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

Cross-view Aggregation Network For Stereo Image Super-Resolution

Zhitao Chen, Tao Lu*, Kanghui Zhao, Bolin Zhu, Zhen Li, Jiaming Wang, Yanduo Zhang
Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology
2290445045czt@gmail.com lutxyl@gmail.com

Abstract

Although stereo image super-resolution has been extensively studied, many existing works only rely on attention in a single epipolar direction to reconstruct stereo images. In the case of asymmetric parallax images, these methods often struggle to capture reliable stereo correspondence, resulting in reconstructed images suffering from blurring and artifacts. In this paper, we propose a novel method called Cross-View Aggregation Network for Stereo Image Super-Resolution (CANSSR) and explore the relationship between multi-directional epipolar lines to construct reliable stereo correspondence. Specifically, we propose a multi-directional cross-view aggregation module (MCAM) that effectively captures multi-directional stereo correspondence and obtains cross-view complementary information. Furthermore, we design a channel-spatial aggregation module (CSAM) that aggregates multi-order global-local information in intra-view to reconstruct clearer texture features. In addition, we equip a large kernel convolution in the Feed-forward Network to acquire richer detailed texture information. The extensive experiments conclusively demonstrate that CANSSR outperforms the state-of-the-art method both qualitatively and quantitatively in terms of stereo image super-resolution on the Flickr 1024 and Middlebury datasets.

1. Introduction

Recently, there has been a noticeable surge in the utilization of stereo imaging devices, particularly within the domains of dual-lens smartphones, unmanned systems, augmented reality, virtual reality, autonomous driving, and robotics, etc. Stereoscopic vision has received substantial attention from both academia and industry. However, due to the physical imaging limitations [14] of binocular cameras, low-resolution (LR) stereo images pose significant challenges for practical applications [31]. Therefore, reconstructing high-resolution (HR) images is extremely urgent for the stereo vision task. Compared with single image super-resolution (SISR), stereo image super-resolution (SR) needs to utilize complementary information in cross-views, and lost or occluded details are restored by leveraging complementary information from the another view image. In practice, due to the binocular camera imaging settings, stereo images often exhibit a horizontal or vertical pixel-level offset, known as horizontal parallax and vertical parallax. Several studies [7][31] have demonstrated that the parallax effect between LR images induces sub-pixel displacement, which contains huge spatial dependence information in the stereo vision system. However, these methods only utilize horizontal parallax prior and fail to consider vertical parallax prior in order to improve network performance. Therefore, it is crucial to effectively utilize the multi-directional parallax prior for stereo image SR.

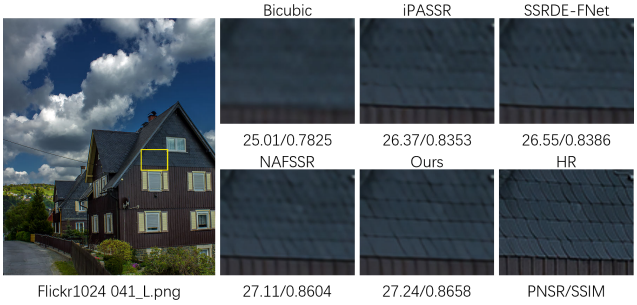

Figure 1. Visual comparison (×4) on Flickr1024 [30] dataset. iPASSR [31], SSRDE-FNet [7], and NAFSSR [6] suffer from blurring artifacts, **Please zoom in for details**.

Recently, deep learning methods have made great progress in stereo image SR. Existing approaches frequently employ attention mechanisms to capture stereo correspondence and spatial dependencies to enhance model performance. To capture stereo correspondence, several studies [7, 31] proposed utilizing parallax attention along the horizontal epipolar line. To obtain spatial dependencies, [4, 6] integrate cross-view information to effectively capture similarity features between stereo images while reducing loss for intra-view and cross-view high-frequency

*Corresponding author

detail information, and won the NTIRE [27, 28] champion with state-of-the-art performance. Lin *et al.* [19] proposed a lightweight transformer architecture to capture long-range dependencies between stereo images.

Although existing methods have achieved commendable performance, these methods excessively prioritize capturing stereo correspondence along the horizontal epipolar line, and do not effectively capture spatial dependencies. In practice, asymmetric parallax is often observed, which can affect the generalization ability of existing methods [7, 31] that assume parallax only exists in the horizontal direction. As shown in Figure 1, some methods [6, 7, 31] based on horizontal parallax prior suffer from blurring and artifacts, often stemming from their incapacity to accurately capture stereo correspondence. Therefore, an intriguing research problem is how to efficiently capture spatial dependencies while incorporating multi-directional parallax priors for stereo correspondence.

To address this issue, we propose a novel method named Cross-View Aggregation Network for Stereo Image Super-Resolution (CANSSR) that exploits multi-directional parallax attention to capture both horizontal and vertical stereo correspondences while enhancing long-range dependence. Specifically, we propose a multi-direction cross-view aggregation module to aggregate the horizontal and vertical stereo correspondences to obtain more reliable complementary information from cross-view. Furthermore, to effectively aggregate the high-frequency detailed information within intra-view, we propose a channel-spatial aggregation module to enhance the ability to capture multi-order global-local information. Finally, we introduce a large-kernel gated feed-forward network to aggregate richer texture information, and a non-linear free activation function [6] is introduced to enhance the non-linear representation ability.

