

NTIRE 2022 Challenge on Stereo Image Super-Resolution: Methods and Results

Longguang Wang, Yulan Guo*, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Liangyu Chen, Xiaojie Chu, Wenqing Yu, Kai Jin, Zeqiang Wei, Sha Guo, Angulia Yang, Xiuzhuang Zhou, Guodong Guo, Bin Dai, Feiyue Peng, Huaxin Xiao, Shen Yan, Yuxiang Liu, Hanxiao Cai, Pu Cao, Yang Nie, Lu Yang, Qing Song, Xiaotao Hu, Jun Xu, Mai Xu, Junpeng Jing, Xin Deng, Qunliang Xing, Minglang Qiao, Zhenyu Guan, Wenlong Guo, Chenxu Peng, Zan Chen, Junyang Chen, Hao Li, Junbin Chen, Weijie Li, Zhijing Yang, Gen Li, Aijin Li, Lei Sun, Dafeng Zhang, Shizhuo Liu, Jiangtao Zhang, Yanyun Qu, Hao-Hsiang Yang, Zhi-Kai Huang, Wei-Ting Chen, Hua-En Chang, Sy-Yen Kuo, Qiaohui Liang, Jianxin Lin, Yijun Wang, Lianying Yin, Rongju Zhang, Wei Zhao, Peng Xiao, Rongjian Xu, Zhilu Zhang, Wangmeng Zuo, Hansheng Guo, Guangwei Gao, Tiejong Zeng, Huicheng Pi, Shunli Zhang, Joohyeok Kim, HyeonA Kim, Eunpil Park, Jae-Young Sim, Jucai Zhai, Pengcheng Zeng, Yang Liu, Chihao Ma, Yulin Huang, Junying Chen

Abstract

In this paper, we summarize the 1st NTIRE challenge on stereo image super-resolution (restoration of rich details in a pair of low-resolution stereo images) with a focus on new solutions and results. This challenge has 1 track aiming at the stereo image super-resolution problem under a standard bicubic degradation. In total, 238 participants were successfully registered, and 21 teams competed in the final testing phase. Among those participants, 20 teams successfully submitted results with PSNR (RGB) scores better than the baseline. This challenge establishes a new benchmark for stereo image SR.

1. Introduction

Stereo image pairs can encode 3D scene cues into stereo correspondences between the left and right images. With the popularity of dual cameras in mobile phones, autonomous vehicles and robots, stereo vision has attracted increasingly attention in both academia and industry. In many applications like AR/VR, and robot navigation, in-

creasing the resolution of stereo images is highly demanded to achieve higher perceptual quality and help to parse the real world.

In recent years, remarkable progress of image super-resolution (SR) have been witnessed with deep learning techniques. Most existing approaches focus on super-resolving single images. However, these methods cannot make full use of the cross-view information in stereo images. Recent CNN-based video SR methods incorporate optical flow estimation and SR in unified networks to exploit temporal information in multiple frames. Nevertheless, these methods usually suffer limited performance on stereo image SR since the disparity can be much larger than their receptive fields.

Stereo image SR aims to reconstruct a pair of high-resolution (HR) stereo images from a pair of low-resolution (LR) observations. Since disparities between stereo images can vary significantly for different baselines, focal lengths, depths and resolutions, it is highly challenging to incorporate stereo correspondence for stereo image SR.

The NTIRE 2022 stereo image SR challenge takes a step forward to establish a benchmark for stereo image SR. It uses the Flickr1024 dataset [1] and employs standard bicubic degradation.

This challenge is one of the NTIRE 2022 associated challenges: spectral recovery [2], spectral demosaicing [3], perceptual image quality assessment [4], inpainting [5], night photography rendering [6], efficient super-resolution [7], learning the super-resolution space [8], super-resolution and quality enhancement of compressed video [9], high dynamic range [10], stereo image super-

*Corresponding author: Yulan Guo (yulan.guo@nudt.edu.cn).

Section 7 provides the authors and affiliations of each team.

NTIRE 2022 webpage: <https://data.vision.ee.ethz.ch/cvl/ntire22/>

Challenge webpage: <https://codalab.lisn.upsaclay.fr/competitions/1598>

Leaderboard: <https://codalab.lisn.upsaclay.fr/competitions/1598#results>

Github: <https://github.com/The-Learning-And-Vision-Atelier-LAVA/Stereo-Image-SR/tree/NTIRE2022>

resolution, and burst super-resolution [11].

2. Related Work

In this section, we briefly review several major works on single image and stereo image SR.

2.1. Single Image SR

Single image SR is a long-standing problem and has been investigated for decades. In the past 10 years, deep learning-based single image SR methods have achieved promising performance.

Dong et al. [12] proposed the first CNN-based SR network called SRCNN to reconstruct HR images from LR inputs. Kim et al. [13] proposed a deeper network with 20 layers (*i.e.*, VDSR) to improve SR performance. Afterwards, SR networks became increasingly deep and complex, and thus more powerful in intra-view information exploitation. Lim et al. [14] proposed an enhanced deep SR network (*i.e.*, EDSR) using both local and residual connections. Zhang et al. [15] combined residual connection [16] with dense connection [17], and proposed residual dense network (*i.e.*, RDN) to fully use hierarchical feature representations for image SR. Subsequently, Zhang et al. [18] further improved SR performance by designing a residual-in-residual network with channel attention. Li et al. [19] suggested to make full use of image features with different scales and proposed a multi-scale residual network (*i.e.*, MSRN). Recently, Transformer has been widely used in computer vision and achieved promising performance. In the area of low-level vision, Liang et al. [20] applied Swin Transformer [21] to image restoration, and designed a SwinIR network to achieve state-of-the-art performance on single image SR. Lu et al. [22] proposed an effective super-resolution Transformer (*i.e.*, ESRT) for SISR, which reduces GPU memory consumption through a lightweight Transformer and feature separation strategy. Readers can refer to recent surveys [23–25] to learn more details about single image SR.

2.2. Stereo Image SR

Compared to single image SR in which only context information within one view is available, stereo image SR can use the additional information provided by the second view (*i.e.*, cross-view information) to improve SR performance. However, since an object is projected onto different locations in a stereo image pair, the cross-view information is hindered to be fully exploited.

To handle this disparity issue, Jeon et al. [26] proposed a network (*i.e.*, StereoSR) to learn a parallax prior by jointly training two cascaded sub-networks. The cross-view information is integrated by concatenating the left image and a stack of right images with different pre-defined shifts. Wang et al. [27, 28] proposed a parallax attention module (PAM)

to model stereo correspondence with a global receptive field along the epipolar line. The proposed PASSRnet achieves better performance than StereoSR and is more flexible with disparity variation. Based on parallax attention mechanism, Ying et al. [29] proposed a stereo attention module and embedded it into pre-trained SISR networks for stereo image SR. Song et al. [30] combined self-attention with parallax attention and proposed a SPAMnet for stereo image SR. Yan et al. [31] proposed a domain adaptive stereo SR network (DASSR) in which the disparity was firstly estimated by using a pretrained stereo matching network and the views were warped to the other side to incorporate cross-view information. Xu et al. [32] incorporated the idea of bilateral grid processing in a CNN framework and proposed a bilateral stereo SR network.

