NTIRE 2023 Challenge on Stereo Image Super-Resolution: Methods and Results

Longguang Wang*, Yulan Guo*†, Yingqian Wang*, Juncheng Li*, Shuhang Gu*, Radu Timofte*, Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, Shijie Zhao, Xuhan Sheng, Yukan Ding, Ming Sun, Xing Wen, Dafeng Zhang, Jia Li, Fan Wang, Zheng Xie, Zongyao He, Zidian Qiu, Zilin Pan, Zhihao Zhan, Xingyuan Xian, Zhi Jin, Yuanbo Zhou, Wei Deng, Ruofeng Nie, Jiajun Zhang, Qinquan Gao, Tong Tong, Kexin Zhang, Junpei Zhang, Rui Peng, Yanbiao Ma, Licheng Jiao, Haoran Bai, Lingshun Kong, Jinshan Pan, Jiangxin Dong, Jinhui Tang, Pu Cao, Tianrui Huang, Lu Yang, Qing Song, Bingxin Chen, Chunhua He, Meiyun Chen, Zijie Guo, Shaojuan Luo, Chengzhi Cao, Kunyu Wang, Fanrui Zhang, Qiang Zhang, Nancy Mehta, Subrahmanyam Murala, Akshay Dudhane, Yujin Wang, Lingen Li, Garas Gendy, Nabil Sabor, Jingchao Hou, Guanghui He, Junyang Chen, Hao Li, Yukai Shi, Zhijing Yang, Wenbin Zou, Yunchen Zhang, Mingchao Jiang, Zhongxin Yu, Ming Tan, Hongxia Gao, Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, Thomas B. Schön, Jingxiang Chen, Bo Yang, XiSheryl Zhang, Chenghua Li, Weijun Yuan, Zhan Li, Ruting Deng, Jintao Zeng, Pulkit Mahajan, Sahaj Mistry, Shreyas Chatterjee, Vinit Jakhetiya, Badri Subudhi, Sunil Jaiswal, Zhao Zhang, Huan Zheng, Suiyi Zhao, Yangcheng Gao, Yanyan Wei, Bo Wang, Gen Li, Aijin Li, Lei Sun, Ke Chen, Congling Tang, Yunzhe Li, Jun Chen, Yuan-Chun Chiang, Yi-Chung Chen, Zhi-Kai Huang, Hao-Hsiang Yang, I-Hsiang Chen, Sy-Yen Kuo, Yiheng Wang, Gang Zhu, Xingyi Yang, Songhua Liu, Yongcheng Jing, Xingyu Hu, Jianwen Song, Changming Sun, Arcot Sowmya, Seung Ho Park, Xiaoyan Lei, Jingchao Wang, Chenbo Zhai, Yufei Zhang, Weifeng Cao, Wenlong Zhang

Abstract

This paper summarizes the 2nd NTIRE challenge on stereo image super-resolution (SR) with a focus on new solutions and results. The task of the challenge is to super-resolve a low-resolution stereo image pair to a high-resolution one with a magnification factor of $\times 4$. 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 3 tracks, including one track on distortion (e.g., PSNR) and bicubic degradation, one track on perceptual quality (e.g., LPIPS) and bicubic degradation, as well as another track on real degradations. In total, 175, 93, and 103 participants were successfully registered for each track, respectively. In the test phase, 21, 17, and 12 teams successfully submitted results with PSNR (RGB) scores better than the baseline. This challenge establishes a new benchmark for stereo image SR.

1. Introduction

In recent years, dual cameras have been widely applied in AR/VR, mobile phones, autonomous vehicles and robots to record and perceive the 3D environment. In these applications, increasing the resolution of stereo images is highly demanded to achieve higher perceptual quality and finer-grained parsing of the real world. To this end, stereo image super-resolution (SR) has been introduced to reconstruct a high-resolution (HR) stereo image pair with finer details

*Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte are the NTIRE 2023 challenge organizers, while the other authors participated in this challenge.

†Corresponding author: Yulan Guo.

Section 7 provides the authors and affiliations of each team.

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

Challenge webpage (Track 1): https://codalab.lisn.upsaclay.fr/competitions/10047

Challenge webpage (Track 2): https://codalab.lisn.upsaclay.fr/competitions/10048

Challenge webpage (Track 3): https://codalab.lisn.upsaclay.fr/competitions/10049

Github: https://github.com/The-Learning-And-Vision-Atelier-LAVA/Stereo-Image-SR/tree/NTIRE2023

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 workshop [1]. This challenge, employed the Flick1024 dataset [2] and widely-applied bicubic degradation to synthesize LR stereo images. Besides, the objective is to minimize the distortion between super-resolved stereo images and the groundtruth. However, degradations in real-world scenarios are more complicated than the bicubic one. In addition, the perceptual quality and the stereo consistency are also critical to the visual effects of stereo images.

Succeeding the previous year, NTIRE 2023 Stereo Image SR Challenge presents three competition tracks. Track 1 is inherited from the NTIRE 2022 challenge, focusing on bicubic degradation and restoration accuracy in terms of PSNR. Track 2 also adopts bicubic degradation to synthesize LR images but focus on restoration accuracy in terms of LPIPS. In track 3, complicated degradations including blur, noise, downsampling, and compression are used for LR image synthesis, with PSNR being employed to measure restoration accuracy.

This challenge is one of the NTIRE 2023 Workshop series of challenges on: night photography rendering [3], HR depth from images of specular and transparent surfaces [4], image denoising [5], video colorization [6], shadow removal [7], quality assessment of video enhancement [8], stereo super-resolution [9], light field image super-resolution [10], image super-resolution ($\times 4$) [11], 360° omnidirectional image and video super-resolution [12], lens-to-lens bokeh effect transformation [13], real-time 4K super-resolution [14], HR nonhomogenous dehazing [15], efficient super-resolution [16].

2. Related Work

In this section, several major works on single image and stereo image SR are briefly reviewed.

2.1. Single Image SR

Single image SR is a challenging task and has received extensive attention. In the past 10 years, deep learning based single image SR methods have emerged and continuously refreshed the best results. Dong *et al.* [17] proposed the first CNN-based SR model (*i.e.*, SRCNN) to learn the mapping between LR and HR images, making the model can directly reconstruct HR images from LR inputs. Kim *et al.* [18] proposed a deeper network (*i.e.*, VDSR) to improve SR performance by using 20 layers and residual learning strategy. After that, CNN-based SR models are blooming and achieve promising results. For instance, Lim *et al.* [19] proposed an enhanced deep SR model (*i.e.*, EDSR) by using both local and residual connections. Zhang *et al.* [20] proposed a residual dense network (*i.e.*, RDN) to fully use hierarchical features by combining residual connection [21] with dense connection [22]. Subsequently, Zhang *et al.* [23] further proposed a residual channel attention network (*i.e.*, RCAN) with channel attention mechanism. Li *et al.* [24] 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.* [25] proposed a second-order attention network (*i.e.*, SAN) for more powerful feature correlation learning, which achieves superior performance.

Recently, Transformer has been widely used in computer vision and further promote the development of single image SR. Liang *et al.* [26] designed a SwinIR model for image restoration by applying Swin Transformer [27], which achieves state-of-the-art performance. Lu *et al.* [28] 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. After that, Transformer developed rapidly in image restoration, with a series of models being proposed. For example, Wang *et al.* [29] proposed a general Ushaped Transformer (*i.e.*, Uformer) for various image restoration tasks. Zamir *et al.* [30] proposed an encoder-decoder Transformer (*i.e.*, Restormer) for image restoration with multi-scale local-global representation learning. These models not only produce promising results on single image SR, but also perform well on other image restoration tasks. More detailed and advanced methods can refer to recent surveys [31–33].

2.2. Stereo Image SR

Motivated by the remarkable progress of deep learning techniques in single image SR, several learning-based stereo image SR methods have been developed. Jeon *et al.* [34] proposed a pioneering network (namely, StereoSR) to incorporate cross-view information by concatenating the left image and a stack of right images with different predefined shifts. Wang *et al.* [35, 36] introduced a parallax attention module (PAM) to model stereo correspondence with a global receptive field along the epipolar line. Song *et al.* [37] further developed a SPAMnet by combining self-attention with parallax attention. Yan *et al.* [38] proposed a domain adaptive stereo SR network (DASSR) to incorporate cross-view information through explicit disparity esti-

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

mation using a pre-trained stereo matching network.

More recently, Wang *et al.* [39] developed an improved version of PASSRnet (*i.e.*, iPASSR) to handle a series of practical issues (*e.g.*, illuminance variation and occlusions) in stereo image SR. Dai *et al.* [40] proposed a feedback network to alternately solve disparity estimation and stereo image SR in a recurrent manner. Ma *et al.* [41] introduced a GAN-based perception-oriented stereo image SR method that can generate visually pleasing and stereo consistent SR results. Guo *et al.* [42] proposed a new Transformer-based parallax fusion model called Parallax Fusion Transformer.

In the NTIRE 2022 Stereo Image SR Challenge, the champion team developed NAFSSR network [43] by using nonlinear activation-free network (NAFNet) for feature extraction and PAM for cross-view information interaction. The runner-up team proposed a SwiniPASSR network by combining the Swin Transformer with PAM. The second runner-up team also adopted a Transformer-based network termed SSRFormer with a Siamese structure.

3. NTIRE 2023 Challenge

The objectives of the NTIRE 2023 challenge on example-based stereo image SR are: (i) to gauge and push the state-of-the-art in SR; and (ii) to compare different solutions.

3.1. Dataset

Training Set. The training set of the Flickr1024 dataset [2] (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: Fidelity & Bicubic Degradation

Degradation Model. 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.

Evaluation Metrics. 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.

• Track 2: Perceptual & Bicubic Degradation

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

$$I^{LR} = I^{HR} \downarrow_s, \tag{2}$$

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

Evaluation Metrics. High-quality stereo image SR results should recover rich and clear details with high stereo consistency. Given a pair of stereo SR results (I^{SR}_{left} and I^{SR}_{right}), LPIPS [44] is used as the metric to evaluate the perceptual quality of separate images. To further evaluate the stereo consistency between an SR image pair, this challeng first use a state-of-the-art stereo matching method [45] to obtain a disparity map D^{HR} from an HR image pair as the groundtruth. Then, a disparity map D^{SR} is estimated from the SR image pair. Mean absolute error (MAE) between D^{SR} and D^{HR} is adopted as the metric to measure the stereo consistency. The final score is calculated as:

$$\begin{aligned} score =& 1 - 0.5 \times \mathcal{L}\left(I^{SR}_{left}, I^{HR}_{left}\right) - 0.5 \times \mathcal{L}\left(I^{SR}_{right}, I^{HR}_{right}\right) \\ & - 0.1 * \mathcal{S}\left(D^{SR}, D^{HR}\right), \end{aligned} \tag{3}$$

where $\mathcal{L}\left(I^{SR}_{left}, I^{HR}_{left}\right)$ represents the LPIPS score of I^{SR}_{left}, $\mathcal{S}\left(D^{SR}, D^{HR}\right)$ calculates normalized MAE between disparity maps D^{SR} and D^{HR}.

• Track 3: Fidelity & Realistic Degradation

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

$$I^{LR} = \mathcal{C}\left(\left(I^{HR} \otimes k\right) \downarrow_s + n\right), \tag{4}$$

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

Evaluation Metrics. PSNR and 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.

Table 1. NTIRE 2023 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)	PSNR (Y)	SSIM (RGB)	SSIM (Y)	Architec	Disparity	Ensemble
1	BSR	M. Cheng, H. Ma, Q. Ma, X. Sun, S. Zhao, X. Sheng	23.8961	25.3143	0.7396	0.7479	Transf	PAM	Data & Model
2	TeamNoSleep	Y. Ding, M. Sun, X. Wen	23.8911	25.3097	0.7358	0.7440	CNN	PAM	Data & Model
3	SRC-B	D. Zhang, J. Li, F. Wang, Z. Xie	23.8830	25.2992	0.7400	0.7479	Transf	PAM	Data & Model
4	webbzhou	Y. Zhou, W. Deng, R. Nie, J. Pu, J. Zhang, Q. Gao	23.8220	25.2402	0.7359	0.7437	CNN	PAM	Data & Model
5	BUPT-PRIV	P. Cao, T. Huang, L. Yang, Q. Song	23.8041	25.2195	0.7356	0.7431	Trans	PAM	Data & Model
6	GDUT_506	J. Chen, H. Li, Y. Shi, Z. Yang	23.7719	25.1913	0.7319	0.7397	Transf	PAM	Data
7	STSR Sharpeners	J. Song, C. Sun, A. Sowmya	23.7560	25.1743	0.7299	0.7383	CNN	PAM	Data
8	Giantpandacv	W. Zou, Y. Zhang, M. Jiang, Z. Yu, M. Tan, H. Gao	23.7424	25.1612	0.7290	0.7369	CNN	PAM	Data & Model
9	LVGroup_HFUT	Z. Zhang, H. Zheng, S. Zhao, Y. Gao, Y. Wei, B. Wang	23.7252	25.1445	0.7309	0.7391	CNN	PAM	Data
10	MakeStereoGreatAgain	G. Li, A. Li, L. Sun, W. Li, Z. Yang	23.7181	25.1370	0.7307	0.7390	CNN	PAM	Data
11	McSR	K. Chen, C. Tang, Y. Li, J. Chen	23.7121	25.1293	0.7306	0.7392	Transf	PAM	Data
12	NUSSZ-STEREO	Y. Wang, G. Zhu, X. Yang, S. Liu, Y. Jing	23.7044	25.1269	0.7284	0.7367	CNN	PAM	Data
13	LongClaw	S. Mistry, S. Chatterjee, V. Jakhetiya, B. Subudhi, S. Jaiswal	23.6800	25.1019	0.7268	0.7354	CNN	PAM	Data
14	SYSU_FVL	Z. He, Z. Qiu, Z. Pan, Z. Zhan, X. Xian, Z. Jin	23.6781	25.0988	0.7307	0.7400	Transf	PAM	Data
15	XY	X. Hu	23.6741	25.0944	0.7265	0.7347	CNN	PAM	Data
16	jingxiangchen1	J. Chen, B. Yang, X. Zhang, C. Li	23.6556	25.0620	0.7261	0.7345	CNN	PAM	Data
17	GarasSjtu	G. Gendy, N. Sabor, J. Hou, G. He	23.4644	24.8873	0.7179	0.7267	CNN	PAM	Data
18	JNU_620	W. Yuan, Z. Li, R. Deng	23.4246	24.8356	0.7155	0.7236	Transf	✗	Data
19	CV_IITRPR	N. Mehta, S. Murala, A. Dudhane	23.1710	24.5894	0.7056	0.7164	CNN	PAM	✗
20	CHASE	Z. Guo, S. Luo	23.1483	24.5657	0.7025	0.7115	CNN	PAM	✗
21	candle	B. Chen, C. He, M. Chen	22.8335	24.2698	0.6926	0.7041	CNN	PAM	✗
-	PASSRnet (Baseline)	-	22.7965	24.2016	0.6801	0.6911	✗	PAM	✗
-	Bicubic (Baseline)	-	21.8358	23.3865	0.6287	0.6443	-	-	-

Table 2. NTIRE 2023 Stereo Image SR Challenge (Track 2) results, rankings, and details from the fact sheets. Note that, score calculated using Eq. 3 is used for the ranking. "Transf" denotes Transformer, "Diffu" denotes Diffusion model, and "PAM" denotes parallax attention mechanism.

