Multi-Stage Edge-Guided Stereo Feature Interaction Network for Stereoscopic Image Super-Resolution

Jin Wan, Hui Yin, Zhihao Liu, Yanting Liu, and Song Wang, *Senior Member, IEEE*

***Abstract*—Stereo image super-resolution (SR) aims to simultaneously increase the resolution of stereo image pairs, which benefits many downstream three-dimensional (3D) multimedia broadcasting and stereo vision-related tasks, such as 3D television broadcasting and stereo matching. A key insight in convolutional neural networks-based stereo image SR is to enforce stereo feature interactions between the two stereo views to explore complementary cross-view features that can facilitate the SR in both views. To fully exploit the cross-view stereo features, in this paper we propose a new multi-stage network, cascaded by several stereo feature interactions, progressively improving the SR quality from coarse to fine. In particular, an edge-guided stereo attention mechanism is proposed to be embedded into each stereo feature interaction to better capture consistent structure details of the cross-views. Followed by stereo feature fusion and reconstruction modules, we finally put together a multi-stage edge-guided stereo feature interaction network (MESFINet) for stereo image SR. Comprehensive experiments on KITTI2012, KITTI2015, Middlebury, and Flickr1024 benchmark datasets show that the proposed MESFINet achieves superior performance against the state-of-the-art stereo image SR methods and can be used to improve the accuracy of stereo matching.**

***Index Terms*—Super-resolution, stereo image, multi-stage network, edge guidance.**

I. Introduction

Stereo image pairs contain rich three-dimensional (3D) geometric information of the real-world scene and play an important role in many 3D display applications, such as stereoscopic 3D broadcasting [1], 3D reconstruction [2], [3], virtual reality [4], *etc.* In practice, many of these applications [2], [3], [4], [5], [6] desire stereo images to be of higher resolution – besides better visual satisfaction, higher-resolution images can also improve the performance of stereo matching, resulting in more accurate disparity maps and depth estimation. However, the resolution of stereo images is usually limited by the adopted stereo-camera hardware and the media-broadcasting bandwidth from the cameras to the data-processing center [7], [8], [9], [10]. One way to solve this problem is to develop effective stereo image super-resolution (SR) algorithms that can recover a high-resolution (HR) stereo image pair from an input low-resolution (LR) stereo image pair.

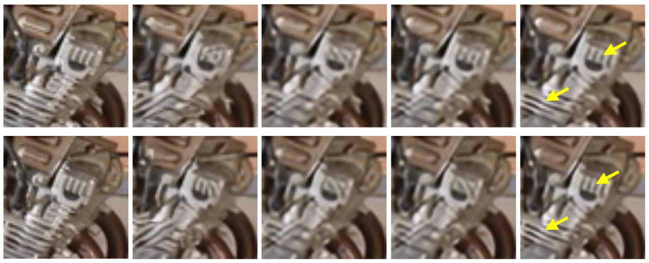

Fig. 1. 2× super-resolution of "motorcycle" in the Middlebury [11] dataset by using StereoSR [12], PASSR [13], iPASSR [14], and our proposed method. Top: left view; Bottom: right view.

Owing to the advancement of deep learning, significant progress has been made on single image SR [15], [16], [17], [18] in recent years. A simple approach for stereo image SR is to perform single image SR to the left-view and right-view images separately. Nevertheless, without considering the correlation and correspondence between the cross views, those approaches do not exploit the full potential of the stereo image pair for maximizing the SR performance. To address this issue, recently several deep learning-based methods [13], [14], [19], [20], [21] have been developed exclusively for stereo image SR by exploring the cross-view correspondence. A key step in these methods is the stereo feature interaction between the two views to transfer the information from one view to the other. Most of these methods only perform stereo feature interaction once in the network and we argue that it may not be sufficient to fully exploit the shared and complementary information between the two views. This motivates us to develop a new multi-stage network and perform stereo feature interaction at each stage, followed by a fusion over all the stages, for high-resolution stereo image reconstruction.

As shown in Fig. 1 (a-c), existing stereo SR methods do not pay enough attention to structural cues (*e.g.,* edges and textures) and generate disastrous structural details (pointed by arrows), which also have a direct bearing on the performance

Manuscript received 10 September 2022; revised 23 December 2022; accepted 19 March 2023. Date of publication 8 May 2023; date of current version 7 June 2023. This work was supported in part by the National Nature Science Foundation of China under Grant 51827813, Grant 61472029, and Grant U1803264; in part by the Research and Development Program of the Beijing Municipal Education Commission under Grant KJZD20191000402; and in part by the Fundamental Research Funds for the Central Universities under Grant 2020YJS031. *(Corresponding authors: Hui Yin; Song Wang.)*

Jin Wan, Hui Yin, and Yanting Liu are with the Beijing Key Laboratory of Traffic Data Analysis and Mining and the Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: jinwan@bjtu.edu.cn; hyin@bjtu.edu.cn; 19112024@bjtu.edu.cn).

Zhihao Liu was with the Beijing Key Laboratory of Traffic Data Analysis and Mining and the Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University, Beijing 100044, China. He is now with the China Mobile Research Institute, Beijing 100053, China (e-mail: liuzhihao@chinamobile.com).

Song Wang is with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29201 USA (e-mail: songwang@cec.sc.edu).

Digital Object Identifier 10.1109/TBC.2023.3264880

of image SR. Furthermore, these structural details of stereo images are essential for related stereo vision tasks (*e.g.,* stereo matching [6]) by providing accurate locations of discriminative image features. Based on this observation, we propose to incorporate the edge priors to guide the stereo feature interactions that are performed at multiple stages. By this way, we can recover high-resolution stereo image pairs with consistent structure details between the left and right views. As shown by the sample results in Fig. 1 (d), the proposed method by combining edge guidance and multi-stage feature interaction can better capture consistent stereo SR image details between the cross views.

More specifically, in this paper, we propose a novel multi-stage edge-guided stereo feature interaction network (MESFINet) for stereo image SR. It consists of multi-stage stereo feature interaction modules, which are designed by using an edge-guided stereo attention mechanism. In this mechanism, we employ an edge-adaptive spatial feature transform to stress the structural details in the stereo feature transformation by modulating the cross-view stereo features with edge priors. Finally, we utilize a stereo feature fusion module to adaptively fuse the interactive stereo features from multi-stage feature interaction modules for high-quality image reconstruction. For experiments, we evaluate the proposed method by conducting a series of ablation studies, as well as comparisons against many existing state-of-the-art methods on KITTI2012 [22], KITTI2015 [23], Middlebury [11], and Flickr1024 [24] benchmark datasets. We also evaluate the proposed method in the task of stereo matching. In summary, the main contributions of this work include:

- We propose a novel multi-stage edge-guided stereo feature interaction network (MESFINet) with several stereo feature interaction modules to fully exploit complementary cross-view information for efficient and high-quality stereo image SR.
- An edge-guided stereo attention mechanism is proposed to stress the correspondence of structural cues between the cross views to reconstruct more local image details in SR stereo images.
- Extensive experiments show that the proposed MESFINet achieves superior performance to existing approaches on various public datasets, in terms of both stereo image SR and stereo matching.

II. Related Work

In this section, we briefly review the related works on single image SR, stereo image SR, and edge guidance.

A. Single Image Super-Resolution

Recent state-of-the-art performances of single image SR are achieved by various convolutional neural networks (CNNs)-based methods that learn the complex mapping relationship between low-resolution (LR) images and high-resolution (HR) images with large-scale training datasets. In [26], a three-layer full convolutional network was used to learn the mapping between pre-amplified LR images and HR images, achieving better performance than traditional SR methods [27], [28], [29]. In [30], [31], [32], skip connection and parameter sharing operations were used in the network to further expand the receptive field for further improving the SR performance. In [33], a sub-pixel convolution method was proposed for HR image reconstruction, by verifying the effectiveness of extracting features from LR images. More recently, inspired by Resnet [34] and Densenet [35], a large number of image SR methods [17], [18], [36], [37], [38], [39], [40], [41], [42], [43] have been proposed by building deeper networks and exploiting richer features from the initial LR image. For example, in [37], a deeper and wider network was built by removing unnecessary batch normalization (BN) and activation functions of SSResnet [15]. In [9], [16], [43], [44], [45], [46], [47], [48], with an attention mechanism, the correlation between spatial and channel features was considered for improving the single image SR performance. All these works only consider a single image for SR, which differs from our proposed work on stereo image SR. As mentioned earlier, while applying a single image SR algorithm to two stereo images separately, the SR performance is limited without considering the correlation between the two views.

