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.

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...
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 2023 [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
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 \( 3 \times 3 \) convolutional layer \( H_{\text{shallow}}(\cdot) \) to extract shallow feature from a LR stereo image pair \( I^{L,R} \). It is described as:

\[
F_{\text{shallow}}^{L,R} = H_{\text{shallow}}(I^{L,R}),
\]

where \( F_{\text{shallow}}^{L,R} \) 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:

\[
F_i^{L,R} = H_{\text{MCAM}}(H_{\text{LGFN}}(H_{\text{CSAM}}(F_{i-1}^{L,R}, F_{i-1}^{R}))),
\]

where \( F_i^{L,R} \) denote the left and right view features of the \( i \)-th layer, respectively, \( H_{\text{MCAM}}, H_{\text{LGFN}}, \) and \( H_{\text{CSAM}} \) 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-
neously, we employ the same convolutional layer denoted by the query, key, and value matrices $Q$, $K$, and $V$ are generated using a combination of a $1 \times 1$ point-wise convolutional layer and a $3 \times 3$ 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:

$$Q^{L,R} = W_{dw}W_{pw}(LN(F^{i}_{i}^{L,R})), \quad (4)$$

$$K^{L,R} = W_{dw}W_{pw}(LN(F^{i}_{i}^{L,R})), \quad (5)$$

$$V^{L,R} = W_{dw}W_{pw}(F^{i}_{i}^{L,R}), \quad (6)$$

where $W_{dw}$, $W_{pw}$, $LN$ represent the $3 \times 3$ depth-wise convolution, the $1 \times 1$ 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 $\Theta_{p}$ 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:

$$s_{i,p} = Q_{p} \Theta_{i,p}^{T}, \quad (7)$$

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 $\Psi \in \mathbb{R}^{H \times W \times C}$. This set $\Psi$ 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:

$$Attention = \sum_{i \in |\Psi|} A_{i,p} \Psi_{i,p}, \quad (8)$$

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 $F^{i}_{i}$ denotes the feature vector of the $i$-th layer in the left view. Finally, the cross-view mapping can be expressed as:

$$F^{i}_{L \rightarrow R} = Attention(Q^{i-1}_{R}, K^{i-1}_{L}, V^{i-1}_{L}) + F^{i-1}_{R}, \quad (9)$$

$$F^{i}_{R \rightarrow L} = Attention(Q^{i-1}_{L}, K^{i-1}_{R}, V^{i-1}_{R}) + F^{i-1}_{L}, \quad (10)$$

where $F^{i}_{L \rightarrow R}, F^{i}_{R \rightarrow L}$ represent the cross-view mapping. $i$ represents the $i$-th feature map. $F^{i-1}_{L}, F^{i-1}_{R}$ 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 \( F_i \), it is described:

\[
F_{\text{CSAM}} = W_{\text{pw}}^1G(SG(W_{\text{dw}}^0W_{\text{pw}}^0(LN(F_i)))) + F_i, \quad (11)
\]

where \( F_{\text{CSAM}} \) represents the features extracted by the Channel-Spatial Aggregation Module. \( W_{\text{pw}}^0, W_{\text{dw}}^0 \) and \( W_{\text{pw}}^1 \) represent a \( 1 \times 1 \) point-wise convolution, \( 3 \times 3 \) depth-wise convolution, and \( 1 \times 1 \) point-wise convolution layer, respectively. \( SG(\cdot) \) and \( G(\cdot) \) 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:

\[
G(X) = CA(X) + MA(X) + X; \quad (12)
\]

\[
MA(X) = X \odot W_{\text{dw}}^0C((\alpha W_{\text{dw}}^2, \beta W_{\text{dw}}^3, \gamma)(W_{\text{dw}}^0X)), \quad (13)
\]

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. \( \alpha, \beta, \gamma \) denote the hyper-parameter to describe multi-order feature weight, and \( \alpha + \beta + \gamma = 1 \). \( W_{\text{dw}}^0, W_{\text{dw}}^2 \) and \( W_{\text{dw}}^3 \) denote \( 5 \times 5, 7 \times 7, \) and \( 5 \times 5 \) 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 \( X \in \mathbb{R}^{C \times H \times W} \), the key process of LGFN can be represented as:

\[
\hat{X} = \varphi(W_{\text{dw}}^2W_{\text{dw}}^1W_{\text{pw}}^0(LN(X))) \odot W_{\text{dw}}^1W_{\text{pw}}^0(LN(X)), \quad (14)
\]

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

4. Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Channels</th>
<th>Blocks</th>
<th>Params</th>
<th>( Q_C )</th>
<th>( K_C )</th>
<th>( V_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>CANSSR-T</td>
<td>48</td>
<td>16</td>
<td>0.55M</td>
<td>24</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>CANSSR-S</td>
<td>64</td>
<td>32</td>
<td>0.92M</td>
<td>32</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>CANSSR-B</td>
<td>96</td>
<td>64</td>
<td>7.47M</td>
<td>48</td>
<td>48</td>
<td>96</td>
</tr>
</tbody>
</table>

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. Datasets

To evaluate the efficiency and effectiveness of our proposed 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 \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \) with a decay of 0. The learning rate starts at \( 1 \times 10^{-3} \) and decays to \( 1 \times 10^{-7} \) 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. \((\text{Left} + \text{Right})/2\)). Meanwhile, the same as the previous method [6], TLSC [5] is used in the inference process.
Table 2. Quantitative Comparison with PSNR/SSIM Metric on Flickr1024, KITTI’2012, KITTI’2015 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], iPASS [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.

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

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 ×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 ×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

<table>
<thead>
<tr>
<th>Method</th>
<th>MCAM</th>
<th>CSAM</th>
<th>LGFN</th>
<th>PSNR</th>
<th>ΔPSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline without PAM</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>23.56</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>23.68</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>23.64</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>23.73</td>
<td>0.17</td>
</tr>
<tr>
<td>Baseline + PAM</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>23.38</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>23.95</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>23.99</td>
<td>0.11</td>
</tr>
</tbody>
</table>

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


[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


