Multi-Level Feature Fusion Network for Lightweight Stereo Image Super-Resolution

Yunxiang Li[1,*], Wenbin Zou[2,*], Qiaomu Wei[3,*], Feng Huang[4,†], Jing Wu[5]
Fuzhou University.[1,4,5] South China University of Technology.[2]
Chengdu University of Information Technology.[3]
1033649629@qq.com, alexzou14@foxmail.com, 1642445844@qq.com,
huangf@fzu.edu.cn, wujing@fzu.edu.cn

Abstract

Stereo image super-resolution utilizes the cross-view complementary information brought by the disparity effect of left and right perspective images to reconstruct higher-quality images. Cascading feature extraction modules and cross-view feature interaction modules to make use of the information from stereo images is the focus of numerous methods. However, this adds a great deal of network parameters and structural redundancy. To facilitate the application of stereo image super-resolution in downstream tasks, we propose an efficient Multi-Level Feature Fusion Network for Lightweight Stereo Image Super-Resolution (MFFSSR). Specifically, MFFSSR utilizes the Hybrid Attention Feature Extraction Block (HAFEB) to extract multi-level intra-view features. Using the channel separation strategy, HAFEB can efficiently interact with the embedded cross-view interaction module. This structural configuration can efficiently mine features inside the view while improving the efficiency of cross-view information sharing. Hence, reconstruct image details and textures more accurately. Abundant experiments demonstrate the effectiveness of MFFSSR. We achieve superior performance with fewer parameters. The source code is available at https://github.com/KarosLYX/MFFSSR.

1. Introduction

Stereo imaging utilizes two cameras to simulate the visual system of human, which has been widely applied in various fields such as augmented reality (AR) [27], virtual reality (VR) [13], and autonomous driving [15]. However, the hardware costs of stereo imaging devices may lead to low resolution (LR). Image super-resolution (SR) can effectively enhance the perceived quality of images by restoring high-frequency details lost during the imaging process using computational optics methods, thus attracting widespread attention.

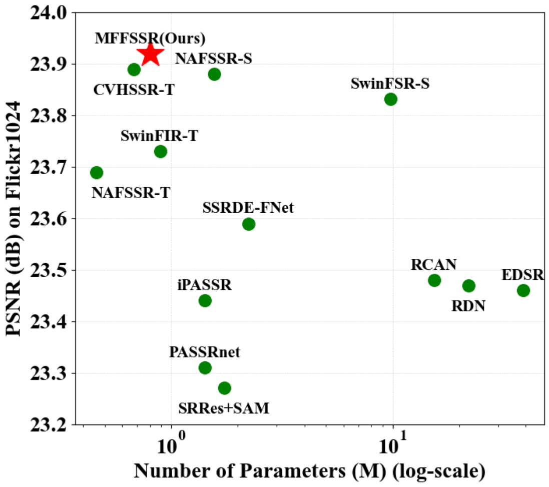

Figure 1. Comparison of the performance and complexity of state-of-the-art methods for 4× stereo SR on the Flickr1024 [34] test set. Our MFFSSR achieves superior performance with fewer parameters.

Recent years, the breakthrough of convolutional neural network (CNN) and Transformer technologies has enabled deep learning-based SR methods to demonstrate powerful performance in single image super-resolution (SISR) tasks [18, 21, 39, 42]. Different from SISR, which can only rebuild a high-resolution (HR) image with intra-view information, stereo images contain additional complementary information from cross-views. By fully leveraging the correlation between left and right perspective images, it is possible to reconstruct higher-quality HR images. However, the disparity effect can introduce uncertainty in the projected positions of objects across different perspective views. Positional variations are more pronounced for objects closer

[*] Equal contributions.
[†] Corresponding author.

to the camera than for those farther away. This complexity makes it challenging to effectively harness details from stereo images. To address this issue, existing methods generally focus on designing complex networks and training strategies. For example, Jeon *et al.* [14] use parallax prior and a two-stage joint network to enhance the spatial resolution of stereo images. Wang *et al.* [31] introduced a parallax-attention mechanism and achieved feature fusion from cross-views based on similarity measurement. Recently, Chu *et al.* [5] refined the Nonlinear Activation Free Network, NAFNet [2], and adapted it to stereo scenarios. Cheng *et al.* [4] suggested a hybrid Transformer and CNN Attention Network among with a three-stage training strategy. They demonstrated excellent results in the NTIRE Stereo Image Super-Resolution Challenge held in 2022 and 2023 respectively.

Despite the capability achieved by the aforementioned methods in extracting information from stereo images, directly cascading feature extraction modules and cross-view feature interaction modules can lead to significant parameter and element redundancy, posing challenges for deployment on edge computing platforms. Therefore, exploring ways to reuse feature information for both intra-view and cross-view feature fusion is crucial for improving the efficiency of stereo image SR networks.

In this work, we develop a Multi-Level Feature Fusion Network for Lightweight Stereo Image Super-Resolution (MFFSSR) to address the above issue. By using the channel separation strategy, we selectively learn intra-view and cross-view information, thereby reducing computational complexity. Specifically, we design a novel Hybrid Attention Feature Extraction Block (HAFEB) to extract multi-level intra-view features. We use Cross-View Interaction Module (CVIM) to extract cross-view information, which has been proven to be effective in [44]. In addition, we embed CVIM within HAFEB and utilize a branching structure to enhance the efficiency of cross-view feature interaction. Through these structures, MFFSSR can effectively integrate multi-level features, achieving high-quality SR with fewer parameters.

