Multi-Scale Visual Perception Based Progressive Feature Interaction Network for Stereo Image Super-Resolution

Anqi Liu, Sumei Li, *Member, IEEE*, Yongli Chang, and Yonghong Hou, *Member, IEEE*

***Abstract*— In recent years, stereo image super-resolution based on convolutional neural network has been extensively researched and achieved impressive performance by introducing complementary information from another view. However, most existing methods still cannot fully capture both intra- and cross-view information due to the neglect of multi-scale information perception, multi-scale binocular alignment and the excitation of large scale to small scale in human vision system. And they generated blurry results due to the consideration of irrelevant information in search for cross-view information. To address these issues, we propose a multi-scale visual perception based progressive feature interaction network (MS-PFINet) for stereo image super-resolution. Specifically, to exploit comprehensive intra- and cross-view information for image reconstruction, we design a two-stream network with multi-branch structure to extract multi-scale features and progressively use cross-view interaction at larger scales to guide that at smaller scales. Moreover, to explore more proper and accurate cross-view information, we propose a feature transformer module (FTM) to search and transfer the most relevant features from another view by hard attention maps and soft attention maps, which are calculated by patch-wise similarity rather than pixel-wise. In addition, in order to encourage a more effective way to transfer texture features for the target view, we propose a perceptual texture matching loss to supervise the accuracy of feature transformer modules. Experimental results show that our proposed method is superior to the state-of-the-art methods in most cases.**

***Index Terms*— Stereo image super-resolution, convolutional neural network, multi-scale, feature transformer, perceptual texture matching.**

I. Introduction

WITH the rapid development of dual-camera mobile phones and 3D devices, the demands for the resolution and definition of stereo image pairs are increasing, thus stereo image super-resolution has achieved extensive attention from industry and academia. Stereo image super-resolution (SR) aims at jointly utilizing low-resolution (LR) stereo image pairs to reconstruct high-resolution (HR) images with more detailed information and clearer textures. As we all know, a pair of stereo image is composed of a left and right image with certain difference, which is caused by horizontal parallax. Therefore, the performance of stereo image SR depends not only on the information in left/right image (i.e., intra-view information) but also on the information between left and right image (cross-view information).

However, in the preliminary stage of stereo image SR, people just used mature single image super-resolution (SISR) methods to reconstruct the left and right image separately, which suffered inferior performance due to the lack of cross-view information. In recent years, several methods have been proposed to incorporate the complementary information from another view. Wang et al. [1] first attempted to introduce a parallax-attention module (PAM) to capture stereo correspondence along the epipolar line, thus different disparity variations in stereo image can be handled. Later, Ying et al. [2] inserted generic stereo attention module (SAM) into arbitrary SISR networks to fully incorporate both the cross- and intra-view information at different stages. Recently, Wang et al. [3] exploited a Siamese network with a symmetric bi-directional parallax attention module (bi-PAM) to super-resolve both left and right image symmetrically. Zhu et al. [4] designed a cross-view block to capture features from both epipolar line and holistic view. Lei et al. [5] proposed an interaction module, which consisted of a series of interaction units, to effectively utilize complementary information between stereo image.

Although the above methods have made successive breakthroughs, there are still some limitations. First, for intra-view information capturing, most methods only focus on mining single-scale context information from the input left/right image [2], [3], [4], [5]. And for cross-view information capturing, they always perform cross-view interaction only once [1], [3], [4] or simply stack multiple interactions [2], [5] throughout the network. In other words, they ignore two facts in human vision system (HVS). One is that humans always perceive scenes by collecting multi-scale spatial information [6]. The other is that humans perform binocular alignment at different scales and the alignment at large scales always stimulates the alignment at small scales [7]. Therefore, the above methods may not be able to fully incorporate both intra- and cross-view information. Second, we observe that most methods [1], [2], [3], [4] extended PAM, which computes the weighted sum of features at all possible disparities in another

Manuscript received 21 September 2022; revised 14 April 2023; accepted 10 July 2023. Date of publication 13 July 2023; date of current version 7 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61971306. This article was recommended by Associate Editor H. Liu. *(Corresponding author: Sumei Li.)*

The authors are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: liuanqi@tju.edu.cn; lisumei@tju.edu.cn; chang_yli@163.com; houroy@tju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3295087.

Digital Object Identifier 10.1109/TCSVT.2023.3295087

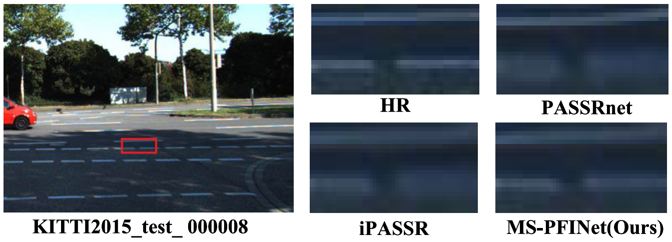


Fig. 1. Visual comparison on the KITTI2015 for 4× SR. We zoom the part in the image with a red rectangle for clearer comparison. We can find that the dividing lines on the road reconstructed by PAM-based methods are more blurry than ours.

view according to feature similarities, to capture reliable cross-view information for stereo image SR. However, not all the features within horizontal epipolar lines can provide beneficial complementary information, the irrelevant features may interfere with reconstruction performance and cause blurry effects. As shown in Fig.1, the dividing lines on the road reconstructed by PAM-based methods are more blurry than ours.

To address the above issues, we propose a multi-scale visual perception based progressive feature interaction network (MS-PFINet) to super-resolve both the left and right view. Firstly, considering the multi-scale information perception, multi-scale binocular alignment and the excitation of large scale to small scale in HVS, we carefully design a two-stream (left and right stream) network with multi-branch structure to fully incorporate both intra- and cross-view information. On one hand, the multi-branch structure with different scales can explore multi-scale intra-view features of different receptive fields. On the other hand, unlike other existing stereo image SR methods, the proposed structure progressively performs interactions at different scales. And the interaction features with larger scales are used to fuse with the intra-view features with smaller scales. Thus, richer intra-view features can be obtained to guide the cross-view interaction. Secondly, to avoid the interference of irrelevant information from cross-view information, inspired by [8], we propose a feature transformer module (FTM) for cross-view interaction. However, different from [8], our proposed FTM searches and transfers the most similar feature from one view to another view within the horizontal epipolar line rather than the entire feature map and further makes an attempt to handle occlusions. Specially, in the FTM, we formulate the stereo features as the queries and keys in a transformer [9] and conduct relevance estimation on them to obtain hard attention maps and soft attention maps, then the hard attention maps are used to transfer the most relevant complementary information from one view to another view, the soft attention map are used to rescale the transferred complementary information and generate valid masks to handle the occluded regions. For relevance estimation, we compute the similarity based on patches rather than that based on pixels in PAM, which could improve the accuracy of information transmission. Thirdly, to supervise the reliability of hard and soft attention maps generated by FTM, according to the left-right consistency, we propose a perceptual texture matching loss to make sure the feature transferred from one view is matched with another view, then the irrelevant features can be further eliminated. And inspired by the popular perceptual loss [10], we calculate the matching loss using textures in feature space rather than textures in pixel space. Thus, our FTM can focus on the perceptually similarity features for transfer, which is conducive to obtain more perceptually realistic SR results.