In this study, we conducted comprehensive experiments to demonstrate the superior performance of our proposed CANSSR method across multiple datasets, including Flickr 1024 [30], KITTI 2012 [11], KITTI 2015 [22], and Middlebury [24]. Our contributions can be summarized as follows:

- We propose a novel cross-view aggregation network for stereo image super-resolution, which is capable of acquiring cross-view correspondence features and exhibits a robust capability to capture spatial dependencies.
- To exploit horizontal and vertical parallax prior, we propose MCAM, which learns stereo correspondence in both directions along the epipolar lines. Meanwhile, in order to capture global-local features, we design CSAM to learn multi-order interactions in intra-view.
- We conducted extensive experiments to demonstrate that CANSSR outperforms state-of-the-art methods while maintaining lower model size.

2. Related work

2.1. Single Image Super-Resolution

The SISR task aims to recover a HR image from a LR image by restoring lost high-frequency details. SRCNN [10] was an early deep learning method proposed for SISR, significantly improving reconstruction performance. Since then, numerous deep learning-based image SR techniques have been introduced, progressively employed more complex convolutional neural networks for high-quality SR image reconstruction. Skip-connections, employed in various methods [15, 18, 25, 33], play a crucial role in accelerating convergence and enhancing the efficiency and quality of reconstruction. Subsequently, various attention mechanisms proposed to enhance the expressive power of neural networks, such as spatial attention [23], second-order channel attention [8], and non-local attention [21, 35]. Recently, the Transformer emerged as a crucial component in low-level vision tasks. IPT [1] introduced a vision transformer that significantly enhances image restoration capacity. Meanwhile, SwinIR [17] introduced the Swin Transformer as a solution to address excessive computational redundancy in IPT [1]. The HAT [3] method achieved state-of-the-art performance by integrating multi-head self-attention from transformers with channel attention, extracting both global and local features. However, these methods are not directly applicable to stereo image SR, since they are unable to utilize cross-view supplementary information.

2.2. Stereo Image Super-Resolution

Stereo image SR reconstructs HR images from degraded pairs of left and right view LR images by leveraging complementary information. Jeon *et al* proposed the StereoSR [14] that leveraged parallax priors to reconstruct stereo images and introduced the illumination and chrominance subnetworks to acquire high-frequency detail information. To address parallax variation, some works [7, 26, 31, 32] proposed based on the parallax attention module (PAM), which effectively interacts with cross-view information along the horizontal epipolar line. To capture effective spatial dependencies, Dan *et al.* proposed DFAM [9], a modified atrous spatial pyramid pooling module designed for estimating disparities and warping stereo features. Lin *et al.* proposed Steformer [19], an efficient stereo image SR method based on Transformer, effectively capturing long-range dependence. Some methods [4, 6] have achieved state-of-the-art performance by exploiting their strong ability to capture spatial dependencies and have won the NTIRE [27, 28] championship. Zou *et al.* proposed CVHSSR[36], which explores the interdependencies between various hierarchies from intra-view and achieved excellent results in the NTIRE 2023 [28] competition. Furthermore, current methods that utilize parallax priors are limited to exploiting horizontal

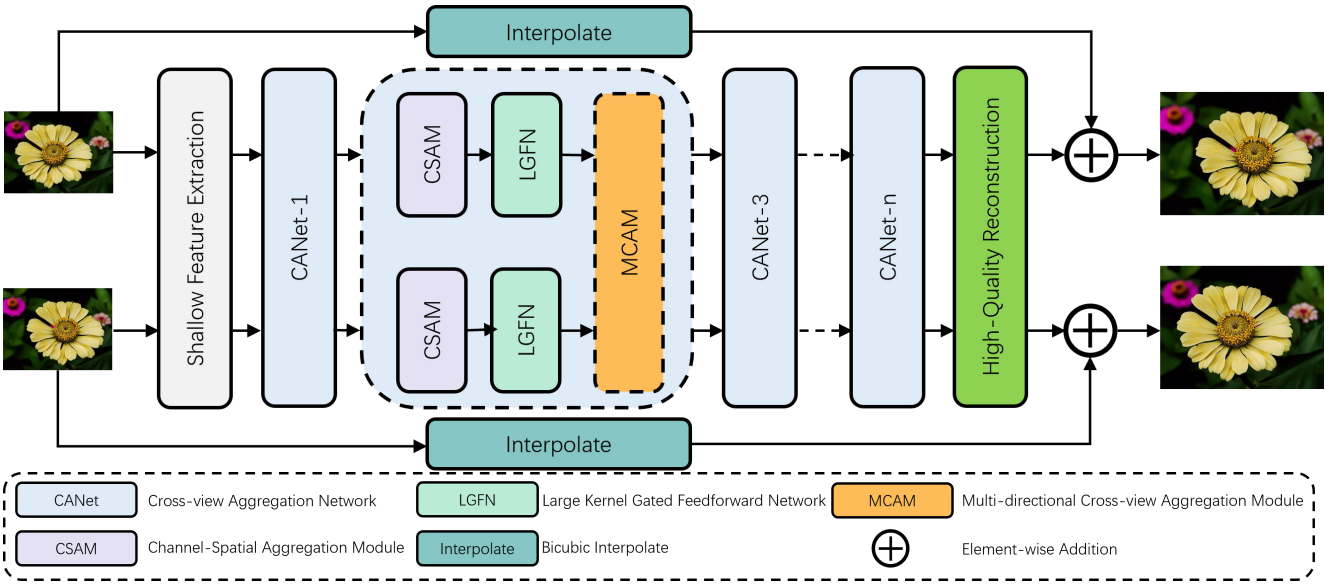

Figure 2. The architecture of our CANSSR for stereo image super-resolution.

parallax and lack a vertically oriented receptive field. In this paper, we design attention with multiple epipolar directions to aggregate more cross-view spatial information and achieve a larger receptive field.