The main contributions of this work are as follows:

- We design HAFEB to extract and fuse multi-level intra-view features. By combining Channel Attention (CA) and Large Kernel Attention (LKA), HAFEB simultaneously reconstructs image details and structures while learning the correlations between local features. Residual connections further facilitate the transmission and fusion of features between different hierarchical levels, thereby preserving the richness and diversity of the extracted features.
- We integrate CVIM into HAFEB using a branching structure, leveraging partial intra-view features for cross-view interaction. Through a channel separation strategy, we optimize the cross-view information sharing mechanism, thus increasing efficiency and reducing computational complexity.
- Based on the designed framework, we propose an effective and lightweight stereo image SR method. As shown in Figure 1, we achieve superior performance with fewer parameters. Extensive experiments confirm the effectiveness of our approach.

2. Related Works

2.1. Single Image Super-Resolution

Recovering HR images from LR images is the aim of image SR, deep learning-based methods achieve this goal by learning the mapping relationship between a large number of LR images and their corresponding HR images. Since Dong *et al.* [9] initially suggested using CNN to achieve image SR task, many deep learning-based methods have emerged, demonstrating excellent performance. Kim *et al.* [16] further improved the effectiveness of CNN in image SR by increasing the depth of the network. By adding dense [36, 41] and residual [19, 40] connections, researchers have optimized the information flow and feature reuse amongst deep neural networks, thus increasing models' robustness and training speed. However, images contain rich multi-level information, and different information contributes differentially to the image SR task. In order to focus more on the important features and structural information in the image, Zhang *et al.* [40] proposed the channel attention mechanism. Since then, various attention mechanisms [7, 8, 22–24, 26, 32] have been proposed as effective means to enhance the performance of image SR.

Recently, Transformer has achieved significant success in the field of computer vision, with Transformer-based SR models achieving state-of-the-art (SOTA) results. However, these models often have a large number of parameters, which limits their practical use. Additionally, single image super-resolution (SISR) can only utilize information within the image itself, without fully exploiting complementary information from other images, thereby restricting further enhancement in potency.

2.2. Stereo Image Super-Resolution

Stereo image SR can make use of information across views to further enhance the resolution effect. Jeon *et al.* [14] proposed the first deep learning-based stereo image SR algorithm, StereoSR, which uses parallax prior and a two-stage joint network enhance the spatial resolution of stereo images. Song *et al.* [29] suggested a Self and Parallax Attention Mechanism (SPAM) to recover HR features while preserve stereo consistency between image pairs. In order to leverage texture-rich single image datasets, Ying *et al.* [38] developed a generic Stereo Attention Module (SAM)

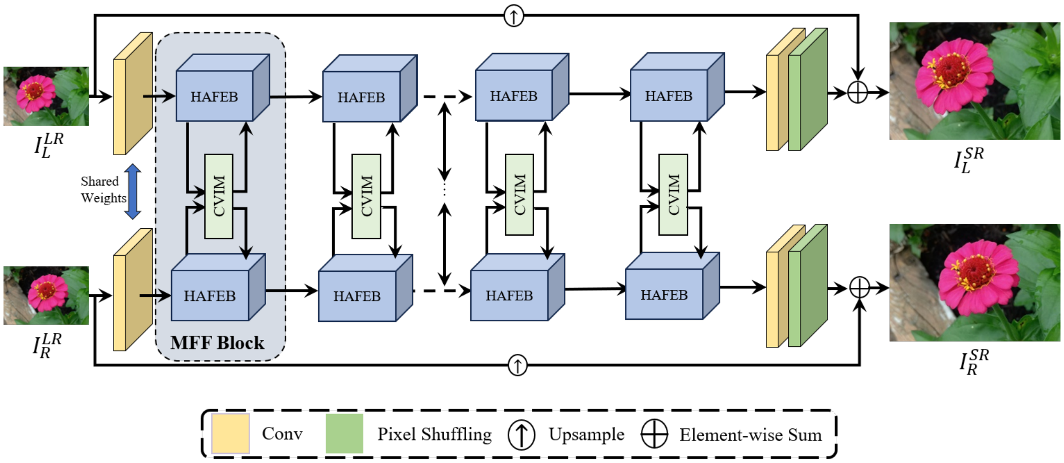

Figure 2. The framework of Multi-Level Feature Fusion Network for Lightweight Stereo Image SR (MFFSSR). HAFEB (shown in Figure 3.) and CVIM (shown in Figure 4.) represent the Hybrid Attention Feature Extraction Block and the Cross-View Interaction Module, respectively. Two HAFEBs with an embedded CVIM compose a MFF Block.

to extend SISR network to a stereo image SR network. Xu *et al.* [37] incorporated the idea of bilateral grid processing into a CNN framework to effectively utilize cross-view information. Wang *et al.* [35] utilized a Bi-directional Parallax Attention Module (BiPAM) to simultaneously interact with information from both left and right perspective images. Additionally, they addressed the issue of inconsistent illumination between left and right perspective images in real-world scenes by improving the loss function. Chu *et al.* [5] used a stack of NAFBlock [2] for intra-view feature extraction and combined it with stereo cross-attention modules for cross-view feature interaction, resulting in excellent performance. Zou *et al.* [44] improved upon their work by designing the CVHSSR, which effectively conveys mutual information between different views. In addition, Transformer-based methods [1, 4, 20] have begun to be applied in the field of stereo image SR, achieving impressive outcomes.

However, the above methods often focus solely on performance while neglecting the potential for application in downstream tasks. To address this issue, we design a lightweight stereo image SR network, redefining the processes of intra-view feature extraction and cross-view feature interaction to enhance the efficiency of the network.

