This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Hybrid Cross-View Attention Network for Lightweight Stereo Image Super-Resolution

Yuqiang Yang* Zhiming Zhang* Yao Du Jingjing Yang Long Bao Heng Sun†
Xiaomi Inc.
{yangyuqiang, zhangzhiming1, duyao3, yangjingjing3, baolong, sunheng3}@xiaomi.com

Abstract

The goal of stereo image super-resolution is to enhance the quality of low-resolution stereo image pairs by utilizing complementary information across views. Although transformer-based methods have shown high efficiency in single-image super-resolution tasks, they have not been fully used in stereo super-resolution tasks. Therefore, it is crucial to incorporate the complementary information of stereo images into the transformer method to improve image details. To address this challenge, we propose a lightweight Hybrid Cross-view Attention Stereo Super-Resolution network (HCASSR), which uses a Transformer-based network for intra-view feature extraction and a cross-view attention module to aggregate stereo image information. We also employ multi-stage training strategies and data ensemble in test-time to improve image quality. Our method has been extensively tested on the KITTI 2012, KITTI 2015, Middlebury, and Flickr1024 datasets, and the experimental results demonstrate that the proposed method outperforms existing works with smaller model size. Additionally, we won 3rd and 2nd place respectively in Track 1 and Track 2 of the NTIRE 2024 Stereo Image Super-Resolution Challenge. Codes and models will be released at https://github.com/YuqiangY/HCASSR.

1. Introduction

Stereo image super-resolution technology is an advanced image processing method whose core goal is to reconstruct a more detailed high-resolution image from a pair of low-resolution stereo views (*i.e.*, left and right views). In many applications like AR/VR and robot navigation, increasing the resolution of stereo images is highly demanded to achieve higher perceptual quality and help to parse the real world. Therefore, this technology has developed rapidly in recent years and has shown great application potential and widespread attention in many fields. Stereo image super-resolution and single image super-resolution have essential similarities, but there is a key difference between them. Single image super-resolution is limited to extracting information from one perspective, while stereo image super-resolution can integrate information from two views with large overlapping areas. This is crucial because information that may be missing from one view may still exist in another view. Therefore, by effectively integrating information from two perspectives, the quality and details of reconstructed images can be significantly improved, which is a key factor in the success of stereo image super-resolution methods.

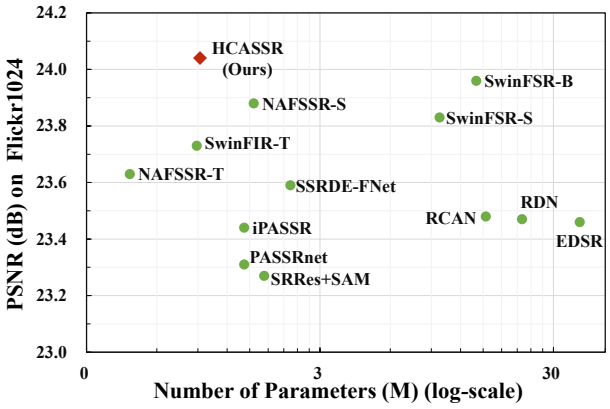

Figure 1. Parameters vs. PSNR of models for 4× stereo SR on Flickr1024 [29] test set. Compared with existing works, the proposed HCASSR achieves higher PSNR with smaller model size.

In previous studies, the application of Transformer architecture in super-resolution methods [3, 18] has proven its significant effectiveness. The reason why the Transformer architecture is effective is because it has a larger receptive field and self-attention mechanism compared to traditional convolutional neural networks, which enables it to better handle long-distance dependencies in images. This powerful feature extraction capability is crucial for stereo image super-resolution, as it requires the integration of information from two perspectives to ensure the preservation of

*Equal contribution.
†Corresponding author.

all valuable details during super-resolution reconstruction. However, the current popular stereo image super-resolution methods [5, 30] are still based on traditional convolutional neural networks. In addition, the Transformer-based methods [4, 32] typically consume much more memory and computing resources than convolutional neural networks, especially when dealing with high-resolution images or large amounts of data.

Given the advantages and limitations of convolutional neural networks and Transformer architectures, researchers have been exploring how to combine them to achieve optimal performance in stereo image super-resolution. Despite various attempts, the optimal hybrid architecture design remains an open issue. Based on the above analysis, we propose an innovative hybrid architecture aimed at fully utilizing the feature extraction capability of Transformer and the efficiency of convolutional neural networks in information exchange between views.

In our method, we use the Transformer-based block as the basic unit to ensure that the most important features are extracted and preserved from each view low-resolution image, and a cross-view attention module to further improve the quality and details of results. As shown in Fig. 1, our HCASSR achieves the best PSNR result, while the model size is only within 1 M. In brief, our contributions can be summarized as follows:

- We propose a lightweight hybrid cross-view attention stereo image super-resolution network which uses a Transformer-based network for intra-feature extraction and a cross-view attention module to complement stereo image information.
- We use the permuted self-attention to replace the self-attention module in the Transformer architecture to reduce network parameters and computational complexity.
- Extensive experiments demonstrate the efficiency and effectiveness of the proposed method both in metrics and in visual quality. As a result, we won 3rd and 2nd place respectively in Track 1 and Track 2 of the NTIRE 2024 Stereo Image Super-Resolution Challenge [28].

2. Related Work

2.1. Single Image Super-Resolution

Single image super-resolution is a fundamental and classic task in the field of computer vision, which aims to reconstruct high-resolution images from a given low-resolution image. In early research, super-resolution technology mainly relied on external images or sample databases to generate high-resolution images.

