This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3236845

Steformer: Efficient Stereo Image Super-Resolution with Transformer

Jianxin Lin, Lianying Yin, Yijun Wang

***Abstract*—With the rapid development of stereoscopic vision applications, stereo image processing techniques have attracted increasing attention in both academic and industrial communities. In this paper, we study the fundamental stereo image super-resolution (SR) problem, which aims to recover high-resolution stereo images from low-resolution (LR) stereo images. Since disparities between stereo images vary significantly, convolutional network-based stereo image SR methods show a limitation in capturing long-range dependencies. To address this problem, this paper proposes to leverage the capability of self-attention in Transformers to efficiently capture reliable stereo correspondence and incorporate cross-view information for stereo image SR. Our model, named Steformer, consists of three parts: cross attentive feature extraction, cross-to-intra information integration and high-quality image reconstruction. In particular, the cross attentive feature extraction module employs residual cross Steformer blocks (RCSB) for long-range cross-view information extraction. Then, the cross-to-intra information integration module exploits cross-view and intra-view information using cross-to-intra attention mechanism (C2IAM). Finally, residual Steformer blocks (RSB) are designed for feature pre-processing in high-quality image reconstruction. Extensive experiments show that Steformer achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations, while the total number of parameters can be reduced by up to 40.71%.**

***Index Terms*—Stereo Image Processing, Image Super-Resolution, Transformer.**

I. Introduction

Single image super-resolution (SISR) aims at reconstructing natural and realistic textures for a high-resolution (HR) image from its degraded low-resolution (LR) counterpart. SISR has been an active area [17, 22, 51, 50] for a long time because it offers the promise of overcoming resolution limitations in many applications, such as medical imaging [14], satellite imaging [34], and so on. Recently, with the wide use of dual-lens smartphones, unmanned aerial vehicles, and autonomous robots, the stereoscopic vision has attracted increasing attention from researchers. Therefore, studying stereo image super-resolution is quite important to the new emerging applications. There have been several works demonstrating that the spatial dependency contained in LR stereo images could help to improve super-resolution performance [40, 42]. However, different objects at different depths and different stereo images can have significantly different parallaxes. In addition, incorporating information from both overlapped and non-overlapped regions could be crucial for stereo image SR performance boosting.

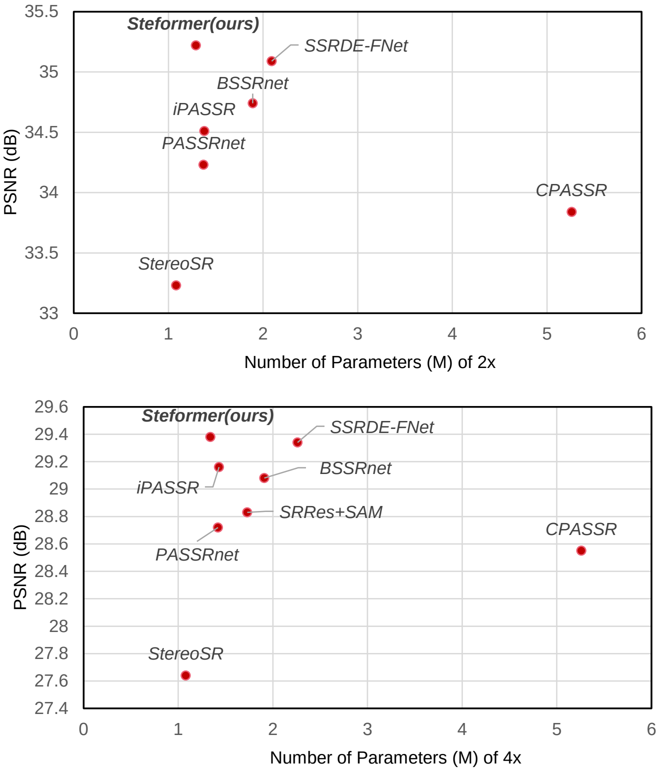


Fig. 1: PSNR results on $Middlebury$ [33] testing set v.s the total number of parameters of different stereo image SR (×2 and ×4) methods.

Due to the ill-posed nature, SR is a highly challenging problem that usually requires strong image priors for effective restoration. Since convolutional neural networks (CNNs) perform well at learning generalizable priors from large-scale data, several CNN based works [15, 40, 47, 45, 4, 42, 8] have been developed for stereo image SR. Although the performance is significantly improved compared with single image-based solutions, the convolution operator has a limited receptive field, thus preventing it from modeling long-range pixel dependencies in stereo images. In addition, the convolution kernels are content-independent, thus lacking the flexibility to model the varying parallax relationship in stereo images. Recently, Transformers [39] have shown a significant performance on natural language processing tasks [2, 9, 30, 31, 32], high-level and low-level vision problems [3, 11, 24, 37, 21, 5, 49]. Transformer-based network structures are naturally

Jianxin Lin, Lianying Yin and Yijun Wang are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: linjianxin@hnu.edu.cn; yin2110@hnu.edu.cn; wyjun@hnu.edu.cn).

Yijun Wang is the corresponding author.

good at capturing long-range dependencies in the data by the global self-attention. However, few efforts are made to explore its role in stereoscopic vision to address the limitations of CNN based methods.

In this paper, we propose an efficient Transformer based structure for stereo image super-resolution, namely Steformer, which consists of three modules: cross attentive feature extraction, cross-to-intra information integration, and high-quality image reconstruction. Cross attentive feature extraction module is composed of several residual cross Steformer blocks (RCSB), each of which employs multi-Dconv interactive attention (MDIA) layers for long-range cross-view information extraction, reducing the impact of the hardship that parallax varies with content position. After extracting hierarchical features with a consequence of RCSBs, a 1×1 convolutional layer for feature aggregation and a residual connection for providing a shortcut bypassing the abundant low-frequency information are employed to facilitate the flow of information. The cross-to-intra information integration module exploits cross-view and intra-view information using cross-to-intra attention mechanism (C2IAM). C2IAM firstly utilizes parallax-attention between the stereo image features to capture the cross-view correlation information as well as convert both left and right features to the other side. Then, interactive-attention between original and predicted monocular image features are applied to effectively aggregate intra-view supplementary information, hence providing comprehensive features for high-quality image reconstruction. In high quality reconstruction module, residual Steformer blocks (RSB), which use multi-Dconv self-attention (MDSA) to capture the internal feature correlation, are utilized for high-quality feature pre-processing. Finally, RCSBs are used to reconstruct high-quality stereo images jointly to adequately incorporate different levels of cross-view information.

This study conductes comprehensive experiments and demonstrate the superior performance of our Steformer on $Flickr1024$ [40], $KITTI2012$ [12], $KITTI2015$ [27] and $Middlebury$ [33] datasets. For example, Steformer derives a numerical gain of 0.14 dB for PSNR with 38.28% fewer network parameters than the state-of-the-art method on the $Middlebury$ testing set as shown in Fig. 1. We expect our work will encourage further research to explore Transformer-based architectures for stereo image SR.

Our contributions can be summarized as follows:

- This work proposes Steformer, an efficient Transformer architecture for stereo image super-resolution.
- The residual Steformer block (RSB) and residual cross Steformer block (RCSB) along with multi-Dconv interactive attention (MDIA) are proposed to better extract long range cross-view information and reconstruct high quality stereo super-resolution images.
- We introduce Cross-to-Intra Attention Mechanism (C2IAM) to further capture cross-view and intra-view information in stereo images.
- Steformer significantly exceeds SOTA methods on various stereo SR datasets with much fewer network parameters.

In summary, we propose a Transformer based stereo image super-resolution model, Steformer. Steformer can address the limitation of CNNs in capturing long-range dependencies and is computationally efficient to handle high-resolution images. The remainder of this paper is organized as follows. Section II reviews the related work while Section III introduces the proposed method. The experimental results are presented in Section IV. Section V discusses the dilemma of stereo image super-resolution task and room for improvement. Finally, section VI presents a summary for the proposed framework.