3. Method

In this section, we begin by introducing the architecture of cross-view aggregation network (CANSSR) in section 3.1. Next, we detail the core composition of CANSSR in Section 3.2, Section 3.3, and Section 3.4, respectively.

3.1. Overview

To reconstruct high-quality stereo images, this paper proposes a novel cross-view aggregation network for stereo image super-resolution. As shown in Figure 2, we propose the CANSSR, a symmetric structure comprising three parts: (1) Multi-directional Cross-view Aggregation Module (MCAM), (2) Channel-Spatial Aggregation Module (CSAM), and (3) Large Kernel Gated Feed-forward Network (LGFN). Both CSAM and LGFN are weight-sharing networks, utilizing identical parameters to extract high-frequency information from the left and right views. In addition, MCAM is utilized to fuse cross-view features along horizontal and vertical epipolar lines.

Given a pair of images $I^{L,R} = (I^L, I^R)$, $I^L, I^R \in \mathbb{R}^{H\times W\times 3}$ represent the degraded left and right view LR images. Firstly, we employ a 3×3 convolutional layer $H_{shallow}(\cdot)$ to extract shallow feature from a LR stereo image pair $I^{L,R}$. It is described as:

$$F^{L,R}_{shallow} = H_{shallow}(I^{L,R}), \tag{1}$$

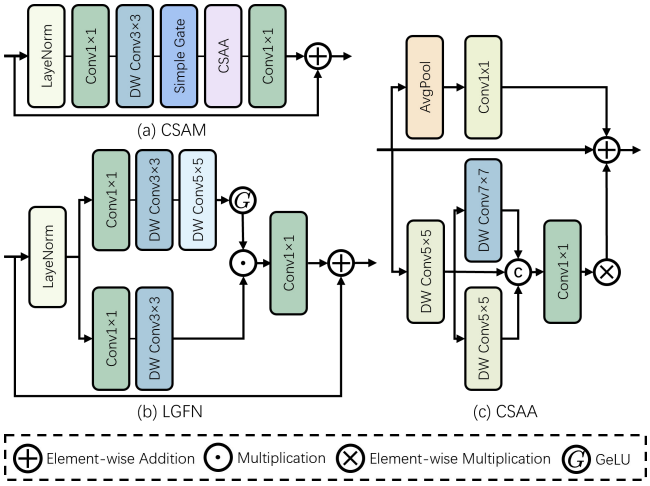

Figure 3. (a) Channel-Spatial Aggregation Module (b) Large Kernel Gated Feedforward Network (c) Channel-Spatial Aggregation Attention

where $F^{L,R}_{shallow}$ represents the left and right view shallow features. After stacking N CSAM, LGFN, and MCAM blocks, we obtain deep high-frequency features. It is described as:

$$F^L_i, F^R_i = H_{MCAM}(H_{LGFN}(H_{CSAM}(F^L_{i-1}, F^R_{i-1}))), \tag{2}$$

where F^L_i, F^R_i denote the left and right view features of the i-th layer, respectively.H_{MCAM}, H_{LGFN}, and H_{CSAM} denote MCAM, LGFN, and CSAM block, respectively.

Finally, the left and right view images are upsampled by the Pixel Shuffle module respectively to obtain the SR im-

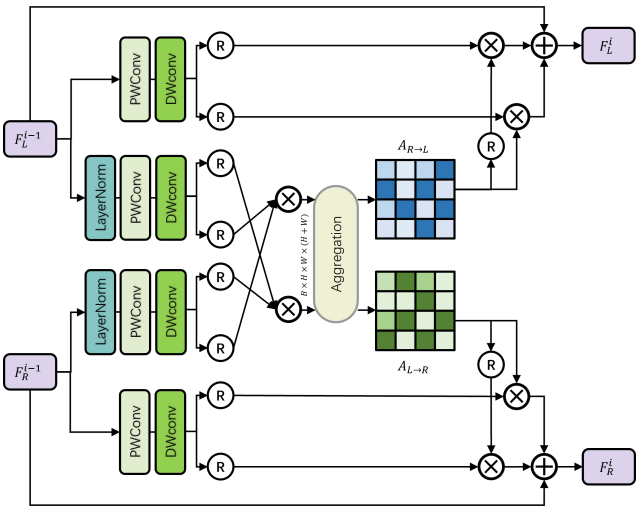

Figure 4. Multi-directional cross-view aggregation module

ages. The following formula can describe this:

$$F_{SR}^{L}, F_{SR}^{R} = H_{upsample}(F_i^L, F_i^R), \tag{3}$$

where F_{SR}^{L}, F_{SR}^{R} represent the upsampled left and right view SR images, respectively.

3.2. Multi-directional Cross-view Aggregation Module

In practice, the binocular system sometimes cannot maintain the horizontal direction and experiences frequent occlusions within real scenes, leading to the occurrence of asymmetric parallax. The assumption of horizontal priors often fails to capture accurate stereo correspondence, resulting in reconstructed images suffering from artifacts and blurring effects. To address this issue, we propose MCAM, attention along the horizontal and vertical epipolar directions to capture more realistic stereo correspondence. We improve upon CVIM [36] by designing multi-directional parallax attention in MCAM.

Specifically, this input feature F_{i-1}^{L}, F_{i-1}^{R} utilizes layer normalization as the initial processing step. Subsequently, the query, key, and value matrices Q, K, and V are generated using a combination of a 1×1 point-wise convolutional layer and a 3×3 depth-wise convolutional layer. Simultaneously, we employ the same convolutional layer denoted as $Q, K \in \mathbb{R}^{H\times W\times C}$, where C represents the number of channels. The computation is expressed as:

$$Q^{L,R} = W_{dw}W_{pw}(LN(F_i^{L,R})), \tag{4}$$

$$K^{L,R} = W_{dw}W_{pw}(LN(F_i^{L,R})), \tag{5}$$

$$V^{L,R} = W_{dw}W_{pw}(F_i^{L,R}), \tag{6}$$

where W_{dw}, W_{pw}, LN represent the 3×3 depth-wise convolution,the 1×1 convolutional layer,and the layer normalization, respectively. After obtaining features for q and k, we rotate them to derive the attention feature map $A \in \mathbb{R}^{H\times W\times(H+W)}$ via the **Aggregation** operation.