With the development of deep learning-based methods, super-resolution methods based on Convolutional Neural Networks (CNN) have developed rapidly. SRCNN [7] is the pioneering work of deep learning used in super-resolution reconstruction. The author explains the structure of three-layer convolution into three steps: image patch extraction and representation, non-linear mapping, and reconstruction. VDSR [14], SRDenseNet [23], SRResNet [16], EDSR [19] and RDN [34] further improve performance by using deeper and wider residual blocks. RCAN [33] combines the attention module into residual blocks, giving varying degrees of attention to information from different channels, achieving state-of-the-art performance. NAFNet [2] proposes a simple baseline and achieves state-of-the-art performance.

In addition to CNN-based methods, Transformer-based methods have also emerged in the field of image super-resolution since Transformer has shown great advantages in the field of natural language processing. Compared with CNN, Transformer has been proven to be highly effective in modeling long-range dependencies. Transformer eliminates prior knowledge about locality in the convolutional module, allowing the model to have a larger receptive field. This design means that the model can capture image features on a global scale, rather than being limited to local regions. However, removing this prior knowledge also means that the model needs more data during training to learn sufficient prior knowledge. In practice, IPT [1] demonstrates that even the simplest Transformer can surpass the performance of CNN with sufficient data in low-level tasks. SwinIR [18] reintroduces locality first and adopts a shifted window self-attention module. SRFormer [35] proposes a permuted self-attention (PSA) for image super-resolution tasks, which can handle large window self-attention while maintaining a lower computational cost. HAT [3] combines self-attention, channel attention, and a new overlapping cross-attention to activate more pixels in the receptive field of the Transformer model for image reconstruction.

2.2. Stereo Image Super-Resolution

Single image super-resolution and stereo image super-resolution are two important research directions in the field of computer vision, both dedicated to reconstructing high-resolution images from low-resolution images. However, there is a key difference between these two tasks: stereo image super-resolution processes a pair of images with parallax, namely the left and right views, and there is redundant information between these two views, which can be used to improve the quality of reconstructed images. Traditional single image super-resolution methods are often not directly applicable to stereo image super-resolution tasks because they do not take into account the disparity information in stereo images. To address this issue, researchers have developed various communication modules to facilitate information exchange between the left and right views. The introduction of these communication modules significantly improves the performance of stereo image super-resolution, as they enable two views to share information, thereby improving the

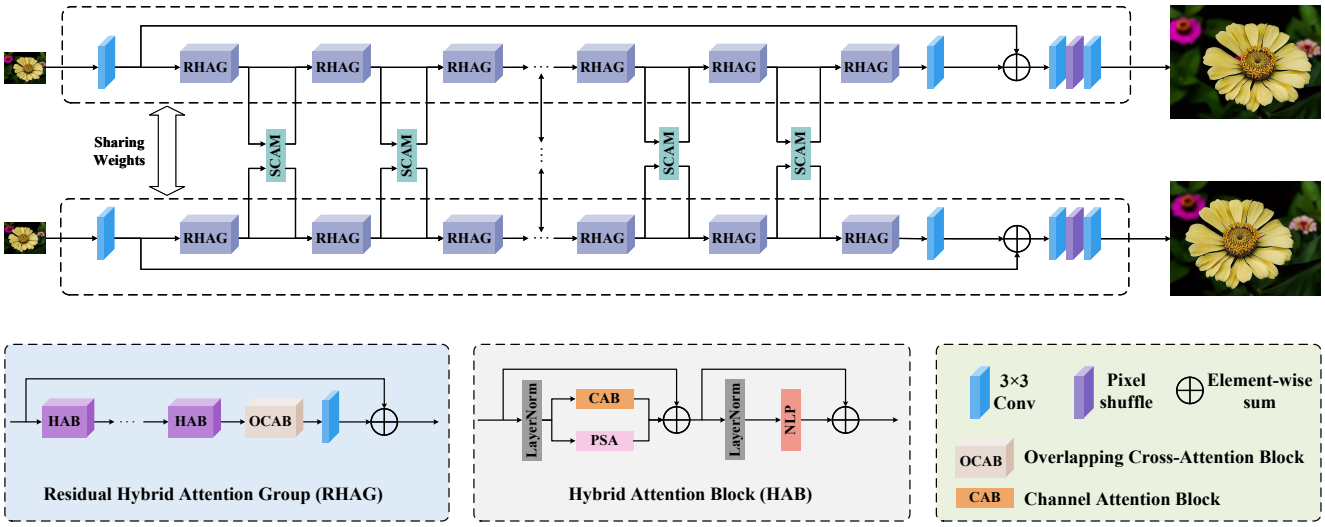

Figure 2. The proposed HCASSR method.

overall quality of reconstruction results.

StereoSR [12] learns a mapping between continuous parallax shifts and a high-resolution image by jointly training two cascaded sub-networks for luminance and chrominance, respectively. PASSRnet [25] introduces a cross attention module that specifically models remote dependencies in the polar direction, while reducing computational and memory costs by limiting the scope of attention mechanisms. iPASSR [30] further improves on PASSRnet [25] by introducing the biPAM module, which can aggregate information from two views after each residual block and effectively handle occlusion problems through a compact bidirectional disparity structure. At the same time, SSRDE-FNet [6] focuses on modeling the differences between two views, proposing a unified architecture to estimate both disparity and super-resolution results. By distorting the deep features of one view based on disparity and using them to improve the reconstruction results of the other view, it further enhances the performance of stereo image super-resolution. NAFSSR [5] which inherits a strong and simple image restoration model, NAFNet [2], significantly improves the final performance by simplifying the cross-attention module, allowing for dense information exchange after each block of the convolutional super-resolution block, and wins the 1st place in the NTIRE 2022 Stereo Image Super-resolution Challenge [26].