More recently, Wang et al. [33] modified PAM [27] to be bidirectional and symmetric, and 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. [34] proposed a feedback network to alternately solve disparity estimation and stereo image SR in a recurrent manner. Ma et al. [35] proposed a GAN-based perception-oriented stereo image SR method that can generate visually pleasing and stereo consistent details. Xu et al. [36] tackled the stereo video SR problem by simultaneously utilizing both cross-view and temporal information.

3. NTIRE 2022 Challenge

The objectives of the NTIRE 2022 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

The Flickr1024 dataset [1] is used in the challenge. Flickr1024 has 1024 pairs of RGB images with 800 for training, 112 for validation and 112 for testing purposes. The manually collected high quality images in Flickr1024 have diverse contents and rich details. In this challenge, we use Flickr1024 for both training and validation, and collect another 100 LR stereo image pairs (with private groundtruth HR images) for test.

3.2. Track and Competition

Track: Bicubic degradation. Standard bicubic degradation (Matlab *imresize* function with default settings) is used to synthesize LR stereo images from HR ones for both training, validation and test sets.

Challenge phases

(1) **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.

(2) 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.

Evaluation protocol. The quantitative metrics are Peak signal-to-noise ratio (PSNR) in deciBels [dB] and the structural similarity index (SSIM). These full-reference measures are calculated in the RGB and Y (luminance) channels, respectively. Results are averaged over all images (for both left and right images).

4. Challenge Results

Among the 238 registered participants, 21 teams successfully participated the final phase and submitted their results, codes, and factsheets. Table 1 reports the final test results, rankings of the challenge, and major details from the factsheets of 20 teams with PSNR (RGB) scores outperforming the baseline. These methods are briefly described in Section 5 and the team members are listed in Appendix 7.

Architectures and main ideas. All the proposed methods are based on deep learning techniques. Transformers (particularly SwinIR) are used in 16 solutions as the basic architecture. To exploit cross-view information, parallax-attention mechanism (PAM) are adopted in 14 solutions to capture stereo correspondence.

Restoration fidelity. The top 2 methods, (*i.e.*, The Fat, The Thin and The Young team and the BigoSR team), achieved similar PSNR scores (with a difference less than 0.08dB). The BUAA-MC2 entry, which ranks 6th, is only 0.21dB behind the best PSNR score of The Fat, The Thin and The Young team.

Data Augmentation. Widely applied data augmentation approaches such as random flipping are used for most solutions. In addition, random horizontal shifting, random RGB channel shuffling and Cutblur [37] 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 several 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. For model ensemble, the results produced by multiple models are averaged for better results.

Conclusions. By analyzing the settings, the proposed methods and their results, we can conclude that: 1) The proposed methods improve the state-of-the-art in stereo image SR. 2) Transformers are increasingly popular in stereo image SR

tasks and produce significant performance improvements over CNNs. 3) Cross-view information lying at varying disparities is critical to the stereo image SR task and helps to achieve higher performance. 4) Benefited from bags of tricks including delicate data augmentation strategies, several single image SR solutions also produces competitive results.

5. Challenge Methods and Teams

5.1. The Fat, The Thin and The Young Team

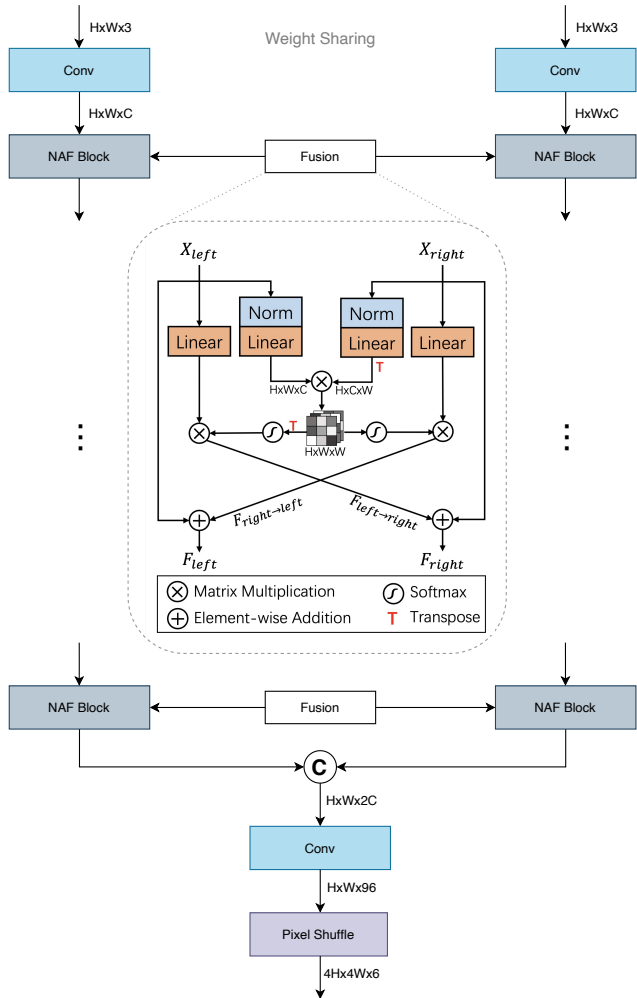


Figure 1. The Fat, The Thin and The Young Team: The network architecture of the proposed Nonlinear Activation Free Stereo image SR network (NAFSSR).

This team proposed a Nonlinear Activation-Free Network (NAFNet) for image restoration [38]. By using the modules in NAFNet for feature extraction, they further extended NAFNet to NAFSSR for stereo image SR, by adding cross attention modules to incorporate cross-view informa-

Table 1. NTIRE 2022 Stereo Image SR Challenge results, final rankings, and details from the factsheets. Note that, PSNR (RGB) is used for the final ranking. “Transf” denotes Transformer, “PAM” denotes parallax attention mechanism, and “DConv” represents deformable convolutions.