In summary, the main contributions of the proposed MS-PFINet are as follows:

1) Considering the multi-scale information processing mechanism of HVS, we design a two-stream network with multi-branch structure, which achieves progressive interaction based on multi-scale feature extraction to obtain reliable and comprehensive intra- and cross-view information.

2) We propose FTM to avoid the interference of irrelevant information, which utilizes both hard attention and soft attention calculated by patch-wise similarity to transfer the most similar feature in the horizontal epipolar line of the stereo image pair.

3) The perceptual texture matching loss is proposed to improve the accuracy of FTM and make texture transfer more effective.

II. Related Work

In this section, we briefly review the previous convolutional neural network (CNN)-based works of SISR and stereo image SR.

A. Single Image Super-Resolution (SISR)

In recent years, with the rapid development of deep learning, CNN-based methods have been widely studied [11], [12], [13]. The seminal work SRCNN [14] only had three layers but got prominent improvements compared with traditional algorithms. To get higher performance, many deep and complex SR networks have emerged. For example, VDSR [15] increased the network layer to 20 and introduced residual learning to avoid gradient vanishing. Lim et al. [16] proposed EDSR and removed the batch normalization layer in the original residual block [17] to boost the reconstruction performance. Zhang et al. [18] combined residual learning and dense skip connection [19] to construct a residual dense network (RDN) to fully utilize hierarchical features. Li et al. [20] presented FilterNet to adaptively filter the redundant low-frequency information and learn more useful features. Liu et al. [21] designed a hierarchical feature exploitation network equipped with cross convolution to better preserve the structural information.

Meanwhile, with the advent of multi-scale learning [22], [23], it has been widely adopted in SISR to further improve the model performance. For example, MSRN [24] stacked the multi-scale residual blocks to adaptively detect image features in different scales. Hu et al. [25] devised a multi-scale cross module to fuse multi-scale complementary information as well as to help information flow. Qin et al. [26] carefully designed a multi-scale feature fusion residual block, and Li et al. [27] proposed a multi-scale dense cross block to effectively extract multi-scale features and fuse them. Wu et al. [28] presented a multi-grained attention block to fully utilize the advantages of multi-scale and attention mechanism in SR tasks. Unlike the above methods, which achieve multi-scale learning on the block/module level and directly fuse all the features at the end of the block/module, we achieve multi-scale learning on the structure level and progressively aggregate the multi-scale features based on the proposed multi-branch structure. More importantly, in addition to extracting multi-scale intra-view features, our proposed multi-branch structure is also used to realize multi-scale cross-view interaction and the guidance of large scale to small scale.

More recently, with the prospering of attention mechanism, it has been widely introduced into many SR networks and has been turned out to be effective. For example, Zhang et al. [29] firstly applied SEblock [30] in SISR and proposed a residual channel attention network to rescale channel-wise features. Dai et al. [31] proposed a second-order attention module to make the discriminative ability of the network more powerful. Zhang et.al [32] designed a multi-context attention block to focus on more informative contextual features, and a refined attention block to explore fine-grained cues for HR image reconstruction. Zhu et al. [33] proposed an expectation-maximization attention mechanism, making it possible to capture the long-range dependencies directly on the HR-size feature map. Since SISR has achieved promising performance for single image reconstruction, it can capture enriched features within a single view, which is also the research basis of stereo image SR.

B. Stereo Image Super-Resolution

Compared with SISR, stereo image SR should focus on capturing not only intra-view information but also cross-view information. Therefore, how to find and effectively integrate reliable cross-view information is one of the pivotal issues for stereo image SR. The pioneer CNN-based stereo image SR was StereoSR [34]. For cross-view information integration, they moved the right image horizontally by different pixels to generate 64 duplicate images, and then cascaded them with the left image for reconstruction. Nevertheless, the maximum parallax range in StereoSR was confined to 64 pixels, information beyond this parallax range cannot be considered. So StereoSR is not suitable for different stereo images with large disparities.

To address the issue, several methods based on disparity estimation were proposed. Yan et al. [35] explicitly used the disparity prior estimated by pre-trained depth prediction network to warp one view to another and adaptively incorporate cross-view information for stereo image restoration. Dan et al. [36] proposed a disparity feature alignment module to exploit the disparity information for feature alignment and fusion. Dai et al. [37] studied stereo image SR and disparity estimation in a unified framework and interacted them with each other to improve their performance. Though disparity estimation based methods have achieved better performance, they rely on the accuracy of disparity estimation, and inaccurate disparity estimation will result in the inability to obtain reliable cross-view information, thereby affecting the reconstruction performance.

To capture cross-view information more flexibly, Wang et al. brought up PASSRnet [1], in which a parallax attention module (PAM) was designed to learn feature similarities with all possible disparities. Thus, global stereo correspondence can be captured and high flexibility can also be maintained. Then, based on PAM, several methods have been brought up. Ying et al. [2] proposed a stereo attention module (SAM) and inserted it into SISR network to explore powerful intra-view features while integrating cross-view information at different stages. Concurrently, Song et al. [38] proposed a self and parallax attention module (SPAM) to aggregate the information from its own image and the counterpart stereo image. Later, Wang et al. [3] proposed a Siamese network (iPASSR) equipped with a symmetric bi-directional parallax attention module (biPAM), which could fully exploit symmetric cues in stereo image pairs to improve the performance of SR. More recently, BSSRnet [39] introduced the idea of bilateral grid into a CNN framework to restore missing details at different locations while preserving stereo consistency. Zhu et al. [4] designed a cross-view capture network to extract both local and global cross-view information to improve the performance. Different from the above PAM-based methods, Lei et al. [5] proposed IMSSRnet, in which an interaction module composed of a series interaction units was designed to comprehensively utilize complementary features between different views. CPASSRnet [40] adopted encoder-decoder architecture to super-resolve both views at multiple scale factors with a single model. Therein, a cross parallax module was presented to alleviate the difficulties of handling stereo image pairs with irregular epipolar lines and large disparities.