With the rapid development of Transformer-based image super-resolution methods, the field of stereo image super-resolution has also ushered in new developments. Researchers are trying to combine the Transformer's superior ability to extract image features with the cross-attention mechanism in the CNN network. SwiniPASSR [13] explores the use of disparity attention networks in the Transformer structure and adopts a progressive training strategy, demonstrating that better results can be achieved based on a single view Transformer backbone. HTCAN [4] introduces a hybrid Transformer and CNN attention network, which employs a two-stage approach to reconstruct stereo images, and wins the 1st place in Track 1 of the NTIRE 2023 Stereo Image Super-resolution Challenge [27].

Therefore, our method is based on a hybrid network of Transformer and CNN, which improves performance while further reducing the number of parameters and the computational complexity.

3. The Method

In this section, more details about the proposed HCASSR are provided. To improve stereo image resolution more effectively and efficiently, we first improve the structure of NAFSSR [5] and introduce the Residual Hybrid Attention Group (RHAG) to better utilize the global information of the image since it can activate more pixels and improve the performance. In addition, data augmentation and multi-stage training strategies are also training and testing-free methods. To improve the performance of the model without increasing the number of parameters and calculations, we use various popular data augmentation methods, such as flip, RGB channel shuffle, and so on. We also use different loss functions in the training and fine-tuning stages, such as Charbonnier Loss [15] and L2 loss [8].

3.1. Network Design

Fig. 2 gives an overview of our proposed HCASSR framework, where we receive the low-resolution stereo images as inputs and super-resolves both left and right view images. More specifically, inspired by NAFSSR [5], our HCASSR

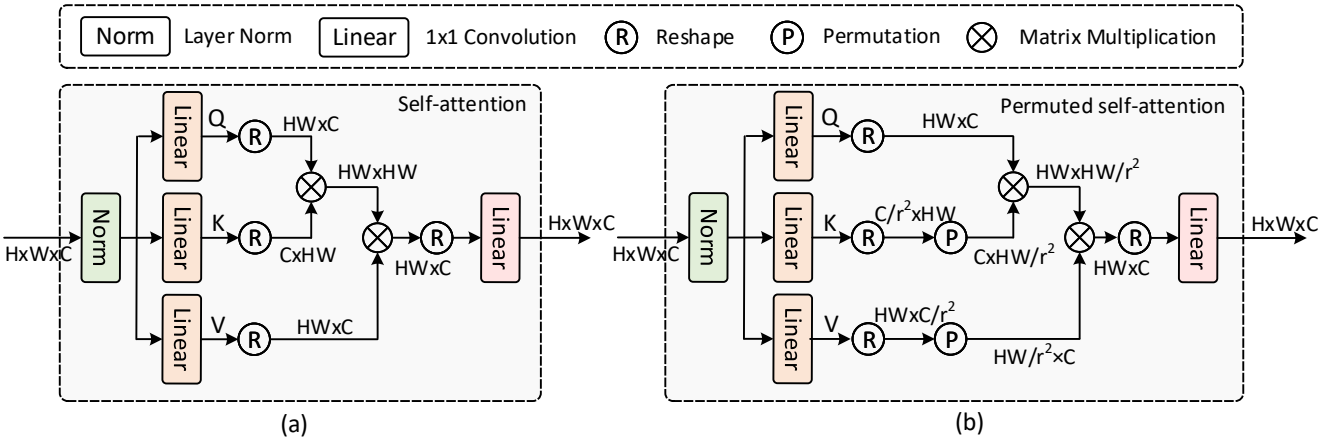

Figure 3. Comparison between (a) self-attention and (b) permuted self-attention. To avoid spatial information loss, [35] proposes to reduce the channel numbers and transfer the spatial information to the channel dimension.

also consists of three modules: intra-view feature extraction, cross-view feature fusion, and reconstruction modules.

Intra-view feature extraction and reconstruction. HCASSR has two weight-sharing branches to extract the intra-view features of the left and the right view images respectively. Firstly, the shallow features are extracted by a 3×3 convolution layer, which provides a preliminary spatial mapping of the inputs. Then, the shallow features are fed into N consecutive RHAG of HAT [3], which is beneficial to activate more pixels and aggregate more global information. We set N as 6. After feature extraction, the features are upsampled by a scaling factor of 4 using a 3×3 convolution layer and a pixel shuffle layer [22]. Every RHAG contains multiple Hybrid Attention Blocks (HAB), an Overlapping Cross-Attention Block (OCAB), and a 3×3 convolution layer, where the window size is set to 16. HAB combines different types of attention mechanisms to activate more pixels for better reconstruction. The module consists of two key components: Window-based Multi-head Self-Attention (W-MSA) and Channel Attention Block (CAB). In addition, to improve the performance and efficiency of the model, we replace the W-MSA module with Permuted Self-Attention (PSA) layer[35] to transfer the spatial information to the channel dimension.