B. Stereo Image Super-Resolution

The core of stereo image SR lies in the usage of the spatial complementary information between the left and right views. Recently, various CNNs-based methods have been developed for stereo image SR with remarkable success. In [12], the disparity prior was used to improve the spatial resolution of stereo images, which, however, is limited by a fixed maximum disparity (64 in the original paper). In [13], [49], a parallax attention module (PAM) was proposed to learn the stereo consistency of the global receptive field along the epipolar line. In [20], a self and parallax attention mechanism (SPAM) was proposed to integrate the intra-view image and its corresponding stereo image information, and a training loss function was designed to strengthen the stereo consistency constraint. In [19], it directly inserted two stereo attention modules (SAM) into the pre-trained SRResnet [15] and performs stereo feature transformation at multiple stages. In [50], a domain-adaptive stereo image SR network DASSR was proposed to integrate pre-explicitly predicted disparity into the entire pipeline by using feature modulation dense blocks. In [8], several interactive modules were used in the network to utilize the cross-view information. In [14], symmetric bi-directional PAM (biPAM) and inline occlusion processing schemes were proposed for the middle stage of the network to further improve the stereo image SR performance. More recently, in [51], cross-view space features were extracted in a global and local manner. In [52], the disparity estimation task was combined in stereo image SR to boost reconstruction performance.

Different from the aforementioned methods, we propose to embed the edge priors in stereo feature interaction to highlight the structural details of stereo feature transformation and conduct multi-stage stereo feature interaction to achieve high-quality stereo image SR.

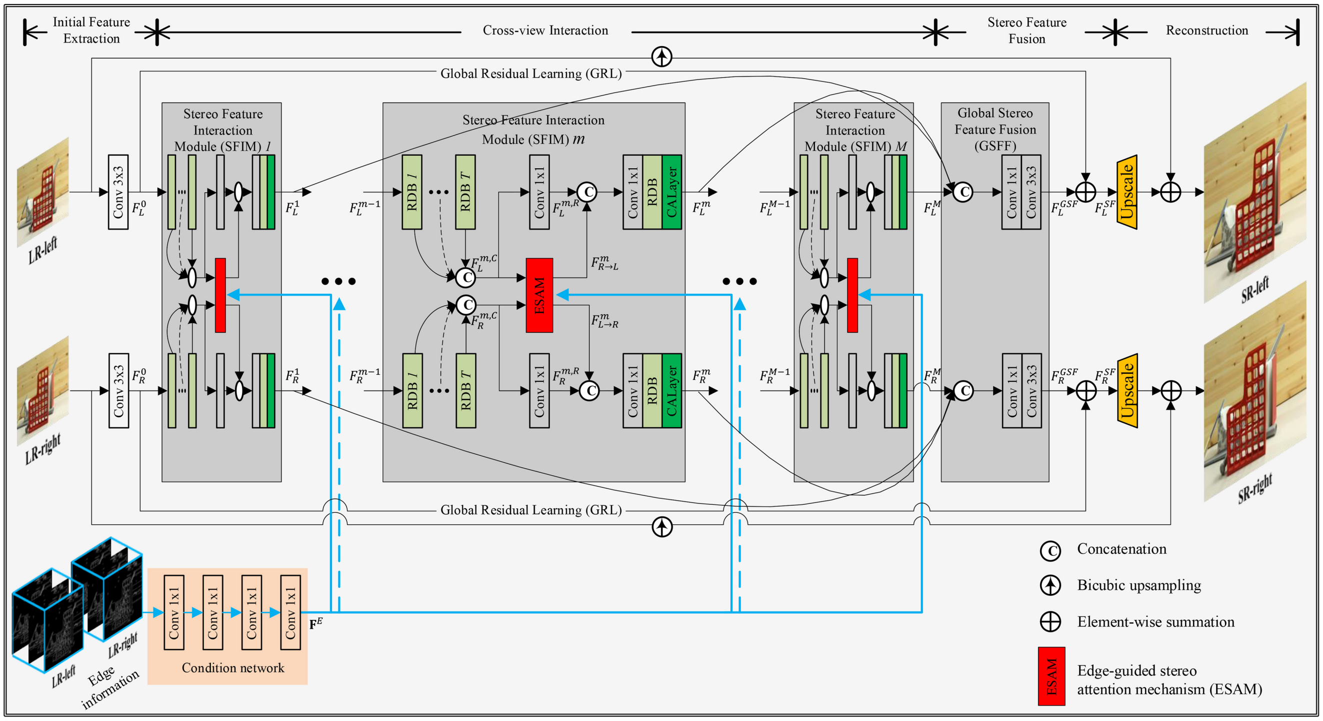

Fig. 2. An overview of the proposed multi-stage edge-guided stereo feature interaction network (MESFINet).

C. Edge Guidance

Edge information has been proven to be useful for many computer-vision tasks. In [53], edge information produced by an edge sub-network was integrated into the stereo-matching network to recover missing details of disparity. In [54], two tasks of edge detection and salient object detection were jointly learned to achieve better image-segmentation results. Most recently, in [55], edge priors were used to assist object structure recovery to improve the performance of camouflaged object detection. In [56], the available edge information was exploited to fill in missing pixels for generating high-resolution images. Our approach differs from these works in that the proposed edge-guided stereo attention mechanism leverages the edge priors to guide multi-stage feature interactions between cross-views in the stereo image SR domain.

III. Proposed Method

In this section, we first briefly introduce the overall network architecture. Then we explain the details of the proposed stereo feature interaction module and the edge-guided stereo attention mechanism. Finally, we describe the proposed stereo feature fusion module in detail.

A. Network Structure

As shown in Fig. 2, the proposed multi-stage edge-guided stereo feature interaction network (MESFINet) estimates the SR stereo images $\mathbf{I}^{SR} = \{I_L^{SR}, I_R^{SR}\}$ from the LR stereo images $\mathbf{I}^{LR} = \{I_L^{LR}, I_R^{LR}\}$. Correspondingly, $\mathbf{I}^{HR} = \{I_L^{HR}, I_R^{HR}\}$ denotes the underlying HR stereo images. In detail, MESFINet consists of four steps: 1) *initial feature extraction*, 2) *cross-view interaction*, 3) *stereo feature fusion*, and 4) *reconstruction*. Note that, the two branches for LR-left and LR-right are weight-sharing to extract the features inside the left and right views.

In Step 1), the edge priors of the left and right views, *i.e.*, multi-scale edge probability maps (in our experiments, scale = 5), are detected by sending $\mathbf{I}^{LR}$ to an edge detection network [25], as shown in Fig. 3. It can be observed that edges maintain high stereo consistency between the two views, which is a prerequisite for our work. A conditional subnetwork with four convolutional layers takes edge probability maps of two views as input to generate edge-guided features $\mathbf{F}^E = \{F_L^E, F_R^E\}$ that is shared by the cross-view interaction part. And we stint the receptive field of the conditional network by using 1×1 kernels for all convolutional layers to reduce interference from smooth regions to edge regions. Concurrently, a 3×3 convolutional layer is used to map the input stereo images to the high-dimensional initial stereo features $\mathbf{F}^0 = \{F_L^0, F_R^0\}$. Then, $\mathbf{F}^0$ goes through global residual learning to facilitate the feature learning and is sent as input to the cross-view interaction part.