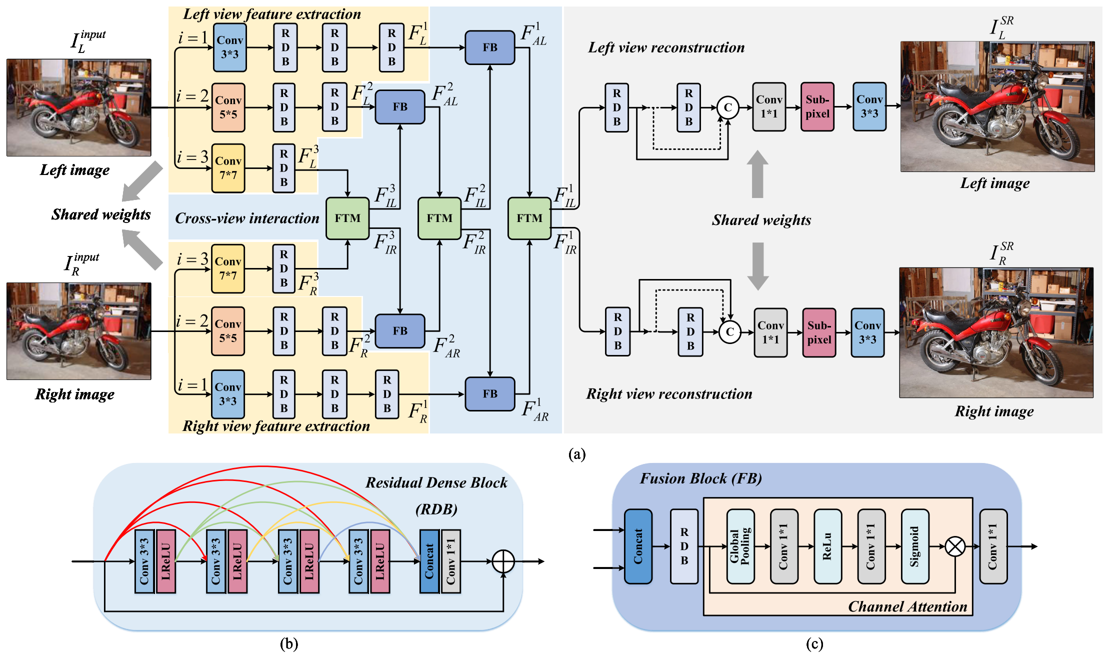
Although the above methods have made a great progress, they ignore the multi-scale information perception, multi-scale binocular alignment and the guidance of large scale on small scale in HVS, thus they still cannot fully capture both intra- and cross-view information for stereo image reconstruction. In addition, the reconstruction results of PAM-based methods are blurry because some irrelevant features were introduced in search for cross-view information. Thus, in this paper, for enriched intra- and cross-view information capturing, a two-stream network with multi-branch structure is designed to explore multi-scale features and progressively use cross-view interaction at larger scales to guide that at smaller scales. And for accurate cross-view information capturing, a feature transformer module is proposed to only transfer the most relevant features.

III. Proposed Method

In this section, we first describe the overall structure of our proposed MS-PFINet and then introduce our feature transformer module (FTM) and perceptual feature matching loss in detail.

A. Network Architecture

The overall architecture of our MS-PFINet is illustrated in Fig.2(a), which is a two-stream network with multi-branch structure. The network is highly symmetrical and the weights of the left and right stream are shared. Given a pair of LR stereo image I_L^{input} and I_R^{input}, feature extraction, cross-view



(a)

(b)

(c)

Fig. 2. (a) The overall structure of our proposed MS-PFINet (b) Residual dense block (c) Fusion block.

interaction, and image reconstruction is performed sequentially to reconstruct the HR left image I_L^{SR} and right image I_R^{SR} simultaneously.

For feature extraction, considering the multi-scale perceptual characteristic of HVS [6], we design a three-branch structure with different scales to capture enriched intra-view features in left/right stream. On each branch, we first use a convolution layer with different kernel sizes (3×3, 5×5, 7×7) to extract initial features at different scales, and then use cascaded residual dense blocks (RDB) [18] for deep intra-view feature extraction. Different branches have different numbers of RDBs, thus features with different receptive field can be obtained for cross-view interaction at different scales. Within each RDB, as shown in Fig.2(b), we use four convolution layers with kernel size 3×3 and the growth rate of the channel is 16. It's noteworthy that we use RDB as basic blocks to extract intra-view features for the reason that RDB can fully use hierarchical features from all the layers and generate features with large receptive fields, which have been demonstrated to be conducive to accurate stereo correspondence estimation. The output of each branch can be described as follows,

$$F_L^i = H_{RDBs}^i\left(H_{(2i+1)\times(2i+1)}\left(I_L^{input}\right)\right)(i=1,2,3), \quad (1)$$

$$F_R^i = H_{RDBs}^i\left(H_{(2i+1)\times(2i+1)}\left(I_R^{input}\right)\right)(i=1,2,3), \quad (2)$$

where F_L^i and F_R^i represent the left and right view features extracted on the i-th branch. H_{RDBs}^i indicates the operation of the cascaded RDBs on the i-th branch, and when $i = 1, 2, 3$, the number of RDBs is 3, 2, 1 respectively. $H_{(2i+1)\times(2i+1)}$ signifies the first convolution layer with kernel size $(2i+1)\times(2i+1)$ on the i-th branch. Thus, rich intra-view features can be extracted from three scales (3, 5, and 7).

For cross-view interaction, considering the binocular alignment at different scales and the guiding effect of large scales on small scales [7], we progressively apply FTM to achieve interaction at different scales and use fusion block (FB) to implement the feature fusion from different scales and the guidance of larger scale on smaller scale, thereby more reliable and comprehensive cross-view information can be explored for image reconstruction. Specifically, the interaction features from FTM on the $(i+1)$-th branch and the intra-view features from the i-th branch are aggregated using FB. Then the aggregated features are further taken as the input of FTM on the i-th branch,

$$F_{IL}^i, F_{IR}^i = H_{FTM}^i\left(F_{AL}^i, F_{AR}^i\right)(i=3,2,1), \quad (3)$$

$$F_{AL}^i = \begin{cases} H_{FB}^i\left(f_{IL}^{i+1}, f_L^i\right) & i=2,1 \\ F_L^3 & i=3 \end{cases}, \quad (4)$$

$$F_{AR}^i = \begin{cases} H_{FB}^i\left(f_{IR}^{i+1}, f_R^i\right) & i=2,1 \\ F_R^3 & i=3 \end{cases}, \quad (5)$$

where F_{IL}^i and F_{IR}^i denote the interaction features for the left and right view, which have already integrated the complementary information between two views on the i-th branch. H_{FTM}^i represents the operation of FTM on the i-th branch. F_{AL}^i and F_{AR}^i indicate the aggregated left and right view features which is a fusion of the interaction features from the $(i+1)$-th branch and the intra-view features from the i-th branch. H_{FB}^i represents the operation of FB on the i-th branch, the structure is shown in Fig.2(c), which contains an RDB for aggregated feature extraction, a channel attention module for feature selection, and a convolution layer with kernel size 1×1 for further fusion and dimensionality reduction.