At each location p in the intra-view spatial dimensional feature map Q, we can obtain a vector $Q_p \in \mathbb{R}^{C/t}$. Simultaneously, we can extract a cross-view set $\Theta_{i,p} \in \mathbb{R}^{C/t}$ as the i-th element of Θ_p by acquiring another view feature vector from K with rows and columns at the same position as p. The equation can define this aggregation operation:

$$s_{i,p} = Q_p\Theta_{i,p}^{\mathsf{T}}, \tag{7}$$

where $s_{i,p} \in S$ represent the score of correlation degree between feature Q_p and $\Theta_{i,p}$, $i = [1, ..., |\Theta_p|]$, $S \in \mathbb{R}^{H\times W\times(H+W)}$. Subsequently, we apply the softmax function to compute the attention feature map A across the entire channel dimension of S.

For another feature $V \in \mathbb{R}^{H\times W\times C}$ generated by convolution, at each position p in the spatial dimensional feature map of V, we can obtain a vector $V_p \in \mathbb{R}^{C}$ and a set $\Psi \in \mathbb{R}^{(H+W)\times C}$. This set Ψ is a collection of feature vectors in V aligned with the same row or column as p. The feature aggregation operation embeds cross-view information:

$$Attention = \sum_{i\in|\Psi_p|} A_{i,p}\Psi_{i,p}, \tag{8}$$

where $Attention$ represents the mapping between cross-views (e.g., left view to right view embedding), $A_{i,p}$ is a score at channel i and position p in A, and F_L^n denotes the feature vector of the n-th layer in the left view. Finally, the cross-view mapping can be expressed as:

$$F_{L\to R}^{i} = Attention(Q_R^{i-1}, K_L^{i-1}, V_L^{i-1}) + F_R^{i-1}, \tag{9}$$

$$F_{R\to L}^{i} = Attention(Q_L^{i-1}, K_R^{i-1}, V_R^{i-1}) + F_L^{i-1}, \tag{10}$$

where $F_{L\to R}^{i}, F_{R\to L}^{i}$ represent the cross-view mapping. i represents the i-th feature map. F_L^{i-1}, F_R^{i-1} represent the feature map of the left and right views, respectively.

3.3. Channel-Spatial Aggregation Module

Although many existing stereo image SR methods prioritize the global reconstruction of cross-view information, they often overlook the significance of intra-view details for image reconstruction. The intra-view contains abundant texture detail information, and is crucial for reconstructing local texture details.

To address this problem, we propose a CSAM that efficiently exploits global-local modeling by channel-spatial

aggregation attention. As shown in Figure 3(c), the channel-spatial aggregation attention consists of multi-order spatial attention and simple channel attention [6]. We employ multi-order spatial attention to efficiently aggregate spatial dependencies and capture different scale local information. Simultaneously, we introduce simple channel attention [6] to capture global information.

Specifically, given an intra-view feature $F_i \in \mathbb{R}^{H\times W\times C}$, it is formulated:

$$F_{CSAM} = W^1_{pw}\mathcal{G}(SG(W^0_{dw}W^0_{pw}(LN(F_i)))) + F_i, \quad (11)$$

where F_{CSAM} represents the features extracted by the Channel-Spatial Aggregation Module. W^0_{pw}, W^0_{dw} and W^1_{pw} represent a 1×1 point-wise convolution, 3×3 depth-wise convolution, and 1×1 point-wise convolution layer, respectively. $SG(\cdot)$ and $\mathcal{G}(\cdot)$ present the SimpleGate [2] function and channel-spatial aggregation attention, respectively.

To aggregate global-local information, we design a CSAA that explores the global-local modeling to enhance texture representation. As shown in Figure 3(c), it is described:

$$\mathcal{G}(X) = CA(X) + MA(X) + X, \quad (12)$$

$$MA(X) = X \odot W^0_{dw}\mathcal{C}((\alpha W^4_{dw}, \beta W^3_{dw}, \gamma)(W^2_{dw}X)), \quad (13)$$

where X represents the input feature map. $CA(\cdot), MA(\cdot)$ represent the simplified channel attention [2] and the multi-order spatial aggregation attention, respectively. α, β, and γ denote the hyper-parameter to describe multi-order feature weight, and $\alpha+\beta+\gamma = 1$. W^2_{dw}, W^3_{dw}, and W^4_{dw} denote 5×5, 7×7, and 5×5 depth-wise convolution, respectively.

3.4. Large Kernel Gated Feedforward Network

We introduce the details of LGFN shown in Figure 3(b), utilizing the gate mechanism and GeLU function to activate the two parallel linear layers. Meanwhile, we equip a large kernel convolution for LGFN in one path to construct a multi-order receptive field to enhance parallel path feature representation.

Specifically, we employ depth-wise convolution with varying kernel sizes to weight the feature maps for more effectively capturing intra-view spatial information. Given an input intra-view feature $X \in \mathbb{R}^{\mathbb{C}\times\mathbb{H}\times\mathbb{W}}$, the key process of LGFN can be represented as:

$$\hat{X} = \varphi(W^2_{dw}W^1_{dw}W^0_{pw}(LN(X))) \odot W^1_{dw}W^0_{pw}(LN(X)), \quad (14)$$

where $\hat{X}$ is the feature maps extracted by the LGFN module. φ is denoted as the GeLU non-linear function. $\odot$ denotes element-wise multiplication.