Rank	Team	Authors	Score (↑)	LPIPS (↓)	Dispary Error (↓)	Architec	Disparity	Ensemble
1	SRC-B	D. Zhang, J. Li, F. Wang, Z. Xie	0.8622	0.1386	0.0098	Transf	PAM	Data & Model
2	SYSU_FVL	Z. He, Z. Qiu, Z. Pan, Z. Zhan, X. Xian, Z. Jin	0.8538	0.1451	0.0107	CNN	PAM	✗
3	webbzhou	Y. Zhou, W. Deng, R. Nie, J. Pu, J. Zhang, Q. Gao	0.8496	0.1493	0.0106	CNN	PAM	✗
4	SSSL	S. H. Park	0.8471	0.1519	0.0099	CNN	✗	✗
5	Giantpandacv	W. Zou, Y. Zhang, M. Jiang, Z. Yu, M. Tan, H. Gao	0.8351	0.1637	0.0121	CNN	PAM	Data
6	DiffX	Y. Wang. L. Li	0.8303	0.1686	0.0110	Diffu	PAM	✗
7	LongClaw	S. Mistry, S. Chatterjee, V. Jakhetiya, B. Subudhi, S. Jaiswal	0.7994	0.1992	0.0143	CNN	PAM	✗
8	BUPT-PRIV	P. Cao, T. Huang, L. Yang, Q. Song	0.7992	0.1994	0.0140	Transf	PAM	Data & Model
9	McSR	K. Chen, C. Tang, Y. Li, J. Chen	0.7960	0.2026	0.0142	Transf	PAM	Data
10	LVGroup_HFUT	Z. Zhang, H. Zheng, S. Zhao, Y. Gao, Y. Wei, B. Wang	0.7958	0.2028	0.0141	CNN	PAM	✗
11	NUSSZ-STEREO	Y. Wang, G. Zhu, X. Yang, S. Liu, Y. Jing	0.7958	0.2028	0.0143	CNN	PAM	✗
12	jingxiangchen1	J. Chen, B. Yang, X. Zhang, C. Li	0.7928	0.2057	0.0144	CNN	PAM	✗
13	JXNU_SR	J. Zeng	0.7904	0.2082	0.0139	CNN	PAM	✗
14	IR-SDE	Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjolund, T. B. Schon	0.7896	0.2086	0.0185	Diffu	PAM	✗
15	JNU_620	W. Yuan, Z. Li, R. Deng	0.7841	0.2144	0.0143	Transf	✗	Data
16	CV_IITRPR	N. Mehta, S. Murala, A. Dudhane	0.7837	0.2149	0.0142	CNN	PAM	✗
17	Chengzhi-Group	C. Cao, K. Wang, F. Zhang, Q. Zhang	0.7743	0.2241	0.0157	Transf	Disparity	✗

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 175 registered participants, 21 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 21 teams. These methods are briefly described in Section 5 and the team members are listed in Section 7.

It can be observed that the top 5 teams successfully outperform the winner method in NTIRE 2022 (*i.e.*, NAFSSR [43] 23.7873), further boosting the performance of stereo image SR. Moreover, the accuracy of the top 2 methods are very close with a minor PSNR difference of 0.005. In addition, although the SRC-B team produces inferior PSNR

Table 3. NTIRE 2023 Stereo Image SR Challenge (Track 3) 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)	PSNR (Y)	SSIM (RGB)	SSIM (Y)	Architec	Disparity	Ensemble
1	IPIU	K. Zhang, J. Zhang, R. Peng, Y. Ma, L. Jiao	22.3531	24.1146	0.5484	0.6303	CNN & Transf	PAM	Data & Model
2	Team OV	H. Bai, L. Kong, J. Pan, J. Dong, J. Tang	21.9490	23.5708	0.6030	0.6311	CNN	PAM	Data
3	SRC-B	D. Zhang, J. Li, F. Wang, Z. Xie	21.8351	23.4341	0.5980	0.6258	Transf	PAM	Data & Model
4	Giantpandacv	W. Zou, Y. Zhang, M. Jiang, Z. Yu, M. Tan, H. Gao	21.8026	23.4194	0.5916	0.6207	CNN	PAM	Data & Model
5	webbzhou	Y. Zhou, W. Deng, R. Nie, J. Pu, J. Zhang, Q. Gao	21.7676	23.3841	0.5885	0.6207	CNN	PAM	Data & Model
6	LVGroup_HFUT	Z. Zhang, H. Zheng, S. Zhao, Y. Gao, Y. Wei, B. Wang	21.7396	23.3545	0.5885	0.6216	CNN	PAM	Data
7	NTU607-stereo	Y. Chiang, Y. Chen, Z. Huang, H. Yang, I. Chen, S. Kuo	21.6973	23.3211	0.5811	0.6153	CNN	PAM	Data
8	SYSU_FVL	Z. He, Z. Qiu, Z. Pan, Z. Zhan, X. Xian, Z. Jin	21.5162	23.1408	0.5811	0.6153	Transf	✗	Data
9	zzuli	X. Lei, J. Wang, C. Zhai, Y. Zhang, W. Cao, W. Zhang	21.4845	23.1094	0.5732	0.6034	Transf	PAM	Data
10	JNU_620	W. Yuan, Z. Li, R. Deng	21.4829	23.1094	0.5732	0.6034	Transf	✗	Data
11	MakeStereoGreatAgain	G. Li, A. Li, L. Sun, W. Li, Z. Yang	21.1203	22.8100	0.5541	0.5937	CNN	PAM	Data
12	Chengzhi-Group	C. Cao, K. Wang, F. Zhang, Q. Zhang	20.7199	22.5314	0.5752	0.6083	Transf	Disparity	✗

results than the top 2 teams, it achieves the highest SSIM score of 0.7400.

4.2. Track 2: perceptual & Bicubic Degradation

Among the 93 registered participants, 17 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 17 teams. These methods are briefly described in Section 5 and the team members are listed in Section 7.

Different from track 1 that focus on minimizing the distortion between super-resolved results and the groundtruth, this track aims to reconstruct stereo images with high perceptual quality and stereo consistency. From Table 2 it can be observed that high LPIPS performance are not always consistent with high stereo consistency. For example, the webbzhou team produces better LPIPS result than the SSSL team but suffers a slight larger disparity error.

4.3. Track 3: Fidelity & Realistic Degradation

Among the 103 registered participants, 12 teams successfully participated the final phase and submitted their results, codes, and fact sheets. Table 3 reports the final test results, rankings of the challenge, and major details from the fact sheets of 12 teams. These methods are briefly described in Section 5 and the team members are listed in Section 7.

As we can see, all methods suffer a notable performance drop on realistic degradations as compared to the standard bicubic one. In addition, although the Team OV team produces inferior PSNR results than the IPIU team, it achieves the highest SSIM score of 0.6030.

4.4. Summary

Architectures and main ideas. All the proposed methods are based on deep learning techniques. Transformers (particularly SwinIR) and the winner method in the NTIRE 2022 challenge (*i.e.*, NAFSSR) are widely used as the basic architecture. Motivated by the high-quality of images synthesized by diffusion models, two solutions in track 2 develop diffusion-based methods, which sheds some lights on future works. To exploit cross-view information, the idea of parallax-attention mechanism (PAM) are 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. Ensemble strategy (including both data ensemble and model ensemble) is adopted in most solutions to further boost the final SR performance. For data ensemble, the inputs are flipped and the resultant SR results are aligned and averaged for enhanced prediction [46]. For model ensemble, the results produced by multiple models are averaged for better results. Note that, several works observe that ensemble strategy may reduce the perceptual quality of images for track 2. As a result, most methods in track 2 do not employ an ensemble strategy.

Conclusions. By analyzing the settings, the proposed methods and their results, it can be concluded that: 1) The proposed methods improve the state-of-the-art performance in stereo image SR. 2) With recent renaissance of CNNs (*e.g.*, ConvNeXt and 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) Tricks including delicate data augmentation strategies, data and model ensemble strategies are critical to the final performance.

5. Challenge Methods and Teams

5.1. BSR - Track 1★

Despite the success of Transformer-like networks in single image SR, directly extending these structures to stereo image SR task has two limitations. First, current Transformers developed for single image SR cannot make use of the complementary stereo information, which is critical for stereo image SR task. Second, Transformers rely

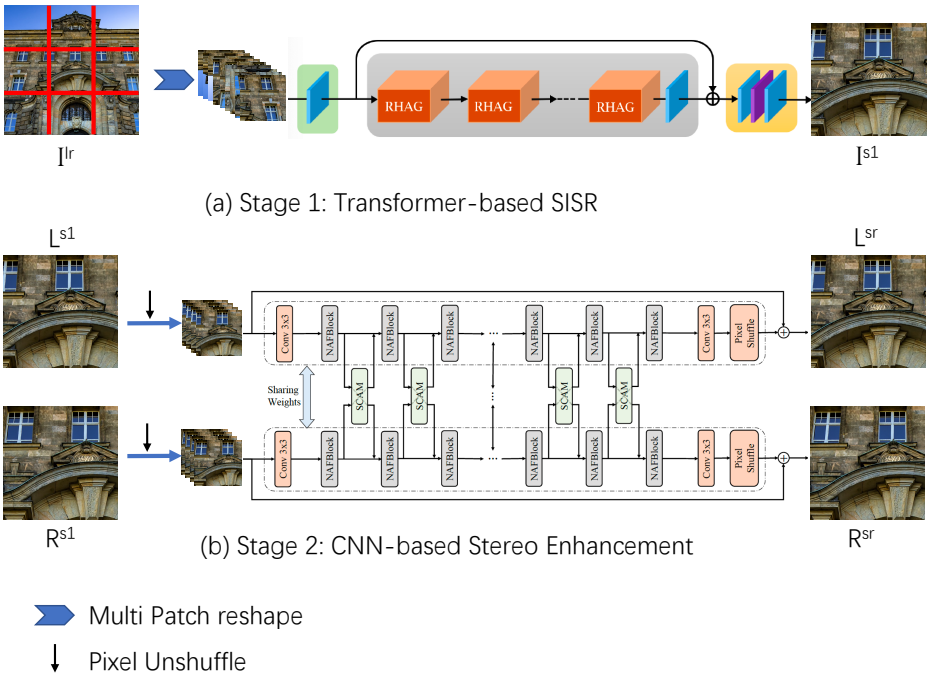


Figure 1. BSR: The network architecture of the proposed HTCAN.

heavily on sufficient data to achieve superior performance against CNN counterparts. In this challenge, this team proposed 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 [47], as illustrated in Fig. 1. In addition, as multi-stage strategies have been demonstrated to be effective in image restoration tasks, a three-stage training strategy was adopted.

(1) Stage 1

Network Architecture. In the first stage, HAT-L [48] was used for single image SR. The GeLU activation in the model was replaced with SiLU activation and the window size was enlarged to 24×24.

Training Settings. The network was first trained with a Charbonnier loss and then fine-tuned with an MSE loss using an Adam optimizer. The batch size was set to 32 and the patch size was set to 48×48. The learning rate was initialized with 2×10^{-4} and reduced by half at the 300K, 500K, 650K, 700K, and 750K iteration. Training was stopped after 800k iterations. To enlarge the receptive field of HAT-L, multi-patch training strategy [49] was adopted.

Data Augmentation. Data augmentation was performed through RGB channel shuffing, horizontal/vertical flipping, rotation, and Mixup.

Ensemble Strategy. Three different models were trained for model ensemble. For each model, data ensemble was employed through random rotation and horizontally/vertically flipping.

(2) Stage 2

Network Architecture. Stage 2 aims to extract and incorporate stereo information in a stereo image pair. To this end, NAFSSR-L [43] ($4\times$ SR) was used at this stage. After stage 1, the output images are super-resolved results with a scaling factor of 4. Then, these results were pixel unshuffled with a factor of 4 and fed to the model at this stage. The input channel of the first convolutional layer is also changed accordingly, resulting in a model termed UnshuffleNAFSSR. In this way, the memory consumption can be reduced and the receptive field can be enlarged.

Training Settings. The training of UnshuffleNAFSSR consists two phases. In the first phase, the NAFSSR-L model is trained on the Flickr1024 dataset using the default settings. In the second phase, pre-trained model was used to initialize UnshuffleNAFSSR for training. The network was first trained with a Charbonnier loss and then fine-tuned with an MSE loss using an AdamW optimizer. The batch size was set to 32 and the patch size was set to 30×90. The learning rate was initialized as 5×10^{-4} and updated using a cosine annealing strategy. The minimum learning rate was set to 1×10^{-7}.

Data Augmentation. Data augmentation was performed through RGB channel shuffling and horizontal/vertical flipping.

Ensemble Strategy. Two models were trained for model ensemble. For each model, data ensemble was em-

ployed through horizontal/vertical flipping and exchange of left/right views.

(3) Stage 3

After stage 2, the resultant models were further used for further ensemble. The process in stage 3 is the same as stage 2.

5.2. TeamNoSleep - Track 1★

This team proposed to progressively reconstruct HR images step by step and the overall architecture is shown in Fig. 2. The proposed network contains two stages. For the first stage, $4\times$ NAFSSR [43] was employed to train 8 different models with different structures, different model sizes, and different data augmentation strategies. At the second stage, models obtained at the first stage were merged to refine the results. The architecture of this stage was a RCAN model with a $1\times$ scale factor. In addition, a robust and practical data augmentation method was adopted to further improve the performance.

Baseline Model and Variants. NAFSSR [43] was used as the baseline model. To obtain model variants with diverse structures, Dropout and Dropout2d operations [50] were employed, as shown in Fig. 3. Moreover, this team designed an attention feature selective unit based SCAM (ASUSCAM) to aggregate features at different levels, which is used to replace the original SCAM in NAFBlocks except the first one.

Data Augmentation. Multiple dataset augmentation strategies were employed, including RGB channel shuffling, horizontal/vertical shifting, left-right view exchange, and image rescaling. Specifically, they downscaled HR images in the training set with different scale factors ($0.9\times$, $0.95\times$, and $0.85\times$) and then downsampled them to obtain LR images for training.

Model Ensemble. Two kinds of model ensemble methods were used to aggregate SR results produced by different models. Specifically, a three-layer CNN was developed to take these SR results as input for spatial aggregation. Then, another three-layer CNN with a softmax layer was introduced to achieve aggregation along the channel dimension.

Training Settings. The Adam optimizer and an MSE loss were employed for optimization. Patches of size 30×90 were used for the training of stage 1 while 128×128 patches were used to train the second-stage models.

5.3. SRC-B - Track 1★, 2★, 3★

Inspired by SwinFIR [51], HAT [52], and NAFSSR [43], this team proposed a SwinFIRSSR by using Swin Transformer [27] and fast Fourier convolution [53], as shown in Fig. 4. HAT uses the Residual Hybrid Attention Group (RHAG) to activate more pixel in image SR Transformer to improve the performance. To improve the representation ability of model, this team replaced the 3×3 convolution in RHAG with a fast Fourier convolution and a residual module to fuse global and local features, namely Spatial-frequency Block (SFB). They also followed NAFSSR to aggregate features from left and right views using stereo cross-attention modules (SCAM).