For image reconstruction, similar to feature extraction, we also use RDB as basic blocks. Firstly, the interaction features F_{IL}^1 and F_{IR}^1 from the first branch are fed into four cascaded RDBs for deep feature mining, respectively. After

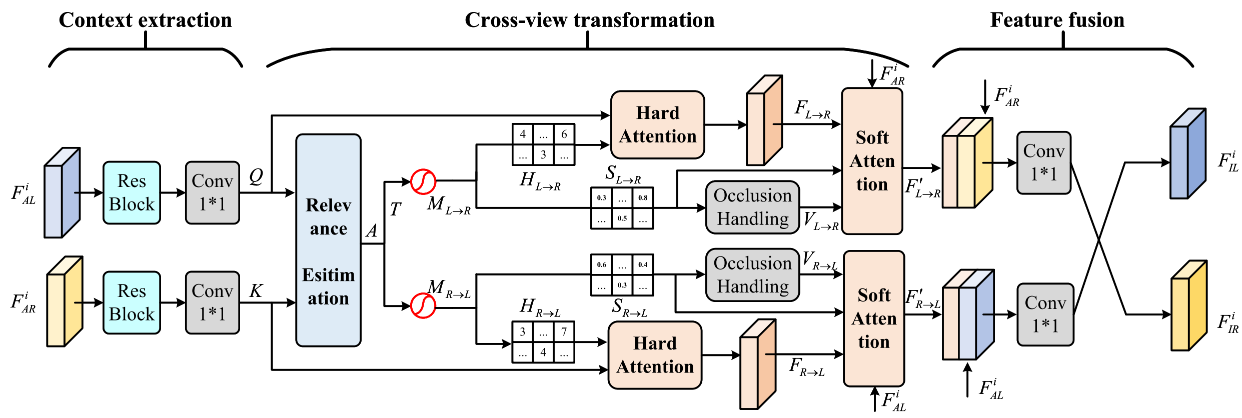


Fig. 3. The proposed feature transformer module (FTM).

that, the output features of all RDBs are further concatenated in channel dimensions to retain information at different levels. Then a convolution layer with kernel size 1×1 is used for feature fusion, a sub-pixel layer and a convolution layer with kernel size 3×3 are used to generate HR left image I_L^{SR} and right image I_R^{SR}.

B. Feature Transformer Module (FTM)

To provide more proper and accurate cross-view information for reconstruction, we propose FTM for the interaction between left and right view. The architecture of our FTM is illustrated in Fig.3, which consists of three parts: context extraction, cross-view transformation, and feature fusion.

For context extraction, given two stereo feature maps $F^i_{AL} \in \mathbb{R}^{H\times W\times C}$ and $F^i_{AR} \in \mathbb{R}^{H\times W\times C}$ as the input of the i-th FTM in Fig.1, they are first fed into a shared transition residual block H_{res} to alleviate training conflict [1], and then separately fed into a convolution layer $H_{1\times 1}$ with kernel size 1×1 to generate $Q, K \in \mathbb{R}^{H\times W\times C}$,

$$Q = H_{1\times 1}\left(H_{res}(F^i_{AL})\right), \tag{6}$$

$$K = H_{1\times 1}\left(H_{res}(F^i_{AR})\right). \tag{7}$$

For cross-view transformation, it mainly contains three operations: relevance estimation for capturing reliable stereo correspondence, hard attention for transferring the most relevant features, and soft attention for further emphasizing the importance of transferred features at different positions. More specially, we take the transfer of the features from the right view to the left view as an example.

1) Relevance Estimation: Unlike PAM [1], SAM [2] and so on, we compute the similarity of the patches along the horizontal epipolor line between Q and K instead of that of pixels, thus local contextual information brought by the patch-wise similarity can be noticed to improve the accuracy of relevance estimation. We first unfold Q and K into patches, denoted as $q^{l,m} \in \mathbb{R}^{P\times P\times C}$ $(l \in [1, H], m \in [1, W])$ and $k^{l,n} \in \mathbb{R}^{P\times P\times C}$ $(l \in [1, H], n \in [1, W])$, where $q^{l,m}$ represents the patch centered at (l, m) in Q, $k^{l,n}$ represents the patch centered at (l, n) in K, $P \times P$ stands for the patch size and here we set $P = 3$. Then for each patch $q^{l,m}$ in Q and $k^{l,n}$ in K, we estimate the relevance $A \in \mathbb{R}^{H\times W\times W}$ between them by inner product, which is further fed into a softmax function to generate relevance map $M_{R\to L} \in \mathbb{R}^{H\times W\times W}$,

$$A\,(l, m, n) = \left\langle q^{l,m}, k^{l,n}\right\rangle, \tag{8}$$

$$M_{R\to L} = Softmax\,(A)\,, \tag{9}$$

where $A\,(l, m, n)$ measures the contribution of $k^{l,n}$ to $q^{l,m}$, $\langle\cdot,\cdot\rangle$ denotes the inner product operation. $Softmax\,(\cdot)$ represents a softmax function. Then the relevance map is further used to obtain hard-attention maps and soft attention maps.

2) Hard Attention: Unlike the parallax attention mechanism that takes a weighted sum of features at all positions along epipolor lines to capture complementary information, we utilize hard attention to only search and transfer the most relevant patch in K for each patch in Q. First we calculate the hard attention map $H_{R\to L} \in \mathbb{R}^{H\times W}$ from the relevance map $M_{R\to L}$ to record the position of the most relevant patch,

$$H_{R\to L}\,(l, m) = \arg\max_n M_{R\to L}\,(l, m, n)\,, \tag{10}$$

where $M_{R\to L}\,(l, m, n)$ represents the matching possibility between $k^{l,n}$ and $q^{l,m}$. $H_{R\to L}\,(l, m)$ is a hard index, which is used to select the most relevant patch at the corresponding position from K for each $q^{l,m}$ in Q. Then the new transferred right view features $F_{R\to L}$ are formed,

$$f^{l,m}_{R\to L} = k^{l,H_{R\to L}(l,m)}, \tag{11}$$

where $f^{l,m}_{R\to L}$ represents the patch centered at (l, m) in $F_{R\to L}$, which is from the patch centered at $(l, H_{R\to L}\,(l, m))$ in K. The transferred features $F_{R\to L}$ are further used in soft attention.