II. Related work

A. Single Image SR

In recent decades, it has already been demonstrated that data-driven CNN based architectures surpass traditional single image super-resolution (SISR) approaches. Specifically, the seminal CNN-based work, SRCNN [10], employs a relatively shallow network to learn a mapping from low-resolution (LR) to high-resolution (HR). Kim et al. [17] found that increasing the network depth can significantly improve the reconstruction quality and proposed a very deep super-resolution network (VDSR) for SISR. Inspired by VDSR, a deep recursive network [18] which improves performance without introducing new parameters was soon proposed. Lim et al. [22] proposed an enhanced deep super-resolution network (EDSR) by using simplified residual blocks. Zhang et al. [51] proposed a very deep residual dense network (RDN), in which residual dense block (RDB) is employed as the essential block to fully exploit the hierarchical features of all convolutional layers. More recently, the concept of attention mechanism was introduced to SISR. Zhang et al. [50] proposed a very deep residual channel attention network (RCAN) to better utilize the rich low-frequency information contained in the LR input and features. Kim et al. [19] introduced several specially designed attention mechanisms by proposing a residual attention module (RAM) and an SR network using RAM (SRRAM). Anwar et al. [1] proposed a densely residual laplacian network (DRLN), using a pyramid level to weigh the different sub-band features. To fully utilize the hierarchical features on the residual branches, Liu et al. [23] proposed a novel residual feature aggregation (RFA) network, which combines several residual modules and forwards features directly on each local residual branch by adding skip connections. Although deeply stacked convolutional neural networks can provide considerable performance boosts for SISR, their huge parameters and computational load are impractical for real-world applications. Park et al. [28] designed a lightweight model to balance the computational load and reconstruction performance. More recently, Zou et al. [54] argued that most CNN-based methods ignore the importance of frequency information which can reflect the semantic information of the images in different wavebands, they used WT to separate the different frequency information of the image and used a multi-branch network to recover this information. Li et al. [20] thought that deep learning-based methods ignore the relationship between L1 and perceptual minimization, they proposed a real-world image super-resolution by exclusionary dual-learning (RWSR-EDL) to troubleshoot feature diversity in perceptual and L1 cooperative learning. Simply processing stereo images separately does not work well since the correlation between the left and right views is ignored, which has prompted the community to explore a method that can fuse information from the other view to achieve a better stereo super-resolution performance.

B. Stereo Image SR

The pioneering work for stereo image SR based on CNNs was presented by Jeon et al. [15], in which a StereoSR network was proposed to learn a parallax prior from stereo image datasets by jointly training two-stage networks with pre-setting the maximum parallax. Wang et al. [40] noted that the parallax attention mechanism can correlate information in the global range of dual-view image polar directions without pre-setting the maximum parallax, which has more flexibility and robustness. Song et al. [36] combined the parallax attention mechanism with the self-attention mechanism to enhance the utilization of non-local context information within a single view based on setting up the association between a stereo image pair. Ying et al. [47] proposed a stereo image super-resolution algorithm based on a generic stereo attention module (SAM). Xu et al. [45] presented the BSSRnet that introduces the concept of bilateral filtering to stereo image SR. Chen et al. [4] proposed a cross parallax attention stereo super-resolution network (CPASSRnet), which can produce multi-scale results simultaneously without a maximum disparity limit or epipolar line limit. Wang et al. [42] proposed a network named iPASSR which further leverages the symmetry in stereo image SR based on PASSRnet [40]. iPASSR utilizes the Bi-directional Parallax Attention Module (BiPAM) to interact with the information of stereo input images simultaneously. Ma et al. [26] proposed a perception-oriented stereo image super-resolution method based on StereoSR [15] by using the feedback guidance to improve perceptual performance. zhu et al. [53] proposed a cross view capture network (CVCnet) to fully capture the global contextual features from cross view images. More recently, Dai et al. [8] considered that stereo super-resolution reconstruction and parallax estimation can be mutually reinforcing, and constructed the HR disparity using HR features produced by the SR process to refine the SR image reconstruction. With the enormous structure, its specially designed network costs large computational and memory resources. The aforementioned methods are CNN-based architecture which has a limiting receptive field. By contrast, we propose an efficient Transformer-based network with great performance on extensive experiments. At the same time, Chu et al. [7] won the 1st place on the NTIRE 2022 Stereo Image Super-resolution Challenge [41] and proposed a CNN-based baseline, which achieves comparable results with Steformer, by adding cross attention module to NAFNet [6] (a network without nonlinear activation functions).

C. Vision Transformers

Inspired by the success of Transformer [39] in the field of natural language processing (NLP), there have been numerous attempts to explore the benefits of Transformer in both high-level and low-level computer vision problems such as object detection [3, 24], image recognition [43, 21, 48], segmentation [24, 44, 52], super-resolution [43, 21, 46], denoising [43, 21, 5] and deraining [5]. However, the Transformer design faces great challenges in processing high-resolution images, especially in SR problems, as the computational complexity of self-attention increases quadratically with the number of image patches, resulting in a large number of parameters and huge computational costs. Recently, some methods have tried to alleviate this dilemma. Liu et al. [24] limited self-attention computation to non-overlapping local windows and employed a shifted windowing scheme for greater efficiency. Liang et al. and Wang et al. [21, 43] used the Swin Transformer design for different image super-resolution and restoration tasks. Zamir et al. [49] further boosted the efficiency of the Transformer by modifying the multi-head self-attention and feed-forward network. Although the Transformer-based method has shown great performance on different tasks including SISR, there has not been a Transformer-based method that is specially designed for stereo image SR tasks which combine dual-views images. To address this problem, this paper proposes to leverage the capability of self-attention in Transformers to efficiently capture reliable stereo correspondence and incorporate cross-view information for stereo image SR.

III. Approach

A. Network architecture

As shown in Fig. 2, Steformer accepts a pair of stereo LR images and super-resolves them. To effectively exploit and aggregate the intra-view and cross-view information, the pipeline is symmetrically constructed and mainly consists of **(1) cross attentive feature extraction**, **(2) cross-to-intra information integration** and **(3) high quality image reconstruction**.

Steformer first applies a convolution to obtain low-level feature embedding from a low-resolution stereo image pair. Then, the cross attentive feature extraction module uses residual cross Steformer blocks (RCSB) to extract hierarchical features involving long-range cross-view information. Next, the cross-to-intra information integration module exploits cross-view information of stereo features and aggregates intra-view supplementary information using the cross-to-intra attention mechanism (C2IAM). In high quality reconstruction module, residual Steformer blocks (RSB) are utilized for high-quality feature pre-processing, and RCSBs are used to reconstruct high-quality stereo images jointly to incorporate different levels of cross-view information. Finally, a convolution layer is applied to the refined features to generate a residual image to which bicubic interpolated images are added to obtain the left and right super-resolution images.

B. Cross attentive feature extraction

Given a low-resolution stereo image pair $(I^{LR}_{left}, I^{LR}_{right}) \in \mathbb{R}^{H\times W\times 3}$, a 3×3 convolutional layer $H_F(\cdot)$ is first applied to extract initial shallow features $(F_{L_0}, F_{R_0}) \in \mathbb{R}^{H\times W\times C}$ as

$$F_{L_0/R_0} = H_F(I^{LR}_{left/right}). \qquad (1)$$

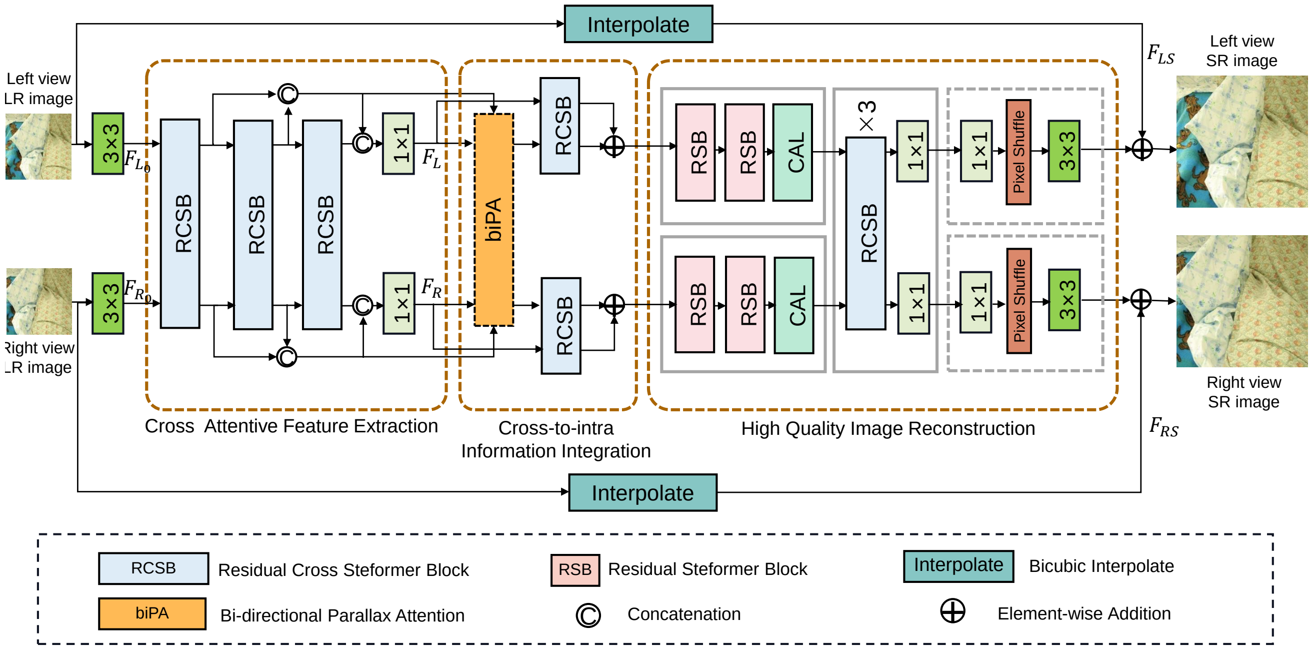

Interpolate
Left view LR image
Right view LR image
3×3
F_{L_0}
F_{R_0}
RCSB
1×1
F_L
F_R
biPA
RSB
CAL
×3
Pixel Shuffle
F_{LS}
F_{RS}
Left view SR image
Right view SR image
Cross Attentive Feature Extraction
Cross-to-intra Information Integration
High Quality Image Reconstruction
RCSB Residual Cross Steformer Block
RSB Residual Steformer Block
Interpolate Bicubic Interpolate
biPA Bi-directional Parallax Attention
© Concatenation
⊕ Element-wise Addition

Fig. 2: Steformer network architecture. Steformer consists of three modules: (1) cross attentive feature extraction module for long-range cross-view information extraction; (2) cross-to-intra information integration module for cross-view correlation and supplementary information aggregation; (3) high quality image reconstruction module for SR images reconstruction, the gray dotted rectangle is the process of upsampling the features into high resolution image.