As shown in Fig. 3(b), taken an input feature map $\mathbf{F}_{in} \in \mathbb{R}^{H \times W \times C}$ and a token compression factor r, we first divide $\mathbf{F}_{in}$ into P non-overlapping square patches $\mathbf{F} \in \mathbb{R}^{PS^2 \times C}$, where S is the size of each patch. Then, the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are generated by three 1×1 convolution layers L_Q, L_K, L_V:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = L_Q(\mathbf{F}), L_K(\mathbf{F}), L_V(\mathbf{F}). \quad (1)$$

Here, the channel dimensions of $\mathbf{Q}$ are the same as $\mathbf{F}$, like Fig. 3(a). However, L_K and L_V reduce the channel dimension of $\mathbf{F}$ to C/r^2, producing $\mathbf{K} \in \mathbb{R}^{PS^2 \times C/r^2}$ and $\mathbf{V} \in \mathbb{R}^{PS^2 \times C/r^2}$, which is different from Fig. 3(a). Next, in

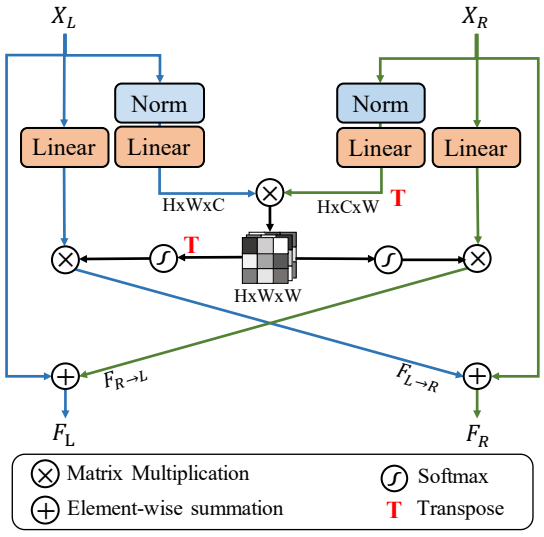

Figure 4. The SCAM module [5].

order to allow more tokens to participate in the self-attention calculation and avoid an increase in calculation costs, we arrange the spatial tokens in $\mathbf{K}$ and $\mathbf{V}$ into channel dimensions, obtaining permuted tokens $\mathbf{K}_p \in \mathbb{R}^{PS^2/r^2 \times C}$ and $\mathbf{V}_p \in \mathbb{R}^{PS^2/r^2 \times C}$.

After that, the reshaped $\mathbf{K}_p$, $\mathbf{V}_p$ and uncompressed Q are performed PSA computation. The formulation can be written as follows:

$$\mathrm{PSA}(\mathbf{Q}, \mathbf{K}_p, \mathbf{V}_p) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_p^T}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V}_p, \quad (2)$$

where $\mathbf{B}$ is the aligned relative position embedding, which can be obtained by interpolating the original embedding defined in [18] since the patch size of $\mathbf{Q}$ does not match

that of $\mathbf{K}_p$. $\sqrt{d_k}$ is a scalar as defined in [9]. In addition, the above equation can be easily converted into a multi-head version by dividing the channels into multiple groups. In this way, PSA transfers the spatial information to the channel dimension. The patch size of $\mathbf{K}_p$ and $\mathbf{V}_p$ can be compressed to $\frac{S}{r} \times \frac{S}{r}$ while their channel dimensions remain unchanged to ensure the expressive ability of the attention map. Compared with the original W-MSA module, PSA can use a larger window size with even fewer computations and achieve better performance with a deeper network.

Cross-view feature fusion. The stereo image super-resolution needs to process dual-view images. How to use dual-view information to complement each other also determines whether the model can achieve good results. Following [5], between the left and right branches, we also insert a Stereo Cross Attention Module (SCAM) after each RHAG to aggregate features extracted from the two views, as shown in Fig. 4. Given stereo features produced by the previous RHAG, SCAM performs bidirectional cross-view attention and generates interaction features that are fused with input features from the same view. Based on Scaled DotProduct Attention [24], the SCAM computes the dot products of the query with all keys and applies the softmax function to obtain the weight of the values. While in the stereo image super-resolution task, there is no vertical displacement between the pixels corresponding to the left and right views, and they are generally on the same horizontal line. Therefore, the SCAM module only performs a dot product on all tokens on the same horizontal line in the left and right images without calculating vertical weights, which is more effective for aggregating the cross-view features.

3.2. Training Strategies

Loss Function. The choice of loss function has a great impact on the performance of the model. As with most low-level visual tasks and competition methods, we use the Charbonnier loss [15] in the training phase and finetune the model with $\mathcal{L}_2$ loss [8], which helps optimize our HCASSR for better results. The Charbonnier loss function $\mathcal{L}_C$ is as outlined below:

$$\mathcal{L}_C = \sqrt{(I_L^{SR} - I_L^{HR})^2 + \varepsilon} + \sqrt{(I_R^{SR} - I_R^{HR})^2 + \varepsilon} \quad (3)$$

where I_L^{SR} and I_R^{SR} are the super-resolved left and right images respectively, I_L^{HR} and I_R^{HR} are the ground truth. As described in [15], the Charbonnier loss is more stable than $\mathcal{L}_1$ loss since it introduces a regularization term ε, which is set to 1×10^{-6} in our training phase. In order to further improve model performance, we change to $\mathcal{L}_2$ function in the fine-tuning stage:

$$\mathcal{L}_2 = \|I_L^{SR} - I_L^{HR}\|^2 + \|I_R^{SR} - I_R^{HR}\|^2 \quad (4)$$

The $\mathcal{L}_2$ loss is closer to the definition of PSNR and can help the model achieve better quantitative results.

Table 1. Ablation studies of different components. We report the results on Flickr 1024 [29] validation datasets. Note that, the MACs is calculated on a stereo image pair of size 320×192.

RHAG	PSA	Params	MACs	PSNR
✗	✗	0.88M	78.77G	23.6254
✓	✗	1.01M	148.78G	23.7963
✓	✓	0.92M	131.79G	23.8247

Table 2. Ablation studies of training strategies. We report the results on Flickr 1024 [29] validation datasets.