Rank	Team	Authors	PSNR (RGB)	PSNR (Y)	SSIM (RGB)	SSIM (Y)	Transf?	Disparity	Ensemble
1	The Fat, The Thin and The Young	L. Chen, X. Chu, W. Yu	23.7873	25.2033	0.7360	0.7438	✗	PAM	Data & Model
2	BigoSR	K. Jin, Z. Wei, S. Guo, et al.	23.7126	25.1305	0.7295	0.7379	✓	PAM	Data & Model
3	NUDT-CV&CPLab	B. Dai, F. Peng, H. Xiao, et al.	23.6007	25.0166	0.7287	0.7366	✓	PAM	✗
4	BUPT-PRIV	P. Cao, Y. Nie, L. Yang, Q. Song	23.5983	25.0100	0.7217	0.7296	✓	✗	Data & Model
5	NKU_caroline	X. Hu, J. Xu	23.5770	24.9978	0.7263	0.7352	✓	PAM	Data
6	BAAA-MC2	M. Xu, J. Jing, X. Deng, et al.	23.5733	24.9861	0.7267	0.7349	✓	Optical Flow	Data
7	No War	W. Guo, C. Peng, Z. Chen	23.5664	24.9864	0.7233	0.7330	✓	PAM	Data & Model
8	GDUT_506	J. Chen, H. Li, J. Chen, et al.	23.5601	24.9789	0.7239	0.7325	✓	PAM	Data
9	DSSR	G. Li, A. Li, L. Sun	23.5533	24.9711	0.7242	0.7322	✓	PAM & DConv	Data
10	xiaozhazha	D. Zhang, S. Liu	23.5490	24.9570	0.7203	0.7290	✓	✗	Data & Model
11	Zhang9678	J. Zhang, Y. Qu	23.5150	24.9346	0.7183	0.7263	✓	PAM	✗
12	NTU607QCO-SSR	H. Guo, Z. Huang, W. Chen, et al.	23.5090	24.9190	0.7186	0.7265	✓	✗	Data
13	supersmart	Q. Liang	23.4896	24.9058	0.7227	0.7331	✓	PAM	Data
14	LIMMC_HNU	J. Lin, Y. Wang, L. Yin, et al.	23.4381	24.8550	0.7199	0.7283	✓	PAM	✗
15	HIT-IL	R. Xu, Z. Zhang, W. Zuo	23.4066	24.8165	0.7144	0.7225	✓	✗	Data
16	Hansheng	H. Guo, G. Gao, T. Zeng	23.2918	24.7072	0.7101	0.7194	✓	PAM	Model
17	VIP-SSR	J. Kim, H. Kim, E. Park, J. Sim	23.2910	24.7146	0.7103	0.7207	✗	PAM	Data
18	phc	H. Pi, S. Zhang	23.2323	24.6584	0.7071	0.7182	✗	PAM	✗
19	qylen	J. Zhai, P. Zeng, Y. Liu, C. Ma	23.2241	24.6480	0.7086	0.7179	✓	PAM	✗
20	Modern_SR	Y. Huang, J. Chen	22.8370	24.2836	0.6820	0.6925	✗	DConv	✗
-	PASSRnet (Baseline)	-	22.7965	24.2016	0.6801	0.6911	✗	PAM	✗
-	Bicubic (Baseline)	-	21.8358	23.3865	0.6287	0.6443	-	-	-

tion. In this report, we briefly introduce their solution and readers can refer to [39] for more details.

As shown in Fig. 1, NAFSSR has two branches with shared weights to process left and right views, respectively. Several attention modules are inserted between the left and right branches to interact cross-view information. Similar to biPAM [33], the attention module calculates the correlation of features along the horizontal epipolar line, and then fuses the features by performing correlation operation.

In addition to the network design, a series of effective tricks were introduced to boost the SR performance. Specifically, in the training phase, random cropping, random horizontal and vertical flipping, random horizontal shifting and random RGB channel shuffling were performed for data augmentation. In the testing phase, four models were used for ensemble, and a series of test-time augmentation approaches, including horizontal and vertical flipping, RGB channel shuffling, and left-right view exchanging, were performed.

Moreover, this team addressed the training/test inconsistency issue described in [40], *i.e.*, the training is performed on image patches while testing is performed on full image. The local-SE module in [40] was adopted in their solution and introduced a 0.1 dB PSNR improvement. Besides, the stochastic depth strategy [41] and the skip-init strategy [42] were used to handle the over-fitting issue and facilitate the training process.

5.2. The BigoSR Team

The BigoSR team developed a SwiniPASSR network by combining the Swin Transformer [21] with the parallax-attention mechanism [28]. To use the cross-view informa-

tion from paired LR images, they employed biPAM [33] in their network. SwiniPASSR consists of three parts including feature extraction, cross-view interaction and reconstruction, as illustrated in Fig. 2. Within the SwinIR-like framework, a biPAM module is plugged into the middle of consecutive residual swin Transformer blocks (RSTBs) to model cross-view information while handling occlusion and boundary issues. To keep semantic structure consistency with convolution-based biPAM module, a layer normalization and a patch unembedding module are used before biPAM.

During the training phase, to facilitate the learning of stereo correspondence, a multi-stage training strategy was employed. In the first stage, stereo image pairs in the training set were divided into separate images and a Swin Transformer based network was trained for the single image SR task. At this stage, the network aims to learn structured information of images and model local spatial relationship. In the second stage, the biPAM module was plugged into middle of RSTBs to model stereo correspondence between a stereo image pair. In the third stage, the input patch size were further enlarged from 24×24 to 48×48 to help biPAM to aggregate cross-view information at a larger range. In the last stage, the stereo losses in the overall loss function were increased by 10 times for fine-tuning to encourage the network to focus more on cross-view information.

5.3. The NUDT-CV&CPLab Team

Inspired by SwinIR, the NUDT-CV&CPLab team proposed a Transformer-based network architecture (namely, SSRFormer) for stereo image SR, as shown in Fig. 3. SSRFormer is a Siamese network architecture with two branches

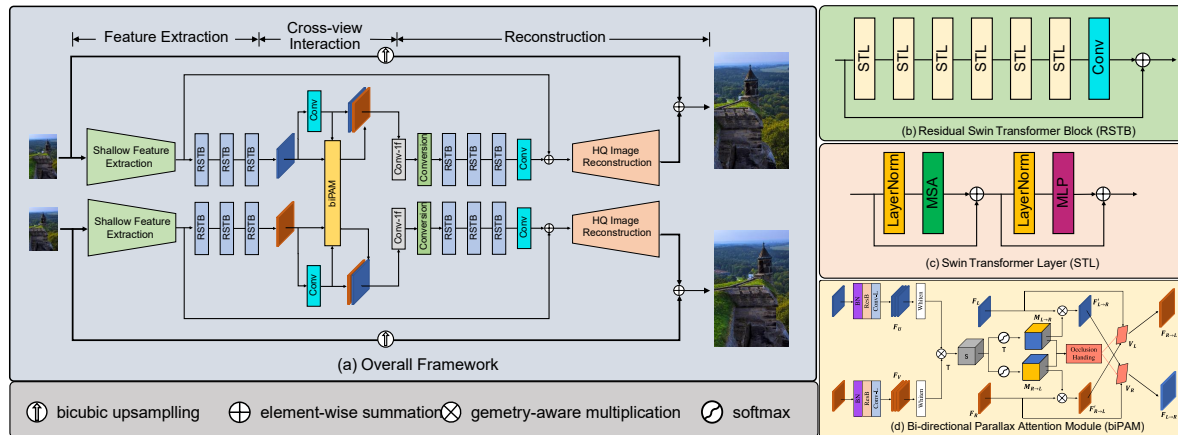


Figure 2. The BigoSR Team: The network architecture of the proposed SwiniPASSR.