In Step 2), we include a sequence of M stereo feature interaction modules (SFIMs) to explore complementary cross-view information. In this multi-stage structure, SFIM takes stereo features generated by the previous stage and edge features as input to perform bidirectional feature interactions between left and right views which will be elaborated in Section III-B, and the interactive stereo features produced by m-th SFIMs are denoted as $\mathbf{F}^m = \{F_L^m, F_R^m\}, m = 1, 2, \ldots, M$. We thus have

$$\begin{aligned}\mathbf{F}^m &= \mathcal{H}_{SFIM,m}\left(\mathbf{F}^{m-1}, \mathbf{F}^E\right)\\ &= \mathcal{H}_{SFIM,m}\left(\mathcal{H}_{SFIM,m-1}\left(\ldots\left(\mathcal{H}_{SFIM,1}\left(\mathbf{F}^0, \mathbf{F}^E\right)\right)\ldots\right)\right), \quad (1)\end{aligned}$$

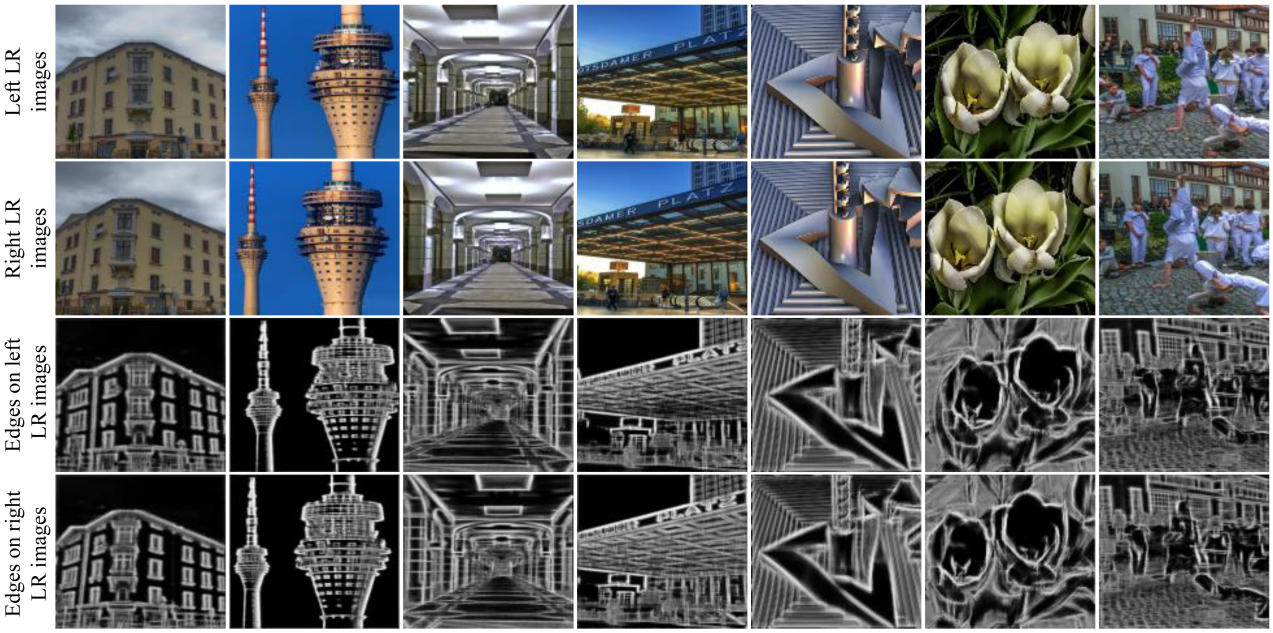

Fig. 3. Examples of edge probability maps produced by BDCN [25]. First row: left LR images. Second row: right LR images. Third row: edge maps of the left LR images. Fourth row: edge maps of the right LR images.

where $\mathcal{H}_{SFIM,m}(\cdot)$ indicates the feature-interactive operation performed by the m-th SFIM.

In Step 3), stereo feature fusion (SFF) is exploited to make full use of hierarchical stereo features from all the preceding layers in a global manner. It is composed of global stereo feature fusion (GSFF) and global residual learning (GRL). Specifically, we can write SFF as

$$\mathbf{F}^{SF} = \mathcal{H}_{SFF}\left(\mathbf{F}^0, \mathbf{F}^1, \ldots, \mathbf{F}^M\right), \tag{2}$$

where $\mathcal{H}_{SFF}(\cdot)$ denotes a composite function of SFF. $\mathbf{F}^{SF} = \{F_L^{SF}, F_R^{SF}\}$ represents the output stereo feature maps of SFF. More details about SFF will be discussed in Section III-D.

In Step 4), the acquired global stereo features are sent to the reconstruction part to amplify the feature maps and reconstruct the features. Concretely, a sub-pixel convolutional layer of ESPCN [33] and a 3×3 convolutional layer are used to map LR stereo features to SR stereo images.

We adopt pixel-wise $L1$ loss in our work. Given a training set $\{\mathbf{I}_i^{HR} = \{I_{L,i}^{HR}, I_{R,i}^{HR}\}, \mathbf{I}_i^{LR} = \{I_{L,i}^{LR}, I_{R,i}^{LR}\}\}_{i=1,\ldots,N}$, where N is the number of training pairs, the loss function with the updated parameters $\boldsymbol{\Theta}$ is

$$\mathcal{L}^{SR}(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{H}_{MESFINet}\left(\mathbf{I}_i^{LR} \mid \Phi\right) - \mathbf{I}_i^{HR}\right\|_1 \tag{3}$$

where Φ represents the edge priors on which the condition can be applied. $\mathcal{H}_{MESFINet}(\cdot)$ indicates the entire function of the proposed MESFINet.

B. Stereo Feature Interaction Module

As shown in Fig. 2, the proposed stereo feature interaction module (SFIM) contains several residual dense blocks (RDBs) [17], an edge-guided stereo attention mechanism (ESAM), and a local feature fusion operation. Specifically, benefiting from RDB can produce abundant local features with a large receptive field, which is verified to be contributed to SR results [17]. In m-th SFIM, we utilize T RDBs to extract the deep features from the feature-maps $\mathbf{F}^{m-1}$ produced by the $(m-1)$-th SFIM. We then concatenate these features and send them to ESAM for the stereo feature transformation. Taking the m-th SFIM as an example, this process can be formulated as

$$F_{R\to L}^m, F_{L\to R}^m = \mathcal{H}_{ESAM}\left(\mathbf{F}^{m,C}, \mathbf{F}^E\right), \tag{4}$$

where $\mathcal{H}_{ESAM}(\cdot)$ denotes the ESAM, $F_{R\to L}^m$ and $F_{L\to R}^m$ represent the transformed stereo features of ESAM, and $\mathbf{F}^{m,C} = \{F_L^{m,C}, F_R^{m,C}\}$ denotes the concatenated features from all the RDBs. Details of ESAM will be described in Section III-C. For the intra-view feature reusability, $\mathbf{F}^{m,C} \in \mathbb{R}^{B\times TC\times H\times W}$ is fed into 1×1 convolutional layer to generate the reused features $\mathbf{F}^{m,R} = \{F_L^{m,R}, F_R^{m,R}\} \in \mathbb{R}^{B\times C\times H\times W}$, where B is the batch size, C is the number of channels, H and W are the height and width of the input image, respectively. For different views, the features from the intra-view and transformed features from another view are fed to a local feature fusion operation to aggregate cross-view information. Taking the left view as an instance, the concatenation of intra-view feature $F_L^{m,R}$ and transformed feature $F_{R\to L}^m$ is first sent to the 1 × 1 convolution layer to reduce the channel dimension and then fed into a residual dense block (RDB) and a channel attention layer (CALayer [16]) to obtain the local fusion features. This way, we further have

$$F_L^m = \mathcal{H}_{CA}\left(\mathcal{H}_{RDB}\left(\mathcal{H}_{Conv}\left(\left[F_L^{m,R}, F_{R\to L}^m\right]\right)\right)\right), \tag{5}$$

where F_L^m is the output of the left-view in m-th SFIM and $[\cdot,\cdot]$ denotes the concatenation operation.

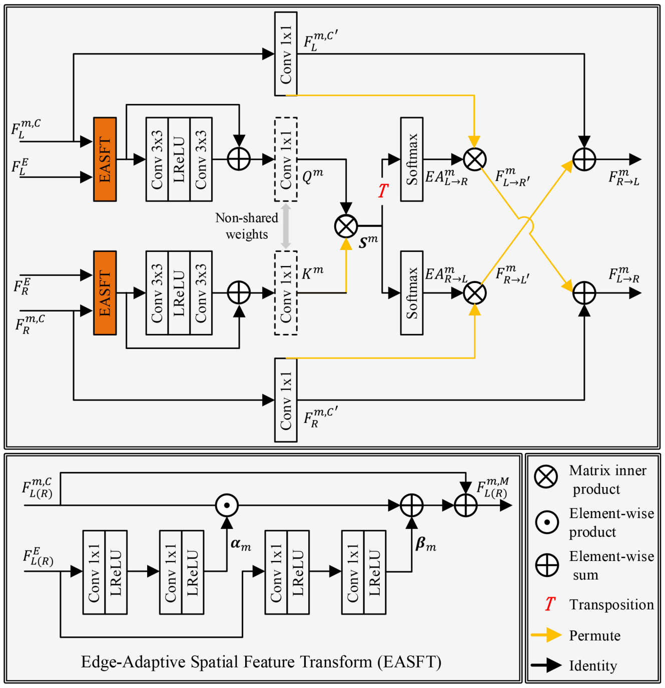

Fig. 4. Our proposed edge-guided stereo attention mechanism (ESAM).