During training, they cropped HR images into sub-images (384×384) and used the Adam [54] optimizer with default parameters to train the model with the Charbonnier L1 loss [55] for 800000 iterations. The initial learning rate was 2×10^{-4} and updated using the MultiStepLR scheduler at iteration [600000, 650000, 700000, 750000]. The batch size was 4 and patch size was 64. Horizontal/vertical flipping, rotation, RGB channel shuffling, and Mixup [56] were used for data augmentation. Inspired by [46], data ensemble and model ensemble were both used to produce the final SR results.

5.4. SYSU_FVL - Track 1, 2★, 3

This team participated in three tracks and proposed three networks, respectively.

Track 1: Similar to most of the methods in NTIRE 2022 Stereo Image SR Challenge, they combined the BiPAM block with SwinIR. Particularly, they applied the BiPAM block after the RSTB in the second half, as shown in Fig. 5. For training, they followed the three-stage training strategy as the runner-up method (SwiniPASSR) in the NTIRE 2022 challenge.

Track 2: The champion method (NAFSSR) in the NTIRE 2022 challenge was employed as the baseline and a LPIPS-based perceptual loss was used for training [57]. The structure of NAFSSR is shown in the Fig. 6. To improve stereo consistency, they performed supervised learning on each SCAM. In addition, they followed PASSRnet to calculate cycle-attention maps and valid masks of SCAM, and then calculated photometric loss, smoothness loss, and cycle loss. During training, they first used an MSE loss for optimization, and then added a stereo consistency loss for fine-tuning. During fine-tuning, the weights of the MSE loss, the perceptual loss, and the stereo consistency loss were set to 1, 1, and 0.01, respectively. It is worth noting that they did not use self-ensemble method as they found that this would destroy the stereo consistency of SR results.

Track 3: They proposed a hybrid attention Transformer, namely HAT (Fig. 7). The proposed HAT combines channel attention and self attention to leverage the former's ability to utilize global information and the latter's powerful presentation capabilities. Furthermore, to better aggregate cross-window information, an overlapping cross-attention module was introduced.

5.5. webbzhou - Track 1, 2★, 3

Self-similarity is critical in exploring non-local textures in single image SR. In the area of stereo SR, cross-view

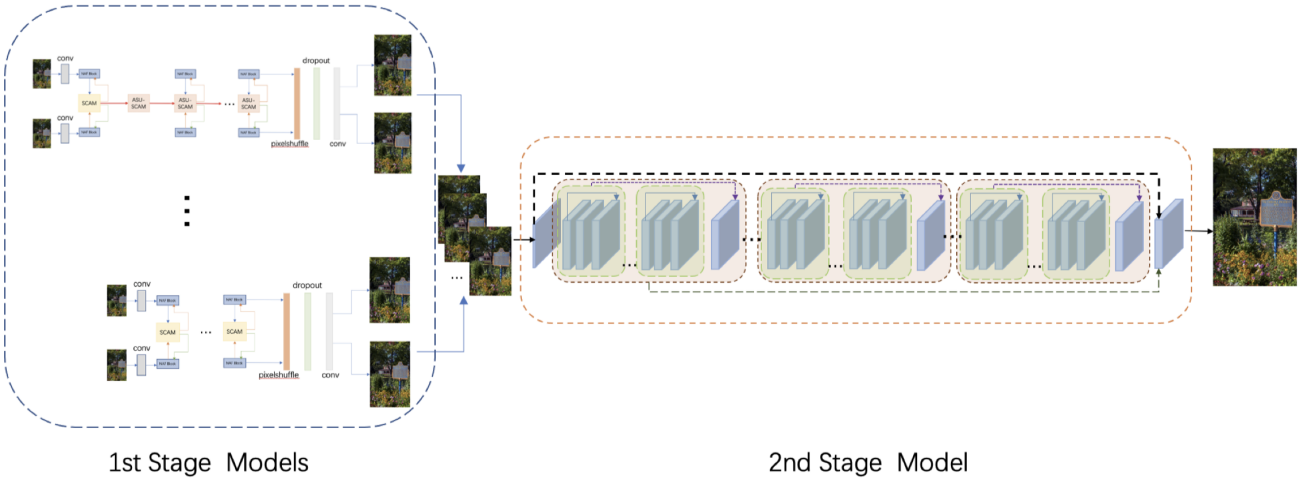


Figure 2. TeamNoSleep-1: Overview of the proposed method. The first stage obtains several $4\times$ SR models and the second stage merge these models to produce the final results.

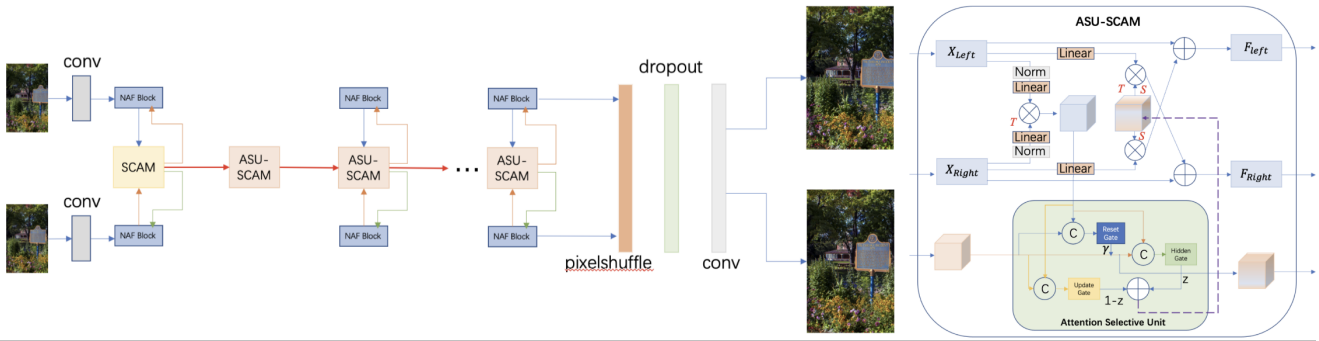


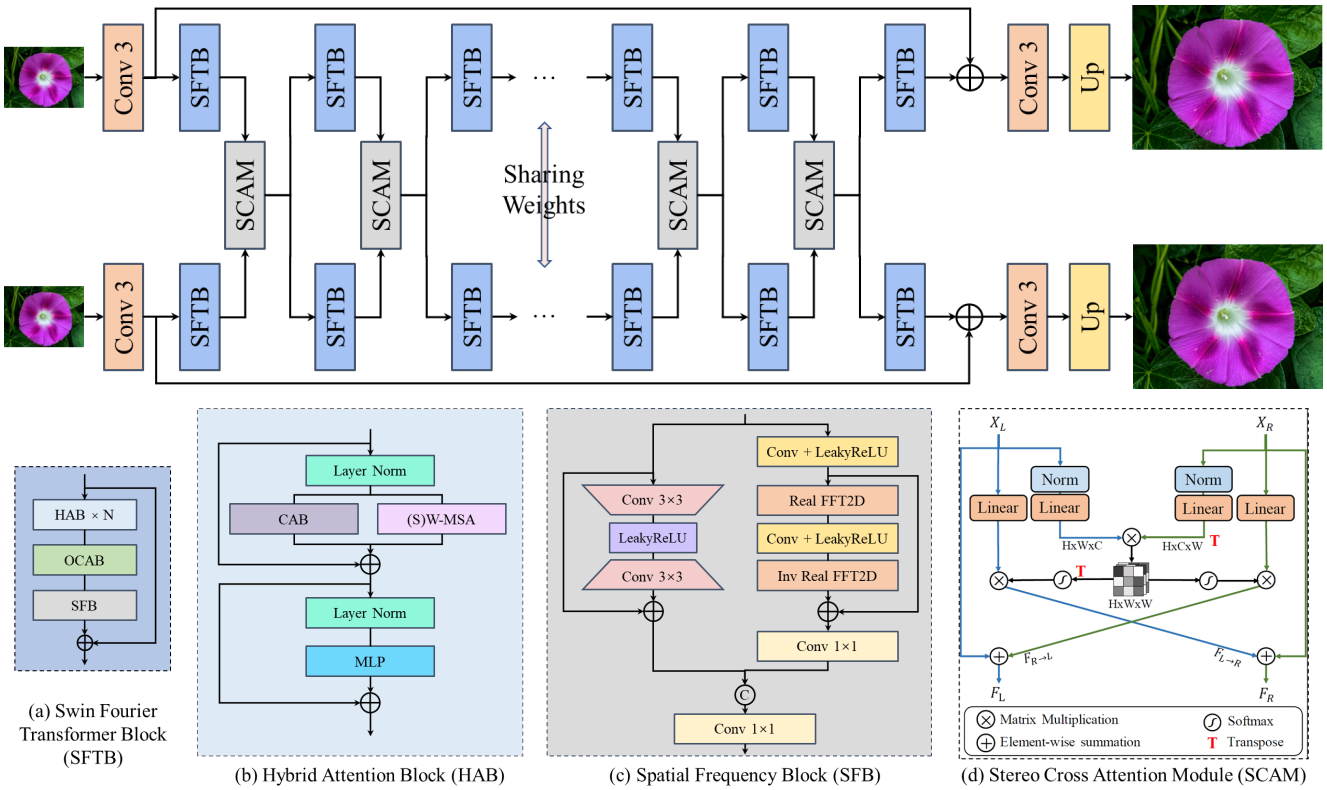
Figure 3. TeamNoSleep-2: The network architecture of the proposed enhanced NAFSSR.

self-similarity is also valuable in restoring textures, in addition to intra-view self-similarity. With this in mind, this team proposed SCGLANet [58], which is composed of several NAFTBlocks and SCGLAM, as show in Fig. 8. Moreover, this team proposed an efficient and sparse attention module named Stereo Cross Global Learnable Attention Module (SCGLAM) to exploit both cross-view and intra-view information. The proposed SCGLAM module can effectively capture intra-view and cross-view similarity with small computational complexity. Moreover, the proposed method hashes features into different buckets and calculates similarities only between buckets, resulting in significant decreases in computational cost. Additionally, they modified the Stereo Cross Attention Module (SCAM) in NAFSSR to a sparse epipolar attention module.

Training Settings. To improve the I/O speed, this team divided the original HR and LR images into patches with a stride of 80. Following NAFSSR, data augmentation was implemented by randomly flipping these patches horizontally and vertically. In addition, they also used random RGB channel shuffling for augmentation. This team has participated in three tracks, the training details for each track are as follows:

- **Track 1:** They trained the models for 400k iterations on eight NVIDIA A40 GPUs with a batch size of 24. These models were optimized by the Adam method with $\beta_1 = 0.9$ and $\beta_2 = 0.9$ and the weight decay was set to 0. The initial learning rate was set to 5×10^{-4}, and true cosine annealing scheduler was chosen as the learning scheme. An L1 loss was first used to train the model, and then an MSE loss was adopted for fine-tuning.

- **Track 2:** They trained the models for 400k iterations on eight NVIDIA A40 GPUs with a batch size of 24. The models were optimized by the Adam method with $\beta_1 = 0.9$ and $\beta_2 = 0.9$ and the weight decay was set to 0. The initial learning rate of the generator and



(a) Swin Fourier Transformer Block (SFTB)

(b) Hybrid Attention Block (HAB)

(c) Spatial Frequency Block (SFB)

(d) Stereo Cross Attention Module (SCAM)

Figure 4. SRC-B: The network architecture of the proposed SwinFIRSSR.

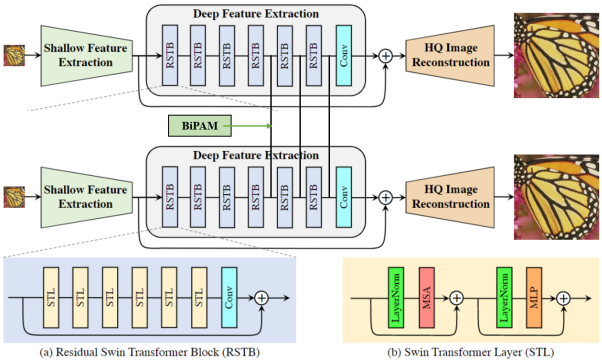


Figure 5. SYSU FVL - Track 1: The network architectural of the proposed SwinIR-BiPAM.

discriminator was set to 1×10^{-4}, and true cosine annealing scheduler was chosen as the learning scheme. Models obtained in track 1 were used for initialization. The overall loss function contains an L1 loss, a GAN loss, and a LPIPS loss.

- **Track 3:** They trained the models for 400k iterations on eight NVIDIA A40 GPUs with a batch size of 24. The models were optimized by the Adam method with $\beta_1 = 0.9$ and $\beta_2 = 0.9$ and the weight decay was set to 0. The initial learning rate was set to 2×10^{-4}, and true cosine annealing scheduler was chosen as the learning scheme. Models obtained in track 1 were used for initialization, and an MSE loss was used to train the models.

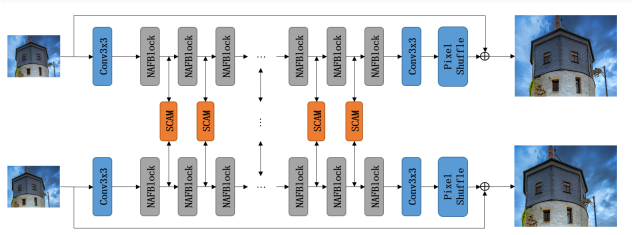


Figure 6. SYSU FVL - Track 2: The network architectural of the proposed SC-NAFSSR.

For track 1 and track 3, model ensemble and data ensemble strategies were applied to produce the final results.

5.6. IPIU - Track 3★

This team combined NAFSSR [59] with LTE [60] and proposed LTFSSR, as shown in Fig. 9. First, data augmen-

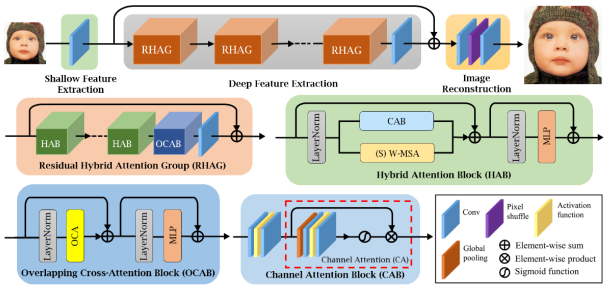


Figure 7. SYSU FVL - Track 3: The network architectural of the proposed HAT.

tation was performed on the training set. Subsequently, the augmented data were used to train multiple models. The outputs of each model were fused to obtain the final results.

During the training phase, six data augmentation methods were utilized, including CutBlur, Blend, RGB channel shuffling, Mixup, CutMix, and CutMixup. this team chose a total of 5 models, including SSRED-FNet [40] and NAFSSR [43], LIIF [60], SwinIR-LTE [61], and RDN-LTE [61]. To fuse the output results of these 5 models, this team first averaged the results produced by these model, and then calculated the MSE between each image and the averaged one. Larger weights will be assigned to the model with smaller MSE values during fusion. The proposed fusion strategy can avoid the final result being skewed by outliers. The training of the models was conducted on 8 Nvidia V100 GPUs using the Adam optimization method and multi-step learning rate decay method.