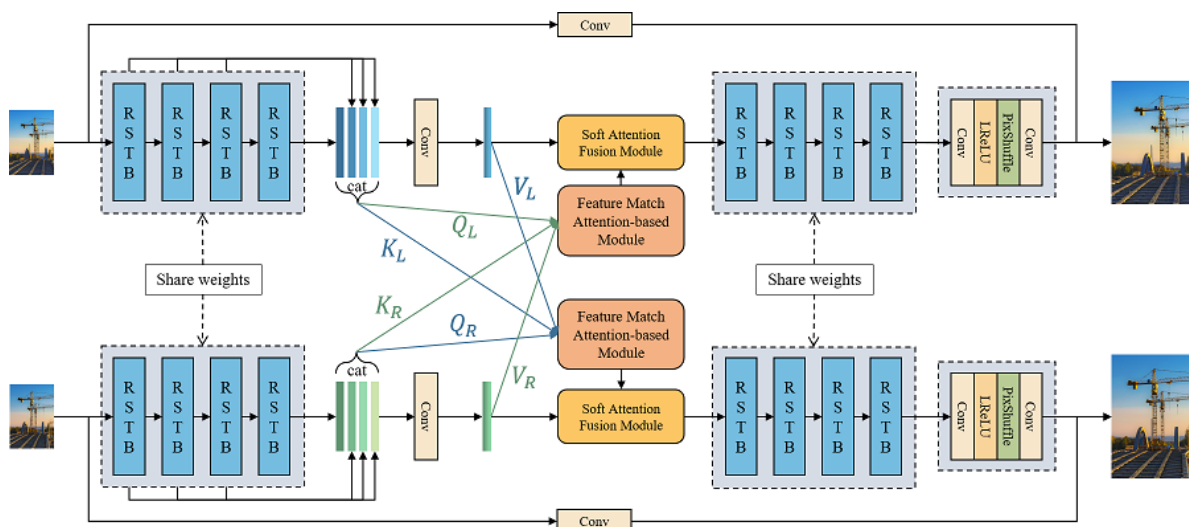


Figure 3. The NUDT-CV&CPLab Team: The network architecture of the proposed SSRFormer.

sharing weights. Specifically, four residual Swin Transformer blocks (RSTBs) are first used as the feature extractor to extract deep features. Then, inspired by parallax attention mechanism, an attention-based feature matching (AFM) module is adopted to extract rich cross-view information without explicitly aligning the left and right images.

During the training phase, 800 pairs of stereo images were used as the training set. HR images were randomly cropped into 192×192 patches, and LR images were cropped accordingly. Random flipping was used for data augmentation. The proposed SSRFormer was first trained for 300,000 iterations on two 2080ti GPUs with batch size of 8 using the L1 loss. Then, the model was further finetuned for 124,000 iterations on four 2080ti GPUs with batch size of 16. An L1 loss was adopted for the first 60000 iterations and an L2 loss was used for the remaining iterations. The learning rate was initialized to 2×10^{-4} and halved at

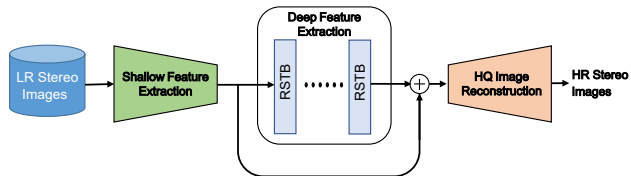


Figure 4. The BUPT-PRIV Team: The network architecture of the proposed SwinIR-impr network.

iteration 250000, 300000, 375000, and 400000.

5.4. The BUPT-PRIV Team

The BUPT-PRIV team developed an improved version of SwinIR [20] to super-resolve left and right images, respectively. The network architecture is shown in Fig. 4. Although the cross-view information is not used in this

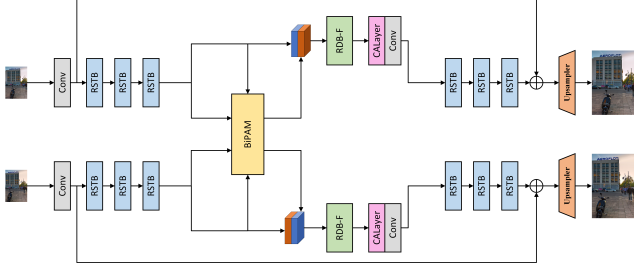


Figure 5. The NKU_caroline Team: The network architecture of the proposed PAMSwIn network.

solution, benefited from the effective data augmentation and test-time augmentation strategies, the proposed method achieves a very competitive SR performance. In addition to the data augmentations originally used in SwinIR, this team further introduced a series of tricks. *In the training phase*, they: 1) adjusted the possibility of selecting training samples to ensure that an image with higher resolution will get a higher possibility to be selected, 2) randomly shuffled RGB channels with a probability of 50%, and 3) trained three models with different combinations of architectures and losses. *In the testing phase*, a series of test-time augmentation approaches was adopted including flipping, self-ensemble, and RGB shuffling. Note that, the window size was set to 16 in this method, which is different from the setting (*i.e.*, 8) in SwinIR.

5.5. The NKU_caroline Team

The NKU_caroline team shares a similar idea with many other teams and developed a PAMSwIn network by combining SwinIR [20] with parallax-attention mechanism [27] for the stereo image SR task. The network architecture of PAMSwIn is shown in Fig. 5. Within the SwinIR framework, a biPAM module is plugged into the middle of residual swin Transformer blocks (RSTBs) to capture cross-view information. Besides, a channel attention layer is employed to exploit correlations between different channels. This team also emphasizes that the order of the input left and right images contains priori information and is critical to the performance. Training with mixed orders of left-right images produced inferior SR performance in their experiments.

During the training phase, a three-stage training strategy was employed. First, the proposed PAMSwIn was trained from scratch for 500K iterations. Then, Cutblur [37] was included for data augmentation to fine-tune the model with the best performance in the first stage for 500K iterations. Note that, the parameters for the biPAM module were fixed at this stage. Finally, a small learning rate was used to further fine-tune the whole model with the highest SR accuracy in the second stage for 500K iterations. During the testing

phase, a self-ensemble strategy was adopted to improve the performance.

5.6. The BUAA-MC2 Team

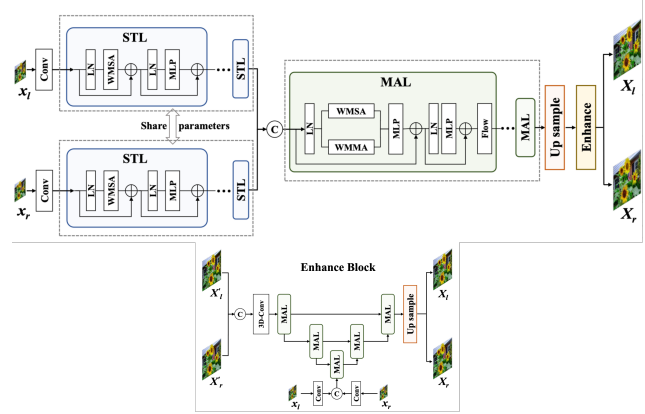


Figure 6. The BUAA-MC2 Team: The architecture of the proposed Stereo Image Super-Resolution Transformer (StereoSRT).

