block computational complexity. To reduce inter-block computational complexity, they employed a single-stage UNet architecture. To achieve savings in terms of intra-block computational complexity, they simplify the baseline from HINet [73] by removing the connections between Gaussian Error Linear Units and Channel Attention to Gated Linear Unit. Figure 21 illustrates the basic blocks of our LightSSR model. The hidden channel of the proposed network was to 48, and the number of blocks was set to 32.

5.15. DVision - Track 1

Despite numerous stereo image SR methods have been developed, these methods suffer huge computational cost. To meet the constraint of this challenge, this team analyzed the computational complexity of existing state-of-the-art stereo image SR models and introduced a improved version NAFSSR in Fig. 22. Specifically, the vanilla convolutional

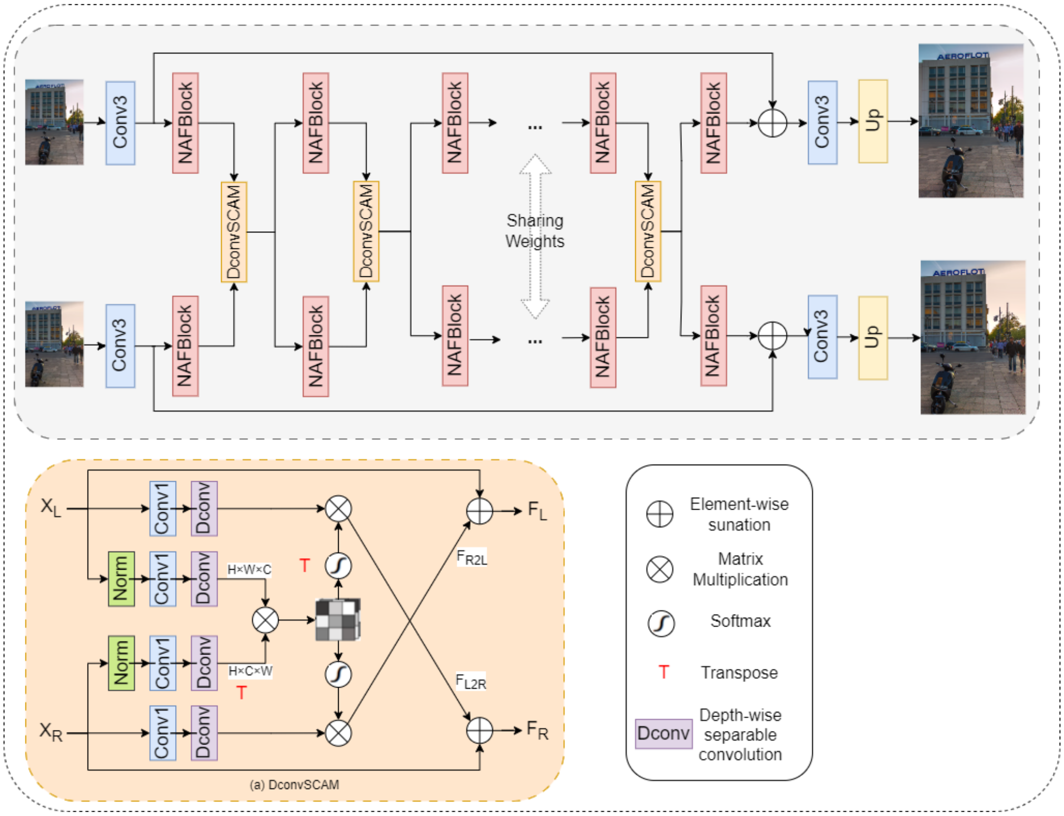


Figure 25. Liz620: The structure of the proposed DconvNAFSSR

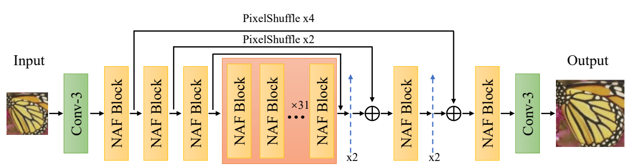


Figure 23. LVGroup_HFUT: The structure of the proposed network.