5.7. Team OV - Track 3★

This team developed a residual-in-residual structured non-linear activation-free network (RIR-NAFNet) to solve the stereo image SR problem. RIR-NAFNet uses the non-linear activation-free block (NAF Block [59]) as the basic block, and adopts the stereo cross-attention module (SCAM [43]) for cross-view features fusion. However, even with a large number of training iterations, simply stacking NAF Blocks and SCAMs to build a deeper network is difficult to achieve significant performance improvements. To address this issue, they adopted the residual-in-residual structure [23] to form an ultra-deep network.

As shown in Fig. 10, RIR-NAFNet first uses a convolutional layer to extract features from the input stereo images, then adopts 12 NAF Groups to enhance cross-view features, and finally enlarges the resolution to obtain the SR results with a long skip connection. Each NAF Group contains a NAFBlock, 7 "NAFBlock-SCAM-NAFBlock" units, and a convolutional layer. On top of this structure, they formed an ultra-deep network with a depth of more than 720 convolutional layers and a width of 96 feature channels.

5.8. BUPT-PRIV - Track 1, 2

This team developed a SwinIR-SCAM network by introducing a stereo cross-attention module (SCAM) to a SwinIR network. As illustrated in Fig. 11, on top of the basic SwinIR model, SCAMs were inserted after every two or three residual Swin Transformer blocks (RSTBs). A three-stage training strategy was adopted. For stage 1, a vanilla SwinIR model was trained for 200k iterations with patch size of 32×32 using an L1 loss. In stage 2, SCAMs were inserted into SwinIR and the resultant model was fine-tuned for another 150k iterations using the same settings. In the final stage, the patch size was enlarged to 32×96 and an L2 loss was used for training. During training, random flipping and random RGB shuffling were used for data augmentation. Data ensemble and model ensemble were adopted to produce the final results.

5.9. candle - Track 1

This team proposed a stereo image SR network with multiple feature extraction blocks termed ME-super. As shown in Fig. 12, input LR stereo images are first fed to cascaded feature extraction blocks. Then, a jointly texture enhanced parallax attention module is adopted to aggregate features from both left and right views. Finally, SR results are reconstructed.

5.10. CHASE - Track 1

Inspired by NAFSSR [43] and iPASSR [39], this team proposed a nonlinear parallax stereo super-resolution (NPSSR) network. Specifically, as illustrated in Fig. 13, input LR images are first fed to cascaded NAFBlocks and SCAMs to extract and aggregate features from left and right views. Then, the resultant features are used to reconstruct the final SR results. During the training phase, a two-stage training strategy was adopted. An L1 loss was used the first stage and an MSE loss was adopted for the second stage.

5.11. Chengzhi-Group - Track 2, 3

This team introduced an enhanced SwinIR network. As shown in Fig. 14, the proposed model consists of three modules, including shallow feature extraction, deep feature extraction, and image reconstruction modules. To capture stereo correspondences, a RAFT-like [45] structure was employed for disparity estimation.

5.12. CV_IITRPR - Track 1, 2

This team proposed a stereo image SR network with a two-branch structure. As shown in Fig. 15, 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

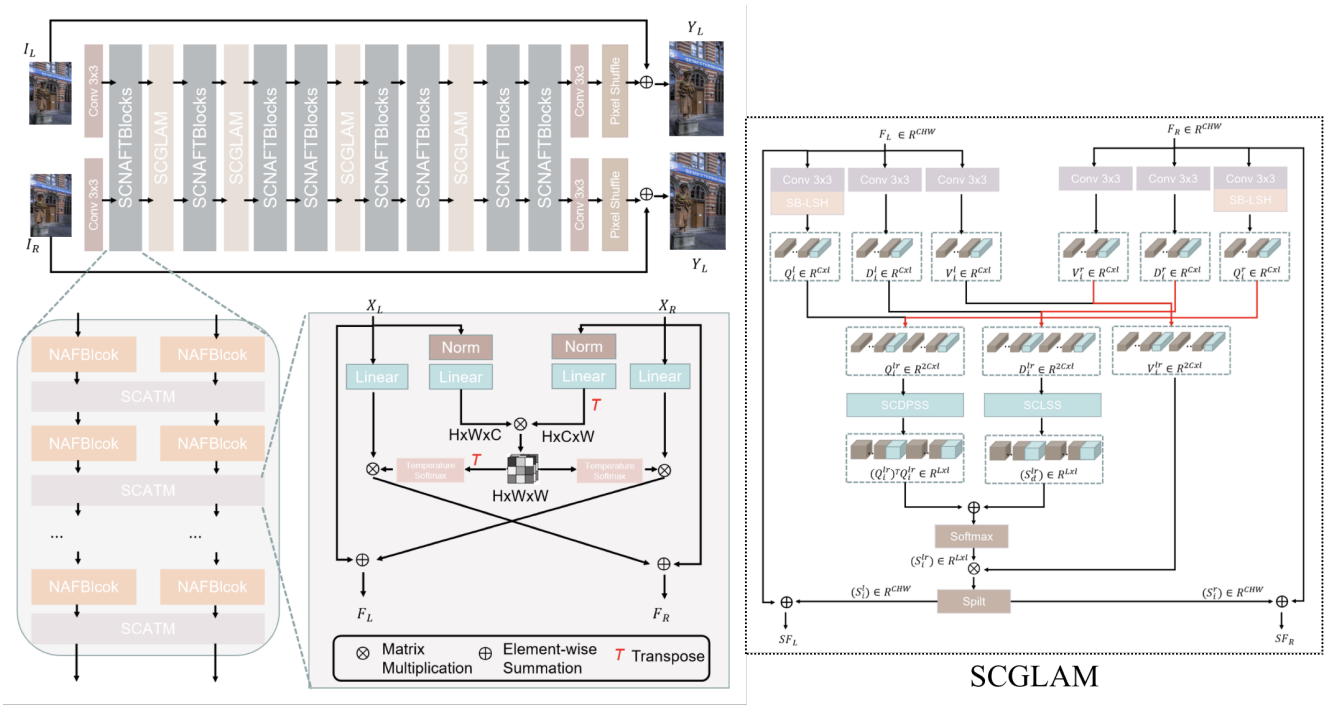


Figure 8. webbzhou: The network architecture of the proposed SCGLANet.

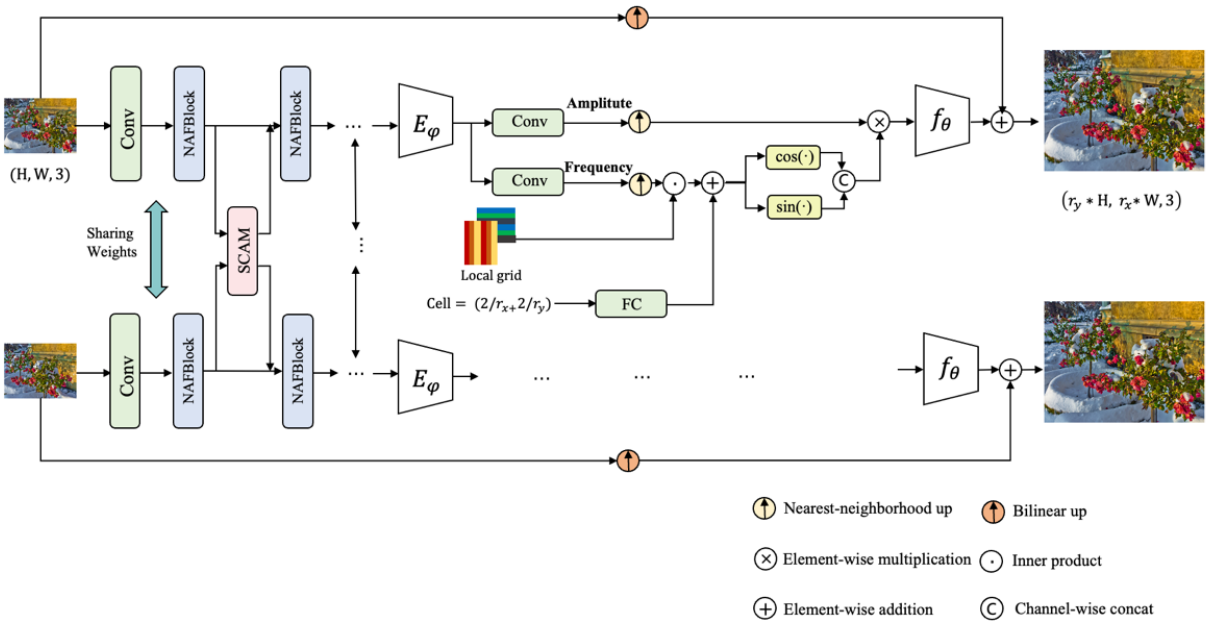


Figure 9. IPIU: The network architecture of the proposed LTFSSR.

at different channels. After that, a bidirectional parallax attention module [39] is employed to aggregate features from both views. Finally, intra-view and inter-view features are collected to produce the super-resolved results.

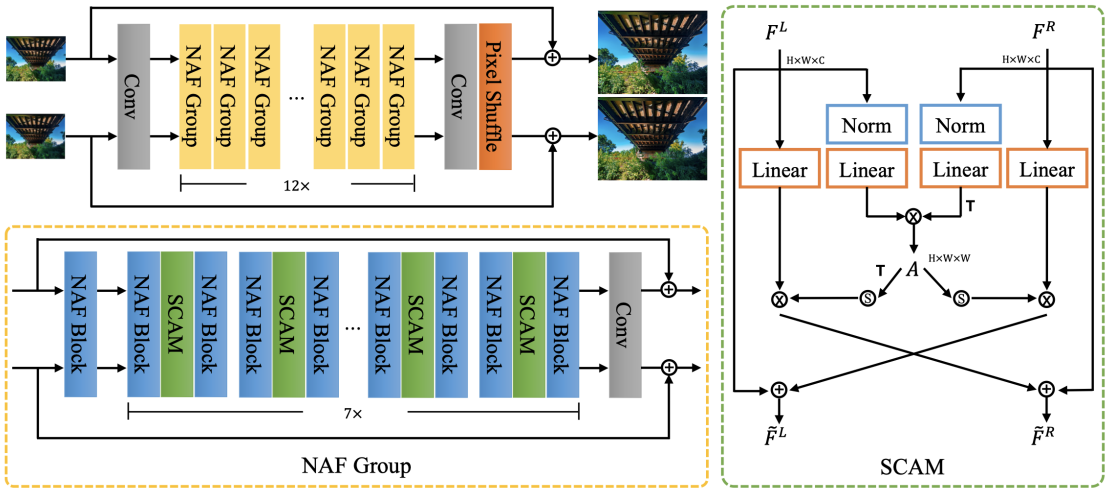


Figure 10. Team OV: The network architectural of the proposed RIR-NAFNet.

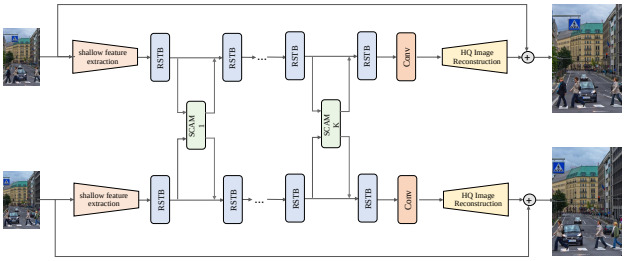


Figure 11. BUPT-PRIV: The network architecture of the proposed SwinIR-SCAM.

5.13. DiffX - Track 2

Motivated by success of diffusion model [62] in synthesizing high-quality images, this team introduced a diffusion model based network termed preconditioned diffusion model with two-way-tied net (PDMTN). Different from current conventional SR diffusion models that use LR images as the conditional inputs, the proposed method employs a SR result predicted by a state-of-the-art SR method (NAFSSR [43]) to encourage the diffusion model to focus on restoring finer perceptual details instead of learning SR from scratch. However, vanilla diffusion model does not take stereo consistency into consideration and could produce severe artifacts. To remedy this, the proposed PDMTN employs a CNN with shared weights to process left and right views, as illustrated in Fig. 16. During the denoising stage, the left and right images are predicted in parallel with two weight-tied UNet branches. Meanwhile, stereo cross-attention modules (SCAM) were plugged to bridge the two branches at each downsampling and upsampling layer to incorporate stereo information.

5.14. GarasSjtu - Track 1

This team proposed Transformer-style network for Stereo Image Super-resolution (TSNSSR), as shown in Fig. 17. The proposed TSNSSR uses two branches with shared weights to process left and right views, and employs stereo cross-attention module to incorporate cross-view information. Motivated by Conv2Former [63] and ConvNeXt [64], a building block termed Conv2FormerXt was developed, which includes a Conv2FormerB and a multi-layer perceptron (MLP). The number of the Conv2FormerXt block was set to 40 and the number of channels was set to 64.

During the training phase, several data augmentation strategies were implemented, including random cropping, random vertical and horizontal flipping, random horizontal shifting, and random RGB channel shuffling. The stochastic depth strategy [65] and the skip-init strategy [66] were utilized to address the over-fitting issue for higher accuracy. The proposed network was trained using the AdamW optimizer for 1×10^5 iterations. The learning rate was initialized as 3×10^{-3} and dropped to 1×10^{-7} with cosine annealing strategy. Batch size was set to 8 and patch size was set to 40×100

During the evaluation phase, data ensemble was performed through flipping horizontally and vertically, shuffling RGB channels, and exchanging left-right views. In addition, since training was conducted on image patches while evaluation was implemented on the whole images, local-SE module [67] was employed to address this inconsistency.

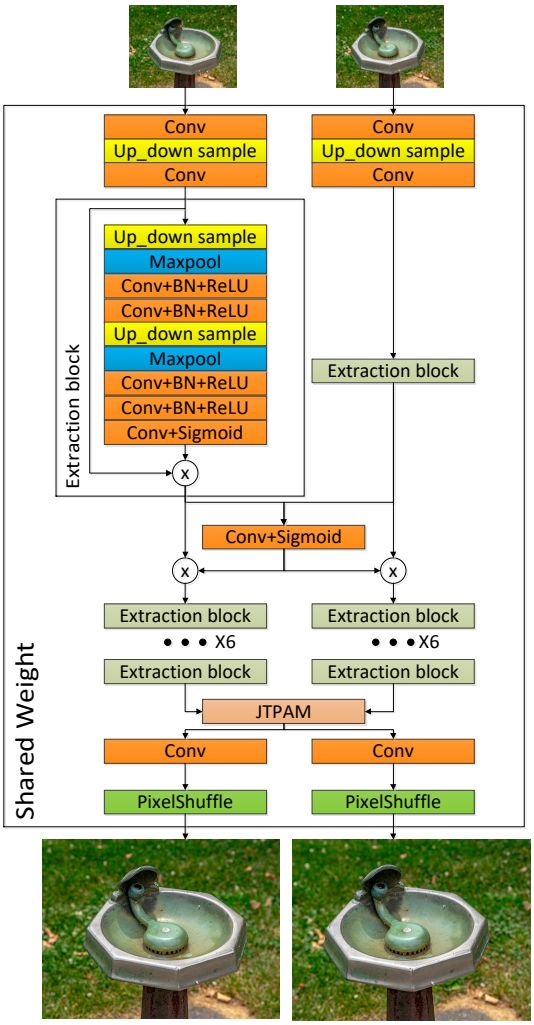


Figure 12. candle: The network architecture of the proposed ME-super network.

