SwinFSR: Stereo Image Super-Resolution using SwinIR and Frequency Domain Knowledge

Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang and Jun Chen
McMaster university, Hamilton, Canada
{chenk59, lil61, liuh127, liy366, tangc61, chenjun}@mcmaster.ca

Abstract

Stereo Image Super-Resolution (stereoSR) has attracted significant attention in recent years due to the extensive deployment of dual cameras in mobile phones, autonomous vehicles and robots. In this work, we propose a new StereoSR method, named SwinFSR, based on an extension of SwinIR, originally designed for single image restoration, and the frequency domain knowledge obtained by the Fast Fourier Convolution (FFC). Specifically, to effectively gather global information, we modify the Residual Swin Transformer blocks (RSTBs) in SwinIR by explicitly incorporating the frequency domain knowledge using the FFC and employing the resulting residual Swin Fourier Transformer blocks (RSFTBs) for feature extraction. Besides, for the efficient and accurate fusion of stereo views, we propose a new cross-attention module referred to as RCAM, which achieves highly competitive performance while requiring less computational cost than the state-of-the-art cross-attention modules. Extensive experimental results and ablation studies demonstrate the effectiveness and efficiency of our proposed SwinFSR.

1. Introduction

Stereo image pairs can encode 3D scene cues into stereo correspondences between the left and right images. With the extensive deployment of dual cameras in mobile phones, autonomous vehicles and robots, the stereo vision has attracted increasing attention in both academia and industry. In many applications such as AR/VR [19,35] and robot navigation [30], increasing the resolution of stereo images is highly demanded to attain superior perceptual quality and optimize performance for downstream tasks [40]. Recently, many deep-learning-based methods [4,21,41,43] have been proposed to address the stereo super-resolution (stereoSR) problem.

In favour of the remarkable capability of the Transformer [37], most recent stereoSR methods [37, 40] are developed

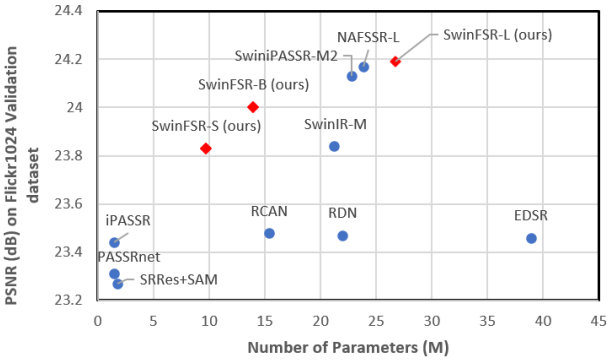

Figure 1. Parameters vs. PSNR of models for 4× stereo SR on Flickr1024 [42] test set. Our SwinFSR families achieve the highest performance.

based on Transformer structure, especially on a variant for image restoration task, i.e., SwinIR [21]. However, there are some common issues with the existing SwinIR based models such as SwiniPASSR [14] and SwinFIR [48]. First, SwiniPASSR does not have a specifically designed mechanism for exploiting features extracted from two views as biPAM [43] is used by default. Second, it focuses on spatial features but not spectral features, thus failing to make full use of large receptive fields to gather global information in a more direct manner. As of SwinFIR [48], it also does not explicitly exploit the interdependence of features extracted from two views due to a lack of cross attention modules. Moreover, SwinFIR cannot estimate epipolar stereo disparity as it requires squared images as inputs.

Inspired by the observation of [36] regarding the effectiveness of the Fast Fourier Convolution (FFC) block in capturing global information, we modify Residual Swin Transformer blocks (RSTBs) in SwinIR by explicitly incorporating the frequency domain knowledge and employ the resulting Residual Swin Fourier Transformer blocks (RSFTBs) for feature extraction. Besides the proposed feature extractor, we also aim to enhance the cross-attention module for effective and efficient informant exchange between

two views. Instead of directly using the off-the-shelf cross-attention modules such as SAM [46], SCAM [4], and biPAM [43], we propose a new cross-attention module named RCAM. Specifically, to balance between efficient inference and accurate learning, we modify the biPAM by removing the need to handle occlusion and redesigning the attention mechanism. Moreover, to address the inflexibility of squared training patches with respect to the epipolar disparity, we modify the local window in the Swin Transformer so that the network can process rectangular input patches. Based on the above innovations, we develop a new stereoSR network, namely SwinFSR. In summary, our SwinFSR has two branches built with RSFTBs to process left and right views, respectively. The two branches share the same weights. RCAMs are inserted between the two branches to exchange and consolidate cross-view information.

Furthermore, various training/testing strategies are adopted to unleash the potential of SwinFSR. In training, we use several effective data augmentation methods to boost SR performance, such as random cropping, flipping, and channel shuffling. We also conduct experiments to find the best possible hyper-parameters, such as dropout rate [18], window size, and stochastic depth [12] of the Swin Transformer based models. As shown in Figure 1, our SwinFSR families have better performance-complexity trade-offs than the existing methods.