Then, the cross features $(F_L, F_R) \in \mathbb{R}^{H\times W\times C}$ are extracted by

$$F_L = H_{CF}(F_{L_0}, F_{R_0}), \tag{2}$$

$$F_R = H_{CF}(F_{R_0}, F_{L_0}), \tag{3}$$

where $H_{CF}(\cdot)$ represents the cross attentive feature extraction module which is constructed by cascading N residual cross Steformer blocks (RCSBs) and a 1×1 convolutional layer which helps to bring the inductive bias of the convolution operation into the Transformer-based network. More concretely, as shown in Fig. 2, intermediate features $F_{L_{1,\dots,N}}, F_{R_{1,\dots,N}}$ and output features F_L, F_R are extracted as

$$F_{L_i}, F_{R_i} = H_{RCSB_i}(F_{L_{i-1}}, F_{R_{i-1}}), i = 1, 2, ..., N, \tag{4}$$

$$F_L^{conc} = \text{Concat}(F_{L_1}, F_{L_2}, ..., F_{L_N}), \tag{5}$$

$$F_R^{conc} = \text{Concat}(F_{R_1}, F_{R_2}, ..., F_{R_N}), \tag{6}$$

$$F_L = H_{aggr}(F_L^{conc}), \tag{7}$$

$$F_R = H_{aggr}(F_R^{conc}), \tag{8}$$

where $H_{RCSB_i}(\cdot)$ represents the i-th residual cross Steformer block and $H_{aggr}(\cdot)$ denotes the 1×1 convolutional layer for dense connection and mapping the features into the original number of channels.

Residual cross Steformer block. As shown in Fig. 3 (b), the residual cross Steformer block (RCSB) is a residual block with cross Steformer layers (CSL) and convolutional layers. Specifically, given the input features $\left(F_{L_{i,0}}, F_{R_{i,0}}\right)$ of the i-th RCSB, the intermediate features extracted by M cross Steformer layers are

$$F_{L_{i,j}}, F_{R_{i,j}} = H_{CSL_{i,j}}(F_{L_{i,j-1}}, F_{R_{i,j-1}}), j = 1, 2, ..., M, \tag{9}$$

where $H_{CSL_{i,j}}(\cdot)$ is the j-th cross Steformer layer in the i-th RCSB. Then, a convolutional layer is employed before being concatenated with input features F_{L_i},F_{R_i} by residual connection. The output of RCSB is formulated as

$$F_{L_i/R_i} = H_{1\times1}(F_{L_{i,M}/R_{i,M}}) + F_{L_{i,0}/R_{i,0}}, \tag{10}$$

where $H_{1\times1}$ is 1×1 convolutional layer.

Cross Steformer layer. The cross Steformer layer (CSL) is designed based on the multi-head self-attention of Transformer layer [39]. As shown in Fig. 3 (b), the core modules of cross Steformer layer are multi-Dconv interactive attention (MDIA) and the Gated Dconv Feed-Forward Network (GDFN) [49].

Due to the heavy computation and memory burden in conventional Transformer design, this study proposes multi-Dconv interactive attention (MDIA) which can explore deep feature correspondences and has linear complexity. Specifically, the input features F_{L_i},F_{R_i} are first processed with layer normalization, then it generates the key and $value$ matrices K and V by employing a 1×1 convolutional layer and a 3×3 depth-wise convolutional layer to enhance the local region feature. Meanwhile, similar to K and V, this study utilizes the same convolutional layers while the input feature incorporates left and right views to generate the $query$ matrix Q. $Q, K, V \in \mathbb{R}^{H\times W\times C}$ are computed as

$$Q = H_{3\times3}(H_{1\times1}(\text{Concat}(LN(F_{L_i}), LN(F_{R_i})))), \tag{11}$$

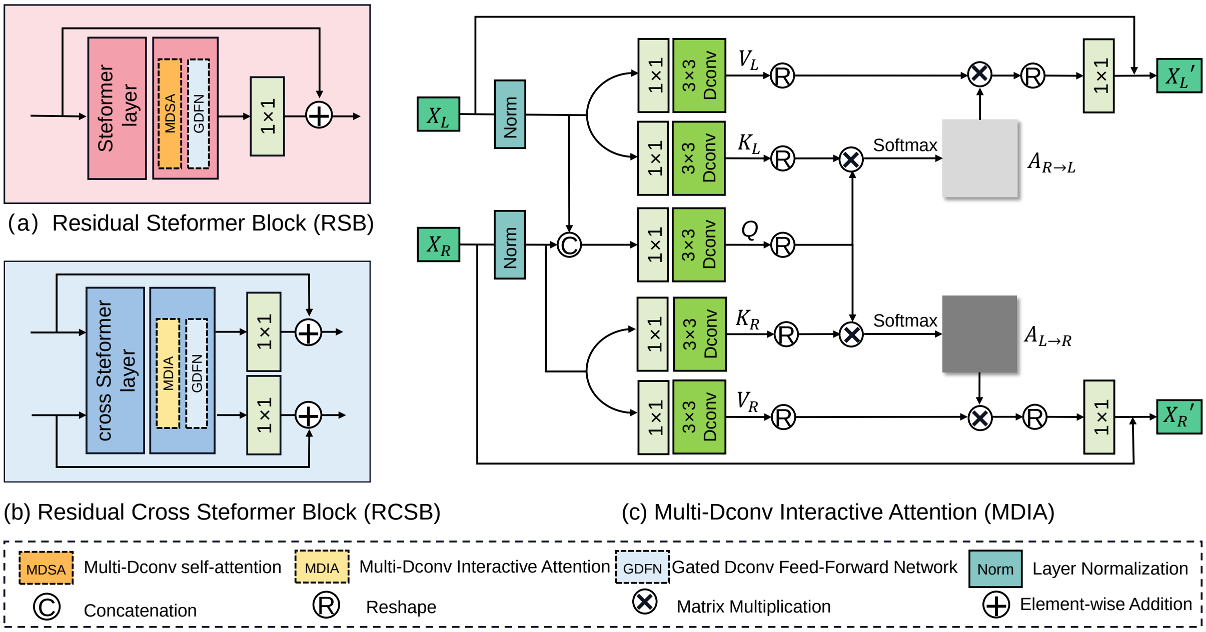


Fig. 3: (a), (b) are the structure of Residual Steformer Block, Cross Residual Steformer Block, (c) is the Multi-Dconv Interactive Attention block in Cross Steformer layer.

$$K_{L/R} = H_{3\times3}(H_{1\times1}(LN(F_{L_i/R_i}))), \tag{12}$$

$$V_{L/R} = H_{3\times3}(H_{1\times1}(LN(F_{L_i/R_i}))), \tag{13}$$

where LN, $H_{1\times1}$ and $H_{3\times3}$ represent the layer normalization, the 1×1 convolutional layer and the 3×3 depth-wise convolutional layer, respectively. Then, Q, K, V are reshaped as size $\mathbb{R}^{HW\times C}$ to obtain attention maps $A_{L\to R}$, $A_{R\to L} \in \mathbb{R}^{C\times C}$ as shown in Fig. 3 (c). The process of multi-Dconv interactive attention is noted as

$$\begin{aligned} &\text{Attention}(Q, K_{L/R}, V_{L/R}) = \text{softmax}(Q \cdot K^T_{L/R}/\omega) \cdot V_{L/R}, \\ &X'_{L/R} = H_{1\times1}(\text{Attention}(Q, K_{L/R}, V_{L/R})) + X_{L/R}, \end{aligned} \tag{14}$$

where ω is a learnable argument; $X_{L/R}$ and $X'_{L/R}$ are the input and output feature maps. Following [49], the Gated Dconv Feed-Forward Network (GDFN) which encodes the information from the location of spatially adjacent pixels and helps to learn local image structure is used to replace the feed-forward network in conventional Transformer design.

The design of CSL can effectively integrate the information from a stereo image pair, since it employs multi-Dconv interactive attention (MDIA) layers for long-range cross-view information extraction, reducing the impact of the hardship that parallax varies with content position.