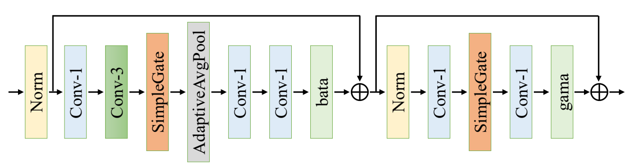


Figure 24. LVGroup_HFUT: The structure of the proposed NAF block.

layers in NAFBlock were updated to omni-dimensional dynamic convolutions to achieve improved performance.

5.16. LVGroup_HFUT - Track 1, 2

This team employed NAFNet as their network architecture. The pipeline is shown in Fig. 23 and the diagram is shown in Fig. 24. First, the input image is fed to a 3×3 convolution and cascaded residual feature blocks (RFB) for feature extraction. Then, bilinear interpolation upsampling as a mechanism to increase the spatial resolution of the feature maps. Concurrently, the upscaled features are passed through another convolutional layer for further refinement. Finally, these refined features are fed into a pixel shuffle layer to obtain the final results.

Training Settings: The proposed network was implemented on PyTorch 2.2.1 and an NVIDIA 4090 GPU. The proposed network was trained for 1500 epochs with a batch size of 32, using AdamW with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for optimization. The initial learning rate was set to 0.001. Random horizontal flip with probability 0.5 was used for data augmentation.

5.17. Liz620 - Track2

This team introduced a model named DconvNAFNet for stereo image SR. Specifically, the champion method (NAFSSR) [52] in the NTIRE 2022 Challenge was used as the baseline. To construct a lightweight and efficient network with improved performance, a feature selective unit (DconvSCAM) based on depth-wise convolution was designed to aggregate different levels of features, as shown in Fig. 25.

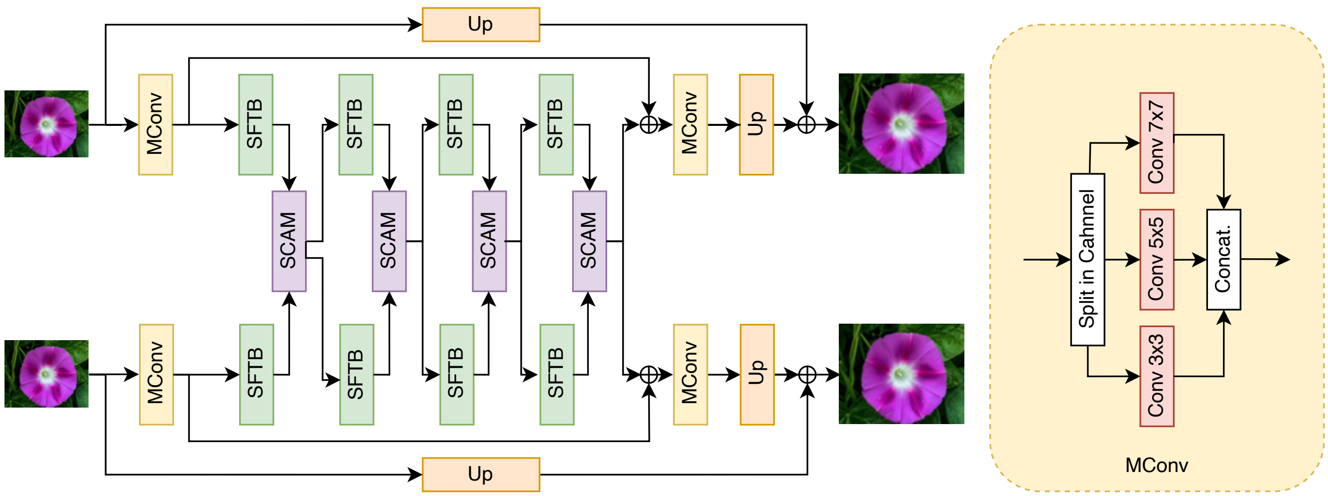


Figure 26. ECNU-IDEALab: The structure of the proposed network.