Base	$\mathcal{L}_2$ Loss	192×192 Patch	Data Ensemble	PSNR
✓	✗	✗	✗	23.8247
✓	✓	✗	✗	23.8690
✓	✓	✓	✗	23.8834
✓	✓	✓	✓	23.9706

Training Patches. For low-level visual tasks, different sizes of training patches have a significant impact on model performance. To accommodate input images of different sizes during the test phase, we use different patches for multi-stage training. Specifically, we first randomly cut the low-resolution input image into a regular 96 x 96 patch for training, and then increase the size of the patch to 192 for fine-tuning. Note that if the original image size is less than 192, we use the reflection-padding operation on the edges.

Data Augmentation. Beginning from Radu *et al*., who propose rotation and flip data augmentation methods based on spatial transformation, various low-level works have used this method. In addition to flip and rotation, we also utilize multiple data augmentations widely used at high-level tasks, which are based on the pixel domain, such as Mixup and RGB channel shuffle. Mixup randomly mixes two images in a certain proportion. RGB channel shuffle randomly shuffles the RGB channels of input images for color enhancement. These data augmentation approaches can be used not only during training, but also to improve model performance during testing. In order to balance the performance and calculation cost, and meet the test requirements of the challenge which limits the computational complexity (*i.e*., number of MACs) to 400 G (a stereo image pair of size 320×180), we use horizontal flipping, vertical flipping, and original input stereo image pair to obtain higher PSNR.

4. Experiments

4.1. Implementation Detail

Dataset. The NTIRE 2024 Stereo Image Super-resolution Challenge [28] uses 800 stereo image pairs for training, 112 pairs for validation, and 100 pairs for testing. Among them, training and validation sets both come from the

Table 3. Quantitative results achieved by different methods on the KITTI 2012 [10], KITTI 2015 [20], Middlebury [21], and Flickr1024 [25] datasets. Here, PSNR/SSIM values achieved on both the left images (i.e., *Left*) and a pair of stereo images (i.e., $(Left + Right)/2$) are reported. The best results are in **bold faces**.

Method	Scale	Params	*Left*			$(Left + Right)/2$			
			KITTI 2012	KITTI 2015	Middlebury	KITTI 2012	KITTI 2015	Middlebury	Flickr1024
VDSR [14]	×4	0.66M	25.54/0.7662	24.68/0.7456	27.60/0.7933	25.60/0.7722	25.32/0.7703	27.69/0.7941	22.46/0.6718
EDSR [19]	×4	38.9M	26.26/0.7954	25.38/0.7811	29.15/0.8383	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RDN [34]	×4	22.0M	26.23/0.7952	25.37/0.7813	29.15/0.8387	26.32/0.8014	26.04/0.8043	29.27/0.8404	23.47/0.7295
RCAN [33]	×4	15.4M	26.36/0.7968	25.53/0.7836	29.20/0.8381	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
StereoSR [12]	×4	1.42M	24.49/0.7502	23.67/0.7273	27.70/0.8036	24.53/0.7555	24.21/0.7511	27.64/0.8022	21.70/0.6460
PASSRnet [25]	×4	1.42M	26.26/0.7919	25.41/0.7772	28.61/0.8232	26.34/0.7981	26.08/0.8002	28.72/0.8236	23.31/0.7195
SRRes+SAM [31]	×4	1.73M	26.35/0.7957	25.55/0.7825	28.76/0.8287	26.44/0.8018	26.22/0.8054	28.83/0.8290	23.27/0.7233
IMSSRnet [17]	×4	6.89M	26.44/-	25.59/-	29.02/-	26.43/-	26.20/-	29.02/-	-/-
iPASSR [30]	×4	1.42M	26.47/0.7993	25.61/0.7850	29.07/0.8363	26.56/0.8053	26.32/0.8084	29.16/0.8367	23.44/0.7287
SSRDE-FNet [6]	×4	2.24M	26.61/0.8028	25.74/0.7884	29.29/0.8407	26.70/0.8082	26.43/0.8118	29.38/0.8411	23.59/0.7352
PFT-SSR [11]	×4	-	26.64/0.7913	25.76/0.7775	29.58/0.8418	26.77/0.7998	26.54/0.8083	29.74/0.8426	23.89/0.7277
SwinFIR-T [32]	×4	0.89M	26.59/0.8017	25.78/0.7904	29.36/0.8409	26.68/0.8081	26.51/0.8135	29.48/0.8426	23.73/0.7400
NAFSSR-T [5]	×4	0.46M	26.69/0.8045	25.90/0.7930	29.22/0.8403	26.79/0.8105	26.62/0.8159	29.32/0.8409	23.69/0.7384
NAFSSR-S [5]	×4	1.56M	26.84/0.8086	26.03/0.7978	29.62/0.8482	26.93/0.8145	26.76/0.8203	29.72/0.8490	23.88/0.7468
CVHSSR-T [36]	×4	0.68M	26.88/0.8105	26.03/0.7991	29.62/0.8496	26.98/0.8165	26.78/0.8218	29.74/0.8505	23.89/0.7484
HCASSR (**Ours**)	×4	0.92M	**26.93**/**0.8140**	**26.11**/**0.8028**	**29.88**/**0.8575**	**27.03**/**0.8200**	**26.85**/**0.8252**	**29.98**/**0.8578**	**24.04**/**0.7550**