4. Experiments

Model	Channels	Blocks	Params	Q_C	K_C	V_C
CANSSR-T	48	16	0.55M	24	24	48
CANSSR-S	64	32	0.92M	32	32	64
CANSSR-B	96	64	7.47M	48	48	96

Table 1. Parameter setting for different scale models, Q_C, K_C,and V_C represent the MCAM query, key and value channel dimension

4.1. Dateset

To evaluate the efficiency and effectiveness of our propose model, we adopt the training set by merging 60 images from Middlebury [24] and 800 images from Flickr1024 [30] following the experimental setting of iPASSR [31]. In addition, we select 5 images from Middlebury [24], 20 images from KITTI2012 [11], 20 images from KITTI2015 [22], and all test images from Flickr1024 [30] to build the test set, following [6, 14, 26, 31]. To NTIRE 2024 [29] Stereo Image Super-Resolution Challenge, we only employ 800 images from Flickr1024 train set. The LR images are generated by bicubic downsampling. We augment the training set and employ random horizontal, rotation, flips, and RGB channel shuffle, following [6].

4.2. Implementation details

To balance efficiency and effectiveness, we propose different model configurations with varying network depths and channel numbers. The specific architecture details of our model setup are provided in Table 1. To cater to various application scenarios, we propose three configurations: CANSSR-T (tiny), CANSSR-S (small) and CANSSR-B (base) models, respectively. The CANSSR-S model has been submitted to the NTIRE 2024 [29] Stereo Image Super-Resolution Challenge.

Training Settings. All the models were optimized using AdamW [20] with specific parameters β_1 = 0.9 and β_2 = 0.999 with a decay of 0. The learning rate starts at 1×10^{-3} and decays to 1×10^{-7} using the cosine annealing strategy, with a batch size of 32. We trained this model for 200,000 iterations and trained it with two NVIDIA RTX 4090 GPU. To solve the overfitting problem, we employ stochastic depth [13] with probabilities of 0.2 and CANSSR-B, respectively. Our network is only trained with the MSE loss function.

To evaluate, we adopt commonly-used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as quantitative metrics for evaluation, which are calculated in the RGB color space between a pair of stereo images (i.e. $(Left + Right)/2$). Meanwhile, the same as the previous method [6], TLSC [5] is used in the inference process.

Method	Scale	*#Para*	Left			(Left + Right)/2			
			KITTI2012 PSNR/SSIM	KITTI2015 PSNR/SSIM	Middlebury PSNR/SSIM	KITTI2012 PSNR/SSIM	KITTI2015 PSNR/SSIM	Middlebury PSNR/SSIM	Flickr1024 PSNR/SSIM
Bicubic	×2	-	28.44/0.8808	27.81/0.8814	30.46/0.8979	28.51/0.8842	28.61/0.8973	30.60/0.8990	24.94/0.8186
VDSR [15]	×2	0.66M	30.17/0.9062	28.99/0.9038	32.66/0.9101	30.30/0.9089	29.78/0.915	32.77/0.9102	25.60/0.8534
EDSR [18]	×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 [34]	×2	22.0M	30.81/0.9197	29.91/0.9224	34.85/0.9488	30.94/0.9227	30.7/0.933	34.94/0.9491	28.64/0.9084
RCAN [33]	×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
SwinIR [17]	×2	11.75M	31.18/0.9233	30.24/0.9262	35.27/0.9509	31.31/0.9262	31.05/0.9363	35.36/0.9513	29.25/0.9157
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 [26]	×2	1.37M	30.68/0.9159	29.81/0.9191	34.13/0.9421	30.81/0.919	30.6/0.93	34.23/0.9422	28.38/0.9038
IMSSRnet [16]	×2	6.84M	30.90/-	29.97/-	34.66/-	30.92/-	30.66/-	34.67/-	-/-
iPASSR [31]	×2	1.37M	30.97/0.9210	30.01/0.9234	34.41/0.9454	31.11/0.924	30.81/0.934	34.51/0.9454	28.6/0.9097
SSRDE-FNet [7]	×2	2.10M	31.08/0.9224	30.10/0.9245	35.02/0.9508	31.23/0.9254	30.9/0.9352	35.09/0.9511	28.85/0.9132
NAFSSR-T [6]	×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 [6]	×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
Steformer [19]	×2	1.29M	31.16/0.9236	30.27/0.9271	35.15/0.9512	31.29/0.9263	31.07/0.9371	35.23/0.9511	28.97/0.9141
CVHSSR-T [36]	×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
CANSSR-T (Ours)	×2	0.52M	31.19/0.9230	30.17/0.9241	35.12/0.9505	31.30/0.9260	30.99/0.9345	35.20/0.9506	29.11/0.9148
CANSSR-S (Ours)	×2	0.90M	31.33/0.9262	30.34/0.9280	35.42/0.9540	31.46/0.9284	31.14/0.9379	35.49/0.9533	29.36/0.9181
CANSSR-B (Ours)	×4	7.40M	31.46/0.9262	30.47/0.9287	35.86/0.9558	31.61/0.9292	31.26/0.9385	35.91/0.9558	29.72/0.9225
Bicubic	×4	-	24.52/0.7310	23.79/0.7072	26.27/0.7553	24.58/0.7372	24.38/0.7340	26.40/0.7572	21.82/0.6293
VDSR [15]	×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 [18]	×4	38.9M	26.26/0.7954	25.38/0.7811	29.15/0.8383	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RDN [34]	×4	22.0M	26.23/0.7952	25.37/0.7813	29.15/0.8387	26.32/0.8014	26.04/0.8043	29.27/0.8404	23.47/0.7295
RCAN [33]	×4	15.4M	26.36/0.7968	25.53/0.7836	29.20/0.8381	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
SRRes+SAM [32]	×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
SwinIR [17]	×4	11.9M	26.61/0.8039	25.76/0.7912	29.51/0.8460	26.71/0.8101	26.50/0.8143	29.63/0.8476	23.81/0.7441
StereoSR [14]	×4	1.42M	24.49/0.7502	23.67/0.7273	27.70/0.8036	24.53/0.7556	24.21/0.7511	27.64/0.8022	21.70/0.6460
PASSRnet [26]	×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
IMSSRnet [16]	×4	6.89M	26.44/-	25.59/-	29.02/-	26.43/-	26.2/-	29.02/-	-/-
iPASSR [31]	×4	1.37M	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 [7]	×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
NAFSSR-T [6]	×4	0.45M	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 [6]	×4	1.54M	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
Steformer [19]	×4	1.34M	26.61/0.8037	25.74/0.7906	29.29/0.8424	26.70/0.8098	26.45/0.8134	29.38/0.8425	23.58/0.7376
CVHSSR-T [36]	×4	0.68M	26.88/0.8105	26.03/0.7991	29.62/0.8496	26.98/0.8165	26.78/0.8218	29.74/0.8505	23.89/0.7484
CANSSR-T (Ours)	×4	0.55M	26.80/0.8069	25.95/0.7951	29.42/0.8439	26.89/0.8130	26.68/0.8179	29.53/0.8491	23.81/0.7437
CANSSR-S (Ours)	×4	0.92M	26.91/0.8106	26.06/0.7993	29.71/0.8500	27.01/0.8166	26.85/0.8220	29.80/0.8506	23.99/0.7519
CANSSR-B (Ours)	×4	7.47M	27.01/0.8142	26.16/0.8036	30.03/0.8590	27.10/0.8201	26.91/0.8260	30.14/0.8598	24.17/0.7598