The BUAA-MC2 team proposed a Stereo Image Super-Resolution Transformer (StereoSRT). As shown in Fig. 6, the input stereo images are first fed into a shallow convolution layer and several Swin Transformer layers (STL) to extract shallow features. Then, the output feature maps are concatenated and passed to several mutual attention layers (MAL) to extract cross-view information. After MALs, HR images are reconstructed using a sub-pixel convolutional layer. Finally, a multi-scale enhancement module consisting of several MALs is adopted to further enhance the quality of the HR images.

During the training phase, an L1 loss was used for SR and an L2 loss was used for enhancement. The initial learning rate was set to 4×10^{-4} . The model was trained with a multi-stage strategy. Specifically, in the first stage, the model was trained only with the STL part (the output of STL was directly fed into the up-sample module) for 200K iterations with a patch size of 64×64 . In the second stage, the MAL part (without the flow module) was optimized for 200K iterations while the parameters of the STL parts were fixed. In the third stage, the whole network was optimized end-to-end for 100K iterations. In the fourth stage, the flow module was added to MAL and optimized for 300K iterations with the STL parts being fixed. Finally, the whole model was fine-tuned with a patch size of 96×96 for 100K iterations.

5.7. The No War Team

Figure 7 illustrates the network architecture proposed by the NO War team. The participants of this team used the modules of iPASSR [33] for feature extraction and cross-view information interaction. In the reconstruction stage,

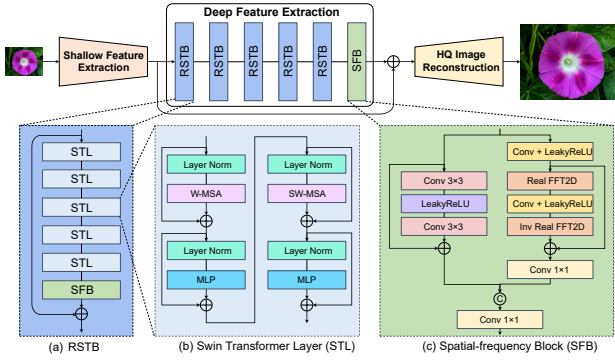


Figure 10. The Xiaozhazha Team: The network architecture of the proposed SwinFIR for stereo image super-resolution.

coarse SR stage, a Bi-Directional Align (BDA) module with pyramid cascading deformable convolutions [43] is used to help the biPAM module [33] to better exploit cross-view information. In the refinement stage, the SR results from the previous stage is fed into a refinement module with 4 groups of residual in residual dense block (RRDB) [44] for enhancement.

During the training phase, the generated LR images were cropped into patches of size 120×120 with a stride of 40. These patches were randomly flipped horizontally and vertically for data augmentation. 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. In the early stage of training, an L1 loss was used to accelerate convergence. Then, an L2 loss was used to obtain higher results in terms of PSNR.

5.10. The Xiaozhazha Team

The xiaozhazha team proposed a network called SwinFIR based on SwinIR [20] and fast Fourier convolution [45], as shown in Fig. 10. SwinFIR consists of three modules, including a shallow feature extraction module, a deep feature extraction module and a high-quality image reconstruction module. The shallow feature extraction and high-quality image reconstruction modules adopt the same configurations as in SwinIR [20]. Since the fast Fourier convolution can extract global features, the participants replaced the 3×3 convolution in SwinIR with fast Fourier convolution and a residual module to fuse global and local features. The proposed spatial-frequency block improves the representation capability of this model.

During the training phase, random horizontal flipping, random vertical flipping, random RGB channel shuffling and mix-up strategy [37] were used for data augmentation. Self-ensemble and multi-model ensemble were adopted to further improve the SR performance.

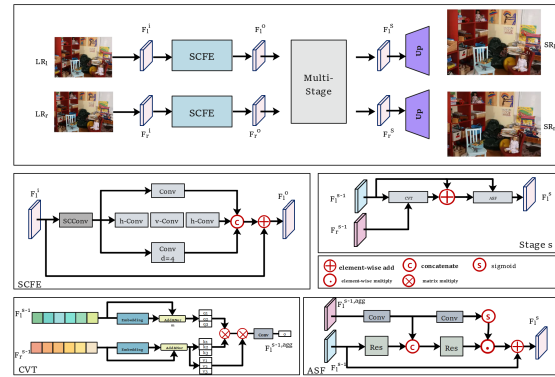


Figure 11. The Zhang9678 Team: The network architecture of the proposed MPTnet.

5.11. The Zhang9678 Team

The Zhang9678 team developed a multi-stage progressive Transformer network (MPTnet) for stereo image SR. The network architecture of the proposed MPTnet is shown in Fig. 11. First, self-calibrated feature extractor (SCFE) is used for feature extraction. Within each SCFE, SCConv [46] and a three-branch structure are employed to achieve large receptive fields. Then, multiple cross-view Transformers (CVTs) and adaptive selective modules (ASFs) are employed to exploit cross-view information. CVT performs information interaction between left and right images along epipolar lines, while ASF aggregates features from different views using a gating mechanism.

5.12. The NTU607QCO-SSR Team

The NTU607QCO-SSR team mainly considers the stereo image super-resolution task as a single image super-resolution task and adopts the state-of-the-art SwinIR [20] as the backbone. As shown in Fig. 12, the model contains convolutional blocks, SwinBlocks, and a pixel shuffling layer. Images are first passed to the 3×3 convolutions and then SwinBlocks are used to extract the global and local features. At the end of the SwinBlocks, image features are passed to a pixel shuffling layer and a 3×3 convolution is used to enlarge the feature maps and reconstruct the SR result. During the training phase, an L1 loss was first used for optimization with 300 epochs. After that, a wavelet-based L1 loss was adopted for fine-tuning. The wavelet-based loss [47,48] uses wavelet transforms to generate sub-images with different scales and frequencies from the original image. Since the resultant sub-images have higher-frequency details, better performance can be achieved.

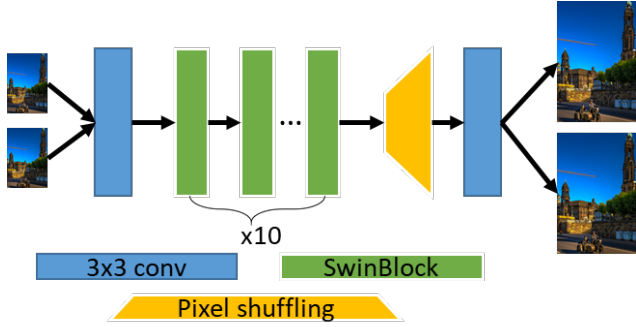


Figure 12. The NTU607QCO-SSR Team: The network architecture of the proposed model.

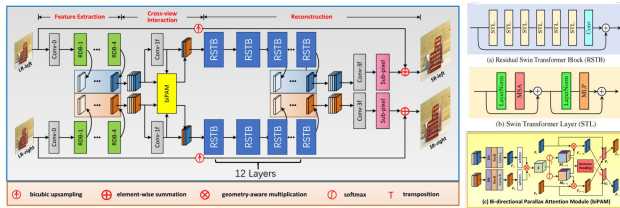


Figure 13. The supersmart Team: The network architecture of the proposed SwinRSTB.