C. Edge-Guided Stereo Attention Mechanism

To leverage the edge priors to stress the details in stereo feature transformation between the two views, we propose an edge-guided stereo attention mechanism (ESAM) embedded in SFIM to enhance the stereo consistency, as shown in Fig. 4.

Specifically, we first send the edge probability maps obtained by the pre-trained BDCN [25] to a conditional subnetwork to produce edge-guided features $\mathbf{F}^E = \{F_L^E, F_R^E\} \in \mathbb{R}^{B\times C\times H\times W}$. Given $\mathbf{F}^E$ and the input cross-view features $\mathbf{F}^{m,C} = \{F_L^{m,C}, F_R^{m,C}\} \in \mathbb{R}^{B\times TC\times H\times W}$, we then apply an edge-adaptive spatial feature transform (EASFT) block to learn new edge-aware representations of the cross-view features. As shown at the bottom of Fig. 4, taking the left view as an example, each EASFT block provides tailored parameters to modulate the left-view feature $F_L^{m,C}$ of multiple stages by sharing the same left-view edge-guided feature F_L^E. In detail, F_L^E is fed into two 1×1 convolutional layers to generate two modulation parameters $\boldsymbol{\alpha}_m$ and $\boldsymbol{\beta}_m$, which have the same feature dimensions as $F_L^{m,C}$. Inspired by [57], these two parameters are used as the scale factor and shift factor for modulating $F_L^{m,C}$. The operation can be written as

$$F_L^{m,M} = (\boldsymbol{\alpha}_m + \mathbf{I}) \odot F_L^{m,C} + \boldsymbol{\beta}_m, \tag{6}$$

where $\odot$ means the element-wise product and $F_L^{m,M}$ denotes the modulated feature of the left view and $\mathbf{I}$ is a tensor with the value of 1 and has the same dimension as $F_L^{m,C}$. The computation of the modulated right-view feature $F_R^{m,M}$ is the analogous process, just replace the input edge-guided feature and intra-view feature of EASFT with those in the right view.

Afterwards, $F_L^{m,M}$ and $F_R^{m,M}$ are first sent to a ResBlock [37], and go through two 1×1 convolutional layers with non-shared weights to generate a query feature map $Q^m \in \mathbb{R}^{B\times C\times H\times W}$ and key feature map $K^m \in \mathbb{R}^{B\times C\times H\times W}$. We use horizontal axial attention [58] to compute the feature similarity along the epipolar line to obtain a score map $S^m \in \mathbb{R}^{BH\times C\times W\times W}$ for stereo correspondence. S^m and $(S^m)^T$ are sent to a softmax normalization to produce bi-direction edge-guided attention $EA_{R\to L}^m$ and $EA_{L\to R}^m$.

To achieve stereo features transformation guided by the edge information, the left-view (*and* the right-view) features take batch-wise matrix multiplication with the corresponding edge-guided attention maps as

$$\begin{aligned} F_{L\to R'}^m &= EA_{L\to R}^m \otimes F_L^{m,C'} \\ F_{R\to L'}^m &= EA_{R\to L}^m \otimes F_R^{m,C'}, \end{aligned} \tag{7}$$

where $F_L^{m,C'}$ and $F_R^{m,C'}$ are generated by $F_L^{m,C}$ and $F_R^{m,C}$ via a 1×1 convolutional layer, $\otimes$ denotes matrix inner product, and $F_{L\to R'}^m$ and $F_{R\to L'}^m$ represent the transformed stereo features. Afterwards, local residual learning is then utilized to facilitate the information flow, which can be written as

$$\begin{aligned} F_{L\to R}^m &= F_{L\to R'}^m \oplus F_R^{m,C'} \\ F_{R\to L}^m &= F_{R\to L'}^m \oplus F_L^{m,C'}, \end{aligned} \tag{8}$$

where $F_{L\to R}^m$ and $F_{R\to L}^m$ represent the output of the ESAM in the m-th SFIM. $\oplus$ is executed by a shortcut connection and element-wise sum.

D. Stereo Feature Fusion Module

After extracting local stereo features with a set of SFIMs, we further utilize a stereo feature fusion (SFF) module to combine hierarchical stereo features in a global way. As shown in Fig. 2, our proposed SFF is composed of global stereo feature fusion (GSFF) and global residual learning (GRL).

Specifically, GSFF is used to fuse the multi-stage interactive stereo features from all the SFIMs to get the global stereo features as

$$\mathbf{F}^{GSF} = \mathcal{H}_{GSFF}\left(\left[\mathbf{F}^1, \ldots, \mathbf{F}^M\right]\right), \tag{9}$$

where $\mathcal{H}_{GSFF}(\cdot)$ indicates the combination of a 1×1 and a 3×3 convolutional layers. To further enhance the reusability of initial features, GRL is used to process the features before image up-sampling and this strategy has been shown to be effective in [16], [17]. This way, we have the following formula:

$$\mathbf{F}^{SF} = \mathbf{F}^{GSF} \oplus \mathbf{F}^0 \tag{10}$$

where $\mathbf{F}^0$ represents the initial feature maps. The final fused stereo features $\mathbf{F}^{SF}$ are used for reconstructing the SR stereo images.

IV. Experiment

In this section, we first describe the datasets and network training settings. We then verify the effects of different components in the proposed network, and conduct quantitative evaluation and qualitative comparison of the proposed method with several state-of-the-art image SR methods on benchmark datasets. Finally, we apply our method to facilitate stereo matching.

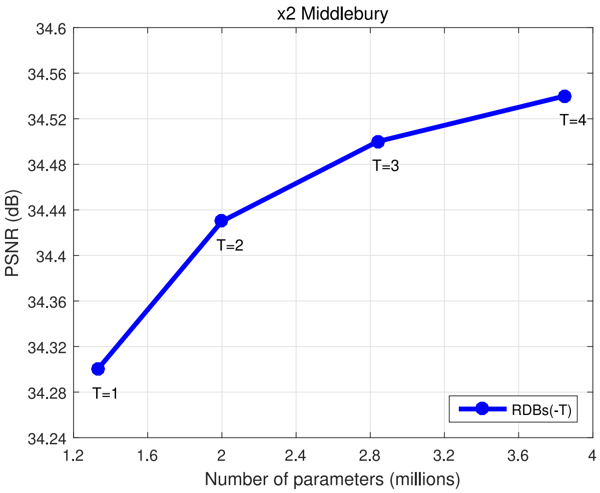

Fig. 5. PSNR performance when using different T, i.e., the number of RDBs, in each SFIM. Here we set M, the number of SFIMs/stages, to be 4 and the results are on Middlebury [11] with ×2 scale.

A. Experimental Settings

Datasets and metrics: In our work, 800 images from the Flickr1024 [24] dataset and 60 images from the Middlebury [11] dataset are used as training data to train the proposed model. For evaluation, we use four widely used datasets: KITTI12 [22], KITTI2015 [23], Middlebury [11], Flickr1024 [24]. In order to make a fair comparison, we use peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) with the same settings as [13], [14], to evaluate our results, and scaling factors of ×2 and ×4 are adopted for training and testing.

Training settings: In each training batch, we use 36 pairs of LR stereo image patches with a size of 30×90 as input, and these patches are generated by bicubic downsampling the corresponding stereo HR images cropped in steps of 20. For data augmentation, we perform random flipping and rotation on all the patches before training. In the optimization process, our models are optimized by Adam [59] with setting $\beta_1 = 0.9$, $\beta_2 = 0.99$. The basic learning rate is set to 2×10^{-4} and halved at every 30 epochs for 80 epochs in total. We implement our models using PyTorch, run them using two NVIDIA GeForce GTX 2080 Ti GPU cards, and evaluate them using MATLAB R2015b. Following the setting in [37], we use a pre-training strategy – when training our ×4 model, we initialize the model parameters with the pre-trained ×2 model – which can accelerate the convergence of the network.

B. Ablation Study

In this subsection, we analyze the performance of the multi-stage stereo feature interactions, the edge guidance, the global stereo feature fusion, the modulation scheme, and the edge probability map.