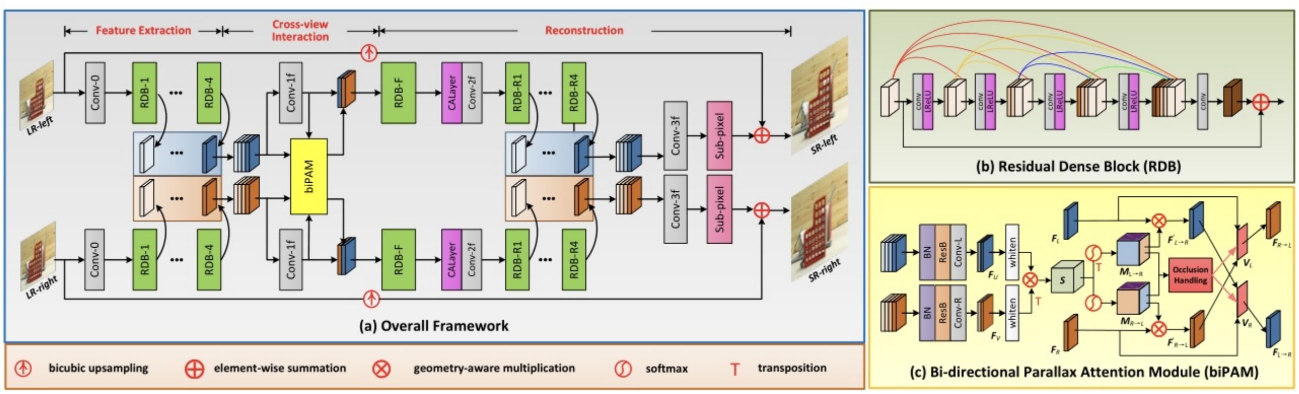


Figure 27. HiYun: The structure of the proposed iPASSR network.

Multiple dataset augmentation strategies were employed, including RGB channel shuffling, horizontal/vertical shifting. During the training process, the model was trained for 400k iterations on a NVIDIA 3090 GPU with a batch size of 16. The model was optimized using the Adam method with β_1=0.9 and β_2=0.9. The initial learning rate was set to 2×10^{-3}, and the cosine annealing scheduler was used.

5.18. ECNU-IDEALab - Track2

This team proposed a multi-convolution parallel stereo image SR method, as shown in Fig. 26. Based on SwinFIRSSR, the proposed MConv module is inserted to improve performance. Specifically, MConv divides the feature channels equally and then uses different convolution kernel sizes to extract multi-scale features. Finally, the features of different scales are concatenated along the channel dimension.

Training Settings: During training, the images were cropped into 32×32 patches with a stride of 20. The proposed model was trained for 800k iterations with a batch size of 64. The learning rate was set to 2×10^{-4} and halved at 600k, 650k, 700k, and 750k iterations. The CharbonnierLossColor loss and the Adam optimizer were employed for optimization.

5.19. HiYun - Track2

This team employed iPASSR as their network, which is illustrated in Fig. 27. The default settings of the original iPASSR were directly used for optimization.

6. Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 62301601, U20A20185, 62372491, and 62301306), the Shenzhen Science and Technology Program (No. RCYX20200714114641140), and the Guangdong Basic and Applied Basic Research Foundation (2022B1515020103). This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2024 sponsors: Meta Reality Labs, OPPO, KuaiShou, Huawei and University of Würzburg (Computer Vision Lab).

7. Organizers, Teams and Affiliations

• NTIRE 2024 Organizers

Challenge: NTIRE 2024 Challenge on Stereo Image Super-Resolution
Chairs: Longguang Wang[1], Yulan Guo[2,3], Yingqian Wang[3], Juncheng Li[4], Zhi Jin[2], Shuhang Gu[5], Radu Timofte[6]
Technical Support: Hongda Liu[2], Yang Zhao[4]
Affiliations:
[1]Aviation University of Air Force
[2]The Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University
[3]National University of Defense Technology
[4]Shanghai University
[5]University of Electronic Science and Technology of China
[6]Computer Vision Lab, University of Würzburg, Germany

• NTIRE 2024 Teams

(1) Davinci - Track 1★, Track 2★

Title: SwinFIRSSR: Stereo Image SR using SwinFIR
Members: Davinci (*davinci7571@gmail.com*), Saining Zhang[1]
Affiliations:
[1]Beijing Institute of Technology