3. Multi-Level Feature Fusion Network

3.1. Overall Framework

The network proposed by us is illustrated in Figure 2. MFFSSR employs a dual-branch network with shared weights to restore images from both left and right perspectives. It consists of three parts: shallow feature extraction, deep feature extraction and interaction, and stereo image reconstruction. The Multi-Level Feature Fusion Block (MFF Block) is the core component of the deep feature extraction and interaction, which consists of two Hybrid Attention Feature Extraction Blocks (HAFEBs) and an embedded Cross-View Interaction Module (CVIM). Detailed information about HAFEB and CVIM will be presented in Section 3.2 and Section 3.3, respectively. Specifically, the operation process of MFFSSR is as follows.

Firstly, given a pair of LR stereo images $I_L^{LR}, I_R^{LR} \in R^{H\times W\times 3}$, a simple convolutional operation is used for them to extract the shallow features $F_L^S, F_R^S \in R^{H\times W\times C}$, where *H*, *W*, and *C* represent the image's height, width, and number of channels, respectively. This process can be described as:

$$F_{L,R}^S = H_{\text{conv}}(I_L^{LR}, I_R^{LR}) \tag{1}$$

where H_{conv} denotes 3×3 convolution operation.

Next, we perform deep feature extraction and interactive fusion using MFF Block on the acquired features. The number of MFF Blocks, denoted by *N*, is flexible and can be adjusted. This process can be described as:

$$F_{L,R}^D = H_{\text{MFF}}^N(H_{\text{MFF}}^{N-1}(\cdots(H_{\text{MFF}}^1(F_{L,R}^S)))) \tag{2}$$

$$F_{L,R}^{i+1} = H_{\text{MFF}}(F_{L,R}^i) \tag{3}$$

where H_{MFF} denotes MFF Block. $F_{L,R}^D, F_{L,R}^{i+1}$ denote the features after deep extraction and interactive fusion and the features obtained after processing by the i-th MFF Block, respectively.

Finally, we utilize the pixel shuffling operation to upsample the output features to the HR size. Furthermore, a global residual structure is used to maintain input image features and increase the performance of SR. This process can be described as:

$$I_L^{SR} = H_{up}(F_L^D) + H_{up}(I_L^{LR}) \tag{4}$$

$$I_R^{SR} = H_{up}(F_R^D) + H_{up}(I_R^{LR}) \tag{5}$$

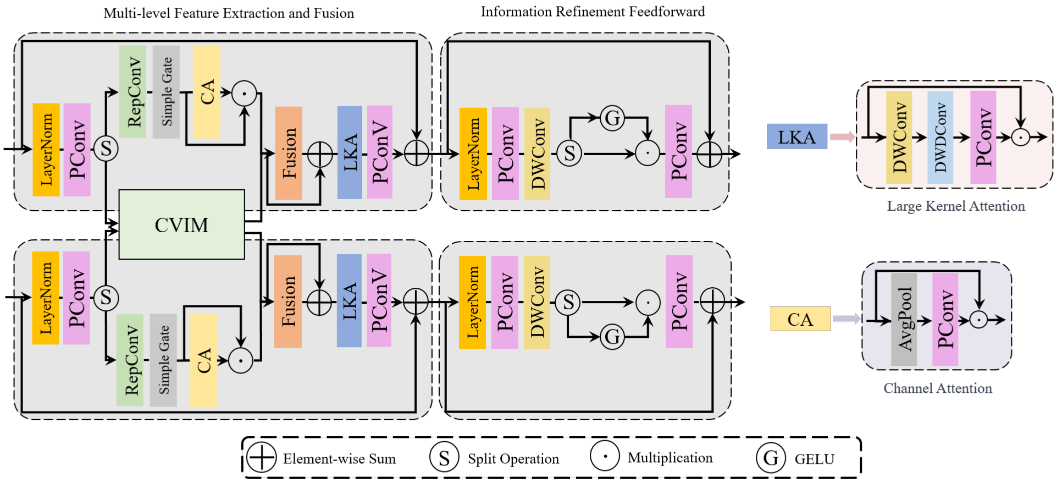

Figure 3. The architecture of our proposed Multi-level Feature Fusion Block (MFF Block). A MFF Block consists of two Hybrid Attention Feature Extraction Blocks (HAFEB) and an embedded Cross-View Interaction Module (CVIM). Each HAFEB has two components: Multi-level Feature Extraction and Fusion (MFEF) and Information Refinement Feedforward (IRF). The HAFEBs for the left and right views are connected to CVIM through the branch structures in MFEF, facilitating the interaction and fusion of features across the views. PConv, RepConv, DWConv and DWDConv in the figure represent point-wise convolution, reparameterized convolution, depth-wise convolution, and depth-wise dilation convolution, respectively.

where H_{up} denotes upsampling operation. I_L^{SR} and I_R^{SR} represent the final left and right perspective images after SR, respectively.

3.2. Intra-View Feature Extraction

Stereo images contain information spanning global, local, and cross-view ranges. Intra-view feature extraction serves as the foundation for cross-view interaction. To efficiently capture and fuse these multi-level features, we introduce the Hybrid Attention Feature Extraction Block (HAFEB).

As shown in Figure 3, the HAFEB consists of two components: (1) Multi-level Feature Extraction and Fusion (MFEF) and (2) Information Refinement Feedforward (IRF). In addition to channel attention and large kernel attention mechanisms, we also employ reparameterized convolution (RepConv) in MFEF. Their comprehensive use significantly enhances the capability and flexibility of HAFEB in feature extraction. Furthermore, we use branch structures to interact partial intra-view features with the embedded Cross-View Interaction Module (CVIM), therefore reducing computational complexity and substantially improving the efficiency of cross-view information fusion.