Our contributions can be summarized as follows:

- Based on a systematic analysis of the issues with the existing methods, we propose a new stereoSR method, SwinFSR. It inherits the advantages of SwinIR and Fast Fourier Convolution and exploits both spatial and spectral features.
- We propose a new cross-attention module, named RCAM, that strikes a good balance between efficient inference and accurate learning. This is realized by modifying biPAM to circumvent occlusion handling as well as redesigning its attention mechanism. It is shown that this modification can help expedite the inference speed without significantly jeopardizing the performance.
- Extensive experimental results demonstrate the effectiveness and efficiency of our proposed approach.

2. Related Works

2.1. Single Image Super-resolution

Single image super-resolution (SISR) aims to generate high-resolution images based on their low-resolution counterparts. SISR has been extensively researched in the fields of image processing and computer vision, and various approaches have been proposed to address this problem. Super-Resolution Convolutional Neural Networks (SRCNN) [8] make the first attempt to bring deep learning to bear upon SISR, and subsequent methods VDSR and EDSR [22, 49] further take advantage of residual and dense connections to achieve improved performances. Attention mechanisms, including channel attention [6, 27, 50] and channel-spatial attention [7, 21, 29], have also been proposed as an effective tool for tackling SISR. Recently, in view of its remarkable ability in natural language processing (NLP), SwinIR [21], a Transformer-based structure has been employed for SISR, achieving state-of-the-art (SOTA) performance.

2.2. Stereo Image Super-Resolution

Stereo image super-resolution (stereoSR) is a challenging task in computer vision that requires generating high-resolution images from stereo image pairs. Convolutional neural networks (CNNs) are commonly used in deep learning-based stereoSR approaches, such as the Single Image Stereo Matching network (SSRN) [26]. It introduces a stereo matching module to establish dense correspondence between low-resolution stereo images and then applies a CNN to enhance the resolution of each image. Attention mechanisms have also been explored in recent works to improve stereoSR. For instance, [41] proposes a parallax attention module (PAM) and builds a PASSRnet for stereoSR to handle varying parallax. [51] designs an attention-based method that can adaptively weigh the stereo features to enhance the resolution of the stereo images. [46] introduces stereo attention modules (SAMs) into pre-trained single image SR (SISR) networks to handle information assimilation. [34] addresses the occlusion issue by using disparity maps regressed by parallax attention maps to assess stereo consistency. [43] develops an iPASSRnet that uses symmetry cues and a Siamese network equipped with a biPAM structure to super-resolve both left and right images. And NAFSSR [4], the winner of the NTIRE 2022 StereoSR Challenge [40], achieves the SOTA by inserting cross-view attention modules (SCAMs) between consecutive NAFblocks [2]. These works have made significant contributions to the stereoSR and have opened up new possibilities for future research in this area.

In this work, we move one step further by introducing a residual stereo cross-attention module (RCAM). In contrast to SAM [46], which requires calculating an occlusion map, our RCAM presents a better solution with high efficiency.

2.3. Vision Transformer

As a recent advance in the field of computer vision, visual Transformers [37] have garnered significant attention for their ability to capture long-range dependencies in images, especially for high-level vision tasks such as image classification [9, 24] and object detection [1, 24, 44]. Moreover, Transformers have also been applied to low-level vi-

sion tasks (see, e.g., [45]). To reduce the computational complexity of self-attention operations in Transformers, a hierarchical visual Transformer called Swin Transformer [1] is proposed using shifted window techniques, which achieved SOTA performance on various tasks such as image recognition, object detection, and segmentation. SwinIR [21] and Swin V2 [23] have implemented some further refinements to make Transformers more efficient. Overall, these works have demonstrated the effectiveness of visual Transformers in a wide range of computer vision tasks.

2.4. Training and Testing Strategies

Regularization methods such as dropout [18] and stochastic depth [12] are widely employed to enhance the model performance in high-level computer vision tasks. Recently, the above regularization methods have been introduced in image restoration tasks. For example, stochastic depth is employed in [4] to address the issue of overfitting to the stereo-training data and improve generalization. Similarly, [18] adjusts the dropout method in their SR tasks. In this work, we will systematically study how the factors such as dropout rate, window size, and stochastic depth can impact PSNR performance in Swin Transformer-based models. Additionally, since test-time augmentation (TTA) [15, 38] is a technique that is frequently used in computer vision competitions to boost performance, we also investigate its capability in the context of stereoSR through an ablation study.

3. Method

In this section, we introduce our method in detail. In Sec 3.1, we first give an overview of the network's architecture. In Sec 3.2, we then present the training and testing strategies.

3.1. Network Architecture

3.1.1 Overall Framework

Figure 2 depicts an outline of our proposed transformer-based Stereo SR network (SwinFSR). SwinFSR takes a low-resolution stereo image pair as input and enhances the resolution of both left and right view images. To be specific, Our SwinFSR has two branches built with RSFTB to process left and right views, respectively. RCAMs described in Figure 4, are inserted between the left and right branches to interact with cross-view information. In essence, SwinFSR is composed of three parts: intra-view feature extraction, cross-view feature fusion, and reconstruction

Intra-view feature extraction and reconstruction. To start, a 3×3 convolutional layer is employed to extract the shallow features from input images. Then, RSFTBlocks are stacked to achieve deep intra-view feature extraction. We will detail the RSFTBlock in Section 3.1.2. Once feature extraction is completed, a Fast Fourier Block (FFB) is applied, followed by a pixel shuffle layer [32] that upsamples the feature by a scale factor of 4. Additionally, to alleviate the burden of feature extraction, we follow [4, 20] to predict the difference between the bilinearly upsampled low-resolution image and the high-resolution ground truth.