3) Soft Attention: To obtain more accurate transferred features, we compute a soft attention map $S_{R\to L} \in \mathbb{R}^{H\times W}$ using the relevance map $M_{R\to L}$,

$$S_{R\to L}\,(l, m) = \max_n M_{R\to L}\,(l, m, n)\,, \tag{12}$$

where $S_{R\to L}\,(l, m)$ denotes the confidence of the transferred features $F_{R\to L}$ at position (l, m). By applying the soft attention map $S_{R\to L}$ to $F_{R\to L}$, the relevant transferred features with higher confidence can be emphasized while the one with lower confidence can be suppressed.

In addition, occluded regions always exist in the stereo image pair, which makes it difficult to find the correspondence between left and right view, so we should further make an attempt to handle occlusions. Inspired by [1], [2], a valid mask $V_{R\to L} \in R^{H\times W}$ is generated according to $S_{R\to L}$,

$$V_{R\to L}(l,m) = \begin{cases} 1, & if\, S_{R\to L}(l,m) > \tau \\ 0, & otherwise. \end{cases}, \tag{13}$$

where the threshold τ is empirically set to 0.01. Then the final transferred right view features $F'_{R\to L}$ can be obtained by,

$$F'_{R\to L} = F_{R\to L} \odot S_{R\to L} \odot V_{R\to L} + F^i_{AL} \odot (1 - V_{R\to L}), \tag{14}$$

where $\odot$ represents the element-wise multiplication. And $F^i_{AL} \odot (1 - V_{R\to L})$ means that we fill the occluded regions of transferred features with the corresponding features from the target view.

For feature fusion, we concatenate the input left view features F^i_{AL} with $F'_{R\to L}$ and feed them into a convolution layer with kernel size 1×1 to fully incorporate cross- and intra-view information, then the final interaction features F^i_{IL} for left views are generated,

$$F^i_{IL} = H_{1\times 1}\left(\left[F^i_{AL}, F'_{R\to L}\right]\right), \tag{15}$$

where $[\cdot]$ stands for concatenation operation in channel dimension. Note that, the interaction features F^i_{IR} for right view is generated following the similar way.

C. Loss Function

We introduce two loss functions for the training of our MS-PFINet. The overall loss function is defined as,

$$L = L_{SR} + \lambda \sum_{i=1}^{3} L^i_{matching}, \tag{16}$$

where L_{SR} and $L^i_{matching}$ represent the SR loss for image reconstruction and our proposed perceptual texture matching loss to supervise the FTM on the i-th branch. λ is the regularization weight to balance the two loss functions and here we set it to 0.001.

The SR loss is defined as the mean absolute error between the super-resolved and ground truth stereo image,

$$L_{SR} = \left\| I_L^{SR} - I_L^{HR} \right\|_1 + \left\| I_R^{SR} - I_R^{HR} \right\|_1, \tag{17}$$

where I_L^{SR} and I_R^{SR} represent the super-resolved left and right image, I_L^{HR} and I_R^{HR} denote their corresponding ground truth images.

In order to supervise the accuracy of the hard attention map and soft attention map during the training process, constraining the transferred features to have similar textures to the target view, a perceptual texture matching loss is designed according to left-right consistency, which also uses the mean absolute error. Inspired by perceptual loss, the matching loss is calculated by features extracted from the shallow layers of the pretrained VGG network [41] to make our FTM pay more attention to perceptually relevant textures for transfer, thus more visually satisfying results can be obtained. Note that, since left-right consistency only holds in non-occluded regions, the perceptual texture matching loss is formulated as follows,

$$\begin{aligned} L^i_{matching} &= \left\| V^i_{R\to L} \odot \left(FT\left(\phi\left(I_R^{input}\right), H^i_{R\to L}\right) \odot S^i_{R\to L} - \phi\left(I_L^{input}\right)\right) \right\|_1 \\ &\quad + \left\| V^i_{L\to R} \odot \left(FT\left(\phi\left(I_L^{input}\right), H^i_{L\to R}\right) \odot S^i_{L\to R} - \phi\left(I_R^{input}\right)\right) \right\|_1, \end{aligned} \tag{18}$$

where FT stands for the feature transfer operation based on the hard attention map. $\phi(\cdot)$ denotes the first two layers of the pretrained VGG network.

IV. Experiments

In this section, we firstly introduce the datasets we used and describe the implementation details during training and testing, then we compare our network with the state-of-the-art single image SR and stereo image SR methods, finally, we present model analysis to demonstrate the effectiveness of our proposed MS-PFINet.

A. Datasets

For training, we use 800 stereo image pairs from the Flickr1024 dataset [42] and 60 stereo image pairs from the Middlebury dataset [43]. Since the spatial resolution of images from the Middlebury dataset is much larger than others, following [1], [2], [3], [4], [5], we perform bicubic downsampling with a scale factor of 2 on them to generate HR images. For testing, we use four benchmark datasets to evaluate the performance of our proposed method. They are 20 stereo image pairs from the KITTI2012 [44] dataset, 20 stereo image pairs from the KITTI2015 [45] dataset, 5 stereo image pairs from the Middlebury dataset [43] and 112 stereo image pairs from the Flickr1024 dataset [42].

B. Implementation Details

During the training phase, we firstly downscale training datasets with the scale factor of 2 and 4 by bicubic interpolation to obtain LR images for 2× and 4× SR respectively, and then crop them into patches of size 30×90 with a stride 20. Their corresponding HR patches are also cropped. In order to make full use of the datasets, inspired by [46], we randomly flip training patches horizontally and vertically for data augmentation.

We apply PyTorch framework to implement the proposed method on a GeForce RTX 2080 Ti GPU. All models are trained with Adam [47] optimizer by setting β_1 to 0.9, β_2 to 0.999 and the batch size is set to 16. The initial learning rate is set to 2×10^{-4}, which is reduced by the factor of 2 after every 30 epochs. The training was stopped after 80 epochs because no more performance is gained with more epochs.

