Cross-View Hierarchy Network for Stereo Image Super-Resolution

Wenbin Zou¹, Hongxia Gao¹; Liang Chen², Yunchen Zhang², Mingchao Jiang³, Zhongxin Yu², Ming Tan²
South China University of Technology,¹ Fujian Normal University,² GAC R&D Center,³
alexzou14@foxmail.com, hxgao@scut.edu.cn, cl_0827@126.com, jiangshaoyu1993@gmail.com, cydiachen@cydiachen.tech, wuyizhizi555@163.com, qsz20211396@student.fjnu.edu.cn

Abstract

Stereo image super-resolution aims to improve the quality of high-resolution stereo image pairs by exploiting complementary information across views. To attain superior performance, many methods have prioritized designing complex modules to fuse similar information across views, yet overlooking the importance of intra-view information for high-resolution reconstruction. It also leads to problems of wrong texture in recovered images. To address this issue, we explore the interdependencies between various hierarchies from intra-view and propose a novel method, named Cross-View Hierarchy Network for Stereo Image Super-Resolution (CVHSSR). Specifically, we design a cross-hierarchy information mining block (CHIMB) that leverages channel attention and large kernel convolution attention to extract both global and local features from the intra-view, enabling the efficient restoration of accurate texture details. Additionally, a cross-view interaction module (CVIM) is proposed to fuse similar features from different views by utilizing cross-view attention mechanisms, effectively adapting to the binocular scene. Extensive experiments demonstrate the effectiveness of our method. CVHSSR achieves the best stereo image super-resolution performance than other state-of-the-art methods while using fewer parameters. The source code and pre-trained models are available at https://github.com/AlexZou14/CVHSSR.

1. Introduction

Stereo image technology has made significant strides, leading to the successful application of stereo images in a variety of 3D scenarios, such as augmented reality (AR), virtual reality (VR), and autonomous driving. However, stereo imaging devices, such as the dual cameras on mobile phones, are subject to certain constraints that can result in the production of stereo image pairs with low resolution (LR). Stereo image super-resolution (SR), which has attracted much attention in recent years, aims to generate high-resolution stereo image pairs from their low-resolution counterparts to significantly enhance their visual perception. Therefore, this research has great potential to enhance the user experience in deploying immersive services.

In recent years, numerous deep learning-based algorithms for stereo image SR have been proposed, following the widespread use of convolution neural network (CNN) based methods in single image super-resolution (SISR) [6, 17, 18, 36, 39] tasks. Unlike SISR, which primarily focuses on finding similar textures within an image, stereo image SR must consider both intra-view and inter-view information, both of which play critical roles in stereo image reconstruction. Existing methods typically develop complex networks and loss functions to effectively fuse information from two viewpoints. For example, Jeon et al. [12] learned the parallax prior in stereo datasets through a two-stage network to recover high-resolution images. Wang et

Figure 1. Comparison on the trade-off between model parameters and PSNR for 2x stereo SR on Flickr1024 [31] test set. Our CVHSSR model family achieves state-of-the-art performance with significantly fewer parameters, indicating its superior efficiency and effectiveness for stereo super-resolution tasks.

*Corresponding author
al. [28] introduced a parallax attention mechanism that incorporates global receptive fields to further improve network performance. Song et al. [25] proposed a self and parallax attention mechanism to reconstruct high-quality stereo image pairs. Recently, Chu et al. [4] used an efficient non-linear activation-free block and cross-view attention module, achieving the best performance and winning first place in NTIRE2022 [27].

Although the existing stereo image SR methods have achieved impressive performance, they have not fully explored the rich hierarchy features of the intra-view, which could affect the information transfer between cross-views. Therefore, an interesting research question remains on how to effectively utilize both global and local features from stereo image pairs to further improve the quality of stereo image SR reconstruction.