Cross-view feature fusion. To engage with information from different views, we incorporate RCAM following every RSFTBlocks. RCAM utilizes stereo features produced by the preceding RSFTBlocks as inputs for conducting bidirectional cross-view interactions and produces interacted features fused with input features from the same view. The details of the RCAM are elaborated in Section 3.1.5.

3.1.2 RSFTBlock.

As shown in Figure 2 (a), the residual Swin Transformer block (RSTB) is a residual block built using Swin Transformer Layers (STL) in Figure 2 (b) and a Fast Fourier Convolution Block in Figure 3. Given the input feature $F_{i,0}$ of the i-th RSFTB, we first extract intermediate features $F_{i,j}$ by L STLs as:

$$F_{i,j} = STL_{i,j}(F_{i,j-1}), j = 1, 2, 3, ..., L, \quad (1)$$

where $STL_{i,j}$ is j-th STL in the i-th RSFTB.

We then feed the feature from L-th STL to FFB to extract frequency domain knowledge. After that, we output the summation of FFB outputs and input features by:

$$F_{i,out} = FFB_i(F_{i,L}) + F_{i,0}, \quad (2)$$

where FFB_i represents the last FFB block in the i-th RSFTB block. And $F_{i,out}$ is the output feature of i-th RSFTB block.

3.1.3 STL Blocks.

As shown in Figure 2 (b), a two-layer multi-layer perceptron (MLP) with fully connected layers and GELU non-linearity between them is used. Prior to using the MSA and MLP, a LayerNorm (LN) layer is attached and a residual connection is employed for both modules. The complete process for the STL block is explained in detail in [21].

3.1.4 Fast Fourier Convolution Blocks (FFB).

Our backbone model SwinIR is mainly composed of residual Swin Transformer blocks (RSTBs) that utilize several Swin Transformer layers to achieve local attention and cross-window interaction. However, in the context of stereo SR, it is advantageous to incorporate both local and global information [11]. To achieve this, we take inspiration from the Fast Fourier Convolution (FFC) [3], which can use the global context in early layers [36]. To this end, we propose

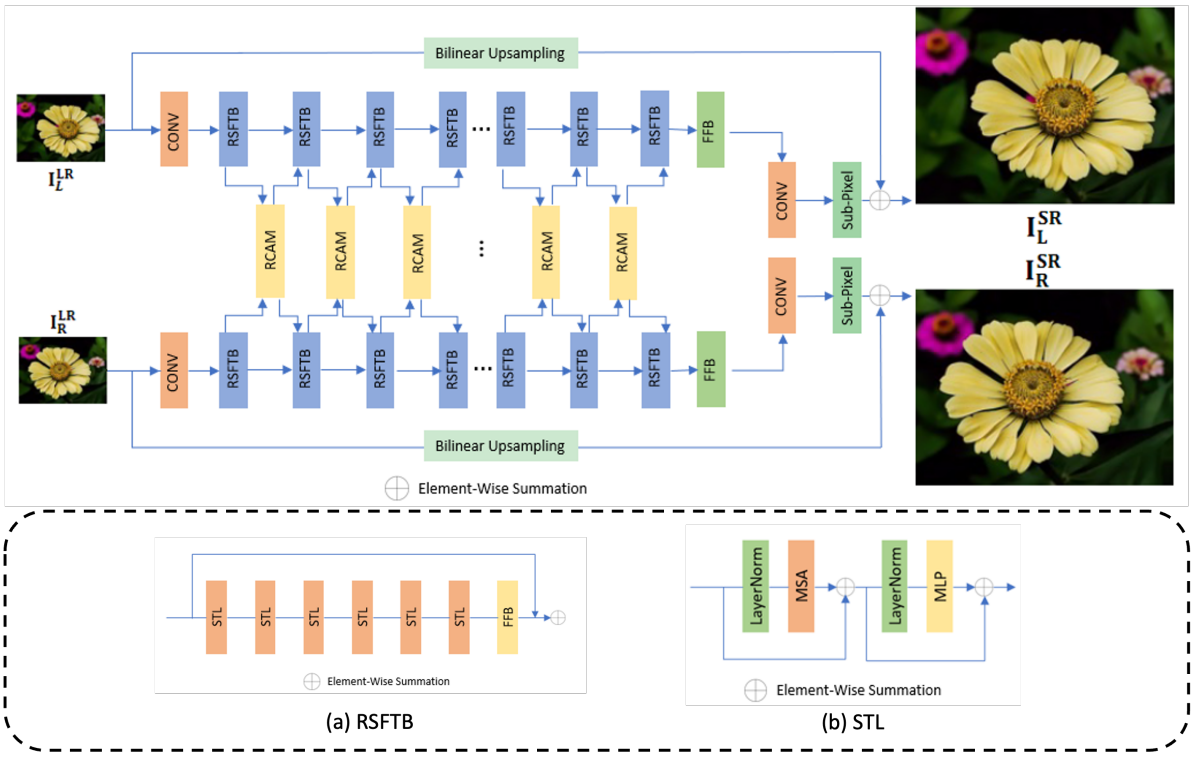

Figure 2. Top: illustration of the proposed SwinFSR Architecture. Bottom: (a) Residual Swin Fourier Transformer Block (RSFTB), (b) Swin Transformer Layer (STL).