1) Multi-Stage Stereo Feature Interaction Modules: We first investigate the impact of the number of RDBs (T) in each SFIM/stage on the network. Figure 5 depicts the PSNR and Parameters trade-off study of SFIM with different RDBs where the number of SFIMs/stages (M) is set as 4. We can see that a larger T brings us better PSNR performance by extracting more features inside the two views. However, a larger T also leads to more parameters.

TABLE I
THE 2× SR PERFORMANCE OF MESFINET WITH DIFFERENT M (THE NUMBER OF SFIMS/STAGES), TRAINED ON MIDDLEBURY [11]

M	#Params.	$(Left + Right)/2$		
		$Middlebury$	$KITTI2012$	$KITTI2015$
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
1	0.59M	33.86/0.9398	30.42/0.9172	30.11/0.9271
2	1.06M	34.24/0.9433	30.47/0.9179	30.21/0.9276
3	1.55M	34.39/0.9458	30.51/0.9185	30.23/0.9279
4	2.00M	**34.44/0.9459**	**30.52/0.9189**	**30.28/0.9287**

TABLE II
ABLATION STUDY OF EDGE GUIDANCE AND GSFF FOR 4× SR ON FLICKR1024 [24]

Different combinations of Edge guidance and GSFF				
Edge guidance	×	✓	×	✓
GSFF	×	×	✓	✓
PSNR	23.38	23.51	23.47	**23.63**
SSIM	0.7268	0.7311	0.7304	**0.7389**
#Params.	1.79M	1.95M	1.84M	2.00M

In addition, we change the SFIM/stage number M of MSFINet to analyze its effect. The results are shown in Table I. It can be observed that, setting $T = 2$, the performance improves steadily as the SFIM/stage number M increases by exploiting and extracting more stereo information for image reconstruction. Considering a balance between the SR performance and the number of parameters, we finally set $T = 2$ and adopt a 4−stage MSFINet as our stereo image SR model.

2) Edge Guidance: We validate the effectiveness of edge guidance by adding EASFT to the baseline, *i.e.,* the cross-view stereo features sent to ESAM are modulated by edge-guided features, and then fed to the subsequent stereo feature transformation. It can be seen from Table II that the guidance of edge priors can improve the baseline performance (PSNR/SSIM of + 0.13 dB / 0.0043 for ×4 SR on Flickr1024 [24]) and we also observe that, without the edge guidance, the performance of the proposed model suffers from a decrease – PSNR and SSIM decrease by 0.16 dB and 0.0085, respectively. Additionally, to further validate the contributions of edge-guided feature interactions on the model performance, we apply different numbers of EASFT from back to front on the baseline, ranging from 0 to 4. As shown by the results in Table III, the more EASFTs, the better performance, which verifies that the more edge-guided feature interactions, the better SR performance. These quantitative results prove the benefit and effectiveness of using edge priors to guide the stereo image SR.

3) Global Stereo Feature Fusion (GSFF): In this study, we show the effectiveness of fusing multi-stage interactive stereo features by adding GSFF to the baseline – the stereo features

TABLE III
ABLATION STUDY OF MESFINET WITH DIFFERENT NUMBER OF EASFTS FOR 4× SR ON FLICKR1024 [24]

#EASFT	0	1	2	3	4
PSNR	23.47	23.56	23.58	23.61	**23.63**
Δ	-	+0.09	+0.11	+0.14	+0.16

TABLE IV
ABLATION RESULTS OF MODULATION SCHEME SELECTIONS FOR 2× SR ON KITTI2015 [23], TRAINED ON MIDDLEBURY [11]

Modulated schemes	#Params.	PSNR/SSIM
Concatenation	2.00M	30.18/0.9275
$\mathbf{F}^M = (\boldsymbol{\alpha}_m + \mathbf{I}) \odot \mathbf{F}^{m,C}$	1.95M	30.19/0.9279
$\mathbf{F}^M = \boldsymbol{\beta}_m + \mathbf{F}^{m,C}$	1.95M	30.21/0.9279
$\mathbf{F}^M = (\boldsymbol{\alpha}_m + \mathbf{I}) \odot \mathbf{F}^{m,C} + \boldsymbol{\beta}_m$	2.00M	**30.28/0.9287**

TABLE V
THE 2× SR PERFORMANCE OF MESFINET WITH DIFFERENT EDGE DETECTORS

Edge detector	$(Left + Right)/2$		
	$Middlebury$	$KITTI2012$	$KITTI2015$
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Canny	35.17/0.9507	31.19/0.9250	30.90/0.9352
Sobel	35.19/**0.9510**	31.20/0.9251	30.91/**0.9354**
BDCN	**35.21/0.9510**	**31.22/0.9253**	**30.92/0.9354**

output by multi-stage SFIMs are concatenated together and sent to a 1×1 convolutional layer and a 3×3 convolutional layer before the reconstruction part. As shown in Table II, with the GSFF module, the proposed model can steadily improve the PSNR from 23.35 dB to 23.47 dB on Flickr1024 [24]. Correspondingly, the performance suffers from a decrease (PSNR: –0.13 dB, SSIM: –0.0079 for ×4 Flickr1024 [24]) if GSFF is removed in our model. These comparisons show that the fusion of multi-stage stereo features indeed improves the performance of stereo image SR.

4) Modulation Scheme: For the edge guidance, we try different modulation schemes with edge priors, as shown in the first column of Table IV, and compare it with our proposed scheme, *i.e.,* Eq. (6), in the proposed ESAM. From the quantitative results in Table IV, it can be seen that the use of custom parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with edge priors is crucial for the spatial modulation of cross-view stereo features, and the proposed modulation scheme, shown at the last row, achieves the best performance.

5) Edge Probability Map: To analyze the impact of the quality of the edge probability maps, we use different edge detectors to generate edge probability maps for image SR, as shown in Table V. Figure 6 shows the edge probability maps of different edge detectors on challenging examples. It can be learned from Table V and Fig. 6 that the better the quality of edge priors, the better the SR performance. Meanwhile, since the difference between different edge probability maps is not obvious, the different detectors have a limited impact on SR performance with a PSNR margin of 0.02 ∼ 0.04 dB on Middlebury [11].

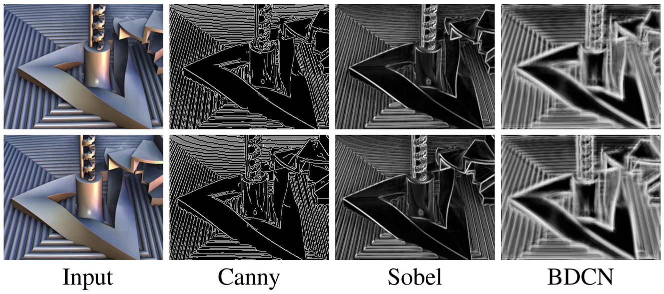

Fig. 6. Examples of edge probability maps produced by Canny, Sobel, and BDCN [25]. Top: left view; Bottom: right view.

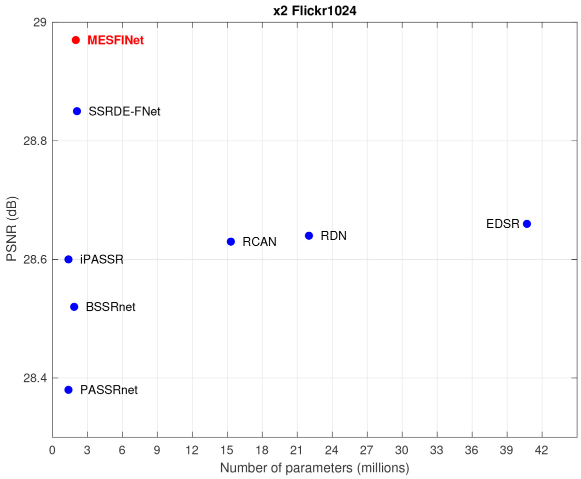

Fig. 7. Trade-off between the 2× SR performance and the number of parameters on Flickr1024 [24].

C. Comparison With State-of-the-Art Methods

In this subsection, we compare our model with 11 state-of-the-art image SR methods: ***1)*** single image SR methods: VDSR [30], EDSR [37], RDN [17], and RCAN [16]; ***2)*** stereo image SR methods[1]: StereoSR [12], PASSRnet [13], SRRes+SAM [19], IMSSRnet [8], BSSRnet [21], iPASSRnet [14], and SSRDE-FNet [52]. Moreover, the results of the IMSSRnet [8] and BSSRnet [21] are directly cited from its paper, and all compared methods are trained on the same training datasets as our method to make a fair comparison.