C. Cross-to-intra information integration

In cross-to-intra information integration module, this study proposes a Cross-to-Intra Attention Mechanism (C2IAM) to further capture cross-view and intra-view information in stereo images. The architecture of C2IAM is illustrated in Fig. 4. Firstly, calculate the bi-directional parallax-attention (biPA) as iPASSR [42] between the stereo image features, which is used for capturing the cross-view correlation information, as well as converting both left and right features to the other side. Then, it adopts an RCSB which computes interactive attention between original and predicted monocular image features to aggregate intra-view supplementary information, providing comprehensive features for high-quality image reconstruction.

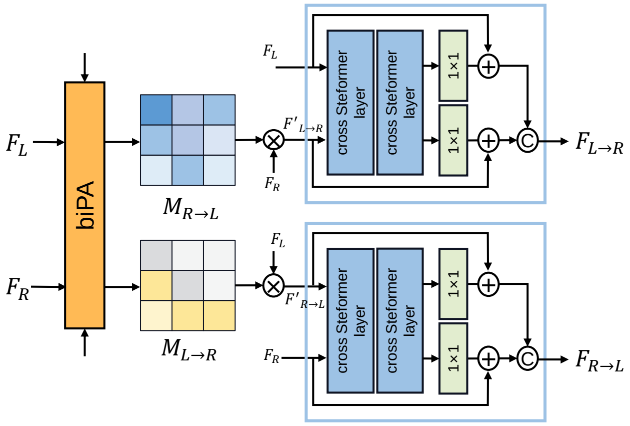


Fig. 4: Cross-to-Intra Attention Mechanism (C2IAM).

Specifically, since hierarchical feature representations contribute to stereo correspondence learning [40], all intermediate cross-features (the up and down arrows pointing to the orange biPA block in Fig. 4) obtained in cross attentive feature extraction module are concatenated with F_R, F_L as the inputs fed to biPA to obtain the parallax-attention maps $M_{L\to R}$ and $M_{R\to L}$. Then, the initial conversion features $F^{'}_{L\to R}$, $F^{'}_{R\to L}$ are generated by multiplying the attention maps and F_L, F_R, respectively. Finally, C2IAM constructs the final conversion feature $F_{L\rightarrow R}$ using an RCSB by the co-action of initial conversion feature $F^{'}_{L\rightarrow R}$ and F_R to get more texture from another view. $F_{R\rightarrow L}$ is generated with the same process.

D. High quality image reconstruction

In high quality image reconstruction module, this study first proposes a **residual Steformer block (RSB)** which cascades M **Steformer layers (SL)** and a 1×1 convolution layer with a residual connection. As shown in Fig. 3 (a), each SL is composed of multi-Dconv self-attention (MDSA) and the Gated Dconv Feed-Forward Network (GDFN) [49]. To fuse the features preliminarily, features $F_{L\rightarrow R}, F_{R\rightarrow L}$ are first fed to N_{RSB} cascaded RSBs. Next, the output features $F^{init}_L, F^{init}_R \in \mathbb{R}^{H\times W\times 2C}$ are fed to a channel attention layer (CAL) [50] to fully exploit contextual information beyond the local region in convolutional kernel. After a 1×1 convolutional layer for feature integration, pre-processed F^{pre}_L, $F^{pre}_R \in \mathbb{R}^{H\times W\times C}$ are obtained.

Then, similar to the cross attentive feature extraction module, this study utilizes RCSB as the base unit for high quality image reconstruction. The fusion features F^{pre}_L and F^{pre}_R are fed to N_{RCSB} cascaded RCSBs, which can adequately incorporate different levels of cross-view information in the reconstruction stage. Then, a convolutional layer and a sub-pixel convolution layer [35] are used to upsample the features and obtain the final super-resolved feature F^{fin}_L and F^{fin}_R. Meanwhile, using bicubic interpolation [16], the input LR stereo image pair is upscaled to the desired resolution, denoted as

$$F_{LS} = H_{\uparrow}\left(I^{LR}_{left}\right), \tag{15}$$

$$F_{RS} = H_{\uparrow}\left(I^{LR}_{right}\right), \tag{16}$$

where $H_{\uparrow}(\cdot)$ represents the bicubic upsampling. Then, to facilitate model learning and reuse higher level features, a long skip connection is performed, which element-wise summed F_{LS}, F_{RS} with F^{fin}_L, F^{fin}_R to produce the super-resolution images respectively:

$$I^{SR}_{left} = F^{fin}_L + F_{LS}, \tag{17}$$

$$I^{SR}_{right} = F^{fin}_R + F_{RS}, \tag{18}$$

E. Loss function

In this section, the loss functions are introduced to optimize Steformer network. The overall loss function can be formulated as

$$\mathcal{L} = \mathcal{L}_{SR} + \lambda(\mathcal{L}_{photo} + \mathcal{L}_{smooth} + \mathcal{L}_{cycle} + \mathcal{L}_{consist}), \tag{19}$$

where $\mathcal{L}_{SR}$, $\mathcal{L}_{photo}$, $\mathcal{L}_{smooth}$, $\mathcal{L}_{cycle}$, $\mathcal{L}_{consist}$ represent the SR loss, photometric loss, smooth loss, cycle loss and consistency loss, respectively. λ is the weight of the regularization term.

SR Loss. This study uses the L_1 distance between the reconstructed images and corresponding ground-truth images as SR loss:

$$\mathcal{L}_{SR} = \left\|I^{SR}_{left} - I^{HR}_{left}\right\|_1 + \left\|I^{SR}_{right} - I^{HR}_{right}\right\|_1, \tag{20}$$

where I^{HR}_{left}, I^{HR}_{right} are high quality ground-truth stereo images.

Smooth Loss. This study employs smooth loss to reduce the amount of undesired noise which implies faulty correspondence in the attention map by minimizing the divergence of neighboring pixel values in textureless regions:

$$\begin{aligned}\mathcal{L}_{smooth} = \sum_{M}\sum_{i,j,k}(&\|M(i,j,k) - M(i+1,j,k)\|_1 \\ &+ \|M(i,j,k) - M(i,j+1,k+1)\|_1),\end{aligned} \tag{21}$$

where $M \in \{M_{L\rightarrow R}, M_{R\rightarrow L}\}$. $M_{L\rightarrow R}(i,j,k)$ represents the correspondence between $I^{LR}_{left}(i,j)$ and $I^{LR}_{right}(i,k)$. $\|M(i,j,k) - M(i+1,j,k))\|_1$ and $\|M(i,j,k) - M(i,j+1,k+1)\|_1$ are used to achieve vertical and horizontal attention consistency [40], respectively.

Residual Losses. Due to the nature of stereo image acquisition, i.e. the cameras are located in different directions and at different angles, the luminous intensity may be different between a pair of stereo images. This problem can lead to an inability to obtain accurate correspondence. To avoid these cases, we follow [42] using the residual images to calculate the photometric loss, the cycle loss, and the consistency loss. Specifically, the low-resolution images are replaced with residual images R_{left} and R_{right}, noted as

$$R^{LR}_{left} = H_{\downarrow}(|I^{HR}_{left} - H_{\uparrow}(I^{LR}_{left})|), \tag{22}$$

$$R^{LR}_{right} = H_{\downarrow}(|I^{HR}_{right} - H_{\uparrow}(I^{LR}_{right})|), \tag{23}$$

where $H_{\downarrow}(\cdot)$ represents the down-sampling. Therefore, the photometric loss and cycle loss can be defined as

$$\begin{aligned}\mathcal{L}_{photo} = &\sum_{p\in V_{L\rightarrow R}} \|R^{LR}_{left}(p) - M_{R\rightarrow L} \otimes R^{LR}_{right}(p)\|_1 \\ &+ \sum_{p\in V_{R\rightarrow L}} \|R^{LR}_{right}(p) - M_{L\rightarrow R} \otimes R^{LR}_{left}(p)\|_1,\end{aligned} \tag{24}$$

$$\begin{aligned}\mathcal{L}_{cycle} = &\sum_{p\in V_{L\rightarrow R}} \|R^{LR}_{left}(p) - M_{L\rightarrow R\rightarrow L} \otimes R^{LR}_{left}(p)\|_1 \\ &+ \sum_{p\in V_{R\rightarrow L}} \|R^{LR}_{right}(p) - M_{R\rightarrow L\rightarrow R} \otimes R^{LR}_{right}(p)\|_1,\end{aligned} \tag{25}$$

where $M_{L\rightarrow R\rightarrow L} = M_{L\rightarrow R} \otimes M_{R\rightarrow L}$, $M_{R\rightarrow L\rightarrow R} = M_{R\rightarrow L} \otimes M_{L\rightarrow R}$, V is the mask of attention map and p represents the valid mask value. Note that $\mathcal{L}_{photo}$ is only calculated in non-occluded regions. To further achieve stereo consistency, we use the residual high quality images $R^{SR}_{left} = H_{\downarrow}(|I^{HR}_{left} - I^{SR}_{left}|)$ and $R^{SR}_{right} = H_{\downarrow}(|I^{HR}_{right} - I^{SR}_{right}|)$ to calculate residual stereo consistency loss. That is

$$\begin{aligned}\mathcal{L}_{consist} = &\sum_{p\in V_{L\rightarrow R}} \|R^{SR}_{left}(p) - M_{R\rightarrow L} \otimes R^{SR}_{right}(p)\|_1 \\ &+ \sum_{p\in V_{R\rightarrow L}} \|R^{SR}_{right}(p) - M_{L\rightarrow R} \otimes R^{SR}_{left}(p)\|_1.\end{aligned} \tag{26}$$

IV. Experiments

In this section, we first introduce the datasets and implementation details. Then, we compare the proposed Steformer to several state-of-the-art single image SR and stereo image SR approaches. Finally, we perform comprehensive ablation studies to validate each component of the proposed Steformer.