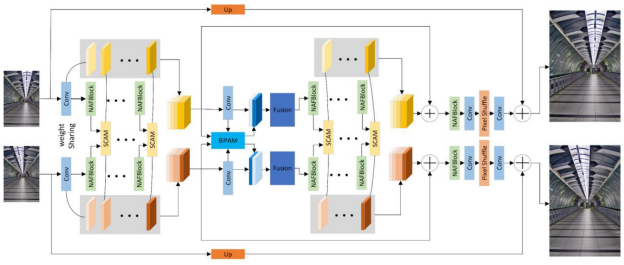


Figure 13. CHASE: The network architecture of the proposed NPSSR network.

5.15. GDUT_506 - Track 1

This team used NAFSSR [43] as the baseline (Fig. 18) and incorporated both FFT loss and Charbonnier loss to boost its performance. This team observed that the FFT loss is able to improve the performance of tiny model and ultimately introduce gains on large models. Besides, Charbonnier loss helps to achieve more stable convergence than the MSE loss when training large models.

The training process of the proposed network contains two stages. In the first stage, the network was trained for $4\times$ SR with a batch size of 4 and a patch size of 30×90. The AdamW optimizer was adopted in this stage, with the initial learning rate being set to 3×10^{-3} and the iteration number being set to 100k. In the second stage, the network was fine-tuned for another 100k iterations, with a batch size of 4, a patch size of 60×180, and a learning rate of 3×10^{-4}. The AdamW method was also employed for optimization. During the training phase, horizontal/vertical flipping and RGB channel shuffling were used for data augmentation.

During the test phase, models trained with different hyper-parameters were used for ensembling and data ensemble was also applied to produce the final results.

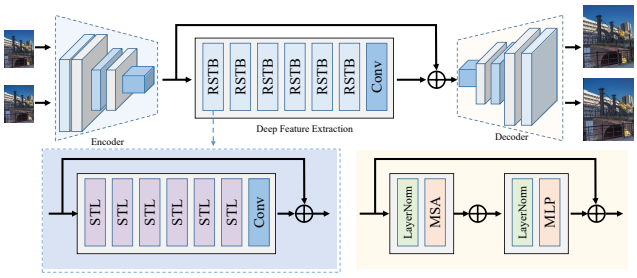


Figure 14. Chengzhi-Group: The network architecture of the proposed enhanced SwinIR network.

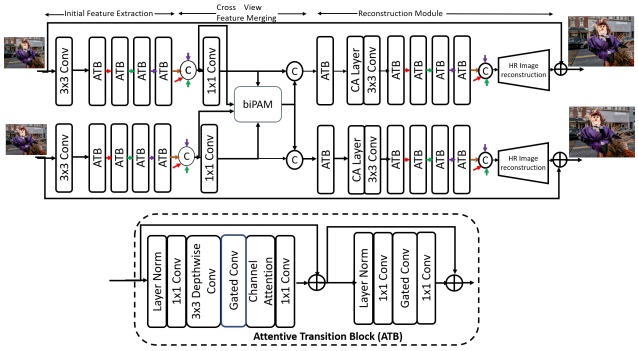


Figure 15. CV_IITRPR: The network architecture of the proposed network.

5.16. Giantpandacv - Track 1, 2, 3

This team proposed a cross-view hierarchical network for stereo image super-resolution (CVHSSR) [68], as shown in Fig. 19. Specifically, CVHSSR consists of cross-hierarchy information mining blocks (CHIMB) and cross-view interaction modules (CVIM). The CHIMBs were de-

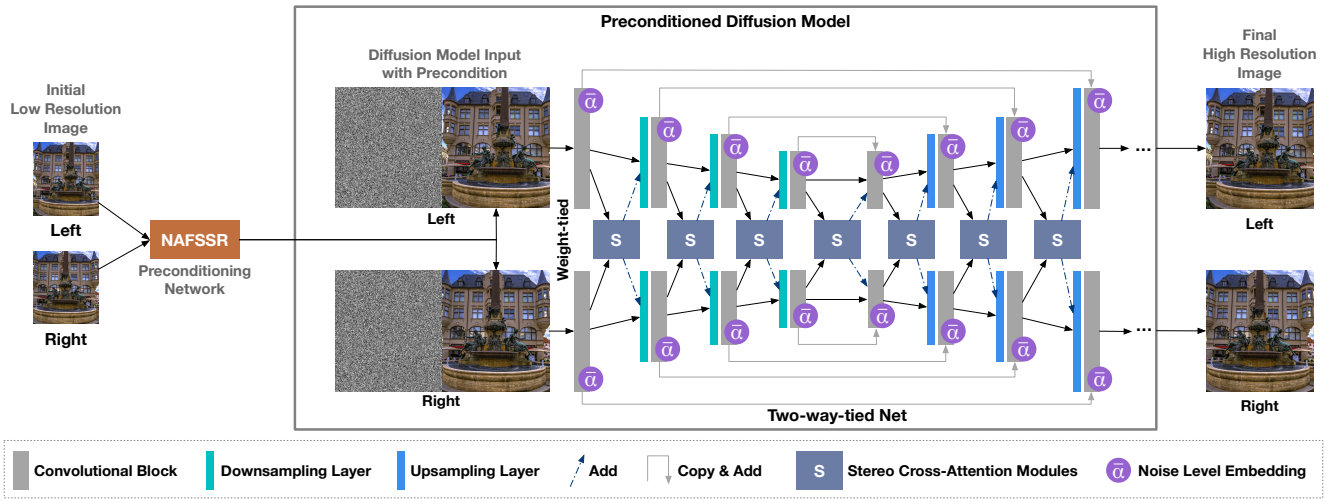


Figure 16. DiffX: The network architecture of the proposed PDMTN.

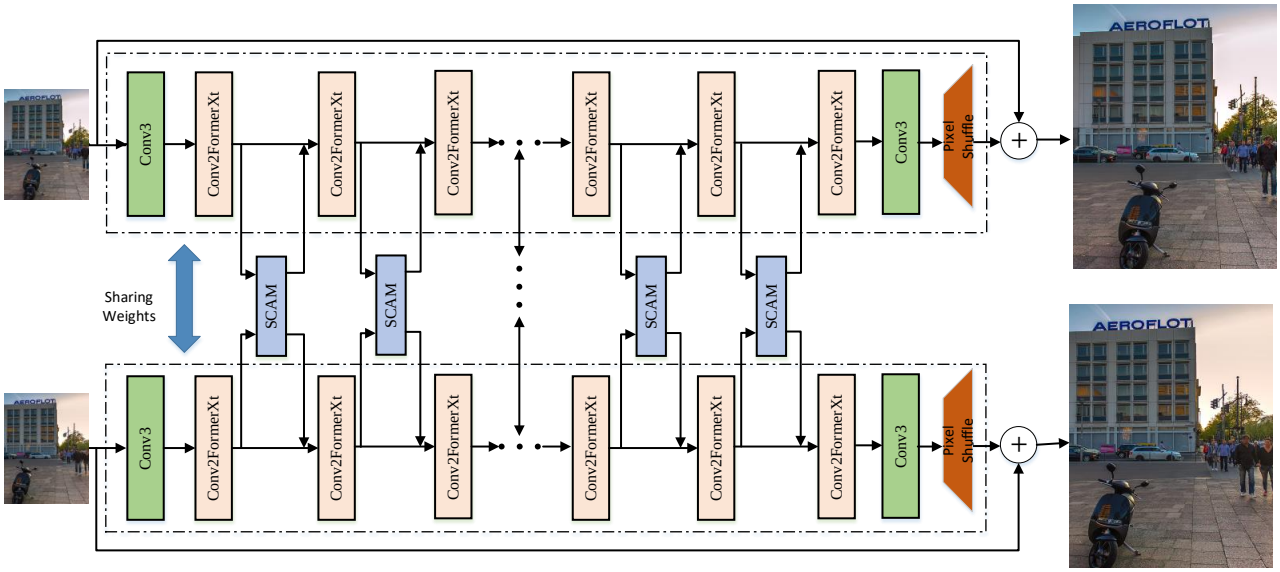


Figure 17. GarasSjtu: The network architecture of the proposed TSNSSR.

signed to extract similar features both locally and globally from the image, which can effectively restore accurate texture details. The CVIMs were mainly used to fuse features from different viewpoints.

Training Settings. This team has participated in three tracks, the training details for each track are as follows:

- **Track 1:** They trained their network using an MSE loss and optimized using the Lion method with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 8. The learning rate was initially set to 5×10^{-4} and decayed with the cosine strategy. The model was trained for 200000 iterations, and finally fine-tuned using both an MSE and a frequency Charbonnier loss for 200000 iterations, with a learning rate of 5×10^{-5}. In addition, the stochastic depth strategy was used for network regularization, with the drop path rate being set 0.3.

- **Track 2:** The network weights in track 1 were used as a pre-trained model for track 2. Then, an adversarial loss was adopted for training. The network was optimized using the Lion method with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 8. The learning rate was initially set to 1×10^{-4} and decayed with the cosine strategy. The training was stopped after 150000 iterations. The stochastic depth strategy was used for

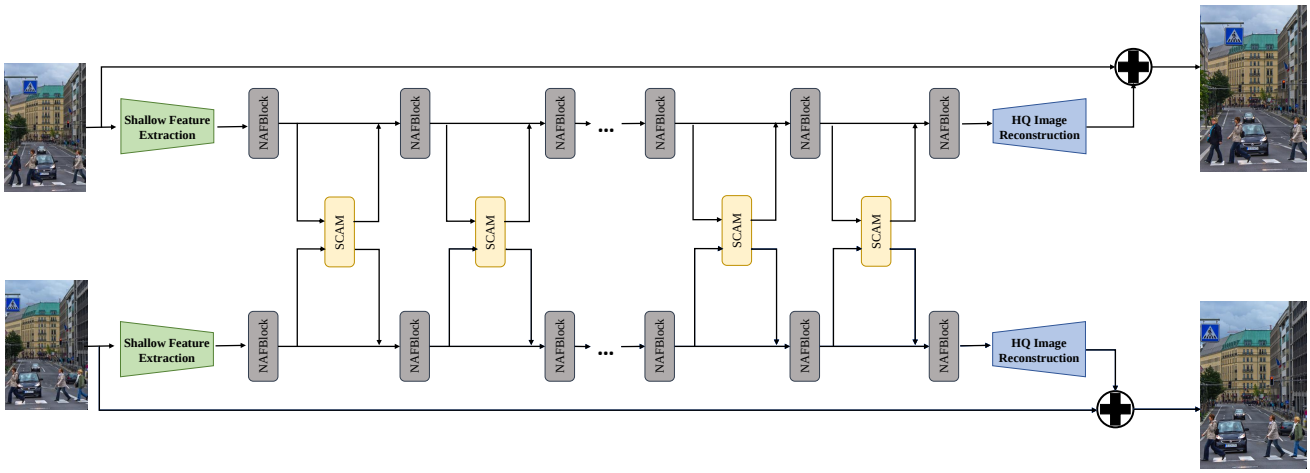


Figure 18. GDUT_506: The network architecture of the proposed Parallax Res-Transformer Network.

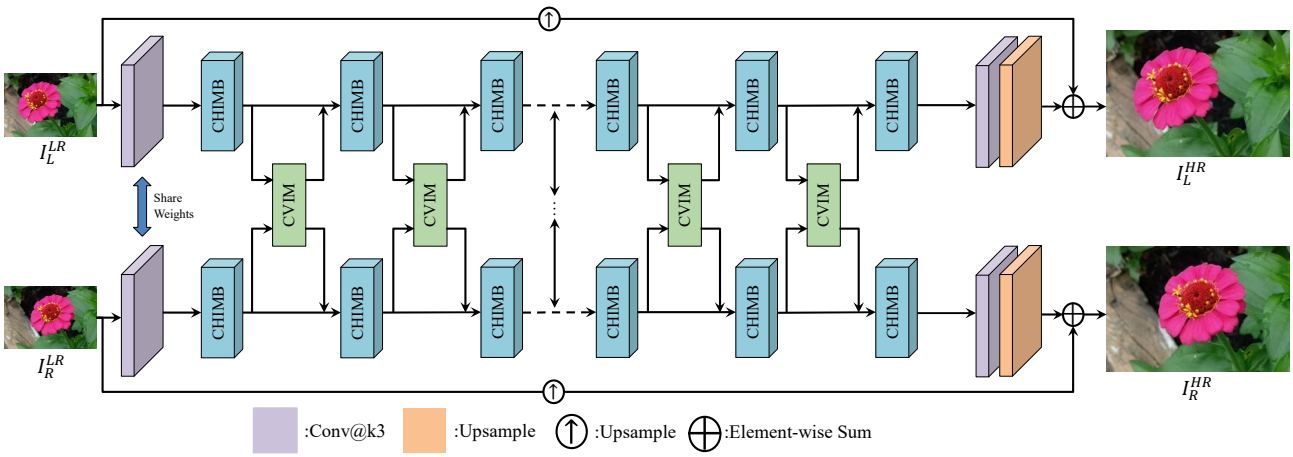


Figure 19. Giantpandacv: The network architecture of the proposed CVHSSR.

network regularization, with the drop path rate being set 0.3.

- **Track 3:** The network was trained using an MSE loss and optimized using the Lion method with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 8. The learning rate was initially set to 5×10^{-4} and decayed with the cosine strategy. The model was trained for 200000 iterations, and finally fine-tuned using both an MSE and a frequency Charbonnier loss for 200000 iterations with a learning rate of 3×10^{-5}. In addition, the stochastic depth strategy was used for network regularization, with the drop path rate being set to 0.2.

For all the tracks, data ensemble strategy was used to improve PSNR. In tracks 1 and 3, they further used model ensemble strategy to produce the final results.

5.17. IR-SDE - Track 2

This team leveraged the diffusion models for realistic image restoration [70]. Specifically, IR-SDE [69] was used as the baseline diffusion framework, which can naturally transform the high-quality image to its degraded counterpart, without relying on explicit degradation models. As shown in Fig. 20, IR-SDE is a mean-reverting SDE in which the forward process is defined as:

$$dx = \theta_t (\mu - x)dt + \sigma_t dw, \tag{5}$$

where θ_t and σ_t are time-dependent positive parameters that characterize the speed of the mean-reversion and the stochastic volatility, respectively. Since it is an Ito SDE, this team further derived a reverse-time SDE:

$$dx = \left[\theta_t (\mu - x) - \sigma_t^2 \nabla_x \log p_t(x)\right] dt + \sigma_t d\hat{w}. \tag{6}$$

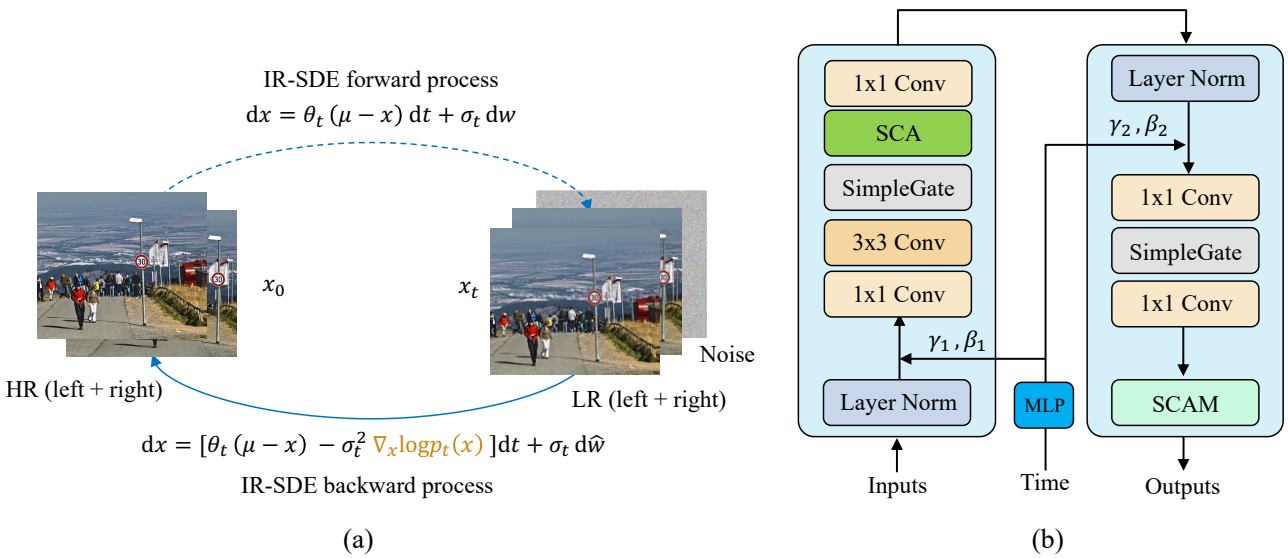


Figure 20. IR-SDE: The network architecture of the proposed Reffusion method. (a) Image restoration based on IR-SDE [69], which uses a mean-reverting stochastic differential equation (SDE) to recover images. (b) The modified NAFBlock. Here "SCA" is the simple channel attention, and "SimpleGate" is an element-wise operation that splits feature channels into two parts and then multiplies them as output.