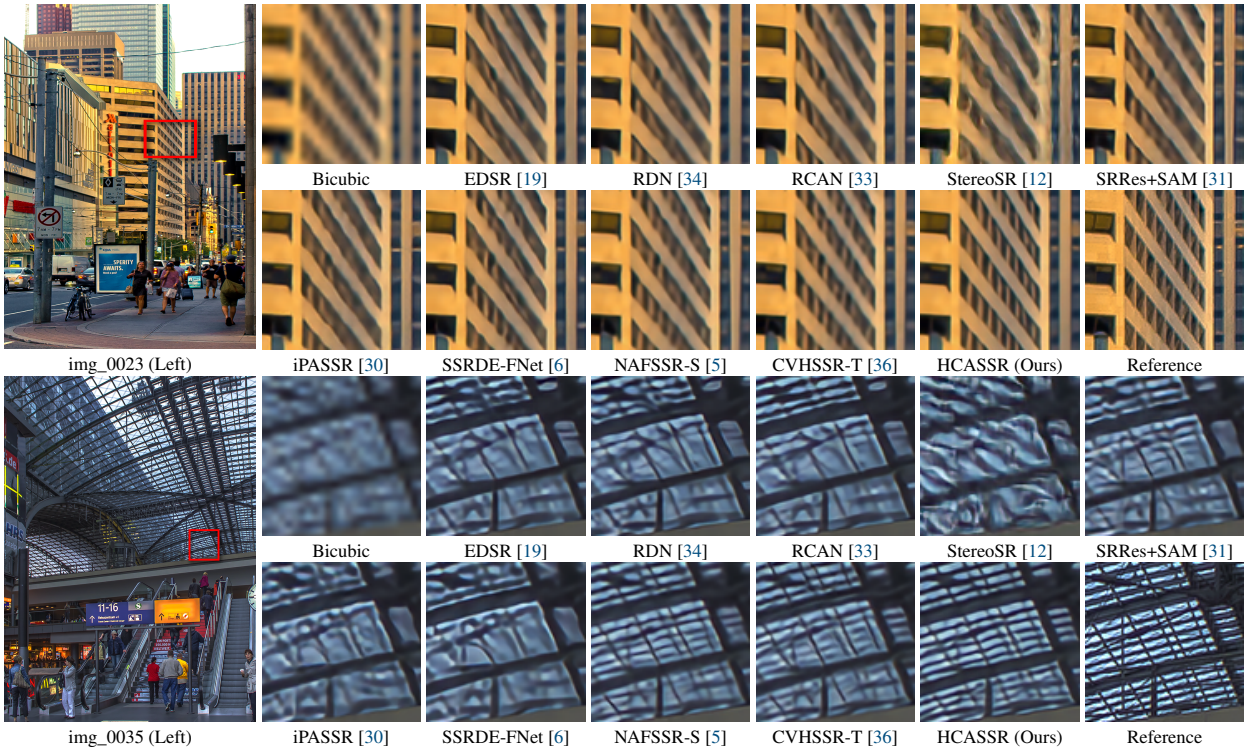

Figure 5. Visual results (×4) achieved by different methods on the Flickr1024 [25] dataset.

Flickr1024 [29] dataset, while the testing set contains an additional set of 100 stereo image pairs. The low-resolution input pairs of Track 1 are produced by downscaling with the standard Bicubic method, while the input pairs of Track 2 is created with complex realistic degradations (*i.e.*, blur, downsampling, additive noise and JPEG compression).

Training Settings. Our proposed HCASSR network is trained in a multi-stage strategy. We first train the model with the Charbonnier loss [15] using the Adam optimizer ($\beta_1 = 0.9$ and $\beta_2 = 0.99$) and stop after 400k iterations. Then the model is fine-tuned with the $\mathcal{L}_2$ loss using the same optimizer. The batch size is set to 16 and the patch size is first set to 96×96 and then 192×192 for fine tuning. The learning rate is initialized with 5×10^{-4} and set to 1×10^{-4} and