A. Datasets

To evaluate the effectiveness of the proposed method, we constructed training set by merging 60 images from $Middlebury$ [33] and 800 images from $Flickr1024$ [40] following the experimental setting of iPASSR [42]. In addition, we collected testing set by selecting 5 images from $Middlebury$ [33], 20 images from $KITTI2012$ [12], 20 images from $KITTI2015$ [27] and all testing images from $Flickr1024$, following [15, 40, 42]. The LR images were generated by bicubic downsampling. In the training set, the generated LR images were cropped into 30×90 patches with a stride of 20, and their HR counterparts were cropped accordingly. Randomly flipping horizontally and vertically was applied for data augmentation. There were 49,020 and 298,143 patches for ×4 SR and ×2 SR training respectively.

B. Implementation details

To trade-off the efficiency and effectiveness of Steformer, the number of RCSBs and RSBs in each module, which are denoted as N_{RCSB} and N_{RSB} respectively, were set to 3 and 2. In each RCSB or RSB, there were two cascaded cross Steformer layers or Steformer layers. Table I and II present the specific architecture of Cross Steformer Layer (CSL) and Steformer Layer (SL) respectively.

The coefficient λ in the loss function was set to 0.1 to achieve the balance of different loss terms. Steformer was implemented in PyTorch [29] and trained with one NVIDIA RTX 3090 GPU. Following the common training strategy [40, 42], we trained Steformer using the Adam optimizer [25] with the momentum terms (β_1, β_2) of (0.9, 0.999) and a batch size of 16 for 120 epochs since more epochs do not provide further improvement. The learning rate was determined experimentally and initially set to 2×10^{-4} and reduced to half for every 30 epochs until it was reduced to 5×10^{-5}.

We utilized the commonly-used peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics to evaluate the performance of the proposed method. To be consistent with the comparison methods [15, 40, 47, 42], we cropped the left borders by 64 pixels when calculating PSNR and SSIM on the left views.

TABLE I: Cross Steformer Layer (CSL) configuration.

Layer name	Input size	Output size	Configuration
Multi-Dconv Interactive Attention (MDIA)			
LayerNormalize1	$H \times W$×64	$H \times W$×64	BiaFree_LayerNorm()
Depth-wise-conv_kv	$H \times W$×128	$H \times W$×128	3×3
$K/K', V/V'$	$H \times W$×128	64×HW 64×HW	chunk(), rearrange
Depth-wise-conv_q, Q	$H \times W$×64	64×HW	3×3, rearrange
att_map_X	64×HW 64×HW 64×HW	64×64	Attention(Q, K, V)
att_map_X'	64×HW 64×HW 64×HW	64×64	Attention(Q, K', V')
conv0	$H \times W$×64	$H \times W$×64	1×1
Gated Dconv Feed-Forward Network (GDFN)			
LayerNormalize2	$H \times W$×64	$H \times W$×64	BiaFree_LayerNorm()
conv1	$H \times W$×64	$H \times W$×340	1×1
Depth-wise-conv1	$H \times W$×340	$H \times W$×340	3×3, chunk()
conv2	$H \times W$×170	$H \times W$×64	1×1

TABLE II: Steformer Layer (SL) configuration.

Layer name	Input size	Output size	Configuration
Multi-Dconv Self Attention (MDSA)			
LayerNormalize1	$H \times W$×64	$H \times W$×64	BiaFree_LayerNorm()
Depth-wise-conv_kv	$H \times W$×128	$H \times W$×128	3×3
K, V	$H \times W$×128	$H \times W$×64 $H \times W$×64	chunk(),rearrange
Depth-wise-conv_q, Q	$H \times W$×64	$H \times W$×64	3×3
att_map_X	64×HW 64×HW 64×HW	64×64	Attention(Q, K, V)
conv0	$H \times W$×64	$H \times W$×64	1×1
Gated Dconv Feed-Forward Network (GDFN)			
LayerNormalize2	$H \times W$×64	$H \times W$×64	BiaFree_LayerNorm()
conv1	$H \times W$×64	$H \times W$×340	1×1
Depth-wise-conv1	$H \times W$×340	$H \times W$×340	3×3,chunk()
conv2	$H \times W$×170	$H \times W$×64	1×1

C. Results

1) Quantitative results: As the quantitative results shown in Table III, Steformer achieves remarkable PSNR and SSIM scores on test sets for ×2 and ×4 SR on both single and stereo image SR tasks. More specifically, with 38.28% fewer network parameters and FLOPs, the PSNR values of our method on the $Middlebury$ dataset are 0.14 dB higher than state-of-the-art method SSRDE-FNet on ×2 stereo image SR, while the total number of parameters can be reduced by 40.71% on ×4 stereo image SR, which indicates that the Steformer architecture is highly efficient on stereo image SR. Compare to the work NAFSSR [7] in the same period, which is a excellent work that won the 1st on NTIRE 2022 Stereo Image Super-resolution Challenge [41], Steformer achieves comparable results for ×4 stereo image SR task with fewer parameters and FLOPs (i.e. +0.02db on $Middlebry$ dataset and -0.05 to -0.01db on other datasets). For the ×2 stereo image SR task, NAFSSR has better performance than Steformer except for the $KITTI2015$ dataset, but Steformer still has fewer parameters and FLOPs than NAFSSR.

The quantitative results demonstrate the superiority of our proposed Steformer in modeling long-range pixel dependencies and varying parallax relationships in stereo images and promoting stereo image SR performance.

2) Qualitative results: Fig. 5 shows the qualitative results for ×4 stereo image SR on $KITTI2015$ and $Middlebury$ dataset. Most CNN-based methods produce blurry images or even incorrect textures since they only use spatial information. In contrast, our proposed method has more details and sharper edges, $i.e.$, in Fig. 5 zoom-in regions, the textures of Steformer is more close to ground-truth. These results demonstrate that

TABLE III: Quantitative Comparison with PSNR/SSIM Metric on $Flickr1024$, $KITTI2012$, $KITTI2015$ and $Middlebury$ Datasets. Higher PSNR/SSIM Values Means Better Performance.