In this paper, we propose a novel method to address the issue of unexplored intra-view hierarchical features in stereo image SR. The proposed method, named Cross-View Hierarchical Network for stereo image SR (CVHSSR), aims to extract rich feature representations from intra-views at different hierarchies and fuse them to enhance the performance of stereo image SR. To achieve this, we introduce two core modules: the cross-hierarchy information mining block (CHIMB) and the cross-view interaction module (CVIM), which explore and fuse similar features from different hierarchies across views. Specifically, the CHIMB module is designed to model and recover intra-view information at various hierarchies, utilizing the large kernel convolution attention and the channel attention mechanism. On the other hand, the CVIM module effectively integrates similar information from different views by utilizing the cross-view attention mechanism. By exploiting these modules, the CVHSSR can incorporate more diversified feature representations from different spatial levels of the two views, resulting in enhanced SR reconstruction quality.

The key contributions of this work are summarized as follows:

- We propose the CHIMB to efficiently extract hierarchy information from intra-views. In contrast to the NAF-Block used in NAFNet [1], CHIMB models global and local information from intra-view by using channel attention and large kernel convolution attention, effectively helping the network to restore the correct texture features.

- We have designed the CVIM to fuse similar information from different views in our method. Unlike the cross-view attention mechanism used in other methods, CVIM utilizes depth-wise convolution to capture similar information from intra-view, then facilitating cross-view information fusion.

- Based on CHIMB and CVIM, we propose a simple yet effective method for stereo image SR. Our approach achieves state-of-the-art performance with fewer parameters, as shown in Figure 1. Extensive experiments confirm the validity of our proposed CVHSSR.

2. Related Work

2.1. Single Image Super-Resolution

Image super-resolution is a regression problem that maps a low-resolution image to its corresponding high-resolution image. Since Dong et al. [8] introduce CNN into the SISR field with their pioneering work SRCNN, CNN-based methods have been proven to achieve impressive performance in SISR tasks. On this basis, Lim et al. [13] further improved network performance by deepening the network and increasing the dimension of intermediate features. With the development of deep learning technology, the researchers use residual and dense connections [14,18,36,40] to control network information flow, thereby obtaining better image reconstruction performance. However, these methods did not consider the importance of different features, leading to redundant network design. Therefore, RCAN [39] introduced channel attention mechanisms to model the interdependence between feature channels and adaptively rescale the features of each channel. Subsequently, various attention mechanisms were proposed to enhance network expression ability, including spatial attention [7,23], second-order attention [6], non-local attention [20,21], and large kernel convolution attention [30].

Recently, the Transformer has achieved great success in the field of vision. Therefore, many researchers [16,17,19,33,38] introduce the Transformer into SISR tasks. With the powerful learning ability of Transformers, Transformer-based methods have achieved state-of-the-art performance in the field of single-image super-resolution. Despite the consistently improved performance, SISR cannot utilize complementary information from different views in stereo image pairs, which limits the performance of stereo image super-resolution.

2.2. Stereo Image Super-Resolution

Instead of SISR method, which only has access to context information from intro, stereo image SR can leverage the additional information provided by the cross-view information to enhance SR performance. However, the presence of binocular disparity between the left and right views in a stereo image pair can pose a significant challenge to the fusion of information across views. Thus, Jeon et al. [12] proposed the first deep learning-based model for stereo image SR (namely, StereoSR). This approach addresses the challenge of fusing complementary features of the left and right views by concatenating the left image and a stack of right images with predefined shifts. On this basis, Wang et
al. [28, 29] introduced a parallax attention module (PAM) to model stereo correspondence by effectively capturing global contextual information along the epipolar line. These methods outperform StereoSR and exhibit greater flexibility in accommodating disparity variation. In pursuit of more refined stereo correspondence, Song et al. [25] extended the parallax-attention mechanism to propose SPAM, which aggregates information from both the primary and cross views to generate stereo-consistent image pairs. Yan et al. [35] proposed a domain adaptive stereo SR network (DASSR) to achieve both stereo image SR and stereo image deblurring tasks. Xu et al. [34] introduced a parallax attention module (PAM) leveraging the symmetry cues present in stereo image pairs. Ma et al. proposed a perception-oriented StereoSR framework, which aims to restore stereo images with improved subjective quality. More recently, Chu et al. [4] developed the NAFSSR network by utilizing nonlinear activation-free blocks [1] for intra-view feature extraction and PAM for cross-view feature interaction, which achieves the champion in the NTIRE 2022 Stereo Image SR Challenge [27].