Given an input tensor $Fin \in R^{H\times W\times C}$, the working process of MFEF can be described as follows:

$$F_{\text{MFEF}} = H^2_{\text{pconv}}(H_{\text{LKA}}(H_{\text{f}}(\kappa(H^1_{\text{pconv}}(LN(F_{\text{in}})))) \\ + \kappa(H^1_{\text{pconv}}(LN(F_{\text{in}}))))) + F_{\text{in}} \tag{6}$$

where $LN(\cdot)$ denotes layer normalization. $H^{(\cdot)}_{\text{pconv}}$, H_{f}, H_{LKA} represent 1×1 point-wise convolution, feature fusion operation, and large kernel attention, respectively. F_{MFEF} is the output feature of MFEF. We use notation $\kappa(\cdot)$ to represent hybrid feature fusion extraction operation. Specifically, given the input feature $X \in R^{H\times W\times C}$, it is firstly split into two parts $X_1 \in R^{H\times W\times\theta}$, $X_2 \in R^{H\times W\times(1-\theta)}(\theta \in [0, 1])$ on channel dimension. λ is a hyperparameter that controls the ratio. We set $\theta = 0.75$ to balance between the parameters and efficiency. More detailed information about it can be found in the ablation study in Section 4.3. Then, X_1 and X_2 are used for further feature extraction and cross-view information interaction, respectively. This process can be described as:

$$\kappa(X) = \delta_{\text{SG}}(H_{\text{rconv}}(X_1)) \odot H_{\text{CA}}(\delta_{\text{SG}}(H_{\text{rconv}}(X_1))) \\ + F_{\text{CVIM}}(X_2) \tag{7}$$

where H_{rconv} and F_{CVIM} represent the RepConv and the output feature of CVIM, respectively. δ_{SG} means Simple-Gate function and $\odot$ denotes element-wise multiplication.

The internal computation process of large kernel attention and channel attention can be described as follows:

$$H_{\text{LKA}}(X) = X \odot (H_{\text{pconv}}(H_{\text{dd7}}(H_{\text{d5}}(X)))) \tag{8}$$

$$H_{\text{CA}}(X) = X \odot (H_{\text{pconv}}(H_{\text{Avg}}(X))) \tag{9}$$

where H_{Avg}, H_{d5} and H_{dd7} represent average pooling operation, 5×5 depth-wise convolution, and 7×7 depth-wise dilation convolution respectively.

Then, IRF employs a non-linear gate mechanism to focus on complementary details across different levels. The

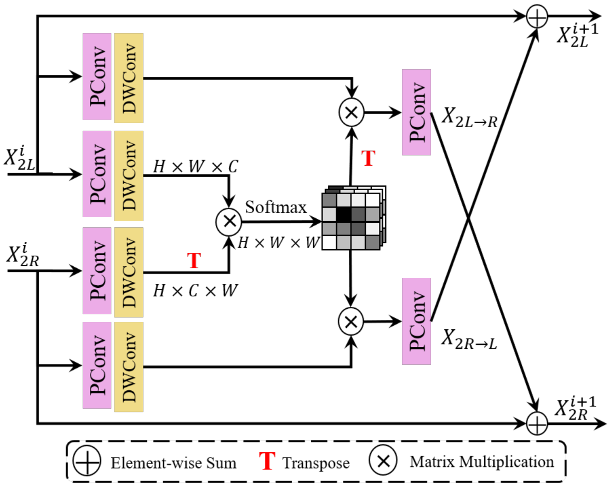

Figure 4. The architecture of Cross-View Interaction Module (CVIM). It is embedded in two Hybrid Attention Feature Extraction Blocks of the parallel branches to achieve efficient cross-view feature interaction. PConv, DWConv in the figure represent point-wise convolution, depth-wise convolution, respectively.

working process of IRF can be described as follows:

$$F_{\text{out}} = H^{4}_{\text{pconv}}(\delta_{\text{NG}}(H^{1}_{\text{d3}}(H^{3}_{\text{pconv}}(LN(F_{\text{MFEF}}))))) + F_{\text{MFEF}} \tag{10}$$

where $H^{(\cdot)}_{\text{d3}}$ and δ_{NG} represent 3×3 depth-wise convolution and non-linear gate function, respectively. The F_{out} denotes the output feature of HAFEB.

3.3. Cross-View Feature Interaction

We refine the Cross-View Interaction Module (CVIM) proposed in CVHSSR [44]. Redundant cross-view feature interaction has little contribution to SR performance improvement but can lead to a significant increase in computational complexity. To improve cross-view interaction efficiency and reduce parameters, we constrain the input feature in dimension. Additionally, layer normalization is removed for better integration into the Hybrid Attention Feature Extraction Blocks. The details of CVIM is as shown in Figure 4. It combines Scaled DotProduct Attention [30], which utilizes queries and keys to generate corresponding weights:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q}\mathbf{K}^{T}/\sqrt{C})\mathbf{V} \tag{11}$$

where $\mathbf{Q} \in R^{H\times W\times C}$ is the query matrix from one view, and $\mathbf{K}, \mathbf{V} \in R^{H\times W\times C}$ are key and query matrices to another view.