At test time, the only unknown part is the score $\nabla_x \log p_t(x)$ of the marginal distribution at time t. This team employed a CNN network to estimate this score to backward from the low-quality image to the high-quality image. Unlike L1 loss that usually produces smooth/blurry results, the proposed Reffusion aims to achieve a competitive perceptual performance as well as the distortion scores (*i.e.*, PSNR). To handle left and right images at the same time in the SDE, this team simply concatenated these two images along the channel dimension. However, these two images were separately processed to generate their features in the network, which were then fused with attentions blocks.

In addition, this team further improved the results by updating the score-network from U-Net to NAFNet [59]. To adaptively insert the scalar time into the network, this team constructed a simple MLP to learn two pairs of scale-shift parameters and apply them to the features with affine transforms. Such a network leads to better learning of score function conditioned on current state x_t, original low-quality image x_t, and time t. This team also employed Stereo Cross Attention modules [43] to fuse stereo features. Since the diffusion model needs the inputs to be the same size as the outputs, this team first upsampled LR images (both left and right) with $4\times$ scale factor, and then ran the Reffusion method on the HR images.

Training Settings. In the training stage, HR images were cropped to 128×128 while LR images were cropped to 32×32 but then upscaled to 128×128 using bicubic interpolation. Random rotation was performed to augment the training data. The proposed Reffusion model was trained using an L1 loss and optimized using Lion [**?**] with β_1=0.9, β_2=0.99 and a batch size of 4. The learning rate was initially set to 3×10^{-5} and decreased with a Cosine scheduler. The diffusion step was set to 100. All models were trained on an A100 GPU for about 3 days.

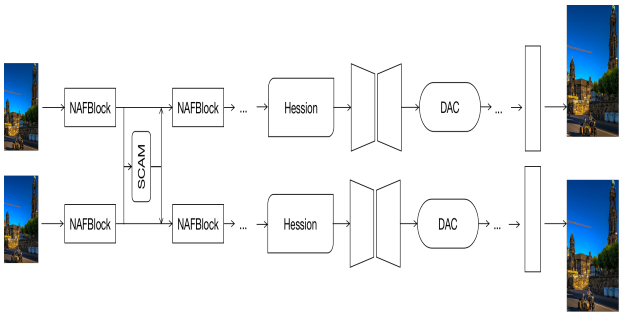


Figure 21. jingxiangchen1: The network architecture of the proposed HNAFSSR.

5.18. jingxiangchen1 - Track 1, 2

This team took NAFSSR as the backbone and made several improvements. As shown in Fig. 21, this team added the Hession filter after the baseline to further extract the high-frequency information of the image. Then, the extracted details are expanded and fed to the encoder and decoder modules. The extracted high-frequency details and the image are aligned through an alignment module.

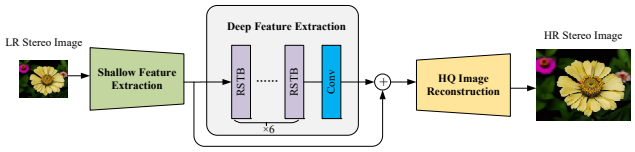


Figure 22. JNU_620: The network architecture of SwinIR.

5.19. JNU_620 - Track 1, 2, 3

This team considered the stereo image SR task as an single image SR task, and developed their method based on SwinIR [26], as shown in Fig. 22. For different tracks, they proposed different losses for network training.

For track 1, they introduced a back-projection (BP) loss for optimization. The BP loss enforced that the downscaled version of the super-resolved images match the LR observations. Therefore, this team leveraged bicubic downsampling to ensure that the projection of the estimated HR image is consistent with the original LR one. The BP loss is defined as

$$L_{BP} = \left|\left|S(\hat{I}_{HR}, s) - I_{LR}\right|\right|_1, \tag{7}$$

where S denotes the bicubic downsampling operation and s represents the downscale factor (*i,e.*, 4). Therefore, the overall loss function at the first stage can be formulated as:

$$L_{total-1} = L_1 + \lambda L_{BP}, \tag{8}$$

where λ is set to 1.0 in this method.

For track 2, they introduced an edge loss for optimization. To generate realistic textures while reducing artifacts, a Laplacian edge detector was used to extract texture information on the luminance component of the super-resolved images. By applying the Laplacian operator to the luminance component of the estimated SR result and the original HR image, their edge images can be obtained and compared. Accordingly, the edge loss can be calculated as:

$$L_{edge} = \left|\left|LAP(\hat{I}_{HR}) - LAP(I_{HR})\right|\right|_1, \tag{9}$$

where $\hat{I}_{HR}$ is the estimated HR image, I_{HR} is the HR image and $LAP(\cdot)$represents the edge extraction module with Laplacian operator. Therefore, the overall loss function can be written as:

$$L_{total-2} = L_1 + \lambda L_{edge}, \tag{10}$$

where λ is set to 1.0 in this method.

For track 3, only an L1 loss was used for training.

During the training phase, HR images were randomly cropped into 192×192 patches, while LR images were cropped accordingly. Random flipping was used for data augmentation. In the testing phase, a test-time data ensemble strategy was adopted to improve the performance.

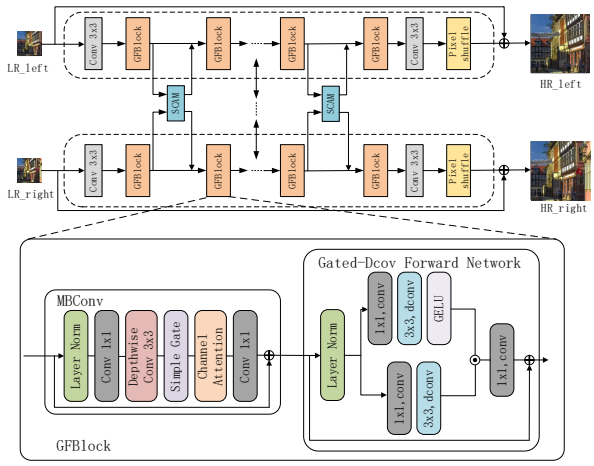


Figure 23. JXNU_SR: The network architecture of the proposed GFNet.

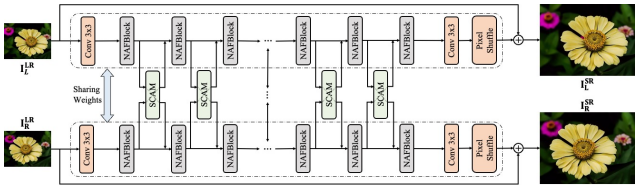


Figure 24. LongClaw: The network architecture of NAFSSR.

Note that, the window size was set to 12 in this method for tracks 1 and 3, which is different from the setting (*i,e.*, 8) in SwinIR.

5.20. JXNU_SR - Track 2

This team proposed a Gated Feature Net (GFNet), as shown in Fig. 23. In their method, PAM was used for cross-view feature interaction, and gated blocks were employed for feature learning. To better control the flow of features, this team improved the subsequent modules by using Gated-Dconv Forward Network to achieve a balance between efficiency and performance. The Gated-Dconv Forward Network consists of deep convolution, GELU activation function, and dot product, which can selectively suppress the flow of less informative features, allowing for better flow of more informative features. In the training phase, this team used the same setting as NAFSSR-S, *i.e.*, setting the number of channels to 64 and the number of module stacks to 32.

5.21. LongClaw - Track 1, 2

This team used the NAFSSR architecture (Fig. 24) with a combination of L1 loss, perceptual loss and texture loss. The perceptual loss was derived from the last layer of a VGG-16 model, and the texture loss was derived from a

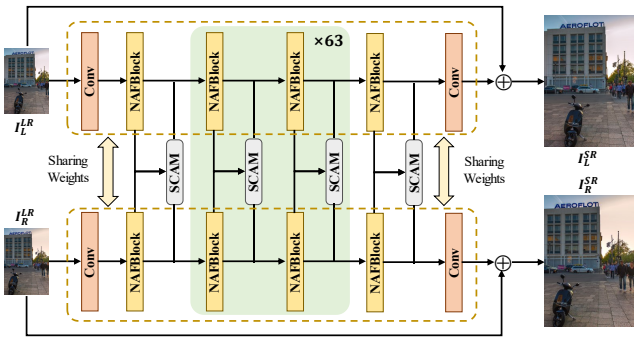


Figure 25. LVGroup_HFUT: The network architecture of the method proposed by the LVGroup_HFUT team.

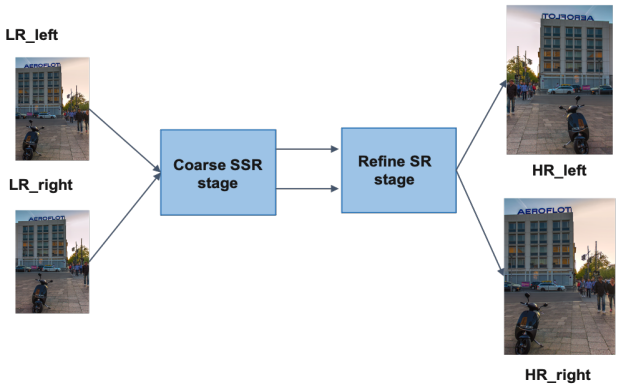


Figure 26. MakeStereoGreatAgain: The network architecture of DSSR.

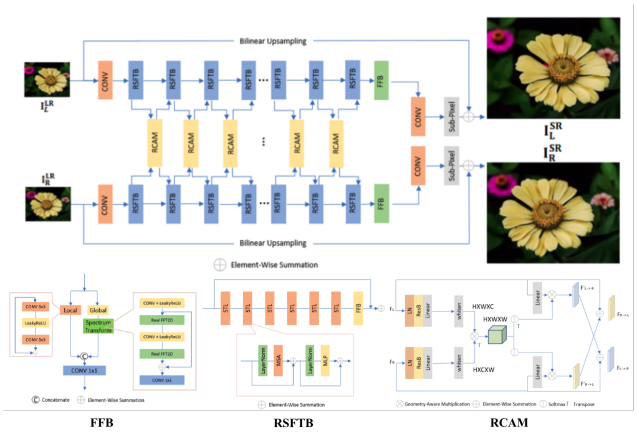


Figure 27. McSR: The network architecture of the proposed SwinFSR.

VGG-19 model. The Gram matrix was calculated for those layers and then used to calculate the MSE loss. The texture loss helps to capture different textures, while the perceptual loss helps to capture different structures of the image. This team used the NAFSSR-L $4\times$ model, which was trained for 240000 iterations with a batch size of 8. The weights assigned to the L1 loss, the perceptual loss and the texture loss were 1, 1×10^{-4}, and 5, respectively.

5.22. LVGroup_HFUT - Track 1, 2, 3

This team used NAFSSR for stereo image SR, as shown in Fig. 25. Only an L1 loss was used for training in all the three tracks. This model was optimized using the AdamW 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. Besides, this model was trained on 40×100 patches with a batch size of 10 for 3×10^5 iterations. These patches were randomly flipped horizontally and vertically for data augmentation. The RGB channels were also randomly shuffled for color augmentation. The stochastic depth with 0.3 probability was employed to overcome the over-fitting issue.

5.23. MakeStereoGreatAgain - Track 1, 3

The overall framework of the proposed method is shown in Fig. 26. This team developed a bi-directional alignment (BDA) module to effectively interact cross-view information. Then, this team designed a Siamese network equipped with a BDA to super-resolve both sides of views in a highly symmetric manner. Next, the super-resolved images were further fed into a refinement module for post-processing. Finally, this team adopted a self-ensemble strategy to further improve the SR performance.

The proposed network was implemented in PyTorch framework. All models were optimized using the Adam method with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 36. The initial learning rate was set to 2×10^{-4} and reduced to half after every 30 epochs. The training was stopped after 100 epochs since more epochs do not provide further consistent improvement.

5.24. McSR - Track 1, 2

This team proposed a SwinFSR [71] based on SwinIR [26]. The overall framework is shown in Fig. 27. Instead of directly using residual Swin Transformer blocks (RSTBs) for feature extraction in SwinIR, they modified them by explicitly incorporating the frequency domain knowledge and proposed a residual Swin Fourier transformer block (RSFTB). Specifically, they introduced a Fast Fourier Convolution Block (FFB) in the RSTBs to extract global features. As shown in Fig. 27, the FFB consists of two branches, *i.e.*, a local spatial convolution on the left and a global fast Fourier convolution spectrum transform on the right [53, 72]. The outputs from these two branches are then concatenated and passed to a convolution to generate the final result. In addition, they further proposed a cross-attention module, namely RCAM, to exploit cross-view in-

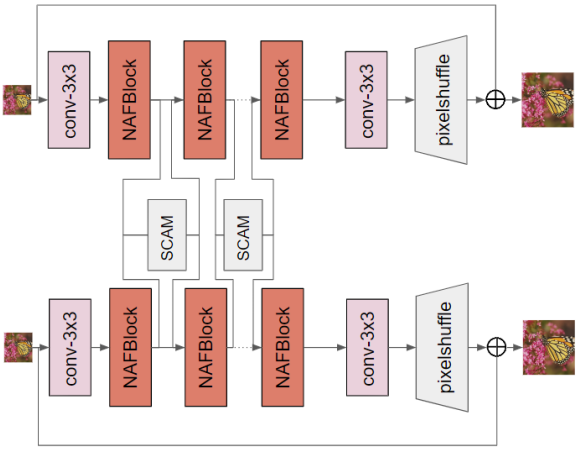


Figure 28. NTU607-stereo: The network architecture of the proposed improved NAFSSR.