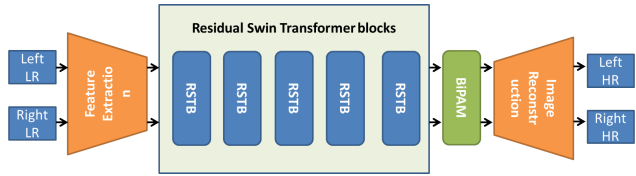


Figure 14. The LIMMC_HNU Team: The network architecture of the proposed PAMswinIR.

5.13. The Supersmart Team

The supersmart team proposed a method called SwinRSTB, as shown in Fig. 13. Since SwinIR [20] is designed for single image SR and cannot incorporate cross-view information, this team combined iPASSR [33] with SwinIR for stereo image SR. In the proposed SwinRSTB network, the RSTB module in SwinIR was used to replace the RGB module in iPASSR.

5.14. The LIMMC_HNU Team

The LIMMC_HNU team developed a PAMswinIR network inspired by SwinIR [20] and iPASSRnet [33]. The network architecture is illustrated in Fig. 14. Different from the solutions of many other teams, they postponed the bi-PAM module until the end of the residual swin Transformer blocks. During the training phase, the loss function in [33] was first used for training to capture cross-view correspondence. Then, only MSE loss was adopted for fine-tuning.

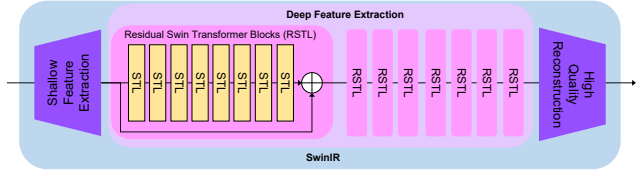


Figure 15. The HIT-IIL Team: Network architecture of SwinIR.

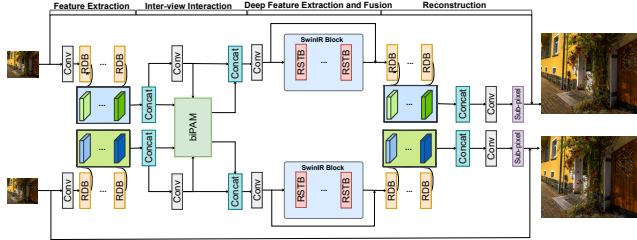


Figure 16. The Hansheng Team: The network architecture of the proposed fine-tuned SwinIR.

5.15. The HIT-IIL Team

The HIT-IIL team employed SwinIR [20] (see Fig. 15) as a basic SISR model and introduced an FFT loss for optimization. The FFT loss measures the difference between outputs and their corresponding HR image in the frequency domain:

$$\mathcal{L}_{FFT}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{F}(\mathbf{y}) - \mathbf{F}(\hat{\mathbf{y}})\|_1, \quad (1)$$

where \mathbf{F} denotes the Fourier transform, \mathbf{y} is the output of the model, and $\hat{\mathbf{y}}$ represents the ground truth image. Therefore, the overall loss function can be written as:

$$\mathcal{L}_{total}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_1 + \lambda * L_{FFT}(\mathbf{y}, \hat{\mathbf{y}}). \quad (2)$$

Compared with the model trained with only L1 loss, the additional FFT loss helps the model converge faster and obtain higher performance.

5.16. The Hansheng Team

The Hansheng team developed a stereo image SR network based on SwinIR [20] and iPASSR [33]. As shown in Fig. 16, the proposed network first performs feature extraction and cross-view interaction using the modules of iPASSR. Then, 6 RSTB blocks are used to aggregate the left and right features, with each block consisting of 6 STL blocks. Next, the resultant features are further fed into several RDBs for reconstruction. During the training phase, input images were cropped into patches of 48×48 , and the window size in STL blocks was set to 8.

5.17. The VIP-SSR Team

The VIP-SSR team improved the performance of iPASSRnet [33] by introducing a hierarchical feature blended-

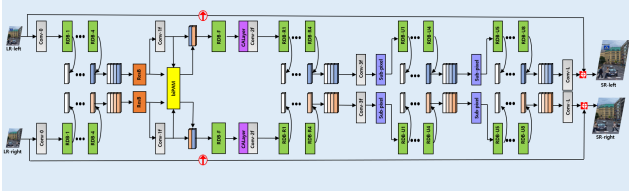


Figure 17. The VIP-SSR Team: The network architecture of the proposed HFB-iPASSR.

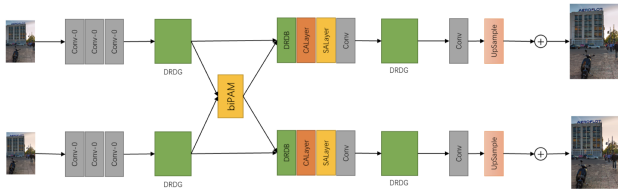


Figure 18. The phc Team: Network architecture of Improved-PASSR.

iPASSR (HFB-iPASSR) network. They first split the pixel-shuffle layer in iPASSR into two pixel-shuffle layers, with each one being followed by a residual dense group (RDG) to relax the discontinuity along pixels caused by the pixel-shuffle operation. They also add a residual block to the first RDG as post-processing to utilize the relation of multi-level features. The architecture of the proposed HFB-iPASSR network is shown in Fig. 17.

5.18. The phc Team

Inspired by PASSRnet [27, 28], the phc team proposed an Improved-PASSR, as shown in Fig. 18. Specifically, the participants introduced a self-attention module to further capture long-range correlations within the image. Meanwhile, the participants deepen the original RDB layers and employ a pixel perception block (PPB) for feature enhancement. In the upsampling block, two sub-pixel layers are used to generate the super-resolved image gradually. Since batch normalization cannot introduce a notable performance improvement, it is removed from the network. During the training phase, the model was optimized for 100 epochs using the Adam method with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a batch size of 32. The initial learning rate was set to 2×10^{-4} and reduced to half after every 40 epochs.

5.19. The qylen Team

The qylen team combined iPASSR with Transformers to achieve improved SR performance. As shown in Fig. 19, the proposed network sequentially performs feature extraction, Transformer-based information fusion, and SR reconstruction. The feature extraction and SR reconstruction parts of this method are similar to those in iPASSR. In the proposed Transformer-based information fusion mod-

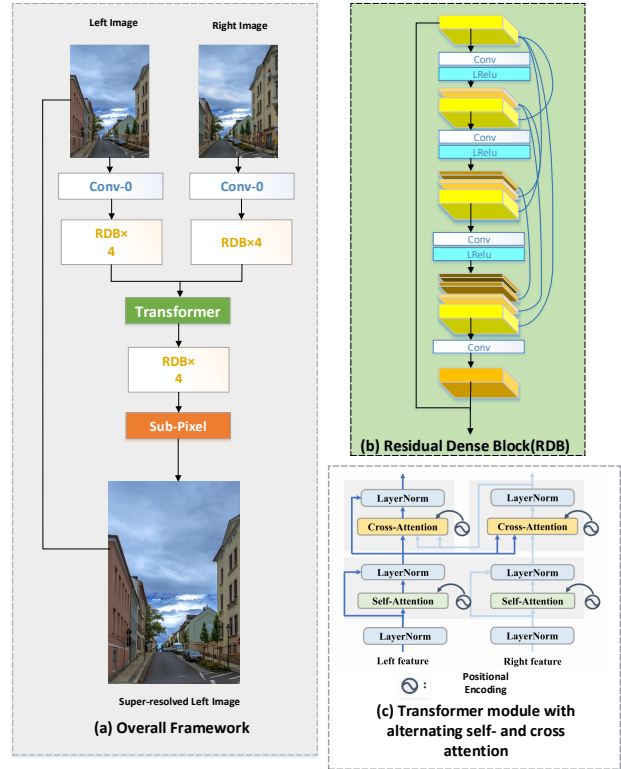


Figure 19. The qylen Team: The network architecture of the proposed iPASSR-Transformer Net.