Although existing methods have achieved superior performance in stereo image SR, they typically focus on modeling cross-view information while neglecting the hierarchical similarity relationships from intra-view. To address this limitation, we propose to leverage both local and global hierarchical feature representations to further improve the performance of state-of-the-art stereo image SR methods.

3. Cross-View Hierarchy Network

3.1. Overall Framework

To avoid complex network designs that require a large number of parameters and computational effort, we adopt a simple weight-sharing two-branch network structure to recover the left and right view images, as illustrated in Figure 2. Our CVHSSR mainly consists of four components: the shallow feature extraction, the cross-hierarchy information mining block (CHIMB), the cross-view interaction modules (CVIM), and stereo image reconstruction. Specifically, the CHIMB is designed to extract similar features both locally and globally from the intra-view image, effectively restoring accurate texture details. The CVIM is mainly used to fuse features from two viewpoints. More details of CHIMB and CVIM are described in Sections 3.2 and 3.3.

Firstly, given an input stereo low-resolution images $I_{LR}^L, I_{LR}^R \in R^{H \times W \times 3}$, CVHSSR first applies a convolution to obtain two-views shallow feature $F_{0}^L, F_{0}^R \in R^{H \times W \times C}$, where $H \times W$ denotes the spatial dimension and $C$ is the number of channels. It can be formulated as:

$$F_{0}^{L,R} = H_{\text{conv}}(I_{LR}^L, I_{LR}^R),$$  \hspace{1cm} (1)

where $H_{\text{conv}}$ denotes $3 \times 3$ convolution operation.

Next, we integrate CHIMB and CVIM into a cross-view hierarchy (CVH) block, which not only extracts deep intra-view features but also fuses information from different view images. Therefore, we stack $N$ CVH blocks to obtain output features that incorporate information from multiple perspectives. It can be expressed as:

$$F_{out}^{L,R} = H_{CVH}^{N}(H_{CVH}^{n-1}(\ldots H_{CVH}^{i}(F_{i}^{L,R})\ldots)),$$
$$F_{i+1}^{L,R} = H_{CVH}(F_{i}^{L,R}) = H_{CH}(F_{i}^{L,R}),$$  \hspace{1cm} (2)

where $H_{CVH}, H_{CH}$ denote CVH block, CVIM, and CHIMB, respectively. $F_{out}^{L,R}, F_{i}^{L,R}$ denote the output of the $N$-th CVH Block and $i$-th CVH Block, respectively.

Finally, we upsample the output features to the HR size using the pixel-shuffle operation. Additionally, we incorporate a global residual path to leverage the input stereo image.
information to further improve the super-resolution performance. It can be expressed as:

\[ I^{SR}_{L} = H_{up}(L^{SR}_R) + H_{up}(I^{SR}_L) = H_{CVHSSR}(I^{SR}_L), \]
\[ I^{SR}_{R} = H_{up}(R^{SR}_L) + H_{up}(I^{SR}_R) = H_{CVHSSR}(I^{SR}_R), \]

where \( H_{up} \) denotes the upsampling operation. \( H_{CVHSSR} \) denotes the proposed CVHSSR network. \( I^{SR}_{L}, I^{SR}_{R} \) denote the final restored left-view and right-view images, respectively.

### 3.2. Cross-Hierarchy Information Mining Block

Many existing methods for stereo image SR focus mainly on modeling cross-view information and do not adequately exploit the hierarchy information from the intra-view image. This leads to difficulties in recovering clear texture details. To address this issue, we have proposed the CHIMB as depicted in Figure 3, which is capable of effectively extracting different hierarchy information features from images.

The CHIMB consists of two parts: (1) The cross-hierarchy information extractor (CHIE); (2) The information refinement feedforward network (IRFFN). The CHIMB incorporates both channel attention and large kernel convolution attention to capture both global and local similarity relationships. The channel attention calculates global statistics of the feature map to enhance the focus on important features. Meanwhile, the large kernel convolution attention utilizes larger kernel convolution to capture long-range dependencies in the intra-view image, thereby enhancing the attention to local information. These two attention mechanisms in combination enable the CHIE to effectively model the hierarchy information contained in the input images and accurately recover texture details.