Method	$Scale$	$\#Params.$	$\#FLOPs$	$Left$			$(Left + Right)/2$			
				$KITTI2012$	$KITTI2015$	$Middlebury$	$Flickr1024$	$KITTI2012$	$KITTI2015$	$Middlebury$
Bicubic	×2	-	-	28.44/0.8808	27.81/0.8814	30.46/0.8979	24.94/0.8186	28.51/0.8842	28.61/0.8973	30.60/0.8990
VDSR [17]	×2	0.66M	10.77G	30.17/0.9062	28.99/0.9038	32.66/0.9101	25.60/0.8534	30.30/0.9089	29.78/0.9150	32.77/0.9102
EDSR [22]	×2	38.63M	208.73G	30.83/0.9199	29.94/0.9231	34.84/0.9489	28.66/0.9087	30.96/0.9228	30.73/0.9335	34.95/0.9492
RDN [51]	×2	21.99M	118.79G	30.81/0.9197	29.91/0.9224	34.85/0.9488	28.64/0.9084	30.94/0.9227	30.70/0.9330	34.94/0.9491
RCAN [50]	×2	15.31M	82.23G	30.88/0.9202	29.97/0.9231	34.80/0.9482	28.63/0.9082	31.02/0.9232	30.77/0.9336	34.90/0.9486
SwinIR [21]	×2	1.32M	7.89G	30.89/0.9206	29.98/0.9237	34.69/0.9475	28.67/0.9091	31.02/0.9235	30.77/0.9341	34.80/0.9478
StereoSR [15]	×2	1.08M	90.11G	29.42/0.9040	28.53/0.9038	33.15/0.9343	25.96/0.8599	29.51/0.9073	29.33/0.9168	33.23/0.9348
PASSRnet [40]	×2	1.37M	6.53G	30.68/0.9159	29.81/0.9191	34.13/0.9421	28.38/0.9038	30.81/0.9190	30.60/0.9300	34.23/0.9422
BSSRnet [45]	×2	1.89M	8.42G	30.99/0.9225	30.05/0.9256	34.73/0.9468	28.53/0.9090	31.03/0.9241	30.74/0.9344	34.74/0.9475
CPASSR [4]	×2	5.26M	7.73G	29.68/0.9079	29.69/0.9193	33.68/0.9433	28.12/0.9017	29.87/0.9113	30.39/0.9295	33.85/0.9436
iPASSR [42]	×2	1.38M	7.44G	30.97/0.9210	30.01/0.9234	34.41/0.9454	28.60/0.9097	31.11/0.9240	30.81/0.9340	34.51/0.9454
SSRDE-FNet [8]	×2	2.09M	27.60G	31.08/0.9224	30.10/0.9245	35.02/0.9508	28.85/0.9132	31.23/0.9254	30.90/0.9352	35.09/0.9511
NAFSSR [7]	×2	1.51M	12.04G	31.19/0.9247	30.17/0.9267	35.46/0.9549	29.24/0.9177	31.33/0.9277	30.98/0.9367	35.51/0.9547
Steformer (ours)	×2	1.29M	7.43G	31.16/0.9236	30.27/0.9271	35.15/0.9512	28.97/0.9141	31.29/0.9263	31.07/0.9371	35.23/0.9511
Bicubic	×4	-	-	24.52/0.7310	23.79/0.7072	26.27/0.7553	21.82/0.6293	24.58/0.7372	24.38/0.7340	26.40/0.7572
VDSR [17]	×4	0.66M	10.77G	25.54/0.7662	24.68/0.7456	27.60/0.7933	22.46/0.6718	25.60/0.7722	25.32/0.7703	27.69/0.7941
EDSR [22]	×4	38.90M	214.86G	26.26/0.7954	25.38/0.7811	29.15/0.8383	23.46/0.7285	26.35/0.8015	26.04/0.8039	29.23/0.8397
RDN [51]	×4	22.04M	119.16G	26.23/0.7952	25.37/0.7813	29.15/0.8387	23.47/0.7295	26.32/0.8014	26.04/0.8043	29.27/0.8404
RCAN [50]	×4	15.36M	82.62G	26.36/0.7968	25.53/0.7836	29.20/0.8381	23.48/0.7286	26.44/0.8029	26.22/0.8068	29.30/0.8397
SwinIR [21]	×4	1.35M	7.89G	26.43/0.7996	25.60/0.7868	29.16/0.8379	23.53/0.7322	26.52/0.8058	26.29/0.8098	29.25/0.8385
StereoSR [15]	×4	1.08M	90.11G	24.49/0.7502	23.68/0.7273	27.70/0.8036	21.70/0.6460	24.53/0.7556	24.21/0.7511	27.64/0.8022
PASSRnet [40]	×4	1.42M	6.72G	26.26/0.7919	25.41/0.7772	28.61/0.8232	23.31/0.7195	26.34/0.7981	26.08/0.8002	28.72/0.8236
SRRes+SAM [47]	×4	1.73M	12.92G	26.35/0.7957	25.55/0.7825	28.76/0.8287	23.27/0.7233	26.44/0.8018	26.22/0.8054	28.83/0.8290
BSSRnet [45]	×4	1.91M	11.25G	26.45/0.8014	25.57/0.7872	29.12/0.8354	23.40/0.7289	26.47/0.8049	26.17/0.8075	29.08/0.8362
CPASSR [4]	×4	5.26M	7.73G	25.38/0.7753	25.05/0.7707	28.47/0.8245	23.12/0.7161	25.50/0.7818	25.63/0.7926	28.55/0.8251
iPASSR [42]	×4	1.43M	7.82G	26.47/0.7993	25.61/0.7850	29.07/0.8363	23.44/0.7287	26.56/0.8053	26.32/0.8084	29.16/0.8367
SSRDE-FNet [8]	×4	2.26M	70.21G	26.60/0.8031	25.73/0.7901	29.27/0.8416	23.55/0.7346	26.69/0.8091	26.46/0.8133	29.34/0.8411
NAFSSR [7]	×4	1.53M	12.26G	26.62/0.8051	25.78/0.7927	29.27/0.8447	23.63/0.7397	26.72/0.8113	26.49/0.8155	29.36/0.8447
Steformer (ours)	×4	1.34M	7.81G	26.61/0.8037	25.74/0.7906	29.29/0.8424	23.58/0.7376	26.70/0.8098	26.45/0.8134	29.38/0.8425

our cross attentive feature extraction module and cross-to-intra information integration module can explore more accurate stereo correspondence, and high quality image reconstruction module can recover more details and alleviate the blurring artifacts in our super-resolved images.

3) Performance on real-captured images: To test the performance of Steformer on real-world scenarios, we conducted experiments by directly applying image SR methods to real-captured images. As shown in Fig. 6, Steformer produces visually pleasing images with clear and sharp edges, whereas other compared methods may suffer from unsatisfactory artifacts. Single image SR methods cannot well recover the missing details by using intra-view information only, while iPASSR lacks the ability of capturing long-range dependencies. In contrast, our Steformer benefits from the ability of capturing cross-view and long-range information, thus producing images with less blurring artifacts. Note that we have difficulty in obtaining the results of SSRDE-FNet [8] since it needs a very large amount of graphics memory coming from the influence of the enormous structure in the network, e.g. processing a real-world stereo image pair with a size of 640K requires more than 40GB graphics memory, which also proves that it is difficult to apply in industry.

4) Benefits to disparity estimation: Stereo image SR task is tightly associated with parallax estimation since the accurate stereo correspondence effectively boosts the SR performance. Therefore, we investigated the ability to find stereo correspondence by performing disparity estimation on super-resolved stereo image pair. We performed downsampling to obtain low-resolution images on the $KITTI2012$ test set. Then, we used the state-of-art stereo SR methods to obtain super-resolved images, and performed parallax estimation [13] on the obtained SR images. We calculated the end-point-error (EPE) and t-pixel error (>t pixel) rate to compare the disparity performance. Lower EPE and t-pixel error mean better performance. As shown in Table IV, the standard Steformer results are slightly inferior to SSRDE-FNet [8]. That is because SSFDE-FNet [8] was specifically designed for parallax estimation and has twice the number of parameters as standard

TABLE IV: Quantitative Comparison Results Achieved by GwcNet [13] on ×4 SR Stereo Images. All These Metrics were Averaged on the Test Set of the $KITTI2012$ Dataset [12].

Method	EPE ↓	>1$px(\%)$ ↓	>2$px(\%)$ ↓	>3$px(\%)$ ↓
RDN [51]	0.4832	8.02	2.52	1.37
RCAN [50]	0.4820	7.90	2.50	1.40
SwinIR [21]	0.4763	7.73	2.44	1.35
iPASSR [42]	0.4673	7.36	2.36	1.31
SSRDE-FNet [8]	0.4556	7.09	2.26	1.24
Steformer (ours)	0.4584	7.17	2.29	1.28
Steformer_L (ours)	0.4553	7.02	2.25	1.25
HR	0.3460	4.55	1.42	0.78