During the testing phase, we utilize the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) to evaluate the quality of the super-resolved images, the higher PSNR and SSIM values, the better quality. To achieve fair comparison with other stereo SR methods, we calculate all the metrics in RGB color space and on the left views with their left borders (64 pixels) being cropped.

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

In this subsection, we compare our proposed MS-PFINet with several state-of-the-art SR methods on 2× and 4× SR, including five SISR methods (VDSR [15], LapSRN [48], EDSR [16], RDN [18], RCAN [29]) and five recent stereo image SR methods (PASSRNet [1], SRResNet+SAM [2], IMSSRnet [5], CVCnet [4], iPASSR [3]). For fair comparison, we generate all the stereo image SR results using their officially released models and obtain all the SISR results using the models released in SRResNet+SAM and iPASSR for evaluation.

For quantitative comparison, we utilize PSNR and SSIM to evaluate the quality of the super-resolved images. The average PSNR and SSIM of each method on the left images from four datasets are presented in Table I. From Table I, we can observe that, compared with stereo image SR methods that perform interaction only once [1], [3], [4] or simply stack multiple interactions [2], [5], our proposed MS-PFINet obtains the highest PSNR/SSIMs on all benchmark datasets for 2× SR and achieves best performance on KITTI2012, KITTI2015 and Flickr1024 dataset for 4× SR. As for Middlebury dataset for 4× SR, only iPASSR performs slightly higher than our proposed MS-PFINet, but the visual quality of our reconstructed image is better than iPASSR, as shown in Fig.5. This is because our proposed MS-PFINet adopts the multi-scale visual perception structure and perceptual texture matching loss, which can help to improve the visual quality without much objective indexes reduction, alleviating the contradiction between objective index and visual quality in image restoration tasks to some extent.

Compared with SISR methods, our proposed MS-PFINet achieves the highest PSNR and SSIM values on KITTI2012, KITTI2015 and Flickr1024 for 2× SR, and outperforms all the methods on KITTI2012 and KITTI2015 for 4× SR. But on Middlebury dataset, the PSNR values of our proposed MS-PFINet are 0.04dB lower than RDN for 2× SR, 0.18dB lower than RCAN for 4× SR. On Flickr1024 dataset, the PSNR value of our proposed MS-PFINet is 0.01dB lower than RCAN for 4× SR. And we find that in Table I, on Middlebury and Flick1024 dataset, all the other stereo image super-resolution methods are inferior to EDSR, RDN, RCAN and the performance gap is even larger than ours. For example, for 2× SR, the PSNR values of iPASSR are 0.45dB lower than RDN and 0.05dB lower than EDSR on Middlebury and Flickr1024 dataset, respectively. For 4× SR, the PSNR values of CVCnet are 0.57dB and 0.18dB lower than RCAN on Middlebury and Flickr1024 dataset, respectively. That is because EDSR, RDN, RCAN have deeper networks and larger number of parameters than our proposed MS-PFINet and other stereo image super-resolution methods (e.g., 22.0 M vs 1.45M). Moreover, since stereo image pairs in Middlebury and Flickr1024 dataset have high quality and varied textures, the large and deep networks strengthen the ability of mining rich and hierarchical intra-view features so that they can better learn the texture structure and boost the performance. Differently, we aim at using a smaller model size to obtain competitive or even higher performance than heavy models (e.g., 0.18dB and 0.14dB higher than RCAN on KITTI2012 for 2× and 4× SR), which indicates the importance of exploiting cross-view information for stereo image SR.

For qualitative comparison, visual comparison for 2× and 4× are shown in Fig.4 and Fig.5 respectively. From Fig.4, we can see that, compared with SISR methods, our proposed MS-PFINet can reconstruct the parallel lines on the windows with the correct direction, recover grooves on parts of the motorcycle without aliasing effect, and recover the letters on the stairs completely. This is because SISR methods can only use limited spatial information in LR images for SR, while we can use complementary cross-view information provided by stereo image to produce faithful details. Compared with stereo image SR, our proposed MS-PFINet can reconstruct the parallel lines more continuously and the edges of grooves and letters more sharply due to the accurate transfer of features. Similarly, from Fig.5, we observe that the vertical guitar strings cannot be recovered by any other methods except ours. In addition, the parallel line on the rails and the stripes on the plants recovered by other methods are more blurry than ours and accompanied by artifacts.

In addition, we also compared the FLOPs and execution time with other stereo image super-resolution methods in Table II. Noting that, for fair comparison, all the compared methods can super-resolve both left and right images simultaneously in one feed. In the table, FLOPs-I and Time-I denote the FLOPs and execution time of a single cross-view interaction module (SAM, biPAM, FTM) in each method. FLOPs-T and Time-T denote the total FLOPs and execution time. We can observe that our proposed method yields the second-best FLOPS and execution time with the best performance on PSNR/SSIM. Specifically, compared with SRRes+SAM, our proposed FTM consumes slightly more FLOPs than SAM but the total FLOPs is much smaller. While compared with iPASSR, the FLOPs of our proposed FTM is much smaller than biPAM but the total FLOPs is slightly larger. This explains that a single cross-view interaction module is only a small part of the whole network. The overall computation and execution time mainly depends on the feature extraction modules and the number of interactions.

D. Model Analysis

In this subsection, we will conduct several experiments to validate the rationality of the network structure and the effectiveness of each key component in the network.

1) Multi-Branch Structure: Firstly, to choose a proper number of branches for our network, we conduct four experiments by gradually increasing the number of branches. The results are illustrated in Table III. The numeric string in the first column signifies the kernel size of the first convolution layer on each branch. For example, '3-5-7' means that there are three branches, and the kernel size of the first convolution layer on each branch is 3×3, 5×5, 7×7, respectively. We can observe that as the number of branches increase, the performance of the network is gradually improved, which demonstrates

TABLE I

AVERAGE PSNR/SSIMS FOR SCALE 2×, 4×. RESULTS MARKED WITH * ARE DIRECTLY COPIED FROM THE CORRESPONDING PAPER, SINCE THEIR MODELS ARE UNAVAILABLE. BEST AND SECOND BEST RESULTS ARE **BOLDED** AND UNDERLINED