CVIM efficiently facilitates the interaction between left and right view information. Given the input stereo partial intra-view features $X^{i}_{2L}, X^{i}_{2R} \in R^{H\times W\times C}$, we can get the cross-view fusion features $X_{2L\to R}$ through the following process:

$$\mathbf{Q}_{\text{L}} = H^{Q_L}_{d3}(H^{Q_L}_{\text{pconv}}(X^{i}_{2L})) \tag{12}$$

$$\mathbf{K}_{\text{R}} = H^{K_R}_{d3}(H^{K_R}_{\text{pconv}}(X^{i}_{2R})) \tag{13}$$

$$\mathbf{V}_{\text{R}} = H^{V_R}_{d3}(H^{V_R}_{\text{pconv}}(X^{i}_{2R})) \tag{14}$$

$$X_{2L\to R} = H^{R}_{\text{pconv}} Attention_{\text{L}\to\text{R}}(\mathbf{Q}_{\text{L}}, \mathbf{K}_{\text{R}}, \mathbf{V}_{\text{R}}) \tag{15}$$

$X_{2R\to L}$ can be obtained through the similar process:

$$\mathbf{Q}_{\text{R}} = H^{Q_R}_{d3}(H^{Q_R}_{\text{pconv}}(X^{i}_{2R})) \tag{16}$$

$$\mathbf{K}_{\text{L}} = H^{K_L}_{d3}(H^{K_L}_{\text{pconv}}(X^{i}_{2L})) \tag{17}$$

$$\mathbf{V}_{\text{L}} = H^{V_L}_{d3}(H^{V_L}_{\text{pconv}}(X^{i}_{2L})) \tag{18}$$

$$X_{2R\to L} = H^{L}_{\text{pconv}} Attention_{\text{R}\to\text{L}}(\mathbf{Q}_{\text{R}}, \mathbf{K}_{\text{L}}, \mathbf{V}_{\text{L}}) \tag{19}$$

The cross and intra view features are finally fused to generate the output features X^{i+1}_{2L} and X^{i+1}_{2R}:

$$X^{i+1}_{2L} = \gamma_L X_{2L\to R} + X^{i}_{2L} \tag{20}$$

$$X^{i+1}_{2R} = \gamma_R X_{2R\to L} + X^{i}_{2R} \tag{21}$$

where γ_L and γ_R are trainable channel-wise scales and initialized with zeros for stabilizing training.

3.4. Loss Function

The loss function defines the optimization objective of the SR network and plays a crucial role in determining how well it performs. Zou *et al.* have already demonstrated the effectiveness of utilizing spatial and frequency domain losses to jointly guide the SR network for image restoration [44].

Specifically, we use the MSE loss to measure the spatial structural difference between the SR images $I^{SR}_{L,R}$ and the HR images $I^{HR}_{L,R}$, which can be described as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left\| I^{HR}_{L,R} - I^{SR}_{L,R}\right\|^{2} \tag{22}$$

Additionally, frequency Charbonnier loss is introduced to guide the learning of high-frequency information in SR images, aiding in better preservation of details and textures. It can be defined as:

$$L_{FC} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left\| FFT(I^{HR}_{L,R}) - FFT(I^{SR}_{L,R})\right\|^{2} + \varepsilon^{2}} \tag{23}$$

where ε is a constant and is set to 10^{-3}. $FFT(\cdot)$ denotes the fast Fourier transform.

In conclusion, the overall loss function can be expressed as:

$$L_{Total} = L_{MSE}(I^{HR}_{L,R}, I^{SR}_{L,R}) + \lambda L_{FC}(I^{HR}_{L,R}, I^{SR}_{L,R}) \tag{24}$$

where λ is a hyperparameter, it is set to 0.01 to control the proportion of the frequency Charbonnier loss function.

Table 1. Quantitative results achieved by different methods on the KITTI2012 [5], KITTI2015 [23], Middlebury [10], and Flickr1024 [20] test sets. Params represents the number of parameters of the networks. 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 and second best results are red and blue.

Method	Scale	Params	$Left$			$(Left + Right)/2$			
			KITTI2012	KITTI2015	Middlebury	KITTI2012	KITTI2015	Middlebury	Flickr1024
VDSR [16]	×2	0.66M	30.17/0.9062	28.99/0.9038	32.66/0.9101	30.30/0.9089	29.78/0.9150	32.77/0.9102	25.60/0.8534
EDSR [19]	×2	38.6M	30.83/0.9199	29.94/0.9231	34.84/0.9489	30.96/0.9228	30.73/0.9335	34.95/0.9492	28.66/0.9087
RDN [42]	×2	22.0M	30.81/0.9197	29.91/0.9224	34.85/0.9488	30.94/0.9227	30.70/0.9330	34.94/0.9491	28.64/0.9084
RCAN [40]	×2	15.3M	30.88/0.9202	29.97/0.9231	34.80/0.9482	31.02/0.9232	30.77/0.9336	34.90/0.9486	28.63/0.9082
StereoSR [14]	×2	1.08M	29.42/0.9040	28.53/0.9038	33.15/0.9343	29.51/0.9073	29.33/0.9168	33.23/0.9348	25.96/0.8599
PASSRnet [31]	×2	1.37M	30.68/0.9159	29.81/0.9191	34.13/0.9421	30.81/0.9190	30.60/0.9300	34.23/0.9422	28.38/0.9038
IMSSRnet [17]	×2	6.84M	30.90/-	29.97/-	34.66/-	30.92/-	30.66/-	34.67/-	-/-
iPASSR [35]	×2	1.37M	30.97/0.9210	30.01/0.9234	34.41/0.9454	31.11/0.9240	30.81/0.9340	34.51/0.9454	28.60/0.9097
SSRDE-FNet [6]	×2	2.10M	31.08/0.9224	30.10/0.9245	35.02/0.9508	31.23/0.9254	30.90/0.9352	35.09/0.9511	28.85/0.9132
PFT-SSR [12]	×2	-	31.15/0.9166	30.16/0.9187	35.08/0.9516	31.29/0.9195	30.96/0.9306	35.21/0.9520	29.05/0.9049
SwinFIR-T [39]	×2	0.89M	31.09/0.9226	30.17/0.9258	35.00/0.9491	31.22/0.9254	30.96/0.9359	35.11/0.9497	29.03/0.9134
NAFSSR-T [5]	×2	0.45M	31.12/0.9224	30.19/0.9253	34.93/0.9495	31.26/0.9254	30.99/0.9355	35.01/0.9495	28.94/0.9128
NAFSSR-S [5]	×2	1.54M	31.23/0.9236	30.28/0.9266	35.23/0.9515	31.38/0.9266	31.08/0.9367	35.30/0.9514	29.19/0.9160
CVHSSR-T [44]	×2	0.66M	31.31/0.9250	30.33/0.9277	35.41/0.9533	31.46/0.9280	31.13/0.9377	35.47/0.9532	29.26/0.9180
MFFSSR (Ours)	×2	0.78M	31.35/0.9255	30.36/0.9281	35.45/0.9533	31.50/0.9285	31.16/0.9380	35.51/0.9531	29.38/0.9198
VDSR [16]	×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 [42]	×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 [40]	×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 [14]	×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 [31]	×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 [38]	×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 [35]	×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 [12]	×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 [39]	×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 [44]	×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
MFFSSR (Ours)	×4	0.84M	26.89/0.8109	26.05/0.7992	29.64/0.8498	26.99/0.8169	26.78/0.8219	29.75/0.8507	23.92/0.7503