a hybrid module including an FFC and a residual module called the Fast Fourier Block (FFB) to enhance the model's ability. As shown in Figure 3, the FFB has two main components: a local spatial conventional convolution operation on the left and a global FFC spectrum transform on the right. The outputs from both operations are concatenated and then subjected to a convolution operation to generate the final result. Here we formalize the operation. Given an input feature of FFB block $F_{i,L}$, we send $F_{i,L}$ into two distinct branches, local and global. In the local branch, H_{local} is utilized and extracts the local features in the spatial domain, and H_{global} is intended to capture the long-range context in the frequency domain. To increase readability, we use F to represent $F_{i,L}$ in the following paragraphs.

$$F_{local} = H_{local}(F), \tag{3}$$

$$F_{global} = H_{global}(F). \tag{4}$$

We then detail the local and global branches. The local branch is CNN based, as shown in Figure 3. Instead of using a single-layer convolution, we insert a residual connection and two convolution layers to increase the expressiveness of the model. The extraction of F_{local} can be also written as,

$$F_{local} = H_{conv}(F) + F \tag{5}$$

where $H_{conv}(\cdot)$ denotes a simple block containing three layers. Specifically, it consists of two 3×3 convolution layers and a LeakyReLU layer.

In the global branch, we use the spectrum transform structure in accordance with [3]. It can transform the conventional spatial features into the frequency domain to extract the global features by 2-D FFT and perform the inverse 2-D FFT operation to produce final spatial domain features for future feature fusion. The H_{global} in Eq. 4 can also be re-written as,

$$F' = \mathcal{C}(F) \tag{6}$$

$$F_{frequency} = \mathcal{C}''(H_{IFFT}(\mathcal{C}'(H_{FFT}(F'))) + F') \tag{7}$$

where $H_{FFT}(\cdot)$ is the channel-wise 2-D FFT operation. $H_{IFFT}(\cdot)$ is the inverse 2-D FFT operation. $\mathcal{C}$, $\mathcal{C}'$ and $\mathcal{C}''$ denote the used three convolution layers in the global branch.

After obtaining the features from both branches, we finally use a single 1×1 convolution layer $\mathcal{C}_f$ to fuse the two features and reduce the number of channels by half.

$$F_{FFB} = \mathcal{C}_f([F_{local}, F_{global}]) \tag{9}$$

where $[\cdot]$ is the concatenation operation.

3.1.5 Cross-View Interaction.

In this section, we show the details of the proposed Residual Cross Attention Module (RCAM). The structure of RCAM is demonstrated in Figure 4. It is based on Scaled Dot Product Attention [37] and inspired by all the previous cross attention modules [34, 41, 43, 46], which computes the dot products of the query with all keys and applies a softmax function to obtain the weights on the values:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{C})V \quad (10)$$

where $Q \in R^{H\times W\times C}$ is a query matrix projected by the source intra-view feature (e.g., left-view), and $K, V \in R^{H\times W\times C}$ are key, value matrices projected by target intra-view feature (e.g., right-view). Here, H, W, and C represent the height, width and number of channels of the feature map. Since stereo images are highly symmetric under epipolar constraint [43], we follow NAFSSR [4] to calculate the correlation of cross-view features along the W dimension. In detail, given the input stereo intra-view features $F_L, F_R \in R^{H\times W\times C}$, we can get layer normalized stereo features $\bar{F_L}$= LN(F_L) and $\bar{F_R}$= LN(F_R). Next, a residual block (Resb) is applied to the process, and the processed feature is separately fed into two 1×1 convolutions and obtain $\hat{F_L}$ and $\hat{F_R}$. We then follow [43] to feed $\hat{F_L}$ and $\hat{F_R}$ to a whiten layer to acquire normalized features to establish disentangled pairwise parallax attention according to the following two equations:

$$\bar{F_L}'(h, w, c) = \hat{F_L}(h, w, c) - \frac{1}{W}\sum_{i=1}^{W}\hat{F_L}(h, i, c) \quad (11)$$

$$\bar{F_R}'(h, w, c) = \hat{F_R}(h, w, c) - \frac{1}{W}\sum_{i=1}^{W}\hat{F_R}(h, i, c) \quad (12)$$

Then a geometry-aware multiplication will be adopted between $barF_L{}'$ and $\bar{F_R}'$:

$$Attention = \bar{F_L}' \otimes \bar{F_R}' \quad (14)$$

The bidirectional cross-attention between left-right views is calculated by:

$$F_{R->L} = Attention(W_1^L\bar{F_L}, W_1^R\bar{F_R}, W_2^R F_R), \quad (15)$$

$$F_{L->R} = Attention(W_1^R\bar{F_R}, W_1^L\bar{F_L}, W_2^L F_L), \quad (16)$$

where W_1^L, W_1^R, W_2^L and W_2^R are projection matrices. Note that we can calculate the left-right attention matrix only once to generate both $F_{R->L}$ and $F_{L->R}$ (as shown in Figure 4). Finally, the interacted cross-view information $F_{R->L}, F_{L->R}$ and intra-view information F_L, F_R are fused by element-wise addition same as NAFSSR [4]:

$$F_{L,out} = \gamma L F_{R->L} + F_L \quad (15)$$

$$F_{R,out} = \gamma R F_{L->R} + F_R \quad (15)$$

where γL and γR are trainable channel-wise scales and initialized with zeros for stabilizing training.