Methods	Scale	Params	KITTI2012 (PSNR/SSIM)	KITTI2015 (PSNR/SSIM)	Middlebury (PSNR/SSIM)	Flickr1024 (PSNR/SSIM)
VDSR [15]	2×	0.66M	30.16/0.908	28.98/0.905	33.19/0.932	27.17/0.878
LapSRN [48]	2×	0.81M	30.01/0.905	28.90/0.903	32.97/0.929	26.79/0.871
EDSR [16]	2×	38.6M	30.87/0.920	29.97/0.923	34.93/**0.949**	28.61/0.910
RDN [18]	2×	22.0M	30.85/0.920	29.93/0.923	**34.94/0.949**	28.58/0.910
RCAN [29]	2×	15.3M	30.91/0.920	30.00/0.923	34.87/0.948	28.56/0.910
PASSRnet [1]	2×	1.37M	30.68/0.916	29.81/0.919	34.14/0.942	28.29/0.904
SRRes+SAM [2]	2×	–	–/–	–/–	–/–	–/–
IMSSRnet* [5]	2×	6.84M	30.90/–	29.97/–	34.66/–	–/–
CVCNet* [4]	2×	0.97M	30.87/0.920	29.93/0.922	34.40/0.945	28.44/0.908
iPASSR [3]	2×	1.37M	31.00/0.921	30.04/0.924	34.49/0.946	28.56/0.911
MS-PFINet (Ours)	2×	1.40M	**31.09/0.923**	**30.12/0.925**	34.90/**0.949**	**28.66/0.913**
VDSR [15]	4×	0.66M	25.90/0.799	24.98/0.760	27.95/0.803	22.71/0.687
LapSRN [48]	4×	0.81M	25.99/0.780	25.06/0.762	28.05/0.805	22.74/0.688
EDSR [16]	4×	38.9M	26.28/0.794	25.40/0.780	29.17/0.837	23.38/0.728
RDN [18]	4×	22.0M	26.25/0.794	25.39/0.780	29.17/**0.838**	23.39/0.729
RCAN [29]	4×	15.4M	26.38/0.796	25.55/0.783	**29.22**/0.837	**23.40**/0.728
PASSRnet [1]	4×	1.42M	26.26/0.791	25.42/0.776	28.62/0.822	23.21/0.717
SRRes+SAM [2]	4×	1.73M	26.36/0.795	25.56/0.781	28.78/0.828	23.18/0.721
IMSSRnet* [5]	4×	6.89M	26.44/–	25.59/–	29.02/–	–/–
CVCnet [4]	4×	0.99M	26.35/0.794	25.55/0.780	28.65/0.823	23.22/0.719
iPASSR [3]	4×	1.42M	26.49/0.798	25.63/0.784	29.09/0.835	23.37/0.729
MS-PFINet (Ours)	4×	1.45M	**26.52/0.800**	**25.65/0.786**	29.04/0.834	23.39/**0.730**

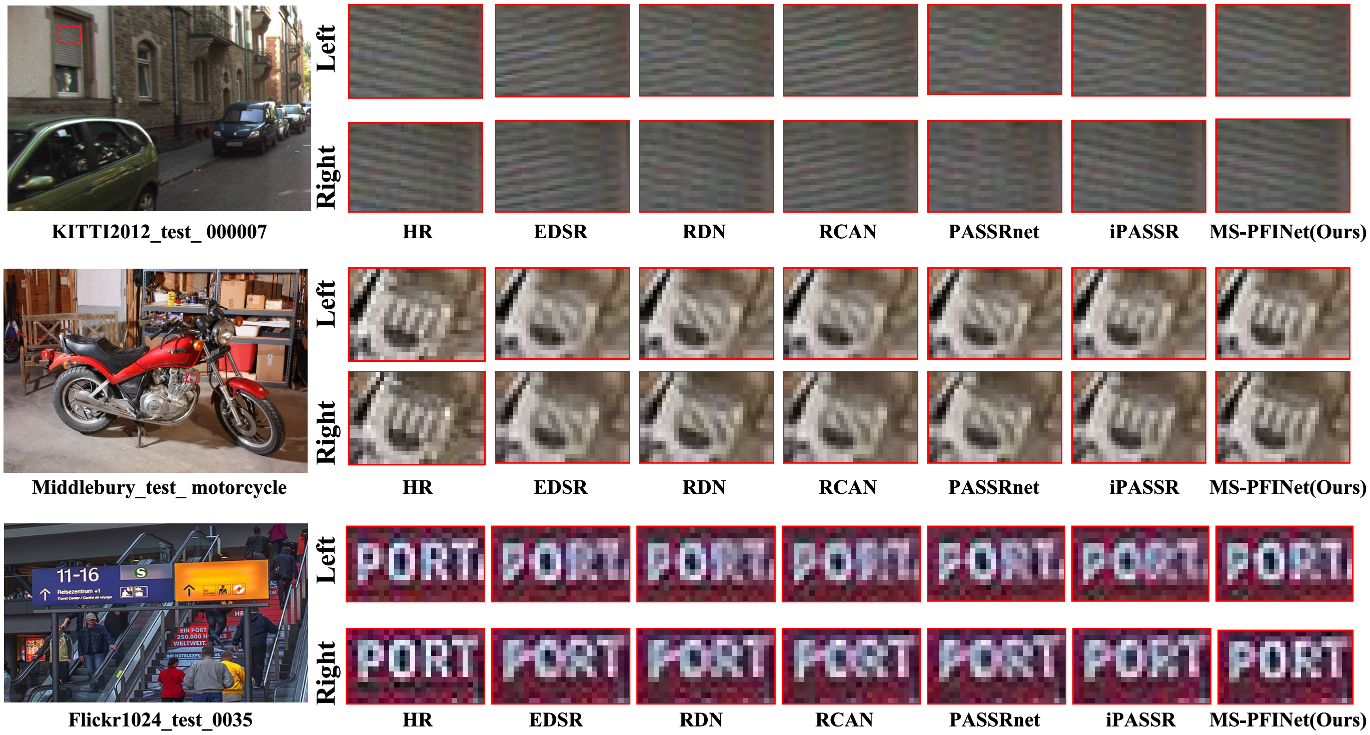


Fig. 4. Visual comparison on the KITTI2012, Middlebury and Flickr1024 datasets for 2× SR.

TABLE II

THE FLOPS AND EXECUTION TIME FOR 4× SR. FLOPS ARE CALCULATED ON A PAIR OF STEREO IMAGE OF SIZE 128 × 128, WHILE TIME/PSNR/SSIM VALUES ARE ACHIEVED ON THE KITTI 2012 DATASET

Methods	Flops-I (G)	Flops-T (G)	Time-I (sec)	Time-T (sec)	KITTI2012 (PSNR/SSIM)
SRRes+SAM [2]	3.25	79.30	0.015	0.215	26.36/0.795
iPASSR [3]	10.28	47.84	0.026	0.134	26.49/0.798
MS-PFINet (Ours)	4.06	51.63	0.018	0.168	26.52/0.800

that by mining features and progressively performing interaction at more scales, more reliable and comprehensive intra-/cross-view information can be utilized to improve the quality of reconstruction. However, at the same time, the number of parameters also increases and when the number of branches exceeds 3, the network does not provide further consistent improvement. As a result, to strike the balance between the cost and reconstruction performance, we choose '3-5-7' as the final model for 2×, 4× SR.