Table 2. Quantitative Comparison with PSNR/SSIM Metric on $Flickr1024$, $KITTI2012$, $KITTI2015$ and $Middlebury$. Higher PSNR/SSIM Values Means Better Performance. The best and second best results are red and blue.

4.3. Result

In this section, we compare the proposed CANSSR(with three different variations) with the existing SR methods. We adopt SISR methods such as VDSR [15], EDSR [18], RDN [34], RCAN [33], SwinIR [17] and stereo image SR methods, for example, StereoSR [14], PASSRNet [26], IMSSRNet [16], iPASSR [31], SSRDE-FNet [7] NAFSSR [6], Steformer [19], and CVHSSR-T [36] are compared with our proposed method. All methods are trained on the same dataset, and evaluates their PSNR and SSIM scores [6].

Quantitative Result. As quantitative results are represented in Table 2, CANSSR outperforms the state-of-the-art methods in terms of PSNR and SSIM scores in the test set for ×2 and ×4 stereo image SR tasks. Specifically, our propose CANSSR-B model only uses 31% of the parameters of NAFSSR-L [6], which exceeds the state-of-the-art method in four standard evaluation datasets in the ×2 stereo image SR task. We propose the CANSSR-B with higher performance than the state-of-the-art method CVHSSR, exceeding 0.02 dB, 0.07 dB, 0.06 dB, and 0.1 dB on KITTI2012 [11], KITTI2015 [22], Middlebury [24], and Flickr1024 [30], respectively. Compared with some lightweight stereo image SR methods, the performance of our propose CANSSR-S method is far better than similar methods, such as NAFSSR-T and Steformer. In the Flickr1024 test set of ×4 tasks, our method exceeds 0.30 dB, 0.41 dB and 0.10 dB, for NAFSSR-T [6], Steformer [19] and CVHSSR-T [36], respectively. This clearly demonstrates the effectiveness and efficiency of our proposed CANSSR network architecture.