Given an input tensor \( F_{in} \in \mathbb{R}^{H \times W \times C} \), the CHIE is formulated as:

\[ F_{CHIE} = W^p_p(\mathcal{H}(\delta_{SG}(W^1_{d3}W^2_p(\text{LN}(F_{in})))))) + F_{in}, \]

where \( W^p_p \) is the \( 1 \times 1 \) point-wise convolution and \( W^s_{d3} \) is the \( 3 \times 3 \) depth-wise convolution. \( F_{CHIE} \) denotes the output feature of CHIE. The \( \text{LN}(\cdot) \) denotes layer normalization. We use the notation \( \delta_{SG}(\cdot) \) and \( \mathcal{H}(\cdot) \) to represent the SimpleGate function and the hybrid attention operation, respectively. Specifically, the SimpleGate first split the input into two features \( X_1, X_2 \in \mathbb{R}^{H \times W \times C/2} \) along channel dimension. Then, it computes the output with the linear gate as \( \delta_{SG}(X) = X_1 \odot X_2 \), where \( \odot \) denotes element-wise multiplication. The hybrid attention \( \mathcal{H}( \cdot ) \) consists of two components: channel attention and large kernel convolution attention. It can be described as follows:

\[ \mathcal{H}(X) = \mathcal{LKA}(X) + \mathcal{CA}(X), \]
\[ \mathcal{CA}(X) = X \odot (W_pH_{Avg}(X)), \]
\[ \mathcal{LKA}(X) = X \odot (W_pW_{dld}W_{d5}(X)), \]

where \( \mathcal{LKA}(\cdot), \mathcal{CA}(\cdot) \) and \( H_{Avg} \) denote the large kernel convolution attention, the channel attention, and the average pooling operation, respectively. \( \odot \) denotes element-wise multiplication. The \( W_{d5} \) and \( W_{dld} \) represent the \( 5 \times 5 \) depth-wise convolution and the \( 7 \times 7 \) depth-wise dilation convolution.

The IRFFN in our pipeline utilizes a non-linear gate mechanism to control the flow of information, allowing each channel to focus on fine details complementary to the other levels. The IRFFN process is formulated as:

\[ F_{out} = W^3_p(\delta_{NG}(W^2_{d3}W^2_p(\text{LN}(F_{CHIE})))) + F_{CHIE} \]

where \( \delta_{NG}(\cdot) \) is the function of non-linear gate mechanism. Similar to SimpleGate, the non-linear gate mechanism divides the input along the channel dimension into two features \( X_1, X_2 \in \mathbb{R}^{H \times W \times C/2} \). The output is then calculated by non-linear gating as \( \delta_{NG}(X) = GELU(X_1) \odot X_2 \), where \( GELU(\cdot) \) denotes the activation function. The \( F_{out} \) denote the output of CHIMB.

### 3.3. Cross-View Interaction Module

The details of the proposed CVIM as shown in Figure 4. This method employs Scaled DotProduct Attention [26], which involves calculating the dot product between the query and keys, followed by the application of a softmax function to generate weights assigned to the corresponding values:

\[ \text{Attention}(Q, K, V) = \text{Softmax}(QK^T/\sqrt{C})V, \]

where \( Q \in \mathbb{R}^{H \times W \times C} \) is query matrix projected by source intra-view feature (e.g. left-view), and \( K, V \in \mathbb{R}^{H \times W \times C} \) are key, value matrices projected by target intra-view feature (e.g. right-view). Here, \( H, W, C \) represents the height, width, and number of channels of the feature map.

The CVIM adopts a novel cross-view attention mechanism that integrates information from both left and right views through the cross-view information mining block.
view images to generate cross-view attention maps. This approach allows for the exploitation of the distinctive information present in each view, leading to more effective feature fusion and better restoration results. In detail, given the input stereo intra-view features $F^i_L, F^i_R \in \mathbb{R}^{H \times W \times C}$, we can get the cross-view fusion features $F_{L\rightarrow R}$ as follows:

\begin{align}
Q_L &= W^Q_d W^Q_p (LN(F^i_L)), \\
K_R &= W^K_d W^K_p (LN(F^i_R)), \\
V_R &= W^V_d W^V_p F^i_R, \\
F_{L\rightarrow R} &= W^P_p Attention_{L\rightarrow R}(Q_L, K_R, V_R),
\end{align}

where $W^Q(\cdot)$ is the $1 \times 1$ point-wise convolution, $W^P(\cdot)$ is the $3 \times 3$ depth-wise convolution. It offers feature refinement from both channel and spatial perspectives.