Table 2. Efficiency evaluations with the state-of-the-art methods. Params represents the number of parameters of the network. *v* represents the variant after reducing parameters. We use Params and FLOPs to evaluate efficiency.

Method	Params	FLOPs
NAFSSR-S [5]	1.54M	36.531G
SwinFIRSSR-*v*	0.89M	48.956G
SCGLANet-*v*	0.75M	28.244G
MFFSSR (Ours)	0.91M	27.415G

4. Experiments

4.1. Implementation Details

In this section, we provide a detailed overview of the datasets, evaluation metrics, and model configurations.

Datasets. Following previous works [6, 35, 37, 38], we used publicly available stereo image datasets for our training and testing. Specifically, we used 800 pairs of images from the Flickr1024 [34] dataset and 60 pairs of images from the Middlebury [28] dataset for training. We use four benchmark test sets: KITTI 2012 [11], KITTI 2015 [25], Middlebury [28], and Flickr 1024 [34] , to fully verify the effectiveness of our model.

Evaluation metrics. We evaluate our model using Peak Signal-To-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) on the RGB color space.

Model Setting. The number of MFF Block and feature channels is flexible and can be changed. In this paper, we adjust the settings in NTIRE 2024 competition to further reduce the number of parameters and improve efficiency. In the ×2 SR model, the number of blocks and the number of channels are set to 16 and 64, respectively. In the ×4 SR model, the number of blocks and the number of channels are set to 24 and 48, respectively.

Training Setting. We augment the training data using random horizontal flipping, rotation, and RGB channel shuffling. We use Lion [3] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$. MFFSSR is training in PyTorch on a server with eight Nvidia A100 GPUs. The learning rate is initially set to 5×10^{-4} and decay the learning rate with the cosine strategy. The total iterations for the model is set to 200,000.

4.2. Comparisons with State-of-the-art Methods

In this section, we compare our proposed MFFSSR with existing image SR methods. These methods include VDSR