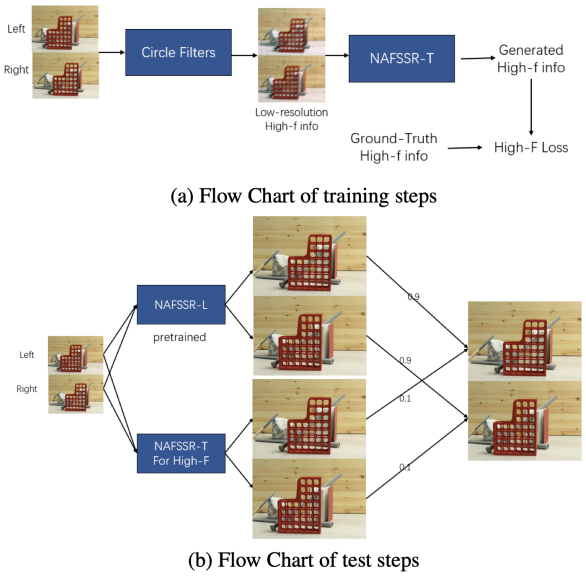


Figure 29. NUSSZ-STEREO: Overview of the proposed method during training phase (a) and test phase (b).

formation.

During the training phase, this team used an L1 loss in track 1 for the convenience. In Track 2, an additional perceptual loss was employed to provide supervision in the feature space.

Training Settings. During training, several data augmentation strategies were adopted, such as random cropping, flipping, and RGB channel shuffling. To address the train/test inconsistency issue and the epipolar stereo disparity in the stereo SR task, the local window in the Swin Transformer was enlarged to process large rectangular patches. Model ensemble was conducted by averaging the results produced by 3 models with the highest PSNR scores during training. Moreover, data ensemble was also utilized through horizontal and vertical flipping.

5.25. NTU607-stereo - Track 3

This team proposed a method based on the NAFSSR [43]. As shown in Fig. 28, NAFSSR-B [43] was used as the backbone and a three-stage strategy was used for training. First, the images were cropped to 30×90 patches for training. Random flipping (vertically and horizontally), RGB channel shuffling, Gaussian noise, blending, and CutBlur [56] were used for data augmentation. The Adam optimizer [73] with $\beta_1 = 0.98$, $\beta_2 = 0.92$, $\beta_3 = 0.99$, and weight decay of 0.002 was utilized to train the network for 100000 iterations. Batch size was set to 8. The learning rate was initialized as 3×10^{-3} and updated using a cosine annealing scheduler. An L2 loss function was used as the loss function.

Second, the AdamW optimizer [74] with weight decay being reduced to 0 was utilized to train the resultant model for another 75000 iterations. Batch size was set to 4. The learning rate was decreased from 3×10^{-5} to 5×10^{-8} using a cosine annealing scheduler.

Third, the images was cropped to 45×120 patches. Only random horizontal flipping was used as for data augmentation. The AdamW optimizer [74] with weight decay being reduced to 0 was utilized to train the resultant model for another 50000 iterations. Batch size was set to 2. The learning rate was decreased from 3×10^{-6} to 5×10^{-8} using a cosine annealing scheduler. All experiments were conducted on four Nvidia RTX 3090 GPUs and it took 2 days to train the model.

During the test phase, data ensemble was used to produce the SR results and TLC [75] was employed to further boost the performance.

5.26. NUSSZ-STEREO - Track 1, 2

Given that the primary challenge of image SR is recovering high-frequency image details, this team proposed a network that processes high-frequency information separately. The overall architecture is shown in Fig. 29. Specifically, high-frequency information were first extracted from the input images and fed to NAFSSR-T. Moreover, an additional focal frequency loss [76] was used to optimize this branch. During the test phase, input LR images were passed NAFSSR-L and NAFSSR-T to produce SR results, respectively. Then, these results were aggregated to obtain the final result:

$$I_{final} = \alpha * I_{original} + \beta * I_{high\_frequency}, \qquad (11)$$

where $\alpha = 0.99$ and $\beta = 0.01$, $I_{original}$ represents images is from NAFSSR-L and $I_{high\_frequency}$ is generated

by NAFSSR-T. The overall loss function was defined as:

$$L_{highf} = L_1 + \lambda * L_{local}, \tag{12}$$

where L_{local} is detailed in [76] and λ is a weighting hyper-parameter.

Training Settings. Images were first cropped into patches of size 30×90, and circular filters were applied to extract high-frequency information. A baseline model (NAFSSR-L [43]) was trained to super-resolve the images while another variant (NAFSSR-T) was used to generate high-frequency details. For NAFSSR-L, the AdamW optimizer and a cosine annealing scheduler were used for training. For NAFSSR-T, the same optimizer and scheduler were adopted.

5.27. SSSL - Track 2

It has been demonstrated that using a single perceptual loss is insufficient to accurately restore diverse textures in images, often generating undesirable artifacts or unnatural results. To remedy this, combinations of perceptual, adversarial, and distortion losses have been studied. However, it remains challenging to find the optimal combinations. Hence, this team proposed a framework that applies optimal objectives for different regions to generate plausible results [77], as illustrated in Fig. 30. Specifically, the framework comprises a predictive model C_ψ and a generative model G_θ, parameterized by ψ and θ, respectively. Model C_ψ infers an LR-sized optimal objective map $\hat{\mathbf{T}}_\mathbf{B}$ for the input LR input x, and G_θ applies it to produce the corresponding SR result:

$$\hat{y}_{\hat{\mathbf{T}}_B} = G_\theta \left(x | \hat{\mathbf{T}}_B \right), \tag{13}$$

$$\hat{\mathbf{T}}_B = C_\psi \left(x \right). \tag{14}$$

The generative model is trained over the proposed objective trajectory, which allows for synthesizing diverse SR results corresponding to different loss combinations. The predictive model is trained using pairs of LR images and corresponding optimal objective maps searched from the objective trajectory.

The network of G_θ consists of an SR branch with 23 basic blocks and a condition branch, as shown in Fig. 30. The condition branch takes an LR-sized target objective map $\mathbf{T}$ and produces shared intermediate conditions that can be transferred to all the SFT layers in the SR branch. The architecture of C_ψ consists of a feature extractor using VGG-19 [78] and a predictor using the Vision Transformer architecture [26].

5.28. STSR Sharpeners - Track 1

This team proposed a hybrid parallax-attention and self-attention network for stereo SR, namely HPSANet. The proposed network is an attempt to fuse a Transformer architecture with cross-view interactions. Instead of densely inserting the cross-view interaction blocks in the network like NAFSSR [43], they used one cross-view interaction for each residual group. The overall network architecture is shown in Fig. 31. HPSANet consist of three parts: 1) shallow feature extraction; 2) deep feature extraction; 3) SR reconstruction. The shallow feature extraction is a 3×3 convolution that increases the number of channels from 3 to 128. The deep feature extraction consists of 6 residual parallax-attention and self-attention blocks (RPSBs) and a 1×1 convolution followed by a long-skip connection with the shallow features. The SR reconstruction is achieved by a pixel-shuffle layer.

The key component of the proposed HPSANet is the residual parallax-attention and self-attention block (RPSB). In RPSB, the features are first fed into one bi-directional parallax attention module (biPAM) with one preceded mobile convolution module (MBC) and one MBC. The biPAM in RPSB is a simplified version of PAM in iPASSR [39]. For the MBC, the structure is similar to that of MBC in NAFSSR [43], but with Leaky ReLU activation and efficient channel attention mechanism being applied. Then, the features are processed by 5 window-based self-attention modules (WSAs), where each WSA is followed by an MBC. The WSA in RPSB is similar to that of WSA in SwinIR [26], but a simple circular shift mechanism [79] was adopted for each WSA and the window size was set to 12.

Training Settings. The training process is divided into two stages. For the first stage, LR images were cropped into patches of size 30×90 with a stride of 20, and HR images are cropped accordingly. The learning rate was initially set to 1.5×10^{-3} and decayed to 1×10^{-7} gradually with a cosine annealing strategy. The number of iterations in this stage was 10^5. For the second stage, the size of the LR patches was increased to 48×120. The learning rate was initially set as 2×10^{-4} and decayed to 1×10^{-7} using a cosine annealing strategy. The number of iterations in this stage was 5×10^4. The models were optimized by the AdamW method with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Meanwhile, data augmentation including random vertical/horizontal flipping and RGB channel shuffling was adopted.

5.29. XY - Track 1

The winner of the 2022 NTIRE Stereo Image SR Challenge [1] (NAFSSR [43]) achieved remarkable SR results. However, it still has a limitation that the physical characteristics of stereo images are not well exploited. Previous stereo image SR methods often assigned physical interpretations to the attention maps in the PAM module, resulting in a better fit for the problem and utilization of the stereo images' characteristics. Inspired by these methods, this team made a simple yet effective improvement to NAF-

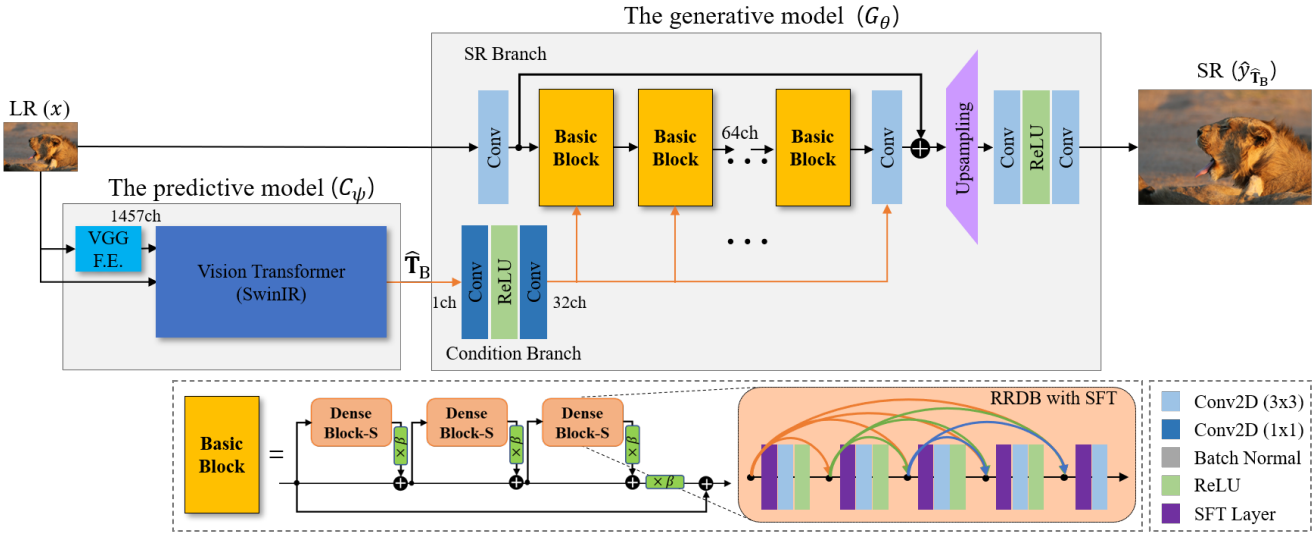


Figure 30. SSSL: The network architecture of the proposed SROOE-ViT.

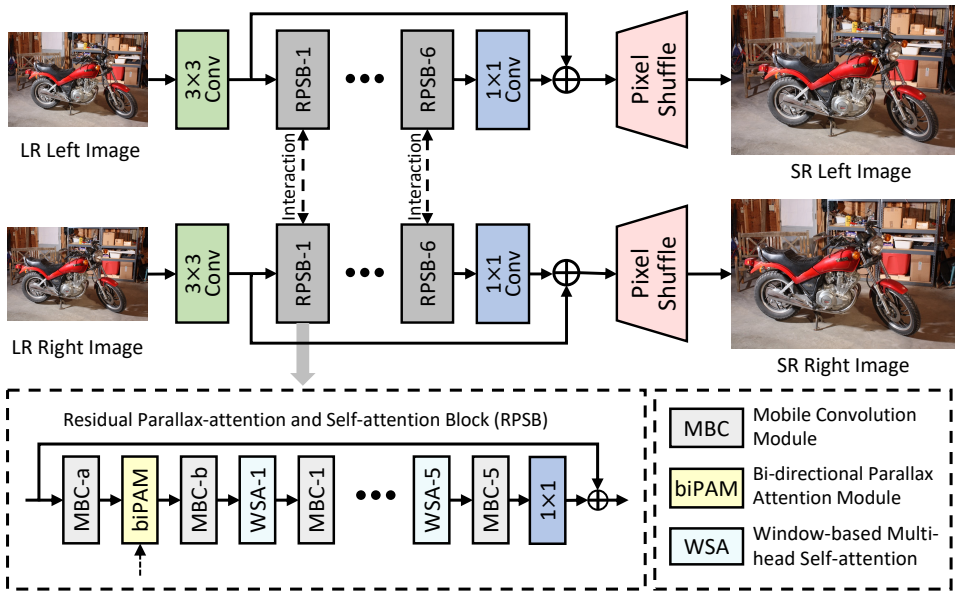


Figure 31. STSR Sharpeners: The network architecture of the proposed HPSANet.

SSR. Specifically, they added a sparse constraint to the attention map in PAM. Note that, this additional constraint does not introduce any computational burden during testing. As shown in Fig. 32, they also replaced the 1×1 convolution with a combination of a 3×3 depth-wise convolution and a 1×1 convolution to improved the performance of the model since 1×1 convolutions may not capture enough spatial information.

Training Settings. During training, 40×100 patches were first cropped from the images with stride of 20. Then, RIR-NAFNet was trained for 600000 iterations on 8 GeForce RTX 3090 GPUs with a batch size of 32 using an L1 loss and an FFT loss [65]. The gradient accumulation strategy was used to enable training with such large batches. Besides, the stochastic depth strategy [23] with a drop rate of 0.3 was adopted to handle the over-fitting issue. Random horizontal/vertical flipping and random RGB channel shuffling were adopted for data augmentation. For inference, the same augmentation strategies are performed for test-times self-ensembling.

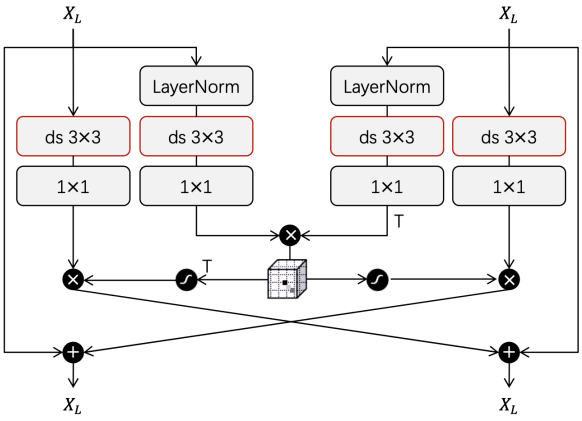


Figure 32. XY: The improved Stereo Cross Attention Module (SCAM) in Sparse NAFSSR.