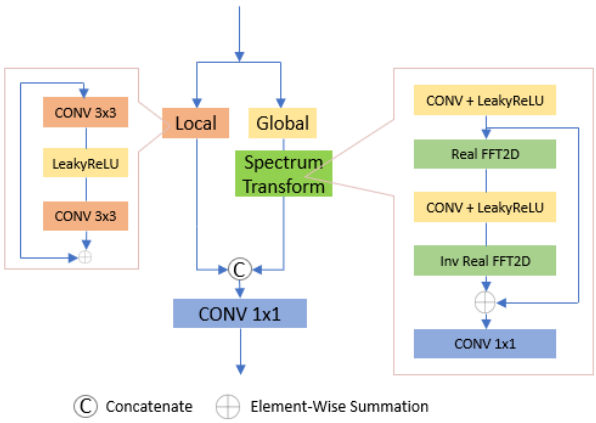

Figure 3. Fast Fourier Convolution Block (FFB).

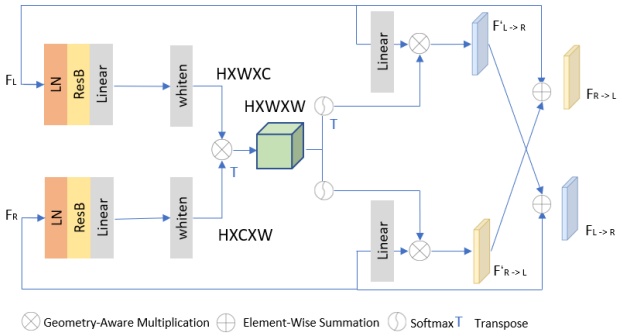

Figure 4. Residual Cross Attention Module (RCAM).

3.2. Training Strategies

Rectangular Training Patches. In stereo image SR tasks, it is common to train models with small squared patches cropped from full-resolution images [43,45]. Due to the fact that disparity of the stereo images existing along the epipolar line, some models use 30×90 rectangular patches to train the stereoSR models [14,48]. We empirically find that the patch size does affect the model performance and we show the experimental results in Table 5. These patches are randomly flipped horizontally and vertically for data augmentation.

Dropout Rate and Stochastic Depth. To further utilize the training data, we adopt stochastic depth [12] and dropout [18] as regularization. The results of using different stochastic depth and dropout rates during model training can be found in Table 6.

Loss Functions. We use the pixel-wise L1 distance between the SR and ground-truth stereo images in the NTIRE

2023 Stereo Image Super Resolution Challenge Track 1 [39]:

$$L_{SR} = ||I_L^{SR} - I_L^{HR}||_1 + ||I_R^{SR} - I_R^{HR}||_1, \quad (8)$$

where I_L^{SR} and I_R^{SR} are respectively the super-resolved left and right images. I_L^{HR} and I_R^{HR} are the ground truths.

For the Challenge Track2, inspired by [47, 52], we adopt a combination of perceptual loss and L1 loss to enhance supervision in the high-level feature space, as outlined below:

$$L_{Final} = L_{SR} + 0.01 * L_{Per} \quad (9)$$

$$L_{Per} = \frac{1}{N} \sum_j \frac{1}{C_j H_j W_j} ||\phi_j(f_\theta(I^{LR})) - \phi_j(I^{HR})||_2^2. \quad (10)$$

The VGG-16 [33], pre-trained on ImageNet, serves as the loss network ϕ. The loss function, expressed in equation 10, uses the left and right low resolution input image I_L^{LR}, I_R^{LR} and their correspondence high resolution ground truth images I_L^{HR}, I_R^{HR}. And the super-resolved images I^{SR}, generated by the SwinFSR model are denoted by $f_\theta(\cdot)$, where $\phi_j(\cdot)$ represents the feature map with a size of $C_j \times H_j \times W_j$. j denote the j-th layer of VGG-16. Moreover, the L2 loss is utilized as the feature reconstruction loss and the perceptual loss function employs N features.

4. Experiments

4.1. Implementation Details

Evaluation Metrics. The evaluation metrics used are peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). These metrics are calculated in the RGB colour space using a collection of stereo images obtained by averaging the left and right views. Table 1 displays the influence of varying architecture, including three different sizes of SwinFSR by modifying the number of blocks. These networks are identified as SwinFSR-S (Small), SwinFSR-B (Big), and SwinFSR-L (Large).

Training Detail. All models are optimized by the Adam [17] with $\beta 1 = 0.9$ and $\beta 2 = 0.9$. The learning rate is set to $1e^{-}4$ and decreased to $1e^{-}5$ with a cosine annealing strategy [25]. If not specified, models are trained on 30×90 patches with a batch size of 1 for $7e^6$ iterations. The window size of the model is 6×15. Data augmentation includes horizontal/vertical flips and RGB channel shuffle are used.