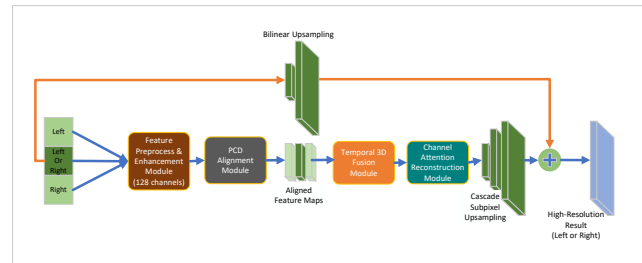


Figure 20. The Modern_SR Team: The network architecture of the proposed Stereo-EDVR.

ule, self-attention is used to model dependencies among pixels within a single view, while cross attention is used to model correspondence of pixels along the epipolar lines between two views. These two attention modules are alternately applied to update the features based on intra-view and cross-view information.

5.20. The Modern_SR Team

The Modern_SR team considered a pair of stereo images as two consecutive frames and developed a Stereo-EDVR network for SR. They aim at seeking a more general SR framework that can be used for different types of SR tasks.

The network architecture of the proposed Stereo-EDVR is shown in Fig. 20. First, a stereo image pair is formulated as a three-frame sequence by duplicating a left or right image. Then, an improved EDVR model with more channels is used to reconstruct an HR left or right image.

6. Acknowledgments

We thank the NTIRE 2022 sponsors: Huawei, Reality Labs, Bending Spoons, MediaTek, OPPO, Oddity, Voyage81, ETH Zurich (Computer Vision Lab) and University of Wurzburg (CAIDAS).

7. Teams and Affiliations

NTIRE2022 team

Title: NTIRE 2022 Challenge on Stereo Image Super-Resolution

Members: Yulan Guo¹ (yulan.guo@nudt.edu.cn), Long-guang Wang¹, Yingqian Wang¹, Juncheng Li², Shuhang Gu³, Radu Timofte⁴

Affiliations:

¹National University of Defense Technology

²The Chinese University of Hong Kong

³The University of Sydney

⁴University of Würzburg, ETH Zürich

(1) The Fat, The Thin and The Young

Title: Nonlinear Activation-Free Network

Members: Liangyu Chen¹ (chenliangyu@megvii.com), Xiaojie Chu², Wenqing Yu¹

Affiliations:

¹MEGVII Technology

²Peking University

(2) BigoSR

Title: SwiniPASSR: Swin Transformer based Parallax Attention Network for Stereo Image Super-Resolution

Members: Kai Jin¹ (jinkai@bigo.sg), Zeqiang Wei², Sha Guo³, Angulia Yang¹, Xiuzhuang Zhou⁴, Guodong Guo⁵

Affiliations:

¹Bigo Technology Pte. Ltd.

²Smart Healthcare Innovation Lab, Beijing University of Posts and Telecommunications

³Peking University

⁴School of Artificial Intelligence, Beijing University of Posts and Telecommunications

⁵Head of Institute of Deep Learning, Baidu Research

(3) NUDT-CV&CPLab

Title: Stereo Image Super-Resolution Transformer

Members: Bin Dai¹ (daicver@gmail.com), Feiyue Peng², Huaxin Xiao¹, Shen Yan¹, Yuxiang Liu¹, Hanxiao Cai¹

Affiliations:

¹College of Systems Engineering, National University of Defense Technology

²College of Liberal Arts and Sciences, National University of Defense Technology

(4) BUPT-PRIV

Title: SwinIR-Impr

Members: Pu Cao (priv@bupt.edu.cn), Yang Nie, Lu Yang, Qing Song

Affiliations:

Pattern Recognition and Intelligent Vision Lab, Beijing University of Posts and Telecommunications

(5) NKU_caroline

Title: PAMSwIn

Members: Xiaotao Hu¹ (1979005820hux@gmail.com), Jun Xu²

Affiliations:

¹College of Computer Science, Nankai University, Tianjin

²School of Statistics and Data Science, Nankai University, Tianjin

(6) BUAA-MC2

Title: StereoSRT: A Stereo Image Super-Resolution Transformer

Members: Mai Xu (MaiXu@buaa.edu.cn), Junpeng Jing, Xin Deng, Qunliang Xing, Minglang Qiao, Zhenyu Guan

Affiliations:

Beihang University

(7) No War

Title: Parallel Interactive Transformer for Stereo Image Super-Resolution

Members: Wenlong Guo (wlguo@zjut.edu.cn), Chenxu Peng, Zan Chen

Affiliations:

Zhejiang University of Technology

(8) GDUT_506

Title: Parallax Res-Transformer Network

Members: Junyang Chen (3117002384@mail2.gdut.edu.cn), Hao Li, Junbin Chen, Weijie Li, Zhijing Yang

Affiliations:

Guangdong University of Technology

(9) DSSR

Title: Deformable Stereo Super-Resolution

Members: Gen Li (*leegeun@yonsei.ac.kr*), Aijin Li, Lei Sun

Affiliations:
Tencent OVB

(10) xiaozhazha

Title: SwinFIR: Rethinking Image Restoration using Swin Transformer and Fast Fourier Convolution

Members: Dafeng Zhang (*594112521@qq.com*), Shizhuo Liu

Affiliations:
SRC-B

(11) Zhang9678

Title: Multi-stage Progressive Transformer for Stereo Image Super-Resolution

Members: Jiangtao Zhang (*zjt9678@qq.com*), Yanyun Qu

Affiliations:
Xiamen University

(12) NTU607QCO-SSR

Title: Transformer-based Super-Resolution with the Edge-aware Loss for Stereo Image Super-Resolution

Members: Hao-Hsiang Yang¹ (*r05921014@ntu.edu.tw*), Zhi-Kai Huang¹, Wei-Ting Chen², Hua-En Chang¹, Sy-Yen Kuo¹

Affiliations:
¹Department of Electrical Engineering, National Taiwan University
²Graduate Institute of Electronics Engineering, National Taiwan University

(13) Supersmart

Title: SwinRSTB
Members: Qiaohui Liang (*qhliang1002@126.com*)
Affiliations:
Personal