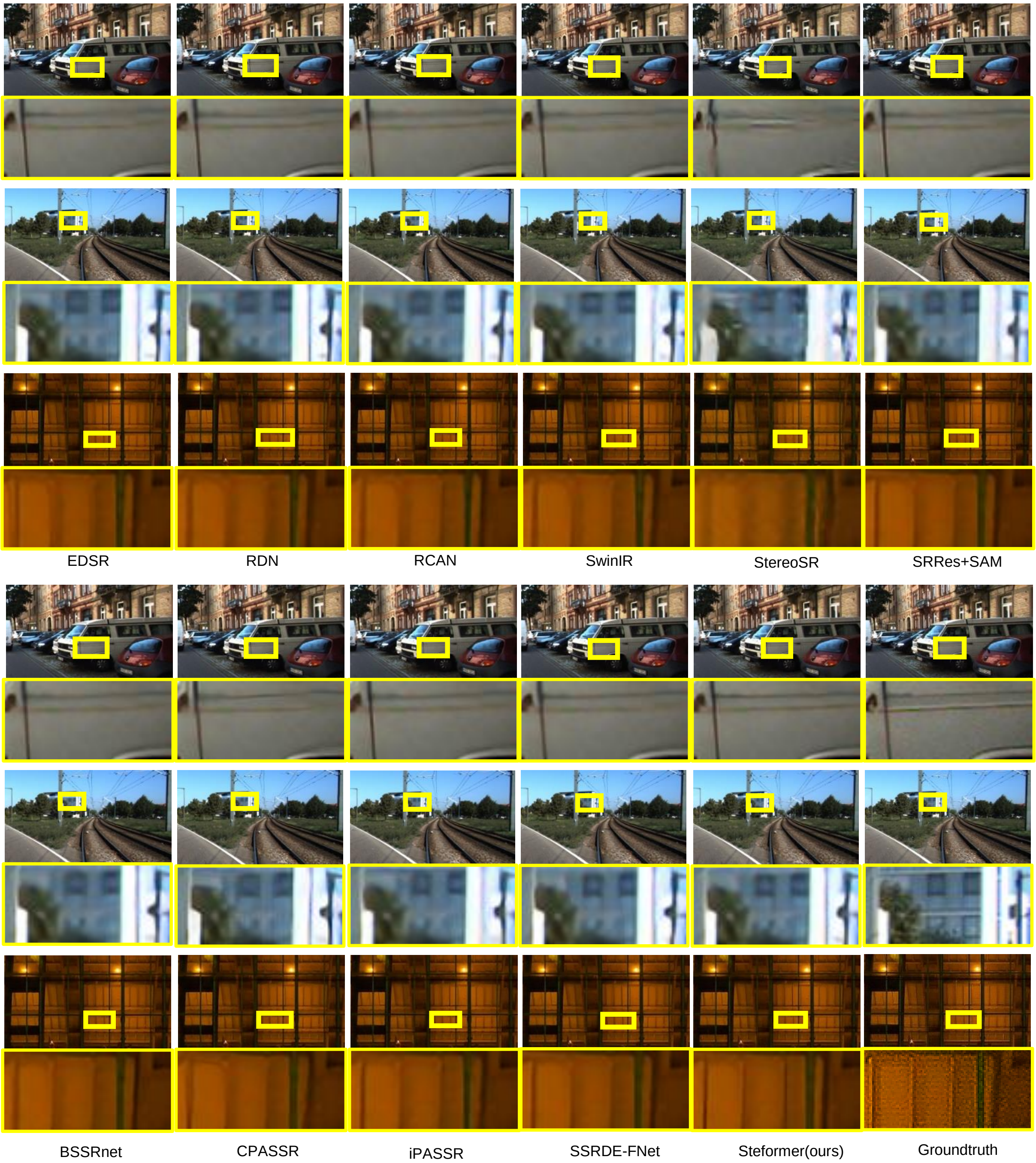


Fig. 5: Visual comparisons for ×4 SR by different methods on the $KITTI2015$ and $Middlebury$ datasets. The yellow rectangle marks zoom-in region.

Steformer. For additional comparison, we also train a large-scale network, namely Steformer_L, in which the number of RCSBs and RSBs in each module were set to 5 and 3. Our large Steformer with a comparable amount of parameters (i.e. 2.25M) has a superior performance. The visual comparison in Fig. 7 demonstrates that Steformer has more accurate disparity estimation performance and closer results to groundtruth.

5) Steformer's ability to obtain long-range dependencies : To demonstrate that Transformer-based Steformer can extract long-range dependencies across two-view images, we visualize the attention map $A_{R\rightarrow L}$ (Fig. 3 (c)). As shown in Fig. 8, a higher score indicates a stronger correlation. The visual attention map where distant locations are attended to demonstrates that the cross-attentive feature extraction module can extract long-range information.

D. Ablation study

In this section, we investigate the effectiveness and necessity of our method in terms of RCSB, RSB, and C2IAM. All the ablation experiments were conducted on the ×4 stereo image SR task on the $KITTI2012$ [12] dataset. The results are shown in Table V.

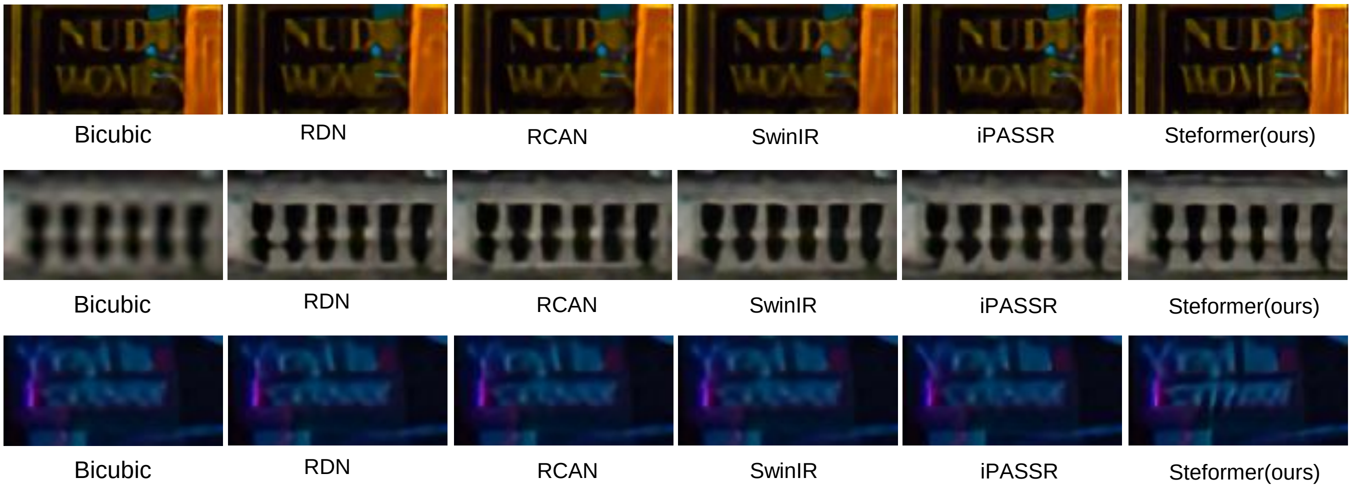


Fig. 6: Visual comparisons for ×4 SR by different methods on real-captured images.

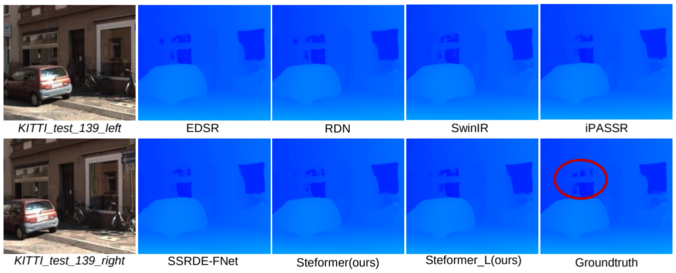


Fig. 7: Visual disparity comparisons for ×4 SR by different methods.

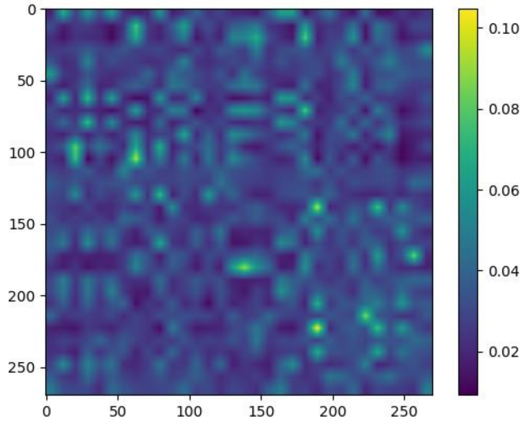


Fig. 8: The visual attention map about $A_{R\to L}$, a higher score indicates a stronger correlation.

TABLE V: Ablation Studies on $KITTI2012$ [12] for ×4 SR.

Method	$\#Params.$	$PSNR$	$SSIM$
Steformer	1.34M	26.70	0.8098
w/o RCSBs in feature extraction	1.10M	26.63	0.8080
w/o RCSBs in reconstruction	1.10M	26.54	0.8044
w/o MDIA in RCSBs	1.34M	26.61	0.8075
w/o RSBs in reconstruction	1.23M	26.61	0.8068
w/o C2IAM	1.22M	26.52	0.8037

RCSBs are used in our Sterformer for both feature extraction and image reconstruction. Firstly, we replaced RCSBs with the same number of Resblocks in feature extraction and image reconstruction modules respectively. Using Resblocks as the network backbone suffers a decrease of 0.07 dB and 0.16 dB in average PSNR respectively compared to using RCSBs. Then, to further exploit the effectiveness of our proposed MDIA in the cross Steformer layer, we replaced the stereo image pair (L, R) with the monocular image pair (L, L) or (R, R). With the equal quantity of parameters, there is a 0.09 dB PSNR reduction. We argue that the cross attention mechanism can treat different channels discriminatingly with learnable weight, which can improve the feature extraction and characterization capabilities of our method. Next, to exploit the effectiveness of RSBs in the reconstruction module, we replaced the RSBs with the same number of Resblocks. Results in Table V demonstrate that Transformer-based RSBs can effectively fuse single-view features to improve global performance. Finally, we replaced the C2IAM with biPAM proposed in [42] and ensured the parameters were similar to our Steformer. When adopting C2IAM, the PSNR value is improved from 26.52 dB to 26.70 dB. A similar phenomenon also appears on the SSIM metric. The ablation experiments demonstrate that our proposed modules are effective on the stereo image SR task.