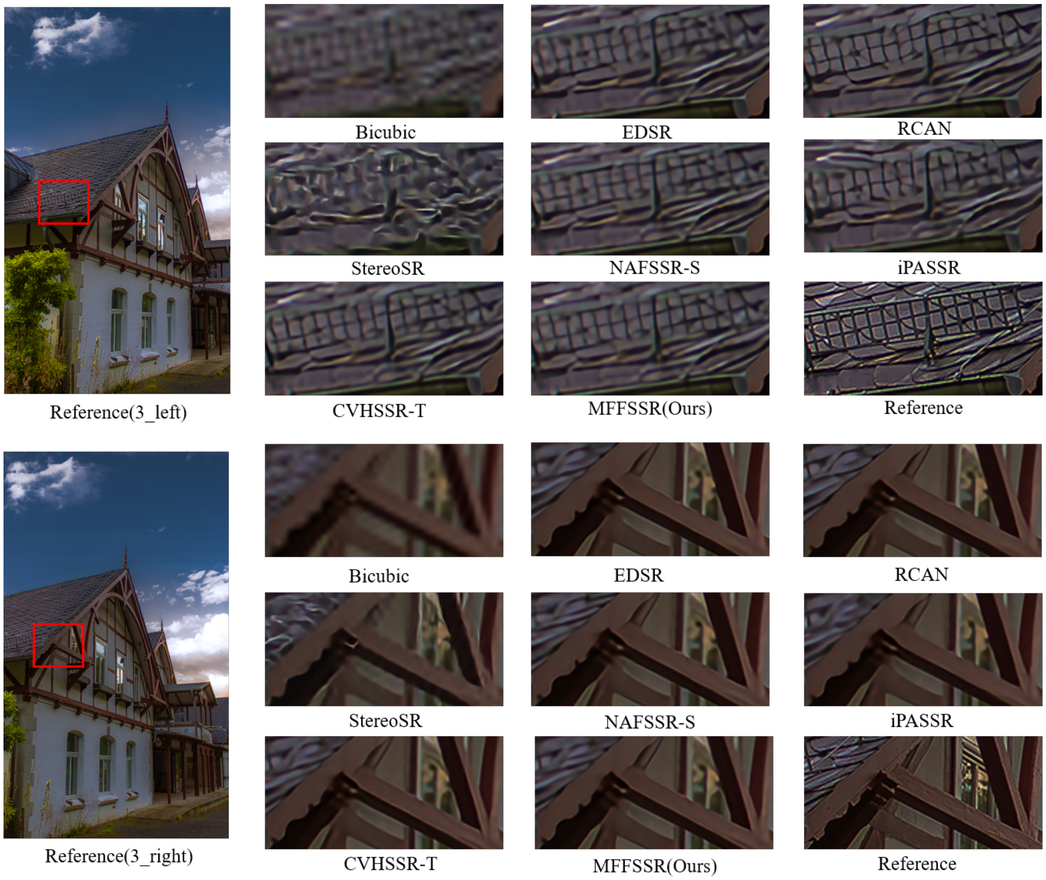

Figure 5. Visual results (×4 SR) achieved by different methods on the Flickr1024 [34] test set.

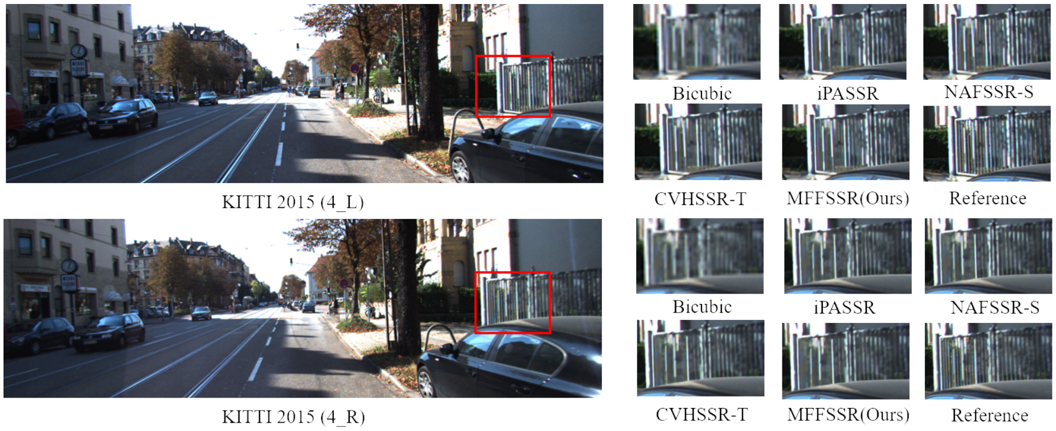

Figure 6. Visual results (×4 SR) achieved by different methods on the KITTI 2015 [25] test set.

[16], EDSR [19], RDN[42], RCAN[40], StereoSR [14], PASSRnet[31], IMSSRnet [17], iPASSSR [35], SSRDE-FNet [6], PFT-SSR [12], SwinFIR [39], NAFSSR [5] and CVHSSR [44] and so on. All models are trained on the same datasets, and these results are from [44].

Quantitative Evaluations. We summarize the SR results of MFFSSR and other SR methods at both ×2 and ×4 upsampling factors in Table 1. Our MFFSSR achieves the best performance with low parameters.

Efficiency Evaluations. We compare the parameters and computational complexity with the state-of-the-art methods from the NTIRE Stereo Image SR Challenge

Table 3. Ablation experiments of different elements in hybrid attention feature extraction block on the Flickr1024 test set [34]. PSNR and SSIM are used to evaluate performance.

	MFFSSR	Net-A	Net-B	Net-C	Net-D
LKA	✔	✘	✔	✘	✔
RepConv	✔	✔	✘	✘	✔
IRF	✔	✔	✔	✔	✘
FFN	✘	✘	✘	✘	✔
PSNR	23.92	23.87	23.91	23.85	23.91
SSIM	0.7503	0.7490	0.7498	0.7476	0.7501
Δ PSNR	0	-0.05	-0.01	-0.07	-0.01

Table 4. Ablation experiments of different weights θ and cross-attention modules on the Flickr1024 test set [34]. Params represents the number of parameters of the network. PSNR, SSIM, Params and FLOPs are used to evaluate performance and efficiency.

	$\theta = 0.250$	$\theta = 0.500$	$\theta = 0.750$	$\theta = 0.750$	$\theta = 0.875$
CVIM	✔	✔	✔	✘	✔
SCAM	✘	✘	✘	✔	✘
PSNR	23.89	23.90	23.92	23.89	23.86
SSIM	0.7489	0.7493	0.7503	0.7487	0.7465
Params	1.52M	1.13M	0.84M	0.80M	0.77M
FLOPs	119.157G	100.191G	89.187G	88.536G	86.671G
Δ PSNR	-0.03	-0.02	0	-0.03	-0.06
Δ Params	+0.68M	+0.29M	0	-0.04M	-0.07M
Δ FLOPs	+29.970G	+11.004G	0	-0.651G	-2.516G

in 2022 and 2023. We achieve better performance than NAFSSR-S [5] while utilizing only 60% of the parameters and 75% of the FLOPs. For fair comparison, we reduce the parameter of SwinFIRSSR [39] and SCGLANet [43] to a level equivalent to that of MFFSSR, resulting in two variant networks: SwinFIRSSR-*v* and SCGLANet-*v*. As shown in Table 2, even in the condition that the parameters are slightly greater than SwinFIRSSR-*v* and SCGLANet-*v*, our computational cost remains lower. The evaluations are tested on 128×128 size as inputs.

Visual Comparison. Figure 5. and Figure 6. display the ×4 SR visualization results of different methods, our model is more visually realistic in perception and provides a clearer restoration of textures and features compared with existing methods. It notably achieves a better restoration effect on fences and buildings.