(2) HiSSR - Track 1★

Title: RIISSR: Reference-based Iterative Interaction for Stereo Image SR
Members: Rongxin Liao[1] (*2023110526@mail.hfut.edu.cn*), Ronghui Sheng[2], Feng Li[1], Huihui Bai[2], Wei Zhang[3], Runmin Cong[3]
Affiliations:
[1]Hefei University of Technology
[2]Beijing Jiaotong University
[3]Shandong University

(3) MiVideoSR - Track 1★, Track 2★

Title: HCASSR
Members: Yuqiang Yang[1] (*yangyuqiang@xiaomi.com*), Zhiming Zhang[1], Jingjing Yang[1], Long Bao[1], Heng Sun[1]
Affiliations:
[1]Xiaomi Inc.

(4) BUPTMM - Track 2★

Title: Efficient CVHSSR
Members: Kanglun Zhao[1] (*kanglunzhao@bupt.edu.cn*), Enyuan Zhang[1], Huiyuan Fu[1], Huadong Ma[1]
Affiliations:
[1]Beijing University of Posts and Telecommunications

(5) webbzhou - Track 1, Track 2

Title: SOAN: Stereo Omnidirectional Aggregation Networks for Lightweight Stereo Image Super-Resolution
Members: Yuanbo Zhou[1] (*webbozhou@gmail.com*), Wei Deng[2], Xintao Qiu, Tao Wang, Qinquan Gao, Tong Tong
Affiliations:
[1]Fuzhou University
[2]Imperial Vision Technology

(6) Qi5 - Track 1

Title: ESwinSSR
Members: Yinghao Zhu[1] (*zhuyinghao0427@gmail.com*), Yongpeng Li[1]
Affiliations:
[1]IQIYI

(7) WITAILab - Track 1

Title: CANSSR: Cross-view Aggregation Network for Stereo Image Super-resolution
Members: Zhitao Chen[1] (*2290445045czt@gmail.com*), Xiujuan Lang[1], Kanghui Zhao[1], Bolin Zhu[1]
Affiliations:
[1]Wuhan Institute of Technology

(8) Giantpandacv - Track 1, Track 2

Title: MIESSR: Efficient Multi-Level Information Extraction Network for Stereo Image Super-Resolution.
Members: Wenbin Zou[1] (*alexzou14@foxmail.com*), Yunxiang Li[2], Qiaomu Wei[3], Tian Ye[4], Sixiang Chen[4]
Affiliations:
[1]South China University of Technology
[2]Fuzhou University
[3]Chengdu University of Information Technology
[4]Hong Kong University of Science and Technology (Guangzhou)

(9) JNU_620 - Track 1, Track 2

Title: Improved Loss for Stereo Image Super-Resolution Based on NAFSSR
Members: Weijun Yuan[1] (*yweijun@stu2022.jnu.edu.cn*), Zhan Li[1], Wenqin Kuang[1], Ruijin Guan[1]
Affiliations:
[1]Jinan University

(10) GoodGame - Track 1, Track 2

Title: CVHSSRPlus
Members: Jian Wang[1] (*jwang4@snapchat.com*), Yuqi Miao[2], Baiang Li[3], Kejie Zhao[4]

Affiliations:
[1]Snap Inc.
[2]Tongji University
[3]Hefei University of Technology
[4]Southern University of Science and Technology

(11) Fly_Flag - Track 1, Track 2

Title: CPFSSR: Combined Permuted Self-Attention and Fast Fourier Convolution for Stereo Image Super-Resolution
Members: Wenwu Luo[1] (*210220105@fzu.edu.cn*), Jing Wu[1]
Affiliations:
[1]Fuzhou University

(12) Mishka - Track 1, Track 2

Title: VSSSR
Members: Yunkai Zhang[1] (*1585832651@qq.com*), Songyan Zhang[2], Jingyi Zhang[1], Junyao Gao[3], Xueqiang You[4]
Affiliations:
[1]Hefei University of Technology
[2]Nanyang Technological University
[3]Tongji University
[4]Zhuhai Zhongke HZ Technology Co.,Ltd