Qualitative Results. As shown in Figure 5, 6, which

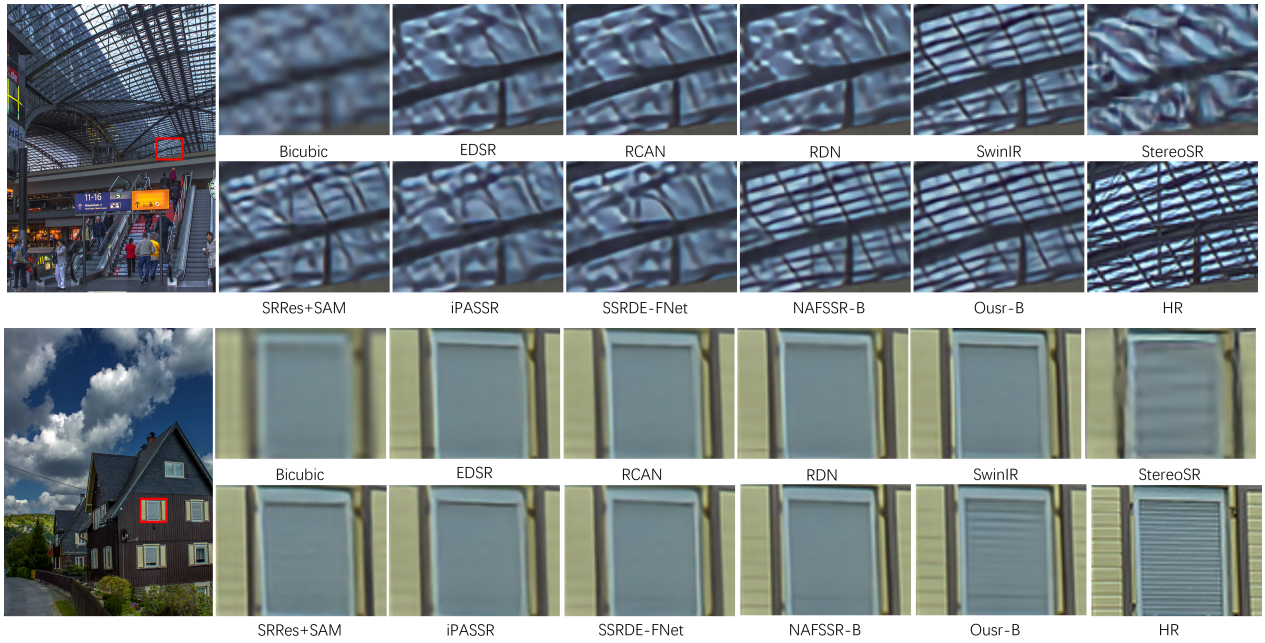

Figure 5. Visual comparisons for ×4 SR by different methods on the Flickr1024 [30] and Middlebury [24] datasets. The red rectangle marks zoom-in region.

Method	EPE ↓	>1px(%) ↓	>2px(%) ↓	>3px(%) ↓	PSNR ↑
RDN	0.8793	15.55	6.23	3.70	26.74
RCAN	0.8737	15.26	6.14	3.64	26.85
SwinIR	0.8646	15.20	6.13	3.60	27.09
iPASSR	0.8546	14.97	6.09	3.57	26.92
SSRDE-FNet	0.8289	14.30	5.84	3.29	27.06
CANSSR	0.7860	14.03	5.56	3.19	27.47
HR	0.6663	11.67	4.61	2.65	∞

Table 3. Quantitative comparison results achieved by GwcNet [12] on ×4 stereo images SR. All these metrics were averaged on the validation Set of the KITTI2012 [11] Dataset.

shows the visual result for $\times 4$ stereo image SR on KITTI2015, Flickr1024, and Middlebury datasets. These images show that our CANSSR method alleviates the blur and artifact problems for the reconstructed images, and the reconstructed images are rich in more texture details and sharper edges. In contrast, other methods may suffer from blurring and artifacts. This further demonstrates the effectiveness of our proposed CANSSR method.

Benefits to disparity estimation. We utilize stereo SR images generated by neural networks to verify the effectiveness of our CANSSR for disparity estimation. Firstly, we employ $\times 4$ downsampling in KITTI2012 [11] validation images, which are partitioned by GwcNet [12]. Then, we tested the KITTI2012 [11] validation set using the state-of-the-art SISR and stereo image SR methods, respectively. End-point error (EPE) and t-pixel error rate ($>$ tpx) were used as quantitative metrics to evaluate the estimated disparity. As shown in Table 3, compared with SSRDE-FNet and iPASSR, our proposed CANSSR increases by 0.0429 and 0.0686, respectively. It proves the effectiveness of our proposed method for improving the disparity estimation results.

4.4. Ablation experiments

To evaluate the effectiveness of our method, we initially remove all the modules proposed for testing and then performed different combination tests on the three modules in the Flickr1024 dataset [30]. As demonstrated by the results in Table 4.

Multi-directional Cross-view Aggregation Module. The model performance greatly correlates with the number of MCAMs. When we set the number of MCAMs to 32 in the small model, compared with the original baseline, the performance of MCAM is improved by 0.39 dB, which effectively proves the performance of MCAM.

Channel-Spatial Aggregation Module. To evaluate the effectiveness of the CSAM module, we found that the improvement was 0.12 dB compared with the baseline. CSAM can effectively capture long-range dependence and obtain the channel-space global information of intra-view to reconstruct high-quality images.

Large kernel Gated Feed-forward Network. We introduce large kernel convolution into the feed-forward network (FFN) to aggregate more effective spatial information. Compared with FFN baseline, our method improves 0.08 dB on the Flickr1024 test set.

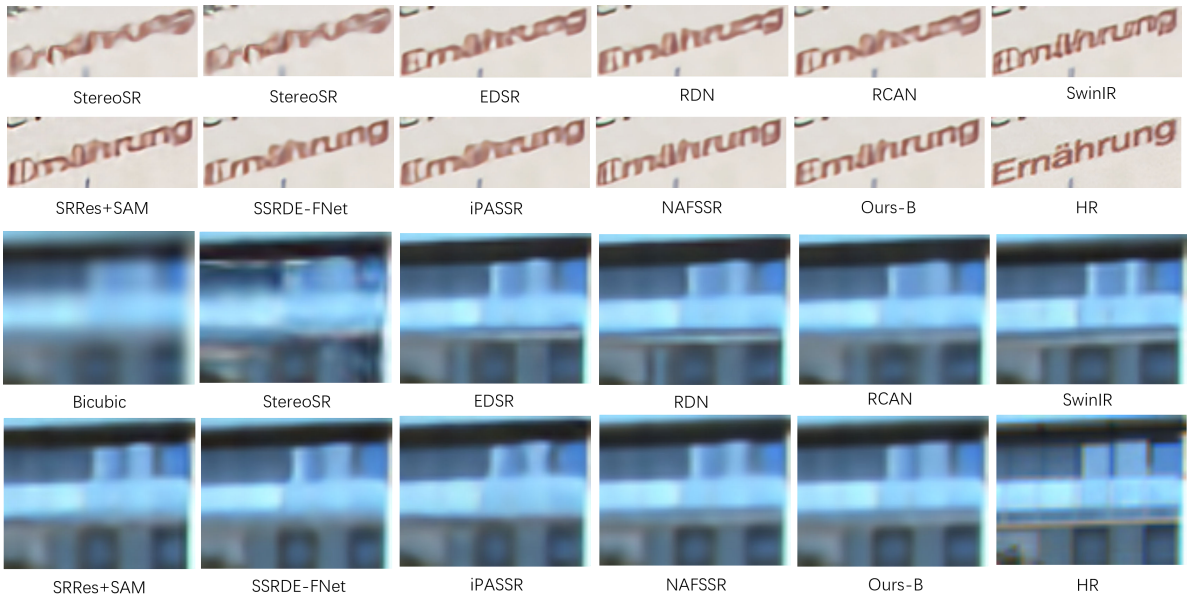

Figure 6. Visual comparisons for ×4 SR by different methods on the Flickr1024 and KITTI2015 datasets.