With an analogous argument, we can obtain the cross-view fusion features $F_{R\rightarrow L}$ as follows:

\begin{align}
Q_R &= W^Q_d W^Q_p (LN(F^i_R)), \\
K_L &= W^K_d W^K_p (LN(F^i_L)), \\
V_L &= W^V_d W^V_p F^i_L, \\
F_{R\rightarrow L} &= W^P_p Attention_{R\rightarrow L}(Q_R, K_L, V_L),
\end{align}

Finally, the interacted cross-view information $F_{L\rightarrow R}, F_{R\rightarrow L}$ and intra-view information $F^i_L, F^i_R$ are fused by element-wise addition:

\begin{align}
F^{L+1}_L &= \gamma_L F_{L\rightarrow R} + F^i_L, \\
F^{R+1}_R &= \gamma_R F_{R\rightarrow L} + F^i_R,
\end{align}

where $\gamma_L$ and $\gamma_R$ are trainable channel-wise scales and initialized with zeros for stabilizing training.

### 3.4. Loss Function

As image super-resolution tasks are primarily focused on restoring high-frequency details, we leverage both spatial and frequency domain losses to jointly guide our network in effectively recovering clear and sharp high-frequency textures. Specifically, given an input two view LR image $I^{L,R}_{LR}$, the proposed model predicts HR stereo image denoted $I^{L,R}_{SR}$.

We optimize our CVHSSR with the following loss function:

\begin{align}
\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{MSE}}(I^{L,R}_{SR}, I^{L,R}_{HR}) + \lambda \mathcal{L}_{FC}(I^{L,R}_{SR}, I^{L,R}_{HR}), \quad (18)
\end{align}

where $I^{L,R}_{HR}$ represents the left-view and right-view HR image, and $\mathcal{L}_{\text{MSE}}$ is the MSE loss:

\begin{align}
\mathcal{L}_{\text{MSE}} &= \frac{1}{N} \sum_{i=1}^{N} ||I^{L,R}_{HR} - I^{L,R}_{SR}||^2, \quad (19)
\end{align}

In addition, $\mathcal{L}_{FC}$ is the frequency Charbonnier loss, defined as:

\begin{align}
\mathcal{L}_{FC} &= \frac{1}{N} \sum_{i=1}^{N} \sqrt{||\text{FFT}(I^{L,R}_{HR}) - \text{FFT}(I^{L,R}_{SR})||^2 + \epsilon^2}, \quad (20)
\end{align}

with constant $\epsilon$ empirically set to $10^{-3}$ for all the experiments. $\text{FFT}(\cdot)$ denotes a fast Fourier transform. The parameter $\lambda$ in Eq. (18) is a hyper-parameter used to control the composition of the frequency Charbonnier loss function. The parameter $\lambda$ is set to 0.01 for all the experiments. More training details of our method are presented in Section 4.

### 4. Experiments

#### 4.1. Implementation Details

In this section, we provide a detailed description of the experimental setting, including the datasets, the evaluation metrics, and the training configurations.

**Dataset.** Following the previous methods [4, 32, 38], we employ the training and validation datasets provided by the Flickr1024 [31]. To be specific, we employ 800 stereo images as the training data, and 112 stereo images as the validation data. We augment the training data with random horizontal, flips, rotations, and RGB channel shuffle. For testing, we use four benchmark datasets: KITTI 2012 [9], KITTI 2015 [22], Middlebury [24], and Flickr1024 [31].

**Evaluation metrics.** We adopted 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., $(Left + Right)/2$).
**Model Setting.** The numbers of the CHIMB blocks and feature channels are flexible and configurable. We construct two CVHSSR networks of varying sizes, which we named CVHSSR-T (Tiny) and CVHSSR-S (Small) by adjusting the number of channels and blocks. Specifically, the number of channels and blocks for CVHSSR-T are set to 48 and 16 respectively. The number of channels and blocks for CVHSSR-S are set to 64 and 32 respectively.