(14) LIMMC_HNU

Title: PAMSwInIR
Members: Jianxin Lin (*linjianxin@hnu.edu.cn*), Yijun Wang, Lianying Yin, Rongju Zhang, Wei Zhao, Peng Xiao.
Affiliations:
College of Computer Science and Electronic Engineering, Hunan University

(15) HIT-III

Title: FFT Loss for Super-Resolution
Members: Rongjian Xu (*1180301003@stu.hit.edu.cn*), Zhilu Zhang, Wangmeng Zuo
Affiliations:
Harbin Institute of Technology

(16) HanSheng

Title: Fine-tuned SwinIR
Members: Hansheng Guo¹ (*hsguo.ai@gmail.com*), Guangwei Gao², Tiejong Zeng¹
Affiliations:
¹The Chinese University of Hong Kong
²Nanjing University of Posts and Telecommunications

(17) VIP-SSR

Title: HFB-iPASSR (Hierarchical Feature Blended iPASSR)
Members: Joohyeok Kim¹ (*kimjh4273@unist.ac.kr*), HyeonA Kim², Eunpil Park¹, Jae-Young Sim^{1,2}.

Affiliations:
¹Department of Electrical Engineering, Ulsan National Institute of Science and Technology
²Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology

(18) phc

Title: Improved-PASSR
Members: Huicheng Pi (*21126334@bjtu.edu.cn*), Shunli Zhang
Affiliations:
Beijing Jiaotong University

(19) qylen

Title: iPASSR-Transformer Net
Members: Jucai Zhai (*jucaizhai@stu.pku.edu.com*), Pengcheng Zeng, Yang Liu, Chihao Ma.
Affiliations:
Peking University

(20) Modern_SR

Title: Stereo-EDVR
Members: Yulin Huang¹ (*815018345@qq.com*), Junying Chen²
Affiliations:
¹City University of Hong Kong
²South China University of Technology

References

- [1] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *ICCVW*, 2019. 1, 2
- [2] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, et al. NTIRE 2022 spectral recovery challenge and dataset. In *CVPRW*, 2022. 1
- [3] Boaz Arad, Radu Timofte, Rony Yahel, Nimrod Morag, Amir Bernat, et al. NTIRE 2022 spectral demosaicing challenge and dataset. In *CVPRW*, 2022. 1
- [4] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy Ren, Radu Timofte, et al. NTIRE 2022 challenge on perceptual image quality assessment. In *CVPRW*, 2022. 1
- [5] Andres Romero, Angela Castillo, Jose M Abril-Nova, Radu Timofte, et al. NTIRE 2022 image inpainting challenge: Report. In *CVPRW*, 2022. 1
- [6] Egor Ershov, Alex Savchik, Denis Shepelev, Nikola Banic, Michael S Brown, Radu Timofte, et al. NTIRE 2022 challenge on night photography rendering. In *CVPRW*, 2022. 1
- [7] Yawei Li, Kai Zhang, Radu Timofte, Luc Van Gool, et al. NTIRE 2022 challenge on efficient super-resolution: Methods and results. In *CVPRW*, 2022. 1
- [8] Andreas Lugmayr, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 challenge on learning the super-resolution space. In *CVPRW*, 2022. 1
- [9] Ren Yang, Radu Timofte, et al. NTIRE 2022 challenge on super-resolution and quality enhancement of compressed video: Dataset, methods and results. In *CVPRW*, 2022. 1
- [10] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Ales Leonardis, Radu Timofte, et al. NTIRE 2022 challenge on high dynamic range imaging: Methods and results. In *CVPRW*, 2022. 1
- [11] Goutam Bhat, Martin Danelljan, Radu Timofte, et al. NTIRE 2022 burst super-resolution challenge. In *CVPRW*, 2022. 2
- [12] 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
- [13] 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
- [14] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR*, 2017. 2
- [15] 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
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2
- [17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017. 2
- [18] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, abs/1812.10477, 2020. 2
- [19] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *ECCV*, pages 517–532, 2018. 2
- [20] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPRW*, pages 1833–1844, 2021. 2, 5, 6, 8, 9
- [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2, 4, 7
- [22] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. Efficient transformer for single image super-resolution. *arXiv preprint arXiv:2108.11084*, 2021. 2
- [23] Zhihao Wang, Jian Chen, and Steven C.H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 2
- [24] Juncheng Li, Zehua Pei, and Tiejong Zeng. From beginner to master: A survey for deep learning-based single-image super-resolution. *arXiv preprint arXiv:2109.14335*, 2021. 2
- [25] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *arXiv preprint arXiv:2107.03055*, 2021. 2
- [26] 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
- [27] 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, 6, 10
- [28] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 2, 4, 10
- [29] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020. 2
- [30] Wonil Song, Sungil Choi, Somi Jeong, and Kwanghoon Sohn. Stereoscopic image super-resolution with stereo consistent feature. In *AAAI*, pages 12031–12038, 2020. 2
- [31] 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
- [32] Qingyu Xu, Longguang Wang, Yingqian Wang, Weidong Sheng, and Xinpu Deng. Deep bilateral learning for stereo image super-resolution. *IEEE Signal Processing Letters*, 28:613–617, 2021. 2

- [33] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *CVPRW*, pages 766–775, 2021. [2](#), [4](#), [6](#), [8](#), [9](#)
- [34] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. In *ACM MM*, 2021. [2](#)
- [35] Chenxi Ma, Bo Yan, Weimin Tan, and Xuhao Jiang. Perception-oriented stereo image super-resolution. In *ACM MM*, pages 2420–2428, 2021. [2](#)
- [36] Ruikang Xu, Zeyu Xiao, Mingde Yao, Yueyi Zhang, and Zhiwei Xiong. Stereo video super-resolution via exploiting view-temporal correlations. In *ACM MM*, pages 460–468, 2021. [2](#)
- [37] 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. [3](#), [6](#), [8](#)
- [38] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv*, 2022. [3](#)
- [39] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *CVPRW*, 2022. [4](#)
- [40] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Revisiting global statistics aggregation for improving image restoration. *arXiv*, 2021. [4](#)
- [41] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. [4](#)
- [42] Soham De and Sam Smith. Batch normalization biases residual blocks towards the identity function in deep networks. *NeurIPS*, 33:19964–19975, 2020. [4](#)
- [43] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. [8](#)
- [44] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*, 2018. [8](#)
- [45] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *NeurIPS*, 33:4479–4488, 2020. [8](#)
- [46] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *CVPR*, pages 10096–10105, 2020. [8](#)
- [47] Hao-Hsiang Yang, Chao-Han Huck Yang, and Yi-Chang James Tsai. Y-net: Multi-scale feature aggregation network with wavelet structure similarity loss function for single image dehazing. In *ICASSP*, pages 2628–2632, 2020. [8](#)
- [48] Wei-Ting Chen, Cheng-Che Tsai, Hao-Yu Fang, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. Contourletnet: A generalized rain removal architecture using multi-direction hierarchical representation. *arXiv preprint arXiv:2111.12925*, 2021. [8](#)