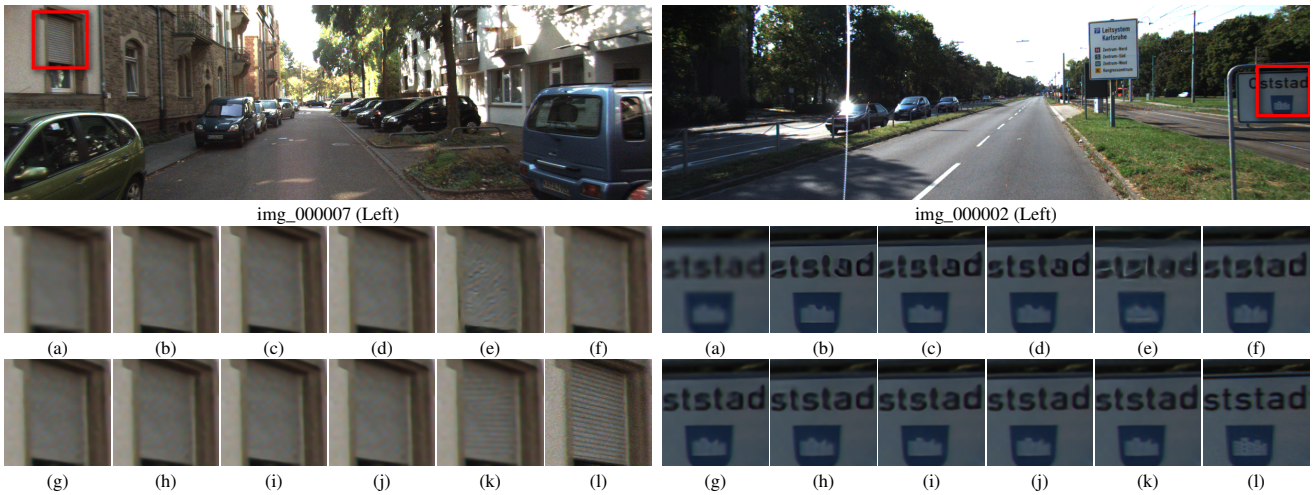

Figure 6. Visual results (×4) achieved by different methods on the KITTI 2012 [10] (left) and KITTI 2015 [20] (right) dataset. (a) Bicubic. (b) EDSR [19]. (c) RDN [34]. (d) RCAN [33]. (e) StereoSR [12]. (f) SRRes+SAM [31]. (g) iPASSR [30]. (h) SSRDE-FNet [6]. (i) NAFSSR-S [5]. (j) CVHSSR-T [36]. (k) HCASSR (Ours). (l) Reference.

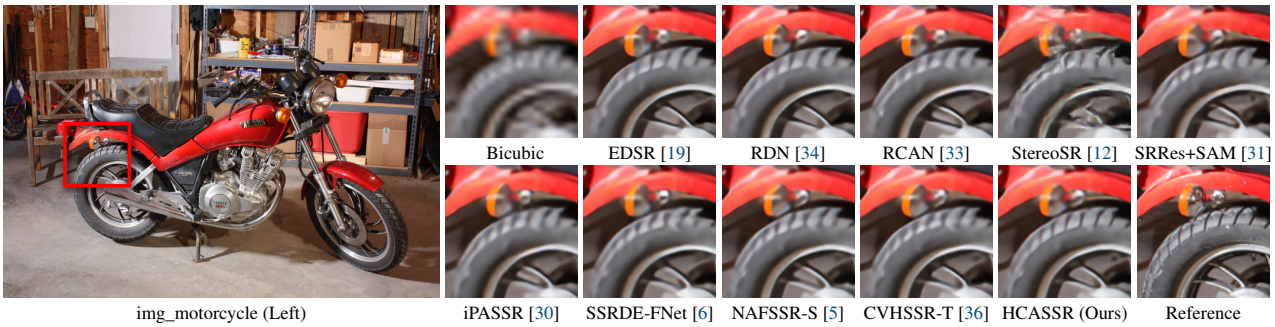

Figure 7. Visual results (×4) achieved by different methods on the Middlebury [21] dataset.

2×10^{-5} respectively for the two fine-tuning processes above. We implement our network with the Pytorch framework and train it with 4 NVIDIA Tesla A100 GPUs. We also use a cosine annealing strategy to update the learning rate. To increase the diversity of the training dataset and avoid an overfitting issue, the data is augmented by horizontal/vertical flipping, random shuffling of RGB channels and Mixup. It is worth noting that the number of parameters is restricted to 1 M and the computational complexity is limited to 400 G (a stereo image pair of size 320 × 180) in the challenge. Therefore, the model ensemble strategy is abandoned and we only use three methods including the original image (the other two are horizontal and vertical flipping) for the data ensemble strategy to improve the performance.

4.2. Ablation Study

In this section, we perform a series of ablation experiments to evaluate the performance of some modules mentioned in Sec. 3.1 and training/testing strategies mentioned in Sec. 3.2. The evaluation is performed with the Flickr1024 [29] validation dataset.

Effectiveness of modules. We use NAFSSR [5] as the baseline to evaluate the RHAG and PSA modules in the network. Note that, we have modified the depth and width in NAFSSR to keep its number of parameters within 1 M. From Tab. 1, we find that replacing the NAFBlock in NAFSSR with the RHAG module in HAT (keeping the number of parameters to about 1 M) can bring a performance improvement of 0.17 dB PSNR. In addition, we introduce the PSA module into the above model, which reduces the number of parameters by 0.09M and the computational complexity by 17G, while improving the performance by almost 0.03 dB PSNR.

Effectiveness of training strategies. In this study, we conducted several experiments on training strategies to improve performance. As shown in Tab. 2, we first use the model trained with the Charbonieer loss [15] and 96 × 96 patch as the baseline. Then, we fine-tune the model with

the $\mathcal{L}_2$ loss, which can bring a performance improvement of 0.0443 dB. We further improved the performance by fine-tuning the model with 192×192 patch size, which resulted in an additional performance gain of 0.0144 dB. Finally, the data ensemble strategy is applied in test time, which includes the original image, horizontal and vertical flipping, and improves the PSNR value by 0.0872 dB.

4.3. Comparison to State-of-the-arts Methods

Settings. Deviating from the description in Sec. 4.1, we use data identical to iPASSR [30] to allow a fair comparison with other methods. In detail, the training set includes 800 images from training set of Flickr1024 [29] and 60 images from Middlebury [21]. Note that, following [30], we perform a bicubic downsampling by a factor of 2 on the images of the Middlebury dataset to generate high-resolution (HR) ground truth images to match the spatial resolution of the Flickr1024 dataset. We then apply bicubic downsampling by a factor of 4 to the HR images to generate low-resolution (LR) inputs. The testing set includes 20 images from KITTI 2012 [10] and 20 images from KITTI 2015 [20], 5 images from Middlebury [21] and 112 images from the testing set of Flickr1024 [25].

Quantitative Evaluations. We compare our HCASSR with existing state-of-the-art super-resolution (SR) algorithms, including SR methods for single images and SR methods for stereo images. Single image SR methods include VDSR [14], EDSR [19], RDN [34], and RCAN [33]. Stereo image SR methods include StereoSR [12], PASSRnet [25], SRRes+SAM [31], IMSSRnet [17], iPASSR [30] and SSRDE-FNet [6], PFT-SSR [11], SwinFIR[32], NAFSSR [5], CVHSSR [36].