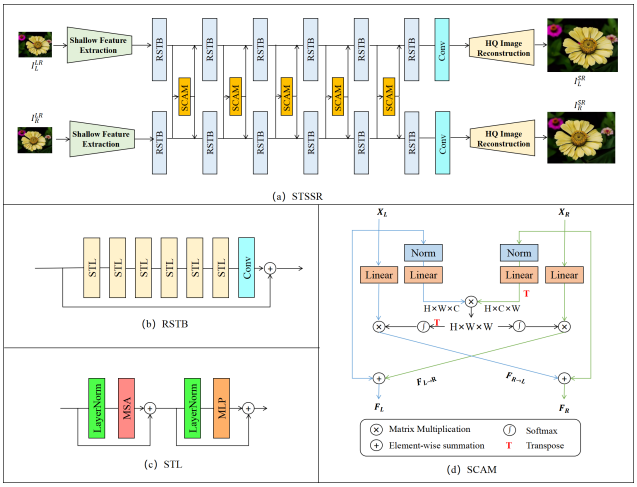


Figure 33. zzuli: The network architecture of the proposed STSSR.

5.30. zzuli - Track 3

This team proposed a two-stage training strategy for Stereo image Super-Resolution by using a modified Swin Transformer (STSSR). The proposed STSSR consists of two SwinIR branches along with stereo cross-attention modules (SCAMs). In the first stage, images from the two views were treated as independent images and fed into SwinIR for training. In the second stage, SCAMs were added after each residual Swin Transformer block (RSTB) to enable information interaction throughout the feature extraction process. Besides, the whole network was fine-tuned to fuse information between the left and right views. Experiments show that the second stage introduced a 0.2 dB PSNR improvement.

6. Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. U20A20185, 61972435), the Guangdong Basic and Applied Basic Research Foundation (2022B1515020103), the Shenzhen Science and Technology Program (No. RCYX20200714114641140, JCYJ20190807152209394), and the Guangdong Key Laboratory of Advanced IntelliSense Technology. 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).

7. Teams and Affiliations

NTIRE2023 team

Title: NTIRE 2023 Challenge on Stereo Image Super-Resolution

Members: Longguang Wang[1] (*wanglongguang15@nudt.edu.cn*), Yulan Guo[2], Yingqian Wang[3], Juncheng Li[4], Shuhang Gu[5], Radu Timofte[6]

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]University of Würzburg, ETH Zürich

(1) BSR - Track 1★

Title: Hybrid Transformer and CNN Attention Network (HTCAN)

Members: Ming Cheng[1] (*chengming.1129@bytedance.com*), Haoyu Ma[1], Qiufang Ma[1], Xiaopeng Sun[1], Shijie Zhao[1], Xuhan Sheng[1]

Affiliations:

[1]ByteDance

(2) TeamNoSleep - Track 1★

Title: Progressive and Robust Method for Stereo Super-Resolution

Members: Yukang Ding[1] (*dingyukang921@163.com*), Ming Sun[1], Xing Wen[1].

Affiliations:

[1]Kuaishou Technology

(3) SRC-B - Track 1★, 2★, 3★

Title: SwinFIRSSR

Members: Dafeng Zhang[1] (*dfeng.zhang@samsung.com*), Jia Li[1], Fan Wang[1], Zheng Xie[1].

Affiliations:

[1]Samsung Research China - Beijing

(4) SYSU_FVL - Track 1, 2☆, 3

Title: SwinIR-BiPAM, SC-NAFSSR, HAT
Members: Zongyao He[1] (*hezy28@mail2.sysu.edu.cn*), Zidian Qiu[1], Zilin Pan[1], Zhihao Zhan[1], Xingyuan Xian[1], Zhi Jin[1].
Affiliations:
[1]Sun Yat-sen University

(5) webbzhou - Track 1, 2★, 3

Title: SCGLANet
Members: Yuanbo Zhou[1] (*webbozhou@gmail.com*), Wei Deng[1], Ruofeng Nie[2], Jiajun Zhang[1], Qinquan Gao[1], Tong Tong[1]
Affiliations:
[1]Fuzhou University
[2]Imperial Vision Technology

(6) IPIU - Track 3★

Title: LTESSR
Members: Kexin Zhang[1] (*22171214672@stu.xidian.edu.cn*), Junpei Zhang[2], Rui Peng[3], Yanbiao Ma[4], Licheng Jiao[1]
Affiliations:
[1]Xidian University

(7) Team OV - Track 3☆

Title: RIR-NAFNet
Members: Haoran Bai[1] (*baihaoran@njust.edu.cn*), Lingshun Kong[1], Jinshan Pan[1], Jiangxin Dong[1], Jinhui Tang[1].
Affiliations:
[1]Nanjing University of Science and Technology

(8) BUPT-PRIV - Track 1, 2

Title: SwinIR-SCAM
Members: Pu Cao[1] (*caopu@bupt.edu.cn),*), Tianrui Huang[1], Lu Yang[1], Qing Song[1]
Affiliations:
[1]Beijing University of Posts and Telecommunications

(9) candle - Track 1

Title: Stereo Image Super-Resolution Network with Multiple Extraction Block (ME-super)
Members: Bingxin Chen[1] (*1559651555@qq.com*), Chunhua He[1], Meiyun Chen[1]
Affiliations:
[1]Guangdong University of Technology

(10) CHASE - Track 1

Title: Nonlinear Parallax Stereo Super-Resolution Network (NPSSR)
Members: Zijie Guo[1] (*gggzj0609@163.com*), Shaojuan Luo[1]
Affiliations:
[1]Guangdong University of Technology

(11) Chengzhi-Group - Track 2, 3

Title: Enhanced-SwinIR
Members: Chengzhi Cao[1] (*chengzhi-cao@mail.ustc.edu.cn*), Kunyu Wang[1], Fanrui Zhang[1], Qiang Zhang[1]
Affiliations:
[1]University of Science and Technology of China

(12) CV_IITRPR - Track 1, 2

Title: Stereo SR Network
Members: Nancy Mehta[1] (*2018eez0017@iitrpr.ac.in*), Subrahmanyam Murala[1], Akshay Dudhane[2]
Affiliations:
[1]Indian Institute of Technology Ropar
[2]MBZUAI Dubai

(13) DiffX - Track 2

Title: Preconditioned Diffusion Model with Two-way-tied Net (PDMTN)
Members: Yujin Wang[1] (*yujin.w96@gmail.com*), Lingen Li[1]
Affiliations:
[1]Shanghai Artificial Intelligence Laboratory

(14) GarasSjtu - Track 1

Title: Transformer Style ConvNet for Stereo Image Super-Resolution
Members: Garas Gendy[1] (*garasgaras@yahoo.com*), Nabil Sabor[2], Jingchao Hou[1], Guanghui He[1]
Affiliations:
[1]Shanghai Jiao Tong University
[2]Assiut University

(15) GDUT_506 - Track 1

Title: Efficient loss for boosting NAFSSR performance
Members: Junyang Chen[1] (*3117002384@mail2.gdut.edu.cn*), Hao Li[2], Yukai Shi[1], Zhijing Yang[1]
Affiliations:
[1]Guangdong University of Technology
[2]Nanjing University of Science and Technology

(16) Giantpandacv - Track 1, 2, 3

Title: Cross-View Hierarchical Network for Stereo Image Super-Resolution
Members: Wenbin Zou[1] (*alexzou14@foxmail.com*), Yunchen Zhang[2], Mingchao Jiang[3], Zhongxin Yu[2], Ming Tan[2], Hongxia Gao[1]
Affiliations:
[1]South China University of Technology

[2]Fujian Normal University
[3]GAC R&D Center

(17) IR-SDE - Track 2

Title: Refusion: Enabling Large-Size Realistic Image Restoration with Latent-Space Diffusion Models
Members: Ziwei Luo[1] (*ziwei.luo@it.uu.se*), Fredrik K. Gustafsson[1], Zheng Zhao[1], Jens Sjölund[1], Thomas B. Schön[1]
Affiliations:
[1]Uppsala University

(18) jingxiangchen1 - Track 1, 2

Title: Deformable Stereo Super-resolution
Members: Jingxiang Chen[1] (*jingxiangchen-09@gmail.com*), Bo Yang[1], XiSheryl Zhang[2], Chenghua Li[2]
Affiliations:
[1]Nanjing University of Information Science and Technology
[2]Chinese Academy of Sciences Institute of Automation

(19) JNU_620 - Track 1, 2, 3

Title: Improved Loss for Super-Resolution Based on SwinIR
Members: Weijun Yuan[1] (*yweijun@stu2022.jnu.edu.cn*), Zhan Li[1], Ruting Deng[1]
Affiliations:
[1]Jinan University

(20) JXNU_SR - Track 2

Title: Gated Feature Net (GFNet)
Members: Jintao Zeng[1] (*2608900429@qq.com*)
Affiliations:
[1]Jiangxi Normal University

(21) LongClaw - Track 1, 2

Title: NAFNet
Members: Pulkit Mahajan[1] (*2019UCS0073@iitjammu.ac.in*), Sahaj Mistry[1], Shreyas Chatterjee[1], Vinit Jakhetiya[1], Badri Subudhi[1], Sunil Jaiswal[2]
Affiliations:
[1]Indian Institute of Technology
[2]K|Lens GmbH

(22) LVGroup_HFUT - Track 1, 2, 3

Title: Stereo Image Super-Resolution with NAFNet
Members: Zhao Zhang[1] (*cszzhang@gmail.com*), Huan Zheng[1], Suiyi Zhao[1], Yangcheng Gao[1], Yanyan Wei[1], Bo Wang[1]
Affiliations:
[1]Hefei University of Technology

(23) MakeStereoGreatAgain - Track 1, 3

Title: Deformable Stereo Super-resolution
Members: Gen Li[1] (*leegeun@yonsei.ac.kr*), Aijin Li[1], Lei Sun[1]
Affiliations:
[1]Xidian University

(24) McSR - Track 1, 2

Title: SwinFSR
Members: Ke Chen[1] (*chenk59@mcmaster.ca*), Congling Tang[1], Yunzhe Li[1], Jun Chen[1]
Affiliations:
[1]McMaster University

(25) NTU607-stereo - Track 3

Title: NAFnet for stereo image super-resolution
Members: Yuan-Chun Chiang[1] (*jack06272@gmail.com*), Yi-Chung Chen[1], Zhi-Kai Huang[1], Hao-Hsiang Yang[1], I-Hsiang Chen[1], Sy-Yen Kuo[1].
Affiliations:
[1]National Taiwan University

(26) NUSSZ-STEREO - Track 1, 2

Title: NAFSSR-DA
Members (Track 1): Yiheng Wang[1] (*303291066@qq.com*), Gang Zhu[1], Xingyi Yang[2], Songhua Liu[2], Yongcheng Jing[3].
Members (Track 2): Yiheng Wang[1] (*303291066@qq.com*), Gang Zhu[1], Songhua Liu[2], Xingyi Yang[2], Yongcheng Jing[3].
Affiliations:
[1]National University of Singapore (Suzhou) Research Institute
[2]National University of Singapore
[3]University of Sydney

(27) XY - Track 1

Title: S-NAFSSR
Members: Xingyu Hu[1] (*huxingyu@hit.edu.cn*).
Affiliations:
[1]Harbin Institute of Technology

(28) STSR Sharpeners - Track 1

Title: HPSANet
Members: Jianwen Song[1,2] (*jianwen.song@unsw.edu.au*), Changming Sun[2,1], Arcot Sowmya[1].
Affiliations:
[1]University of New South Wales
[2]CSIRO Data61

(29) SSSL - Track 2

Title: SROOE-ViT
Members: Seung Ho Park[1] (*rakadian@gmail.com*).
Affiliations:
[1]Seoul National University

(30) zzuli - Track 3

Title: STSSR
Members: Xiaoyan Lei[1] (*xyan lei@163.com*), Jingchao Wang[1], Chenbo Zhai[1], Yufei Zhang[1], Weifeng Cao[1], Wenlong Zhang[2].
Affiliations:
[1]Zhengzhou University of Light Industry
[2]The Hong Kong Polytechnic 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. 2, 20

[2] 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, 3

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

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

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

[23] 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, 10, 21

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

[25] 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

[26] 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, 17, 18, 20

[27] 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, 7

[28] 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

[29] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 2

[30] 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

[31] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, and Jing-Hao Xue. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 2019. 2

[32] Zhihao Wang, Jian Chen, and Steven C.H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2020. 2

[33] Juncheng Li, Zehua Pei, and Tieyong Zeng. From beginner to master: A survey for deep learning-based single-image super-resolution. *arXiv*, 2021. 2

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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. 3, 10, 11, 20

[40] 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. 3, 10

[41] Chenxi Ma, Bo Yan, Weimin Tan, and Xuhao Jiang. Perception-oriented stereo image super-resolution. In *ACM MM*, pages 2420–2428, 2021. 3

[42] Hansheng Guo, Juncheng Li, Guangwei Gao, Zhi Li, and Tieyong Zeng. Pft-ssr: Parallax fusion transformer for stereo image super-resolution. 2023. 3

[43] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPRW*, pages 1239–1248, 2022. 3, 4, 6, 7, 10, 12, 13, 16, 19, 20

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 3

[45] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227, 2021. 3, 10

[46] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, pages 1865–1873, 2016. 5, 7

[47] Ming Cheng, Ma Haoyu, Ma Qiufang, Sun Xiaopeng, Li Weiqi, Zhang Zhenyu, Sheng Xuhan, Shijie Zhao, Junlin Li, and Li Zhang. Hybrid transformer and cnn attention network for stereo image super-resolution. In *CVPRW*, 2023. 6

[48] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023. 6

[49] Lei Li, Jingzhu Tang, Ming Chen, Shijie Zhao, Junlin Li, and Li Zhang. Multi-patch learning: looking more pixels in the training phase. In *ECCVW*, pages 549–560, 2023. 6

[50] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *CVPR*, pages 6002–6012, 2022. 7

[51] 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

[52] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*. 7

[53] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. volume 33, pages 4479–4488, 2020. 7, 18

[54] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[55] 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. 7

[56] 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. 7, 19

[57] Zidian Qiu, Zongyao He, Zhihao Zhan, Zilin Pan, Xingyuan Xian, and Zhi Jin. Sc-nafssr: Perceptual-oriented stereo image super-resolution using stereo consistency guided nafssr. In *CVPRW*, 2023. 7

[58] Yuanbo Zhou, Yuyang Xue, Wei Deng, Ruofeng Nie, Jiajun Zhang, et al. Stereo cross global learnable attention module for stereo image super-resolution. In *CVPRW*, 2023. 8

[59] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33, 2022. 9, 10, 16

[60] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pages 8628–8638, 2021. 9, 10

[61] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *CVPR*, pages 1929–1938, 2022. 10

[62] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 12

[63] Qibin Hou, Cheng-Ze Lu, Ming-Ming Cheng, and Jiashi Feng. Conv2former: A simple transformer-style convnet for visual recognition. *arXiv*, 2022. 12

[64] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv*, 2023. 12

[65] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. 12, 21

[66] Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. *NeurIPS*, 33:19964–19975, 2020. 12

[67] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Revisiting global statistics aggregation for improving image restoration. *arXiv*, 2021. 12

[68] 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*, 2023. 13

[69] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. *arXiv*, 2023. 15, 16

[70] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *CVPRW*, 2023. 15

[71] KE CHEN, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In *CVPRW*, 2023. 18

[72] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 18

[73] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *arXiv*, 2022. 19

[74] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv*, 2017. 19

[75] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *ECCV*, pages 53–71, 2022. 19

[76] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, pages 13919–13929, 2021. 19, 20

[77] Seung Ho Park, Young Su Moon, and Nam Ik Cho. Perception-oriented single image super-resolution using optimal objective estimation. In *CVPRW*, 2023. 20

[78] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 20

[79] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, pages 649–667, 2022. 20