Datasets. To conduct our experiments, we utilize the training and validation datasets provided by the NTIRE Stereo Image SR Challenge [40]. Specifically, we use 800 stereo images from the training set of the Flickr1024 [42] dataset as our training data and 112 stereo images from the validation set of the same dataset as our validation set. The low-resolution images are created by downsampling using the bicubic method. In addition, we follow the dataset splits in [4] to conduct a comparison on KITTI 2012 [10], KITTI 2015 [28], Middlebury [31] and Flickr1024 [42].

Table 1. The performance of different SwinFSRs in size.

Model	#RSFTBs	#Params	PSNR
SwinFSR-S	4	9.76M	23.8319
SwinFSR-B	6	14.01M	23.9630
SwinFSR-L	**12**	**26.75M**	**24.1940**

Table 2. The influence of different cross-attention modules. We here report the results in both PSNR and SSIM for $4\times$SR. TTA represents the test-time augmentation. SwinFSR-L is used to conduct this analysis.

Modules	PSNR		SSIM	
	w/o TTA	w. TTA	w/o TTA	w. TTA
-	23.6921	23.7714	0.7380	0.7397
biPAM [43]	23.8883	24.0510	0.7432	0.7520
SAM [46]	22.3834	22.4366	0.6690	0.6715
SCAM [4]	24.0882	24.1926	0.7564	**0.7616**
RCAM	**24.1233**	**24.1940**	**0.7583**	0.7598

Table 3. The efficiency comparison between several cross-attention modules. We replace the cross-attention module in SwinFSR-L to conduct the analysis. Training time is the cost for $4\times$SR on Flickr1024 [42] training set.

Modules	Params	Time/Epoch	Speedup
SAM [46]	32.72M	1259ms	-
SCAM [4]	25.00M	988ms	%21.5
RCAM	26.75M	1065ms	%15.4

4.2. Ablation Study

Residual Cross-Attention Modules. Here, all the experiments are conducted using SwinFSR-L. To show the effectiveness of RCAM, we substitute the cross-attention module in SwinFSR-L with several SOTA approaches, such as biPAM [43], SAM [46], SCAM [4] and baseline (without cross-attention module.). Table 2 shows the $4 \times$ SR results on Flickr1024 [42]. First, when compared with the baseline that only explored intra-view information, our method is 0.4 dB higher than the baseline in PSNR. Furthermore, compared with biPAM, SCAM, and SAM, our RCAM achieves improvements of 0.235 dB, 0.035 dB, and 1.740 dB, respectively.

In addition, to further show the efficiency of our RCAM, we provide in Table 3 by the number of parameters and training time. It can be observed that our proposed RCAM has fewer parameters and training time than that of SAM. It is worth mentioning that both SCAM and our RCAM do not handle occlusion problems when performing cross-view

Table 4. The influence of different dropout rates. We here report the results in both PSNR and SSIM for 4×SR. TTA represents the test-time augmentation. SwinFSR-S is used to conduct this analysis.

Model	Dropout Rate	PSNR		SSIM	
		w/o TTA	w. TTA	w/o TTA	w. TTA
SwinFSR-S	N/A	23.7304	23.8191	0.7430	0.7451
	0.1	**23.8319**	**23.9240**	**0.7471**	**0.7492**
	0.3	23.8319	23.9230	0.7470	0.7491
	0.5	21.6377	22.4352	0.6365	0.6767

Table 5. The influence of different window sizes and training patch sizes. We here report the results in both PSNR and SSIM for 4×SR. TTA represents the test-time augmentation. SwinFSR-S is used to conduct this analysis.

Patch	Window	PSNR	PSNR w. TTA	SSIM	SSIM w. TTA
32 × 32	4×4	23.52	23.63	0.734	0.738
32 × 32	8×8	23.57	23.65	0.734	0.737
30 × 90	3×9	23.65	23.74	0.739	0.741
30 × 90	**6×15**	**23.83**	**23.92**	**0.747**	**0.749**

integration. Interestingly, we find using SCAM and RCAM does not jeopardize the performance but can help achieve better PSNR and faster training. These outcomes emphasize the importance of a well-designed cross-attention model and the critical impact of integrating both cross-view information and intra-view information.

Test Time Augmentations. Although test-time augmentation (TTA) has been commonly utilized in competitions to enhance performance, its usefulness in stereo SR tasks has not been proven. Here, we use horizontal and vertical flips as our TTA strategy. To evaluate the effectiveness of TTA in this task, we assess each model's inference results using the NTIRE 2023 Stereo Image SR validation dataset [39]. The results, presented in Table 2, 4, 5 and 6, demonstrate that employing TTA is always beneficial. This phenomenon suggests that TTA is indeed effective for stereo SR tasks.

Dropout. According to [18], adding only one line of dropout layer can significantly improve the model performance. We thus follow [18] to put the dropout layer before the last convolution layer. Then, we use SwinFSR-S to investigate the impact of the dropout rate during training. In Table 4, we report results on Flickr1024 [42] validation set. Compare to the SwinFSR-S model without the specific dropout layer, with a 10% dropout rate, the PSNR result can be improved by 0.102 dB. However, when we increase the dropout rate to 30%, the performance does not change. When it comes to 50%, half of the nodes are dropped during the training, which makes the performance decrease by 2.194 dB.