V. Discussion

Since super-resolution reconstruction is an ill-posed inverse process, the super-resolution methods have difficulty in reconstructing the details of the image and are prone to the problem of lack of hierarchy in the reconstructed image as well as artifacts in the recovery of detailed textures. Some of the earlier stereo image SR methods even have worse performance than SISR methods, as the rich local feature layer information within the original low-resolution image is not fully exploited. This has prompted this study to focus on making full use of the intra-image information, like our proposed RSB. The lack of hierarchical information in the image is mainly reflected in the confused depth information of the reconstructed SR image, which also generates the visual discomfort of the result. Parallax is closely related to depth, as objects with a larger depth of field will have larger parallax, and depth can be estimated based on the disparity fields [38]. Therefore, accurate parallax estimation by the stereo image SR method can alleviate the lack of hierarchy in the reconstructed image content. As shown in Table IV and Figure 7, our Steformer can better estimate the disparity between the left and right views than other stereo image SR methods, which demonstrates that our model can produce more visually comfortable SR results.

There is room for improvements in present research. First, since Steformer is trained on x2 or x4 SR tasks respectively, the model is only suitable for one specific resolution. It would be interesting to extend the current model to handle arbitrary resolutions. Second, Better disparity estimation algorithms should be studied since the correlation of the corresponding positions of the left and right view images and the fusion of information are vital to stereo image SR. Finally, the domain gap across different datasets has hindered the generalizability of existing methods. Due to the significant differences in scene types and styles contained in different stereo image datasets, existing stereo image SR methods only achieve superior performance on a few datasets. Future research can focus on exploring the generalization performance of super-resolution algorithms on different types of stereo datasets and solving the domain gap problem.

VI. Conclusion

In this paper, we proposed a Transformer based stereo image super-resolution model, Steformer, which is computationally efficient to handle high-resolution images and conquers the limitation of CNNs in capturing long-range dependencies. We introduced key designs to the core components of Steformer for improved feature aggregation and transformation. Specifically, the cross attentive feature extraction module employs residual cross Steformer blocks (RCSB) with multi-Dconv interactive attention (MDIA) layers for long-range cross-view information extraction, alleviating the problem of significant disparity variation. Then, cross-to-intra information integration module utilizes parallax-attention and interactive-attention to capture cross-view and intra-view information respectively, providing comprehensive features for high-quality image reconstruction. Finally, the high quality image reconstruction module utilizes residual Steformer blocks (RSB) for feature pre-processing, and RCSBs for cross-view information incorporation, thus reconstructing high-quality stereo images jointly.

VII. acknowledgement

The authors thank the anonymous reviewers for their valuable comments. This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 62206089, 62202158), and the Fundamental Research Funds for the Central Universities (Grants HNU: 531118010668, 531118010786).

References

[1] Saeed Anwar and Nick Barnes. “Densely residual laplacian super-resolution”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[2] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[3] Nicolas Carion et al. “End-to-end object detection with transformers”. In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[4] Canqiang Chen et al. “Cross Parallax Attention Network for Stereo Image Super-Resolution”. In: *IEEE Transactions on Multimedia* (2021).

[5] Hanting Chen et al. “Pre-trained image processing transformer”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12299–12310.

[6] Liangyu Chen et al. “Simple baselines for image restoration”. In: *arXiv preprint arXiv:2204.04676* (2022).

[7] 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*. 2022, pp. 1239–1248.

[8] Qinyan Dai et al. “Feedback Network for Mutually Boosted Stereo Image Super-Resolution and Disparity Estimation”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 1985–1993.

[9] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[10] Chao Dong et al. “Learning a deep convolutional network for image super-resolution”. In: *European conference on computer vision*. Springer. 2014, pp. 184–199.

[11] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3354–3361.

[13] Xiaoyang Guo et al. “Group-wise correlation stereo network”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3273–3282.

[14] Jithin Saji Isaac and Ramesh Kulkarni. “Super resolution techniques for medical image processing”. In: *2015 International Conference on Technologies for Sustainable Development (ICTSD)*. IEEE. 2015, pp. 1–6.

[15] Daniel S Jeon et al. “Enhancing the spatial resolution of stereo images using a parallax prior”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1721–1730.

[16] Robert Keys. “Cubic convolution interpolation for digital image processing”. In: *IEEE transactions on*

acoustics, speech, and signal processing 29.6 (1981), pp. 1153–1160.

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. “Accurate image super-resolution using very deep convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1646–1654.

[18] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. “Deeply-recursive convolutional network for image super-resolution”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1637–1645.

[19] Jun-Hyuk Kim et al. “Ram: Residual attention module for single image super-resolution”. In: *arXiv preprint arXiv:1811.12043* 2.1 (2018), p. 2.

[20] Hao Li et al. “Real-World Image Super-Resolution by Exclusionary Dual-Learning”. In: *IEEE Transactions on Multimedia* (2022).

[21] Jingyun Liang et al. “Swinir: Image restoration using swin transformer”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1833–1844.

[22] Bee Lim et al. “Enhanced deep residual networks for single image super-resolution”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 136–144.

[23] Jie Liu et al. “Residual feature aggregation network for image super-resolution”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2359–2368.

[24] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[25] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[26] Chenxi Ma et al. “Perception-Oriented Stereo Image Super-Resolution”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 2420–2428.

[27] Moritz Menze and Andreas Geiger. “Object scene flow for autonomous vehicles”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3061–3070.

[28] Karam Park, Jae Woong Soh, and Nam Ik Cho. “Dynamic Residual Self-Attention Network for Lightweight Single Image Super-Resolution”. In: *IEEE Transactions on Multimedia* (2021).

[29] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).

[30] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[31] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[32] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *arXiv preprint arXiv:1910.10683* (2019).

[33] Daniel Scharstein et al. “High-resolution stereo datasets with subpixel-accurate ground truth”. In: *German conference on pattern recognition*. Springer. 2014, pp. 31–42.

[34] Jacob Shermeyer and Adam Van Etten. “The effects of super-resolution on object detection performance in satellite imagery”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 0–0.

[35] Wenzhe Shi et al. “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 1874–1883.

[36] Wonil Song et al. “Stereoscopic image super-resolution with stereo consistent feature”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12031–12038.

[37] Robin Strudel et al. “Segmenter: Transformer for semantic segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 7262–7272.

[38] Dimitrios Tzovaras, Nikos Grammalidis, and Michael G Strintzis. “Disparity field and depth map coding for multiview 3D image generation”. In: *Signal Processing: Image Communication* 11.3 (1998), pp. 205–230.

[39] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[40] Longguang Wang et al. “Learning parallax attention for stereo image super-resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12250–12259.

[41] Longguang Wang 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*. 2022, pp. 906–919.

[42] Yingqian Wang et al. “Symmetric parallax attention for stereo image super-resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 766–775.

[43] Zhendong Wang et al. “Uformer: A general u-shaped transformer for image restoration”. In: *arXiv preprint arXiv:2106.03106* (2021).

[44] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in Neural Information Processing Systems* 34 (2021).

[45] Qingyu Xu et al. “Deep bilateral learning for stereo image super-resolution”. In: *IEEE Signal Processing Letters* 28 (2021), pp. 613–617.

[46] Fuzhi Yang et al. “Learning texture transformer network for image super-resolution”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 5791–5800.

[47] Xinyi Ying et al. "A stereo attention module for stereo image super-resolution". In: *IEEE Signal Processing Letters* 27 (2020), pp. 496–500.

[48] Li Yuan et al. "Tokens-to-token vit: Training vision transformers from scratch on imagenet". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 558–567.

[49] Syed Waqas Zamir et al. "Restormer: Efficient Transformer for High-Resolution Image Restoration". In: *arXiv preprint arXiv:2111.09881* (2021).

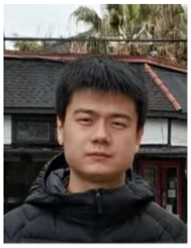
[50] Yulun Zhang et al. "Image super-resolution using very deep residual channel attention networks". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 286–301.

[51] Yulun Zhang et al. "Residual dense network for image super-resolution". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2472–2481.

[52] Sixiao Zheng et al. "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 6881–6890.

[53] Xiangyuan Zhu et al. "Cross view capture for stereo image super-resolution". In: *IEEE Transactions on Multimedia* (2021).

[54] Wenbin Zou et al. "Joint Wavelet Sub-bands Guided Network for Single Image Super-Resolution". In: *IEEE Transactions on Multimedia* (2022).



Jianxin Lin received the B.E. and Ph.D. degrees from University of Science and Technology of China (USTC) in 2015 and 2020. He is currently an associate professor at the School of Computer Science and Electronic Engineering, Hunan University, Changsha, China. He has published over 20 papers on related conferences and journals. He has received top paper award at the ACM Multimedia 2022. His research interests include image and video processing, synthesis and understanding.

Lianying Yin received the B.E. degree from the Yanbian University, China, in 2021. She is currently working toward the M.E. degree with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. Her research interests include computer vision, deep learning, and multimedia analysis.

Yijun Wang received the B.E. and Ph.D. degrees from University of Science and Technology of China (USTC) in 2014 and 2019. She is currently an assistant professor at the School of Computer Science and Electronic Engineering, Hunan University, Changsha, China. She has published over 10 papers on related conferences and journals. Her research interests include multimedia understanding, natural language processing and data mining.