**Training Settings.** Our network was optimized using the Lion method [2] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a batch size of 8. Our CVHSSR was implemented in PyTorch on a PC with four NVIDIA RTX 3090 GPUs. The learning rate was initially set to $5 \times 10^{-4}$ and decayed the learning rate with the cosine strategy. We trained this model for 200,000 iterations. To alleviate the overfitting issue, we use stochastic depth [11] with 0.1 and 0.2 probability for CVHSSR-S and CVHSSR-B, respectively. Moreover, we also use Test-time Local Converter (TLC) [3] to further improve the model performance. The TLC method mainly aims to reduce the discrepancy between the distribution of global information during training and inference by conferring global operations into local operations at inference.

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

In this section, we conduct a comparative analysis between our CVHSSR (with 2 different variations) and existing super-resolution (SR) methods. The comparison involved SISR methods such as VDSR [13], EDSR [18], RDN [40], and RCAN [39], as well as stereo image SR methods like StereoSR [12], PASSRnet [29], SRR+SAM [37], IMSSRNet [15], iPASSR [32], SSRD-FNet [5], NAFFSR [4], SwinFIR [38], and PFT-SRR [10]. All these methods were trained on the same datasets as ours, and their PSNR and SSIM scores were evaluated and reported by [4].

**Quantitative Evaluations.** We present a comparative evaluation of our proposed CVHSSR against existing stereo SR methods at $\times 2$ and $\times 4$ upscaling factors, as summarized in Table 1. Notably, even our smallest CVHSSR-T model outperforms the NAFFSR-S method on all datasets while utilizing 60% fewer parameters. Moreover, our CVHSSR-S model achieves better results than the NAFFSR-B method while requiring 70% fewer parameters. Specifically, our