Window Size and Training Patch Size. According to [48], a larger window size can enhance the performance of stereoSR. Here, we use SwinFSR-S to further investigate the impact of window size. Table 5 reports results on Flickr1024 [42] test set. First, while using the same squared training patch size, a larger window size will improve the performance of SwinFSR-S by 0.049 dB. If further changing the training patch sizes to be rectangular according to the epipolar stereo disparity [43], the performance will be increased by 0.087 dB. Moreover, increasing window size while using rectangular training patches boost the performance by 0.178 dB. Due to the limitation of the GPU resources, we do not further enlarge the window size and training patch size. This shows that the rectangular training patch and larger local window size indeed can help improve the feature extraction ability across stereo images.

Stochastic Depth. As per the research conducted by [4], a deeper stochastic depth can improve the performance of stereoSR. Therefore, we employ SwinFSR-L to examine how stochastic depth affects our Swin Transformer based model. Our results based on the validation set of Flickr1024 [42] are presented in Table 6. During training, incorporating 10% stochastic depth [12] lead to a 0.102 dB improvement in PSNR. When using 20% stochastic depth, the performance of SwinFSR-L improves slightly by 0.1014 dB. However, setting the stochastic depth to 30% results in a performance decrease of 0.121 dB, but it still outperforms the baseline. This suggests that larger models have a tendency to overfit the Flickr1024 training data. However, incorporating stochastic depth can help enhance the overall performance and generalization ability of the networks.

4.3. Comparison with the state-of-the-art methods

To make a fair comparison with previous works, we follow the dataset splits in NAFSSR [4] to train and test our method on four representative datasets, i.e., KITTI 2012 [10], KITTI 2015 [28], Middlebury [31] and Flickr1024 [42]. Specifically, we generate low-resolution images by applying bicubic downsampling to high-resolution (HR) images with a scaling factor of 4. Then we randomly crop 30 × 90 patches from stereo images as inputs. During training, we set all the hyperparameters to the best possible ones given by our ablation studies, such as dropout rate, window size, and stochastic depth. Additionally, we employ hori-

Table 6. The influence of stochastic depth. We here report the results in both PSNR and SSIM for 4×SR. TTA represents the test-time augmentation. SwinFSR-L is used to conduct this analysis.

Model	Stochastic Depth	PSNR		SSIM	
		w/o TTA	w. TTA	w/o TTA	w. TTA
SwinFSR-L	N/A	23.9516	24.0442	0.7518	0.7537
	0.1	24.0786	24.1679	**0.7573**	**0.7591**
	0.2	**24.0928**	**24.1773**	0.7470	0.7491
	0.3	23.9719	24.1035	0.7518	0.7548

Table 7. Comparison with several state-of-the art methods for 4×SR on the KITTI 2012 [10], KITTI 2015 [28], Middlebury [31] and Flickr1024 [42] datasets. The number of parameters is denoted by "Params". Numbers reported for each dataset are in PSNR/SSIM.

Model	#Params	KITTI2012	KITTI2015	Middlebury	Flickr1024
VDSR	0.66M	25.60/0.7722	25.32/0.7703	27.69/0.7941	22.46/0.6718
EDSR	38.9M	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RDN	22.0M	26.32/0.8014	26.04/0.8043	29.27/0.8404	23.47/0.7295
RCAN	15.4M	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
StereoSR	1.42M	24.53/0.7555	24.21/0.7511	27.64/0.8022	21.70/0.6460
SRRes+SAM	1.73M	26.44/0.8018	26.22/0.8054	28.83/0.8290	23.27/0.7233
PASSRnet	1.42M	26.34/0.7981	26.08/0.8002	28.72/0.8236	23.31/0.7195
iPASSR	1.42M	26.56/0.8053	26.32/0.8084	29.16/0.8367	23.44/0.7287
SSRDE-FNet	2.24M	26.70/0.8082	26.43/0.8118	29.38/0.8411	23.59/0.7352
SwiniPASSR-M2	22.81M	-/-	-/-	-/-	24.13/0.7579
NAFSSR-L	23.83M	27.12/0.8194	26.96/0.8257	30.20/0.8605	24.17/0.7589
SwinFSR-S (ours)	9.76M	27.03/0.8143	26.83/0.8213	32.45/0.8891	23.83/0.7471
SwinFSR-B (ours)	14.01M	27.07/0.8151	26.87/0.8222	32.69/0.8910	23.96/0.7510
SwinFSR-L (ours)	26.75M	**27.24/0.8195**	**27.00/0.8257**	**32.73/0.8915**	**24.19/0.7598**

zontal and vertical flips as our test-time augmentation. For the results on Flickr1024, we perform results ensemble by collecting the top three performed models on the validation set and averaging their inference results on the test set as the final results (the same strategy we used in the NTIRE2023 challenge [39]). For the other three datasets, we report the best performance without an ensemble.