The quantitative comparison results are shown in Table 3. Following [30], we report PSNR/SSIM scores on the left images with their left boundaries (64 pixels) cropped, and average scores on stereo image pairs (*i.e.*, (Left + Right) /2) without any boundary cropping. Compared with all networks within 1 M parameters, our method achieves the best results. Specifically, our method surpass previous state-of-the-art model NAFSSR-S [5] by 0.1 dB, 0.09 dB, 0.26 dB, and 0.16 dB at KITTI 2012 [10], KITTI 2015 [20], Middlebury [21] and Flickr1024 [25], respectively, which clearly shows the effectiveness of the proposed HCASSR.

Visual Comparison. We show the visual comparison results for $\times 4$ stereo SR on Flickr1024 [25], KITTI 2012 [10], KITTI 2015 [20] and Middlebury [21]. As shown in Fig. 5, our method produces richer details accurately without obvious artifacts in densely repeated texture areas. Specifically, in the left figure of Fig. 6, only our method successfully restores the horizontal texture on the rolling shutter door. The right figure in Fig. 6 shows the effectiveness of our method in reconstructing font edge details. Our method also recovers clearer tire texture than other methods in Fig. 7.

Table 4. The final results in the NTIRE 2024 Stereo Image Super-Resolution Challenge [28].

Track1		*Track2*	
Rank	PSNR (RGB)	Rank	PSNR (RGB)
1	23.6503	1	21.8724
2	23.6105	2 (**Ours**)	**21.6983**
3 (**Ours**)	**23.6070**	3	21.6702
4	23.5941	4	21.6691
5	23.5896	5	21.5935
6	23.5725	6	21.5655
7	23.5271	7	21.5313
8	23.4851	8	21.5238
9	23.4598	9	21.4970
10	23.4510	10	21.1994
11	23.4270	11	20.7642
12	23.3888	12	20.7518
13	23.1895	13	20.6167
14	23.0977	-	-

4.4. NTIRE Stereo Image SR Challenge

We use the above method to participate in the NTIRE 2024 Stereo Image Super-resolution Challenge [28] Track 1 and Track 2. Different from previous years, this challenge limits the number of parameters to 1 M and the computational complexity to 400 G (a stereo image pair of size 320×180) in testing phase. Specifically, the computational complexity of the data augmentations using in the test-time for the self-ensemble is also taken into account. Therefore, we only use three methods for data augmentation, including the original image, horizontal flipping, and vertical flipping, and do not use a model ensemble strategy. As a result, our final submission ranked 3rd on Track 1 with 23.6070 dB PSNR on the test set and 2nd on Track 2 with 21.6983 dB PSNR on the test set, as shown in Tab. 4.

5. Conclusion

In this work, we propose a lightweight Transformer-based method HCASSR for stereo image super-resolution task, which is stacked by a set of RHAGs with PSA modules for effective intra-view feature extraction. We also insert SCAM in two branches to aggregate intra-view and cross-view features. Additionally, we utilize multi-stage training strategies with data augmentation, hyperparameters and loss functions to improve the performance of the model without loss of efficiency. Extensive experiments demonstrate the efficiency and effectiveness of the proposed method in terms of both metrics and visual quality.

References

[1] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 2, 3

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023. 1, 2, 4

[4] Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Xuhan Sheng, Shijie Zhao, Junlin Li, and Li Zhang. Hybrid transformer and cnn attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1702–1711, 2023. 2, 3

[5] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2022. 2, 3, 4, 5, 6, 7, 8

[6] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1985–1993, 2021. 3, 6, 7, 8

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2

[8] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016. 3, 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6, 7, 8

[11] Hansheng Guo, Juncheng Li, Guangwei Gao, Zhi Li, and Tieyong Zeng. Pft-ssr: Parallax fusion transformer for stereo image super-resolution. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 6, 8

[12] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1721–1730, 2018. 3, 6, 7, 8

[13] Kai Jin, Zeqiang Wei, Angulia Yang, Sha Guo, Mingzhi Gao, Xiuzhuang Zhou, and Guodong Guo. Swinipassr: Swin transformer based parallax attention network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 920–929, 2022. 3

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 2, 6, 8

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017. 3, 5, 6, 7

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[17] Jianjun Lei, Zhe Zhang, Xiaoting Fan, Bolan Yang, Xinxin Li, Ying Chen, and Qingming Huang. Deep stereoscopic image super-resolution via interaction module. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3051–3061, 2020. 6, 8

[18] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 1, 2, 4

[19] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 2, 6, 7, 8

[20] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 6, 7, 8

[21] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014. 6, 7, 8

[22] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4

[23] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proceedings*

of the IEEE international conference on computer vision, pages 4799–4807, 2017. 2

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[25] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019. 3, 6, 8

[26] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte. Ntire 2022 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2022. 3

[27] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, et al. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1372, 2023. 3

[28] Longguang Wang, Yulan Guo, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, and Radu Timofte. Ntire 2024 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2024. 2, 5, 8

[29] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 5, 6, 7, 8

[30] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–775, 2021. 2, 3, 6, 7, 8

[31] 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. 6, 7, 8

[32] 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. 2, 6, 8

[33] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2, 6, 7, 8

[34] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 2, 6, 7, 8

[35] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12780–12791, 2023. 2, 4

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1396–1405, 2023. 6, 7, 8