Quantitative Results: The quantitative results of PSNR and SSIM of 2× and 4× SR are shown in Table VI, and the number of model parameters is also shown in Fig. 7 for a intuitive observation. We can see from Table VI that, our MESFINet achieves the best average results on most datasets for both 2× and 4× stereo image SR against the comparison methods. Specifically, our MESFINet performs better than the popular

[1] SPAMnet [20] and DASSR [50] are not compared with our method because their evaluation scheme is different from all these comparative methods listed in this section, and their codes are not available. And we do not include NAFSSR [60] for comparison because it uses several training tricks to improve performance, such as data augmentation and large image patch size, and requires more memory cost.

TABLE VI

QUANTITATIVE EVALUATION OF STATE-OF-THE-ART SR ALGORITHMS, IN TERMS OF THE AVERAGE PSNR AND SSIM FOR SCALE FACTORS 2× AND 4×. RED INDICATES THE BEST AND BLUE INDICATES THE SECOND BEST PERFORMANCE. '*' INDICATES THAT THE RESULTS OF THE METHODS ARE DIRECTLY CITED FROM THE ORIGINAL PAPERS AND WE DO NOT DEMONSTRATE 2× SR RESULTS OF SRRES+SAM [19] SINCE THEIR MODELS ARE UNAVAILABLE

Method	Scale	#Params.	$Left$			$(Left + Right)/2$			
			$Middlebury$	$KITTI2012$	$KITTI2015$	$Middlebury$	$KITTI2012$	$KITTI2015$	$Flickr1024$
			PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Bicubic	×2	-	30.46/0.8979	28.44/0.8808	27.81/0.8814	30.60/0.8990	28.51/0.8842	28.61/0.8973	24.94/0.8186
VDSR [30]		0.66M	32.66/0.9101	30.17/0.9062	28.99/0.9038	32.77/0.9102	30.30/0.9089	29.78/0.9150	25.60/0.8534
EDSR [37]		40.7M	34.840.9489	30.83/0.9199	29.94/0.9231	34.95/0.9492	30.96/0.9228	30.73/0.9335	28.66/0.9087
RDN [17]		22.0M	34.85/0.9488	30.81/0.9197	29.91/0.9224	34.94/0.9491	30.94/0.9227	30.70/0.9330	28.64/0.9084
RCAN [16]		15.3M	34.80/0.9482	30.88/0.9202	29.97/0.9231	34.90/0.9486	31.02/0.9232	30.77/0.9336	28.63/0.9082
StereoSR [12]		1.08M	33.15/0.9343	29.42/0.9040	28.53/0.9038	33.23/0.9348	29.51/0.9073	29.33/0.9168	25.96/0.8599
PASSRnet [13]		1.37M	34.13/0.9421	30.68/0.9159	29.81/0.9191	34.23/0.9422	30.81/0.9190	30.60/0.9300	28.38/0.9038
IMSSRnet* [8]		6.84M	34.66/-	30.90/-	29.97/-	34.67/-	30.92/-	30.66/-	-/-
BSSRnet* [21]		1.86M	-/-	-/-	-/-	34.73/0.9470	30.98/0.9220	30.04/0.9250	28.52/0.9090
iPASSR [14]		1.37M	34.41/0.9454	30.97/0.9210	30.01/0.9234	34.51/0.9454	31.11/0.9240	30.81/0.9340	28.60/0.9097
SSRDE-FNet [52]		2.10M	35.02/0.9508	31.08/0.9224	30.10/0.9245	35.09/0.9511	31.23/0.9254	30.90/0.9352	28.85/0.9132
MESFINet (ours)		2.00M	35.15/0.9511	31.08/0.9223	30.12/0.9252	35.21/0.9510	31.22/0.9253	30.92/0.9354	28.97/0.9145
Bicubic	×4	-	26.27/0.7553	24.52/0.7310	23.79/0.7072	26.40/0.7572	24.58/0.7372	24.38/0.7340	21.82/0.6293
VDSR [30]		0.66M	27.60/0.7933	25.54/0.7662	24.68/0.7456	27.69/0.7941	25.60/0.7722	25.32/0.7703	22.46/0.6718
EDSR [37]		43.1M	29.15/0.8383	26.26/0.7954	25.38/0.7811	29.23/0.8397	26.35/0.8015	26.04/0.8039	23.46/0.7285
RDN [17]		22.2M	29.15/0.8387	26.23/0.7952	25.37/0.7813	29.27/0.8404	26.32/0.8014	26.04/0.8043	23.47/0.7295
RCAN [16]		15.4M	29.20/0.8381	26.36/0.7968	25.53/0.7836	29.30/0.8397	26.44/0.8029	26.22/0.8068	23.48/0.7286
StereoSR [12]		1.42M	27.70/0.8036	24.49/0.7502	23.67/0.7273	27.64/0.8022	24.53/0.7555	24.21/0.7511	21.70/0.6460
PASSRnet [13]		1.42M	28.61/0.8232	26.26/0.7919	25.41/0.7772	28.72/0.8236	26.34/0.7981	26.08/0.8002	23.31/0.7195
SRRes+SAM [19]		1.73M	28.76/0.8287	26.35/0.7957	25.55/0.7825	28.83/0.8290	26.44/0.8018	26.22/0.8054	23.27/0.7233
IMSSRnet* [8]		6.89M	29.02/-	26.44/-	25.59/-	29.02/-	26.43/-	26.20/-	-/-
BSSRnet* [21]		1.91M	-/-	-/-	-/-	29.13/0.8350	26.46/0.8010	25.59/0.7870	23.37/0.7270
iPASSR [14]		1.42M	29.07/0.8363	26.47/0.7993	25.61/0.7850	29.16/0.8367	26.56/0.8053	26.32/0.8084	23.44/0.7287
SSRDE-FNet [52]		2.24M	29.29/0.8407	26.61/0.8028	25.74/0.7884	29.38/0.8411	26.70/0.8082	26.43/0.8118	23.59/0.7352
MESFINet (ours)		2.05M	29.32/0.8434	26.61/0.8039	25.69/0.7897	29.42/0.8438	26.70/0.8099	26.42/0.8131	23.63/0.7389

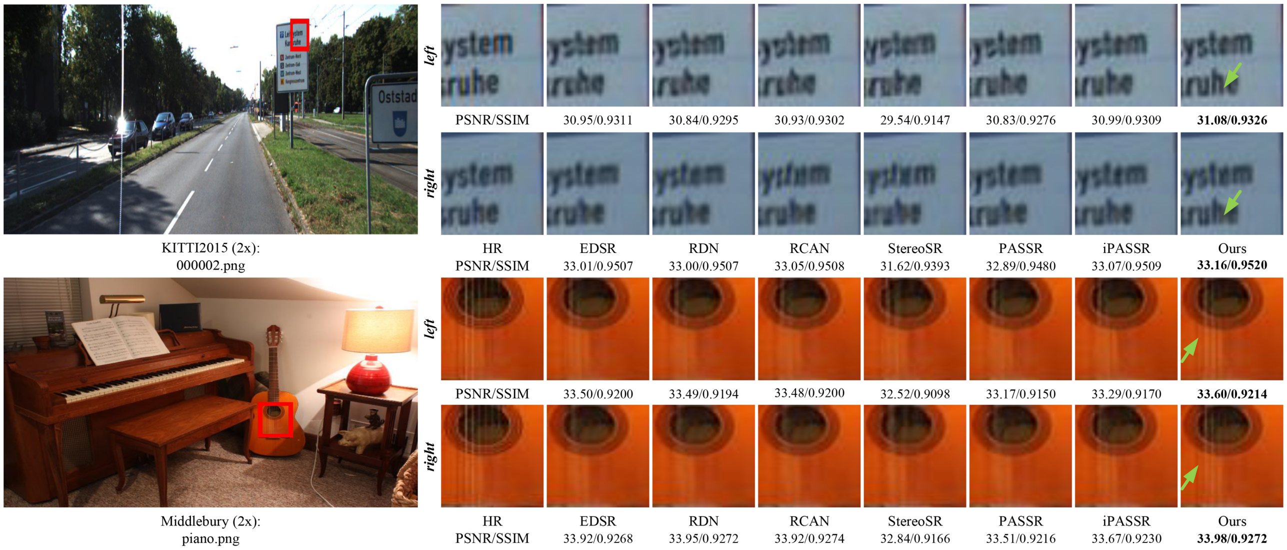

Fig. 8. Visual comparison of 2× SR on the KITTI2015 [23] and Middlebury [11] datasets. (Zoom in for the best view).