Table 7 presents the quantitative comparison of SwinFSR and several SOTA super-resolution methods. Our comparison includes single SR methods such as VDSR [16], EDSR [22], RDN [50], RCAN [49], and SwinIR [21], as well as stereo SR methods including StereoSR [13], PASSRnet [41], SRRes+SAM [46], iPASSR [43], SRRDE-FNet [5], SwiniPASSR [14], and NAFSSR [4]. The evaluation metrics used are PSNR and SSIM, and the dataset used for testing are KITTI 2012 [10], KITTI 2015 [28], Middlebury [31] and Flickr1024 [42]. By checking throughout the table, it can be observed that our method outperforms all the compared approaches on the four datasets. These results further validate the effectiveness of our proposed method.

4.4. NTIRE Stereo Image SR Challenge

We submit a result obtained by the presented approach to the NTIRE 2023 Stereo Image Super-Resolution Challenge Track 1 and 2 [39]. In order to maximize the potential performance of our method, we adopt the stochastic depth [12] with 0.2 probability to improve the model's generality ability. During test time, we adopt horizontal and vertical flips as our TTA strategy. Finally, we average the SR images from the top 3 performance models on the validation set for our final submission. As a result, our final submission achieves 24.1940 dB in PSNR on the validation set and won a ninth place with 23.7121 dB in PSNR on the test set.

5. Conclusion

The goal of this paper is to introduce a novel network called SwinFSR for enhancing the resolution of stereo images. To achieve this, we utilize a series of RSFTBlocks to extract intra-view features with enlarged reception fields and propose residual stereo cross-attention modules (RCAMs) to interact between both intra-view and cross-view features. Additionally, we explore the best possible hyperparameters, such as dropout rate, training patch size, window size, and stochastic depth and found the best values are 10%, 30×90, 6×15 and 20% respectively. Extensive ablation studies demonstrate the effectiveness of the proposed method.

References

[1] H Cao, Y Wang, J Chen, D Jiang, X Zhang, Q Tian, and M Wang. Swin-unet: unet-like pure transformer for medical image segmentation. corr. *arXiv preprint arXiv:2105.05537*, 2021. 2, 3

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 2

[3] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 3, 4

[4] 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. 1, 2, 3, 5, 6, 7, 8

[5] 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. 8

[6] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 2

[7] Tao Dai, Hua Zha, Yong Jiang, and Shu-Tao Xia. Image super-resolution via residual block attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[8] C Dong, CC Loy, K He, and X Tang. Image super-resolution using deep convolutional networks. arxiv e-prints. *arXiv preprint arXiv:1501.00092*, 2014. 2

[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. 2

[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] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 3

[12] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 2, 3, 5, 7, 8

[13] 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. 8

[14] 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. 1, 5, 8

[15] Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems*, 33:4163–4174, 2020. 3

[16] J Kim, J Kwon Lee, and K Mu Lee. Accurate image super-resolution using very deep convolutional networks: Corr. 2015. 8

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Reflash dropout in image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6002–6012, 2022. 2, 3, 5, 7

[19] Vladislav Li, George Amponis, Jean-Christophe Nebel, Vasileios Argyriou, Thomas Lagkas, Savvas Ouzounidis, and Panagiotis Sarigiannidis. Super resolution for augmented reality applications. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6. IEEE, 2022. 1

[20] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 3

[21] 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, 3, 8

[22] 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, 8

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3

[24] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[26] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018. 2

[27] Salma Abdel Magid, Yulun Zhang, Donglai Wei, Won-Dong Jang, Zudi Lin, Yun Fu, and Hanspeter Pfister. Dynamic high-pass filtering and multi-spectral attention for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4288–4297, 2021. 2

[28] 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

[29] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 191–207. Springer, 2020. 2

[30] Krzysztof Okarma, Mateusz Tecław, and Piotr Lech. Application of super-resolution algorithms for the navigation of autonomous mobile robots. In *Image Processing & Communications Challenges 6*, pages 145–152. Springer, 2015. 1

[31] 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 *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 31–42. Springer, 2014. 6, 7, 8

[32] 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. 3

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[34] Wonil Song, Sungil Choi, Somi Jeong, and Kwanghoon Sohn. Stereoscopic image super-resolution with stereo consistent feature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12031–12038, 2020. 2, 5

[35] Fanny Spagnolo, Pasquale Corsonello, Fabio Frustaci, and Stefania Perri. Design of a low-power super-resolution architecture for virtual reality wearable devices. *IEEE Sensors Journal*, 2023. 1

[36] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, and Aleksei Silvestrov. Naejin kong, harshith goka, kiwoong park, and victor lempitsky. 2021. resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 1, 3

[37] 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. 1, 2, 5

[38] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 61–72. Springer, 2019. 3

[39] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, and Radu Timofte. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *CVPRW*, 2023. 6, 7, 8

[40] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Liangyu Chen, Xiaojie Chu, Wenqing Yu, Kai Jin, et al. Ntire 2022 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 906–919, 2022. 1, 2, 6

[41] 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. 1, 2, 5, 8

[42] 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, 6, 7, 8

[43] 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. 1, 2, 5, 6, 7, 8

[44] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 2

[45] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 3, 5

[46] 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, 5, 6, 8

[47] Yankun Yu, Huan Liu, Minghan Fu, Jun Chen, Xiyao Wang, and Keyan Wang. A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2021. 6

[48] 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. 1, 5, 7

[49] 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, 8

[50] 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, 8

[51] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020. 2

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 6