(13) LightSSR - Track 1

Title: LightSSR
Members: Yanhui Guo[1] (*guoy143@mcmaster.ca*), Hao Xu[1]
Affiliations:
[1]McMaster University

(14) DVision - Track 1

Title: NAFSSR-KD+ODConv
Members: Sahaj K. Mistry[1] (*sahajmistry005@gmail.com*), Aryan Shukla[1], Sourav Saini[1], Aashray Gupta[1], Vinit Jakhetiya[1], Sunil Jaiswal[2].
Affiliations:
[1]Indian Institute of Technology Jammu
[2]K | Lens GmbH

(15) LVGroup_HFUT - Track 1, Track 2

Title: Light-NAFNet
Members: Zhao Zhang[1](*cszzhang@gmail.com*), Bo Wang[1], Suiyi Zhao[1], Yan Luo[1], Yanyan Wei[1]
Affiliations:
[1]Hefei University of Technology.

(16) Liz620 - Track 2

Title: DconvNAFNet
Members: Yihang Chen[1](*Ehang@stu.jnu.edu.cn*), Ruting Deng[1], Yifan Deng[1]
Affiliations:
[1]Jinan University.

(17) ECNU-IDEALab - Track 2

Title: MSFIRSSR
Members: Jingchao Wang[1](*jcwang@stu.ecnu.edu.cn*), Zhijian Wu[1], Dingjiang Huang[1]
Affiliations:
[1]East China Normal University.

(18) HiYun - Track 2

Title: iPASSR
Members: Yun Ye[1](*ye.721@buckeyemail.osu.edu*)
Affiliations:
[1]The Ohio State University.

References

[1] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Liangyu Chen, Xiaojie Chu, Wenqing Yu, Kai Jin, et al. NTIRE 2022 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, pages 906–919, 2022. 1

[2] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, et al. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *CVPR*, pages 1346–1372, 2023. 1

[3] Cosmin Ancuti, Codruta O Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[4] Nikola Banić, Egor Ershov, Artyom Panshin, Oleg Karasev, Sergey Korchagin, Shepelev Lev, Alexandr Startsev, Daniil Vladimirov, Ekaterina Zaychenkova, Dmitrii R Iarchuk, Maria Efimova, Radu Timofte, Arseniy Terekhin, et al. NTIRE 2024 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[5] Ren Yang, Radu Timofte, et al. NTIRE 2024 challenge on blind enhancement of compressed image: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[6] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei WU, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[7] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth NTIRE 2024 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[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 (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[9] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2024 challenge on light field image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[10] Longguang Wang, Yulan Guo, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, Radu Timofte, et al. NTIRE 2024 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[11] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2024 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[12] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. NTIRE 2024 challenge on bracketing image restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[13] Nicolas Chahine, Marcos V. Conde, Sira Ferradans, Radu Timofte, et al. Deep portrait quality assessment. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[14] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[15] Jie Liang, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Radu Timofte, Lei Zhang, et al. NTIRE 2024 restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[16] Marcos V. Conde, Florin-Alexandru Vasluianu, Radu Timofte, et al. Deep raw image super-resolution. a NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[17] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[18] Xiaoning Liu, Zongwei WU, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[19] Zeshuai Deng, Zhuokun Chen, Shuaicheng Niu, Thomas Li, Bohan Zhuang, and Mingkui Tan. Efficient test-time adaptation for super-resolution with second-order degradation and reconstruction. In *NeurIPS*, volume 36, 2024. 2

[20] Man Zhou, Keyu Yan, Jinshan Pan, Wenqi Ren, Qi Xie, and Xiangyong Cao. Memory-augmented deep unfolding network for guided image super-resolution. *International Journal of Computer Vision*, 131(1):215–242, 2023. 2

[21] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, pages 12312–12321, 2023. 2

[22] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *CVPR*, pages 10581–10590, 2021. 2

[23] Juncheng Li, Zehua Pei, Wenjie Li, Guangwei Gao, Longguang Wang, Yingqian Wang, and Tieyong Zeng. A systematic survey of deep learning-based single-image super-resolution. *ACM Computing Surveys*, 2024. 2