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Params</th>
<th>KITTI2012</th>
<th>Middlebury</th>
<th>Flickr2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>VDSR [13]</td>
<td>$\times 2$</td>
<td>0.66M</td>
<td>31.17/0.9062 29.50/0.9038 32.66/0.9101</td>
<td>30.30/0.9089 29.78/0.9150 32.77/0.9102</td>
<td>25.60/0.8534</td>
</tr>
<tr>
<td>EDSR [18]</td>
<td>$\times 2$</td>
<td>3.86M</td>
<td>30.83/0.9199 29.94/0.9231 34.84/0.9489</td>
<td>30.96/0.9228 30.73/0.9335 34.95/0.9492</td>
<td>28.66/0.9087</td>
</tr>
<tr>
<td>RDN [40]</td>
<td>$\times 2$</td>
<td>22.0M</td>
<td>30.81/0.9197 29.91/0.9224 34.85/0.9488</td>
<td>30.94/0.9227 30.70/0.9330 34.94/0.9494</td>
<td>28.64/0.9084</td>
</tr>
<tr>
<td>RCAN [39]</td>
<td>$\times 2$</td>
<td>15.3M</td>
<td>30.88/0.9202 29.97/0.9231 34.80/0.9482</td>
<td>31.02/0.9322 30.77/0.9336 34.90/0.9486</td>
<td>28.63/0.9082</td>
</tr>
<tr>
<td>StereoSR [12]</td>
<td>$\times 2$</td>
<td>1.08M</td>
<td>29.42/0.9040 28.53/0.9038 33.15/0.9343</td>
<td>29.51/0.9073 29.33/0.9168 33.23/0.9348</td>
<td>25.96/0.8599</td>
</tr>
<tr>
<td>PASSRnet [29]</td>
<td>$\times 1.37$</td>
<td>3.68M</td>
<td>30.68/0.9195 29.81/0.9191 34.13/0.9421</td>
<td>30.81/0.9190 30.60/0.9300 34.23/0.9422</td>
<td>28.38/0.9038</td>
</tr>
<tr>
<td>IMSSRNet [15]</td>
<td>$\times 2$</td>
<td>0.39M</td>
<td>30.97/0.9210 30.01/0.9234 34.41/0.9454</td>
<td>31.11/0.9240 30.81/0.9340 34.51/0.9454</td>
<td>28.60/0.9097</td>
</tr>
<tr>
<td>iPASSR [32]</td>
<td>$\times 1.37$</td>
<td>3.68M</td>
<td>30.97/0.9210 30.01/0.9234 34.41/0.9454</td>
<td>31.11/0.9240 30.81/0.9340 34.51/0.9454</td>
<td>28.60/0.9097</td>
</tr>
<tr>
<td>SSRED-FNet [5]</td>
<td>$\times 2.10M$</td>
<td>3.68M</td>
<td>30.11/0.9224 30.10/0.9250 34.02/0.9508</td>
<td>31.23/0.9234 30.90/0.9352 35.01/0.9509</td>
<td>28.85/0.9112</td>
</tr>
<tr>
<td>PFT-SRR [10]</td>
<td>$\times 2$</td>
<td>3.15M</td>
<td>31.15/0.9166 30.16/0.9187 35.08/0.9516</td>
<td>31.29/0.9195 30.96/0.9306 35.21/0.9520</td>
<td>29.05/0.9049</td>
</tr>
<tr>
<td>SwinFIR-T [38]</td>
<td>$\times 0.89M$</td>
<td>3.15M</td>
<td>31.09/0.9226 30.17/0.9253 35.00/0.9491</td>
<td>31.22/0.9254 30.96/0.9359 35.11/0.9497</td>
<td>29.03/0.9134</td>
</tr>
<tr>
<td>NAFFSR-T [4]</td>
<td>$\times 0.45M$</td>
<td>3.15M</td>
<td>31.12/0.9224 30.18/0.9253 34.93/0.9495</td>
<td>31.26/0.9254 30.99/0.9355 35.01/0.9495</td>
<td>28.94/0.9128</td>
</tr>
<tr>
<td>NAFFSR-S [4]</td>
<td>$\times 1.54$</td>
<td>3.68M</td>
<td>31.23/0.9236 30.28/0.9266 35.23/0.9515</td>
<td>31.36/0.9266 31.08/0.9367 35.30/0.9514</td>
<td>29.19/0.9160</td>
</tr>
<tr>
<td>NAFFSR-B [4]</td>
<td>$\times 2.77M$</td>
<td>3.68M</td>
<td>31.40/0.9254 30.42/0.9282 35.62/0.9545</td>
<td>31.55/0.9283 31.22/0.9380 36.55/0.9484</td>
<td>29.54/0.9204</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Scale</th>
<th>Params</th>
<th>KITTI2012</th>
<th>Middlebury</th>
<th>Flickr2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVHSSR-T (Ours)</td>
<td>$\times 2.66M$</td>
<td>3.68M</td>
<td>31.31/0.9250 30.33/0.9277 35.41/0.9333</td>
<td>31.46/0.9280 31.13/0.9377 35.47/0.9352</td>
<td>29.26/0.9180</td>
</tr>
<tr>
<td>CVHSSR-S (Ours)</td>
<td>$\times 2.22M$</td>
<td>3.68M</td>
<td>31.42/0.9292 30.41/0.9287 35.73/0.9551</td>
<td>31.57/0.9291 31.22/0.9385 35.78/0.9550</td>
<td>29.56/0.9126</td>
</tr>
</tbody>
</table>

**Table 1.** Quantitative results achieved by different methods on the KITTI2012, KITTI2015, Middlebury, and Flickr2012 datasets. 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.
CVHSSR-S method is 0.19 dB and 0.48 dB higher than NAFSSR-S and SRRDE-FNet, respectively, on the ×4 Flickr1024 dataset for the same amount of parameters. These results demonstrate the effectiveness of our proposed method and its superiority over existing methods in stereo image SR tasks.

**Visual Comparison.** In Figures 5 and 6, we present comparative results of ×4 stereo SR obtained by different stereo SR methods on the Flickr1024 [31], KITTI 2012 [9], and KITTI 2015 [22] datasets. The visual comparison of reconstructed images demonstrates that our proposed CVHSSR-S method achieves sharper and more ac-
curate texture details compared to the NAFSSR-B method which still suffers from over-smoothing fine textures. This validates the superiority and effectiveness of our proposed CVHSSR method.