stereo image SR algorithm iPASSRnet [14] with a comparable number of parameters (the average PSNR gains of ×2 and ×4 SR on Middlebury [11] are 0.70dB and 0.26dB, respectively), and obtains better results than the large and deep single image SR networks, *e.g.*, EDSR [37], RDN [17], RCAN [16], by using fewer parameters. Furthermore, for the

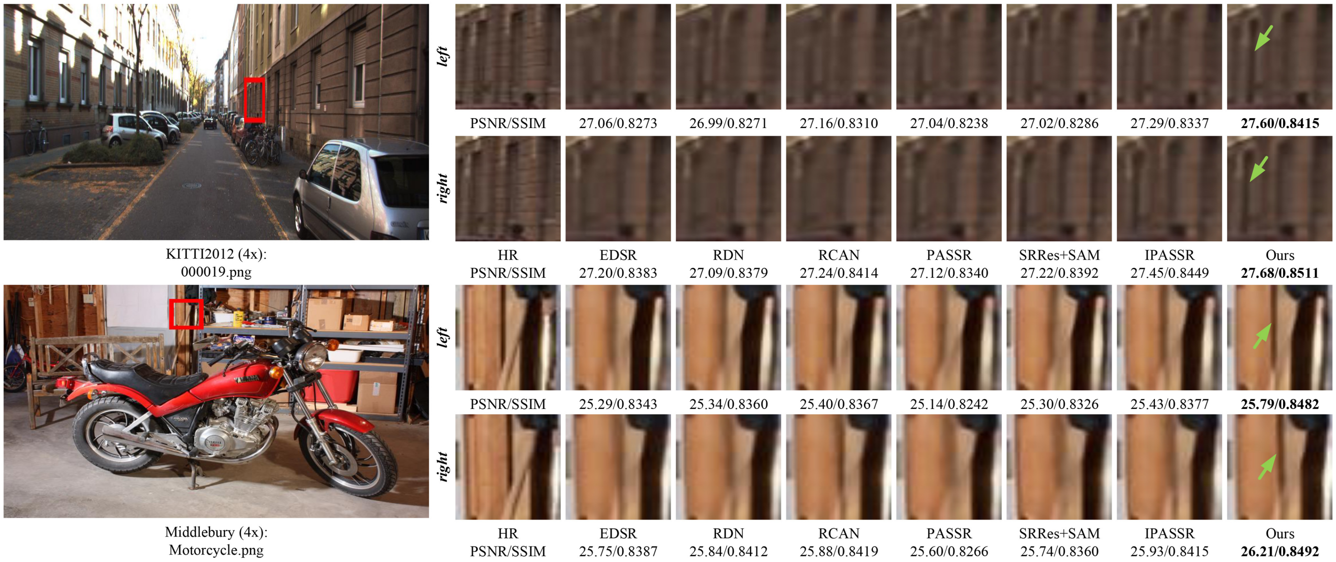

Fig. 9. Visual comparison of 4× SR on the KITTI2012 [22] and Middlebury [11] datasets. (Zoom in for the best view).

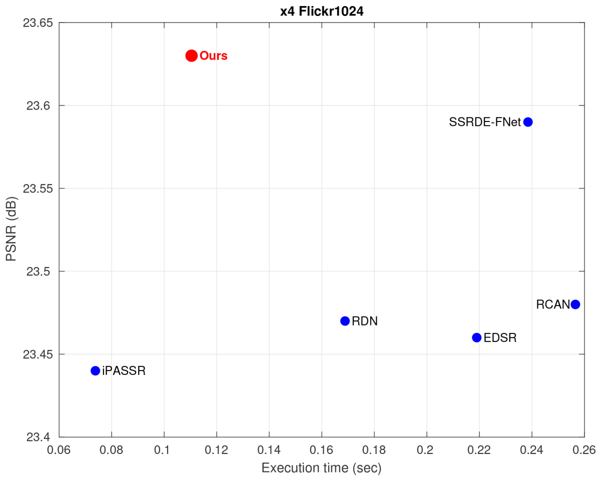

Fig. 10. PSNR *v.s.* running time on Flickr1024 [24] for 4× SR.

TABLE VII
QUANTITATIVE RESULTS ACHIEVED BY LEASTEREO [5] ON 4× SR STEREO IMAGES. ALL THESE METRICS WERE AVERAGED OVER THE TEST SET OF THE SCENEFLOW DATASET [61] AND LOWER VALUES DENOTE BETTER PERFORMANCE

Method	EPE	>1 $pixel$	>2 $pixel$	>3 $pixel$
Bicubic	1.0684	0.0995	0.0515	0.0376
EDSR [37]	0.9270	0.0895	0.0453	0.0332
RDN [17]	0.9349	0.0907	0.0455	0.0331
RCAN [16]	0.9328	0.0908	0.0458	0.0333
SRResnet+SAM [19]	0.9860	0.0979	0.0483	0.0350
iPASSR [14]	0.9343	0.0901	0.0455	0.0331
MESFINet (ours)	0.9135	0.0872	0.0445	0.0328
HR	0.8073	0.0796	0.0405	0.0297

scaling factor ×4, the SSIM of our MESFINet surpasses previous SSRDE-FNet [52] by 0.0027, 0.0017, 0.0013, 0.0037 on Middlebury [11], KITTI2012 [22], KITTI2015 [23] and Flickr1024 [24], respectively. This verifies the effectiveness of the proposed MESFINet.

Qualitative results: Figures 8 and 9 provide the visual comparison of 2× and 4× stereo image SR, respectively. From the zoom-in area in Fig. 8, both the single image SR and the stereo image SR methods produce blurry artifacts and incorrect edges. On the contrary, our MESFINet can restore sharper and clearer edges, *e.g.,* the slogan on the billboard in the image “00002.png” of KITTI2015 [23], and produce more faithful details, *e.g.,* the strings on the guitar in image “piano.png” of Middlebury [11]. A similar phenomenon can be observed in Fig. 9. This is mainly because our MESFINet excels at capturing structural details through edge guidance in multi-stage stereo feature interactions.

Running time: Figure 10 reports the trade-off results between running time (tested with 128×128 input on a single RTX 2080Ti GPU) and PSNR on Flickr1024 [24]. It can be seen that our method achieves comparable running time and the best PSNR value compared to other methods. From Fig. 10, it can be informed that our MESFINet outperforms the state-of-the-art model SSRDE-FNet with up to 2.16× speedup, which indicates that our approach is practical and efficient.

D. Benefits to Disparity Estimation

High-quality stereo image SR can facilitate high-level stereo vision tasks, *e.g.,* stereo matching. We further verify the effectiveness of our proposed MESFINet by using the reconstructed SR stereo image pairs for stereo matching. Here, we use the latest stereo matching algorithm LEAStereo [5] as the evaluation model, and adopt the first 100 images from flythings3D_image validation dataset of SceneFlow [61] for evaluation. The original clean image is used to provide the upper-bound result, and is downsampled 4× to generate LR stereo images for image SR. We then use several state-of-the-art SR methods, *i.e.,* EDSR [37], RDN [17], RCAN [16], SRResnet+SAM [19], iPASSR [14], to reconstruct SR images for stereo matching. In particular, End-Point-Error (EPE) and

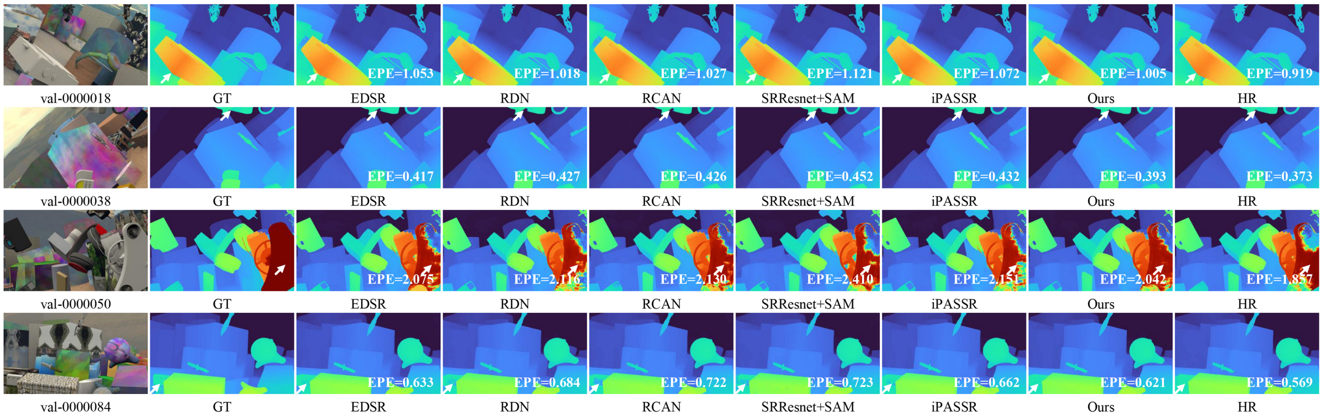

Fig. 11. Stereo matching results of LEAStereo [5] by utilizing 4× SR stereo images obtained by the state-of-the-art image SR methods.