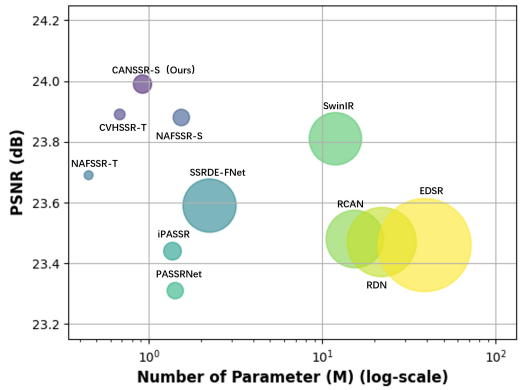

Figure 7. Compare the parameters, PSNR and MACs for 4× stereo image SR on Flickr1024 [30] test set

Method	MCAM	CSAM	LGFN	PSNR	ΔPSNR
Baseline without PAM	✗	✗	✗	23.56	-
	✗	✓	✗	23.68	0.12
	✗	✗	✓	23.64	0.08
	✗	✓	✓	23.73	0.17
Baseline + PAM	✗	✗	✗	23.88	-
	✓	✗	✗	23.95	0.07
	✓	✓	✓	23.99	0.11

Table 4. Ablation Studies by CANSSR-S in ×4 stereo image super-resolution task on Flickr1024 test set.

Runtime Efficiency. We conduct tests on 320×180 images to evaluate the relationship between the computational efficiency and performance of the model. As shown in Figure 7 our method reduces the number of parameters and improves the accuracy compared with the previous NAFSSR [6] method.

4.5. NTIRE Stereo Image SR Challenge

We submitted the results obtained from our proposed approach to the NTIRE 2024 [29] Stereo Image Super-Resolution Challenge. To maximize the performance of our method, we stacked the 16-layer CANet twice and utilized weight sharing to construct a model with a depth of 32, while setting the model width to 64. During the testing phase, we adopted the TLSN [5] strategy. The number of parameters in our model is 0.9221M and MACs is 178.29. As a result, our last submission achieved a PSNR of 23.5725 dB on the test set. We won 6-th in track 1 for Fidelity&Bicubic.

5. Conclusion

In this paper, we propose an efficient stereo image super-resolution model, named CANSSR. In particular, we design a multi-directional cross-view aggregation module to effectively capture multi-directional stereo correspondence and mine cross-view similarity features. Furthermore, we propose channel-spatial aggregation module to enhance global-local information extraction, and large kernel gated feed-forward network to enhance capture spatial dependencies and fine high-frequency information. Extensive experiments demonstrate that CANSSR outperforms current models and achieves state-of-the-art performance.

References

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

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 17–33. Springer, 2022. 5

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 2

[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 Workshop*, pages 1702–1711, 2023. 1, 2

[5] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Revisiting global statistics aggregation for improving image restoration. *arXiv preprint arXiv:2112.04491*, 2(4):5, 2021. 5, 8

[6] 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, 5, 6, 8

[7] 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. 1, 2, 6

[8] 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

[9] Jiawang Dan, Zhaowei Qu, Xiaoru Wang, and Jiahang Gu. A disparity feature alignment module for stereo image super-resolution. *IEEE Signal Processing Letters*, 28:1285–1289, 2021. 2

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 184–199. Springer, 2014. 2

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2, 5, 6, 7

[12] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 7

[13] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Proceedings of the European conference on computer vision (ECCV)*, pages 646–661. Springer, 2016. 5

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1721–1730, 2018. 1, 2, 5, 6

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1646–1654, 2016. 2, 6

[16] 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

[17] 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. 2, 6

[18] 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/CVF Conference on Computer Vision and Pattern Recognition workshops*, pages 136–144, 2017. 2, 6

[19] Jianxin Lin, Lianying Yin, and Yijun Wang. Steformer: Efficient stereo image super-resolution with transformer. *IEEE Transactions on Multimedia*, pages 1–13, 2023. 2, 6

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[21] 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. 2

[22] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 2, 5, 6

[23] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 191–207. Springer, 2020. 2

[24] 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. 2, 5, 6, 7

[25] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In

Proceedings of the IEEE International Conference on Computer Vision, pages 4539–4547, 2017. 2

[26] 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, 5, 6

[27] 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. 2

[28] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Ming Cheng, Haoyu Ma, Qiufang Ma, Xiaopeng Sun, et al. Ntire 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1372, 2023. 2

[29] 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. 5, 8

[30] 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, 2, 5, 6, 7, 8

[31] 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

[32] 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

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

[34] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2472–2481, 2018. 6

[35] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2

[36] Wenbin Zou, Hongxia Gao, Liang Chen, Yunchen Zhang, Mingchao Jiang, Zhongxin Yu, and Ming Tan. Cross-view hierarchy network for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1396–1405, 2023. 2, 4, 6