4.3. Ablation Study

In this section, we conduct various ablation experiments to validate the effectiveness of the proposed structures and parameter settings. All ablation results are obtained using the Flickr1024 [34] test set.

Effectiveness of Hybrid Attention Feature Extraction Block. To further validate the effectiveness of the proposed structures, we investigate the roles of different elements in HAFEB, resulting in four network variants: Net-A (without LKA), Net-B (without RepConv), Net-C (without LKA and RepConv) and Net-D (replace IRF with the simple FFN in NAFSSR [5]). The comparison results are presented in Table 3. LKA plays the most important role in feature extraction process, which offers a wider receptive field. RepConv enhances the network's generalization capability and perceptual capacity for details. The HAFEB leverages a residual structure to fully preserve the high and low-level features obtained from different elements, and integrates them with the cross-view features obtained by CVIM. We can observe that the performance of MFFSSR deteriorates when LKA and RepConv are removed. Removing both results in a decrease in PSNR of 0.07 dB. In addition, IRF more effectively regulate the information flow, achieving the 0.01 dB improvement in PSNR compared to the original FFN. These results demonstrate the crucial role these elements play in our network, indicating their indispensability.

Performance and Efficiency Trade-offs. As shown in Table 4, we compare the effects of (a) different branch weights θ in the multi-level extraction and fusion structure and (b) different cross-attention modules on network performance and efficiency. As the intra-view feature extraction branch weights θ gradually increase, the parameters and FLOPs decrease. However, θ too big leads to insufficient utilization of complementing information from cross-views, causing bad performance. When $\theta = 0.750$, we achieve the best performance, but the increase in parameters and FLOPs is minimal compared to $\theta = 0.875$. Additionally, we compare CVIM with NAFSSR's cross-attention module, SCAM [5]. We obtain a 0.03 dB improvement in PSNR at the cost of 0.04M parameters and 0.651G FLOPs. We believe this trade-off is worthwhile.

4.4. NTIRE Stereo Image SR Challenge

We have submitted the results of our original model to the NTIRE 2024 Stereo Image Super-Resolution Challenge [33]. Our final scores are 23.53 dB PSNR and 21.50 dB PSNR in Track 1 Constrained SR & Bicubic Degradation and Track 2 Constrained SR & Realistic Degradation, respectively, ranking 7th and 9th.

5. Conclusion

In this work, we propose the Multi-Level Feature Fusion Network for Lightweight Stereo Image Super-Resolution (MFFSSR). By improving the process of intra-view feature extraction for stereo images, we introduce the hybrid attention feature extraction block, which can effectively extract multi-level features. Furthermore, we creatively integrate the cross-view interaction module into the intra-view feature extraction structure, which greatly improves effectiveness and decreases the parameters. Extensive experiments demonstrate our proposed model outperforms state-of-the-art methods in stereo image super-resolution.

References

[1] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1764–1774, 2023. 3

[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] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[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, 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. 6, 7

[7] 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

[8] 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 3879–3886, 2019. 2

[9] 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

[10] Chenyang Duan and Nanfeng Xiao. Parallax-based spatial and channel attention for stereo image super-resolution. *IEEE Access*, 7:183672–183679, 2019. 6

[11] 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, 2012. 6

[12] 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, 2023. 6, 7

[13] Zhi-min Guo et al. Research of hand positioning and gesture recognition based on binocular vision. In *2011 IEEE International Symposium on VR Innovation*, pages 311–315, 2011. 1

[14] 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. 2, 6, 7

[15] Bingxi Jia, Jian Chen, and Kaixiang Zhang. Drivable road reconstruction for intelligent vehicles based on two-view geometry. *IEEE Transactions on Industrial Electronics*, 64(5): 3696–3706, 2016. 1

[16] 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, 7

[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, 7

[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

[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

[20] Jianxin Lin, Lianying Yin, and Yijun Wang. Steformer: Efficient stereo image super-resolution with transformer. *IEEE Transactions on Multimedia*, 25:8396–8407, 2023. 3, 6

[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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[22] 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

[23] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5690–5699, 2020. 6

[24] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3517–3526, 2021. 2

[25] 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

[26] 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, 2020. 2

[27] Taragay Oskiper, Mikhail Sizintsev, Vlad Branzoi, Supun Samarasekera, and Rakesh Kumar. Augmented reality binoculars. *IEEE transactions on visualization and computer graphics*, 21(5):611–623, 2015. 1

[28] 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, 2014. 6

[29] 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*, pages 12031–12038, 2020. 2

[30] 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

[31] 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. 2, 6, 7

[32] Li Wang, Jie Shen, E Tang, Shengnan Zheng, and Lizhong Xu. Multi-scale attention network for image super-resolution. *Journal of Visual Communication and Image Representation*, 80:103300, 2021. 2

[33] 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. 8

[34] 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 3852–3857, 2019. 1, 6, 7, 8

[35] 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. 3, 6, 7

[36] Jiu Xu, Yeongnam Chae, Björn Stenger, and Ankur Datta. Dense bynet: Residual dense network for image super resolution. In *2018 25th IEEE International conference on image processing (ICIP)*, pages 71–75, 2018. 2

[37] 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. 3, 6

[38] 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, 6

[39] 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, 6, 7, 8

[40] 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

[41] 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

[42] 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. 1, 6, 7

[43] Yuanbo Zhou, Yuyang Xue, Wei Deng, Ruofeng Nie, Jiajun Zhang, Jiaqi Pu, Qinquan Gao, Junlin Lan, and Tong Tong. Stereo cross global learnable attention module for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1416–1425, 2023. 8

[44] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1396–1405, 2023. 2, 3, 5, 6, 7