$t-pixel$ error rate ($>t\ pixel$) are taken to evaluate the accuracy of stereo matching. As shown in Table VII, different image SR methods lead to different stereo-matching performances. In general, better image SR algorithms can lead to a smaller error of the estimated disparity. And it can also be seen that our method surpasses single image SR and stereo image SR methods in terms of both EPE and $t-pixel$ error rate, since our method can better capture stereo consistent details, *e.g.,* edges and textures, that is important for stereo matching. From the sample results in Fig. 11, our method can produce faithful stereo-matching results by comparing them to the best ones obtained from the original clean image.

V. Conclusion

In this paper, we proposed a multi-stage edge-guided stereo feature interaction network, termed MESFINet, for stereo image super-resolution (SR). Concretely, our MESFINet is cascaded by several stereo feature interaction modules. Enjoying a multi-stage learning strategy, MESFINet progressively enhances the reconstruction quality from coarse to fine. In addition, we proposed an edge-guided stereo attention mechanism and embed it into each stage of stereo feature interaction for capturing more details of SR stereo images. Experimental results have demonstrated the superior performance of our MESFINet over state-of-the-art CNNs-based SR methods on the KITTI2012, KITTI2015, MIddlebury, and Flickr1024 datasets. Impressively, SR stereo results produced by our MESFINet also improve the performance of stereo matching.

References

[1] P. Blanchfield and D. Wang, "Improved tile format of stereoscopic video for 3-D TV broadcasting," *IEEE Trans. Broadcast.*, vol. 60, no. 1, pp. 134–140, Mar. 2014.

[2] C. Zhang, Z. Li, Y. Cheng, R. Cai, H. Chao, and Y. Rui, "MeshStereo: A global stereo model with mesh alignment regularization for view interpolation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2057–2065.

[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2722–2730.

[4] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 899–908.

[5] X. Cheng et al., "Hierarchical neural architecture search for deep stereo matching," in *Proc. Adv. Neural Inform. Process. Syst.*, vol. 33, 2020, pp. 22158–22169.

[6] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[7] M. Cheng et al., "H^2-Stereo: High-speed, high-resolution stereoscopic video system," *IEEE Trans. Broadcast.*, vol. 68, no. 4, pp. 886–903, Dec. 2022.

[8] J. Lei et al., "Deep stereoscopic image super-resolution via interaction module," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3051–3061, Aug. 2021.

[9] J. Wan, H. Yin, Z. Liu, A. Chong, and Y. Liu, "Lightweight image super-resolution by multi-scale aggregation," *IEEE Trans. Broadcast.*, vol. 67, no. 2, pp. 372–382, Jun. 2021.

[10] Q. Jiang et al., "StereoARS: Quality evaluation for stereoscopic image retargeting with binocular inconsistency detection," *IEEE Trans. Broadcast.*, vol. 68, no. 1, pp. 43–57, Mar. 2022.

[11] D. Scharstein et al., "High-resolution stereo Datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.

[12] D. S. Jeon, S.-H. Baek, I. Choi, and M. H. Kim, "Enhancing the spatial resolution of stereo images using a parallax prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1721–1730.

[13] L. Wang et al., "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12250–12259.

[14] Y. Wang, X. Ying, L. Wang, J. Yang, W. An, and Y. Guo, "Symmetric parallax attention for stereo image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2021, pp. 766–775.

[15] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4681–4690.

[16] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 294–310.

[17] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[18] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3867–3876.

[19] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 496–500, Feb. 2020.

[20] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature," in *Proc. AAAI*, vol. 34, Apr. 2020, pp. 12031–12038.

[21] Q. Xu, L. Wang, Y. Wang, W. Sheng, and X. Deng, "Deep bilateral learning for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 28, pp. 613–617, Mar. 2021.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[23] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3061–3070.

[24] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale Dataset for stereo image super-resolution," in *Proc. Int. Conf. Comput. Vis. Workshops.*, Oct. 2019, pp. 3852–3857.

[25] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3828–3837.

[26] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[27] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3194–3205, Jul. 2012.

[28] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006.

[29] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with non-local means and steering kernel regression," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4544–4556, Nov. 2012.

[30] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[31] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1637–1645.

[32] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3147–3155.

[33] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 1874–1883.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[35] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[36] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4549–4557.

[37] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 136–144.

[38] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 624–632.

[39] W. Lai, J. Huang, N. Ahuja, and M. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.

[40] Z. Jiang, H. Zhu, Y. Lu, G. Ju, and A. Men, "Lightweight super-resolution using deep neural learning," *IEEE Trans. Broadcast.*, vol. 66, no. 4, pp. 814–823, Dec. 2020.

[41] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.

[42] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 517–532.

[43] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11065–11074.

[44] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2019.

[45] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, Nov. 2020.

[46] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2024–2032.

[47] B. Niu et al., "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 191–207.

[48] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5689–5698.

[49] L. Wang et al., "Parallax attention for unsupervised stereo correspondence learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2108–2125, Apr. 2020.

[50] B. Yan, C. Ma, B. Bare, W. Tan, and S. C. H. Hoi, "Disparity-aware domain adaptation in stereo image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13176–13184.

[51] X. Zhu, K. Guo, H. Fang, L. Chen, S. Ren, and B. Hu, "Cross view capture for stereo image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 3074–3086, Jun. 2022.

[52] Q. Dai, J. Li, Q. Yi, F. Fang, and G. Zhang, "Feedback network for mutually boosted stereo image super-resolution and disparity estimation," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1985–1993.

[53] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Proc. ACCV*, 2019, pp. 20–35.

[54] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8778–8787.

[55] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," in *Proc. IJCAI*, 2022, pp. 1335–1341.

[56] K. Nazeri, H. Thasarathan, and M. Ebrahimi, "Edge-informed single image super-resolution," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2019, pp. 3275–3284.

[57] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.

[58] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," 2019, *arXiv:1912.12180*.

[59] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

[60] X. Chu, L. Chen, and W. Yu, "NAFSSR: Stereo image super-resolution using NAFNet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2022, pp. 1239–1248.

[61] N. Mayer et al., "A large Dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.

Jin Wan received the B.E. degree from the Changchun University of Science and Technology, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University, China. His research interests include computer vision and machine learning.

Hui Yin received the Ph.D. degree in computer application technology from Beijing Jiaotong University, Beijing, China, where she is currently a Full Professor with the School of Computer and Information Technology. Her current research interests include the machine vision and intelligent information processing and their application in the railway industry.

Zhihao Liu received the B.E. degree from the Beijing Institute of Graphic Communication in 2016, and the Ph.D. degree from Beijing Jiaotong University in 2022. In 2019, he was a visiting Ph.D. student with the University of South Carolina, Columbia, SC, USA. He is currently a Computer Vision Algorithm Engineer with China Mobile Research Institute. His research interests include image and video recognition, restoration, and processing.

Yanting Liu received the B.E. degree from Hebei University, Baoding, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her research interests include deep learning, computer vision, and digital image processing.

Song Wang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, IL, USA, in 2002, where he was a Research Assistant with the Image Formation and Processing Group, Beckman Institute from 1998 to 2002. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, SC, USA, where he is currently a Professor. His research interests include computer vision, image processing, and machine learning. He is currently an Associate Editor for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, and *Pattern Recognition Letters*. He was the publicity or the Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society and is a member of the IEEE Computer Society.