4.3. Ablation Study

In this section, we conduct a set of ablation experiments to evaluate the performance of each proposed module. The evaluation is performed on the Flickr1024 [31] validation dataset.

Effectiveness of Each Operation. To further substantiate the effectiveness of our proposed module, a series of ablation experiments were conducted and the results are presented in Table 2. Initially, the NAFSSR was used as the baseline, and subsequently, the corresponding module was modified continuously to verify the efficacy of the proposed module. As depicted in the table, LKA provided a performance improvement of 0.08 dB to the baseline due to its larger receptive field. However, merely enlarging the receptive field of the network does not fully exploit the hierarchical relationship of the intra-view. Hence, our proposed CHIE provided a superior performance improvement of 0.1 dB to the network. Compared to the simple FFN in NAFSSR [4], our proposed IRFFN more effectively regulated the information flow. Additionally, our proposed CVIM demonstrated a superior ability to fuse similar information from various perspectives and enhance network performance, compared to the traditional PAM [32]. These comparisons unequivocally underscore the effectiveness of our proposed methods.

Effectiveness of $\lambda$ in loss function. To evaluate the influence of different values of $\lambda$ in the loss function, a series of experiments are conducted in this section. The hyperparameter $\lambda$ is utilized to balance the trade-off between the MSE loss function and the frequency Charbonnier loss function. We conducted a range of empirical experiments with six different $\lambda$ values within the range of [0,1], based on previous experience. The impact of different $\lambda$ values on the model performance is illustrated in Figure 7. The results demonstrate that the network achieves optimal performance at $\lambda = 0.01$.

### Table 2. Ablation studies of different components. We report the PSNR (dB) values on Flickr 1024 validation datasets (×4).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>LKA</td>
<td>✘</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>CHIE</td>
<td>✘</td>
<td>✘</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>IRFFN</td>
<td>✘</td>
<td>✘</td>
<td>✔</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>CVIM</td>
<td>✘</td>
<td>✘</td>
<td>✘</td>
<td>✔</td>
<td>✘</td>
</tr>
<tr>
<td>PSNR</td>
<td>23.59</td>
<td>23.67</td>
<td>23.69</td>
<td>23.70</td>
<td>23.72</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.7345</td>
<td>0.739</td>
<td>0.7399</td>
<td>0.7402</td>
<td>0.7413</td>
</tr>
<tr>
<td>$\Delta$PSNR</td>
<td>0</td>
<td>0.08</td>
<td>0.10</td>
<td>0.11</td>
<td>0.13</td>
</tr>
</tbody>
</table>

![Figure 7. Study on the influence of $\lambda$ in loss function. The PSNR performance on the Flickr 1024 dataset (×4). When $\lambda = 0.01$, the model achieves the best results.](image)

4.4. NTIRE Stereo Image SR Challenge

We have submitted the results obtained from our proposed approach to the NTIRE 2023 Stereo Image Super-Resolution Challenge. To enhance the potential performance of our method, we have increased the depth and width of the CVHSSR-Based model. During test-time, we employed self-ensemble and model ensemble strategies. The final submission achieves 24.114 dB PSNR on the validation set and achieves 23.742 dB PSNR on the test set. We won 8th, 5th, and 4th on track 1 Fidelity & Bicubic, track 2 Perceptual & Bicubic, and tack 3 Fidelity & Realistic, respectively.

5. Conclusion

In this paper, we present an efficient stereo image SR method named Cross-View Hierarchy Network (CVHSSR). In particular, we design a cross-hierarchy information mining block that efficiently extracts similar features from intra-views by leveraging the hierarchical relationships of the features. Additionally, we introduce a cross-view interaction module to effectively convey mutual information between different views. The integration of these two modules enables our proposed network to achieve superior performance with fewer parameters. Comprehensive experimental evaluations demonstrate that our proposed CVHSSR method outperforms current state-of-the-art models in stereo image SR.

Acknowledgments

This work was funded by Science and Technology Project of Guangzhou (202103010003), Science and Technology Project in key areas of Foshan (20200001006285), Natural Science Foundation of Guangdong Province (2019A1515011041), Xijiang Innovation Team Project (XJCTD3-2019-04B).
References