[24] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199, 2014. 2

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 2

[26] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pages 136–144, 2017. 2

[27] 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*, pages 1646–1654, 2018. 2

[28] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 2

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[30] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017. 2

[31] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 517–532, 2018. 2

[32] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 2

[33] Dehua Song, Yunhe Wang, Hanting Chen, Chang Xu, Chunjing Xu, and DaCheng Tao. Addersr: Towards energy efficient image super-resolution. In *CVPR*, pages 15648–15657, 2021. 2

[34] Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *ICCV*, pages 4801–4810, 2021. 2

[35] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*, pages 4917–4926, 2021. 2

[36] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, pages 723–731, 2018. 2

[37] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, pages 252–268, 2018. 2

[38] Xiangxiang Chu, Bo Zhang, Hailong Ma, Ruijun Xu, Jixiang Li, and Qingyuan Li. Fast, accurate and lightweight super-resolution with neural architecture search. In *ICPR*, pages 59–64, 2021. 2

[39] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCVW*, pages 1833–1844, 2021. 2

[40] 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*, pages 10012–10022, 2021. 2, 4

[41] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *CVPRW*, pages 457–466, 2022. 2

[42] 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*, pages 5728–5739, 2022. 2, 13

[43] Daniel S. Jeon, Seung-Hwan Baek, Inchang Choi, and Min H. Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *CVPR*, pages 1721–1730, 2018. 2

[44] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *CVPR*, pages 12250–12259, 2019. 2

[45] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2108–2125, 2022. 2

[46] Wonil Song, Sungil Choi, Somi Jeong, and Kwanghoon Sohn. Stereoscopic image super-resolution with stereo consistent feature. In *AAAI*, volume 34, pages 12031–12038, 2020. 2

[47] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *CVPRW*, pages 766–775, 2021. 2, 6

[48] Hansheng Guo, Juncheng Li, Guangwei Gao, Zhi Li, and Tieyong Zeng. Pft-ssr: Parallax fusion transformer for stereo image super-resolution. In *ICASSP*, pages 1–5, 2023. 2

[49] Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, and Steven CH Hoi. Disparity-aware domain adaptation in stereo image restoration. In *CVPR*, pages 13179–13187, 2020. 2

[50] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. In *ACM MM*, 2021. 2

[51] Jin Wan, Hui Yin, Zhihao Liu, Yanting Liu, and Song Wang. Multi-stage edge-guided stereo feature interaction network for stereoscopic image super-resolution. *IEEE Transactions on Broadcasting*, 2023. 2

[52] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPRW*, pages 1239–1248, 2022. 2, 4, 5, 10, 14

[53] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. volume 33, pages 4479–4488, 2020. 2, 4

[54] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33, 2022. 2, 13

[55] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 2

[56] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for sereo image super-resolution. In *ICCVW*, 2019. 2

[57] 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*, 2022. 4

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

[59] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4

[60] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, pages 5835–5843, 2017. 4

[61] Jaejun Yoo, Namhyuk Ahn, and Kyung-Ah Sohn. Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In *CVPR*, pages 8375–8384, 2020. 4

[62] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, pages 1865–1873, 2016. 4

[63] Wenbin Zou, Hongxia Gao, Liang Chen, Yunchen Zhang, Mingchao Jiang, Zhongxin Yu, and Ming Tan. Cross-view hierarchy network for stereo image super-resolution. In *CVPRW*, pages 1396–1405, 2023. 5, 10, 11, 13

[64] Gang Wu, Junjun Jiang, Junpeng Jiang, and Xianming Liu. Transforming image super-resolution: A convformer-based efficient approach. *arXiv*, 2024. 5

[65] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*, pages 12780–12791, 2023. 5

[66] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387, 2023. 7

[67] Yuanbo Zhou, Yuyang Xue, Wei Deng, Ruofeng Nie, Jiajun Zhang, Jiaqi Pu, Qinquan Gao, Junlin Lan, and Tong Tong. Stereo cross global learnable attention module for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1416–1425, 2023. 7

[68] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020. 7

[69] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2024. 10

[70] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-buddy gans for highly detailed image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1412–1420, 2022. 10

[71] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 13

[72] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 13

[73] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *CVPR*, pages 182–192, 2021. 13