Secondly, to better demonstrate the effectiveness of such multi-branch structure with different scales, the experiment '3-3-3' is conducted, which uses convolution layers with the same kernel size 3×3 on all branches to extract a single-scale

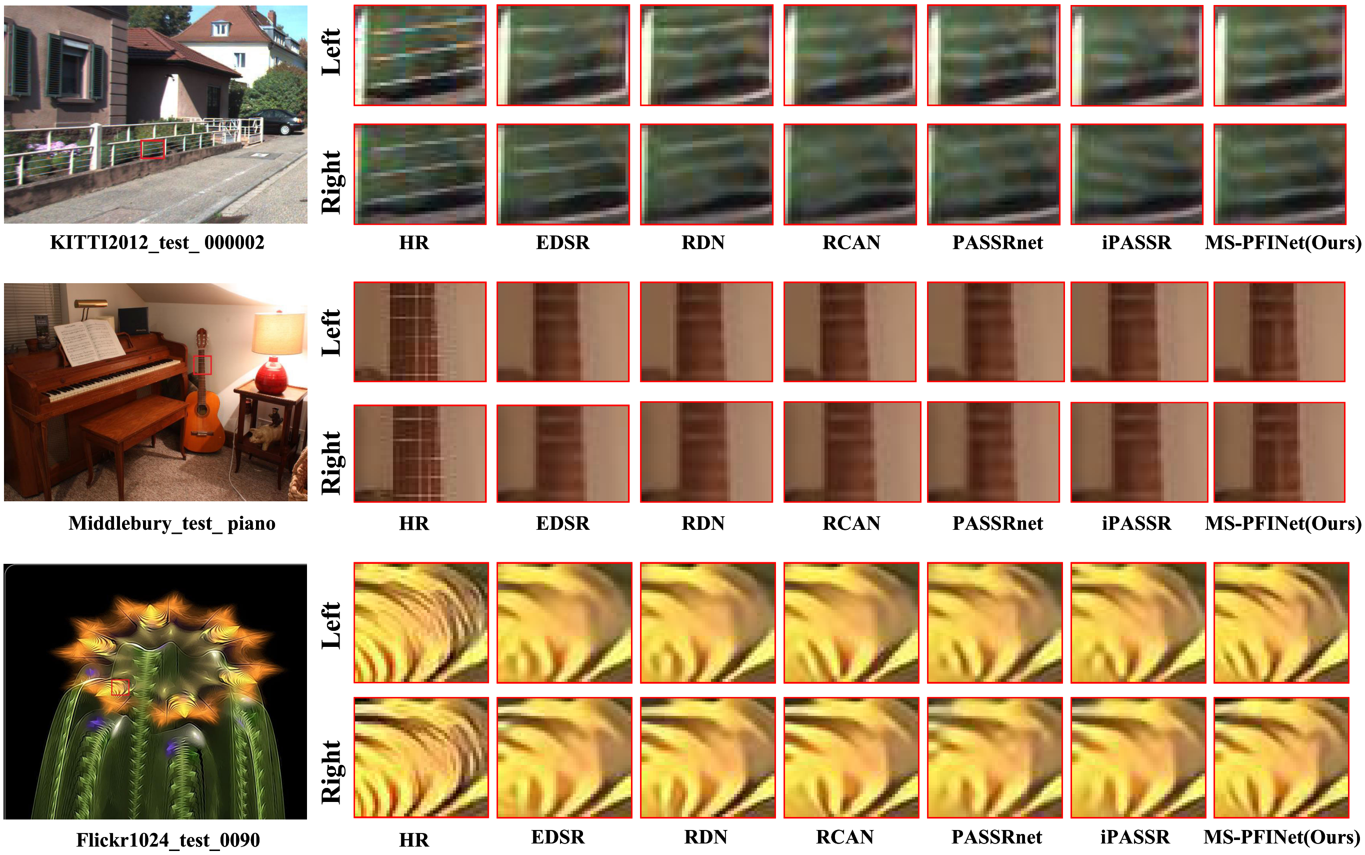


Fig. 5. Visual comparison on the KITTI 2012, Middlebury and Flickr1024 datasets for 4× SR.

TABLE III

THE EFFECTIVENESS OF MULTI-BRANCH STRUCTURE ON BENCHMARK DATASETS FOR 4× SR

Methods	Params (M)	KITTI2012 (PSNR/SSIM)	KITTI2015 (PSNR/SSIM)	Middlebury (PSNR/SSIM)	Flickr1024 (PSNR/SSIM)
3	0.59	26.28/0.793	25.43/0.778	28.79/0.827	23.24/0.722
3-5	1.05	26.46/0.798	25.59/0.785	28.97/0.831	23.35/0.729
3-5-7	1.45	26.52/0.800	25.65/0.786	**29.04/0.834**	23.39/0.730
3-5-7-9	1.81	**26.54/0.800**	**25.67/0.787**	29.01/**0.834**	**23.40/0.731**
3-3-3	1.44	26.48/0.800	25.61/0.786	29.01/0.832	23.37/0.730

TABLE IV

THE COMPARISON RESULTS WITH OTHER METHODS FOR 4× SR

Methods	KITTI2012 (PSNR/SSIM)	KITTI2015 (PSNR/SSIM)	Middlebury (PSNR/SSIM)	Flickr1024 (PSNR/SSIM)
CONCAT	25.16/0.753	24.49/0.735	27.25/0.780	22.18/0.659
SAM [2]	26.42/0.796	25.57/0.782	28.96/0.830	23.31/0.726
CPAM [40]	26.29/0.791	25.46/0.777	28.85/0.824	23.18/0.717
FTM	**26.49/0.799**	**25.61/0.784**	**29.00/0.833**	**23.37/0.729**

feature from the input image. And it has a similar parameter number with ‘3-5-7’. The quantitative results are presented in Table III. From the table, we can observe that the performance of ‘3-3-3’ degraded compared with ‘3-5-7’. In addition, we also give the visual comparison between ‘3-3-3’ and ‘3-5-7’ in Fig. 6. It can be seen that ‘3-5-7’ can reconstruct sharper and clearer edges than ‘3-3-3’, having an obviously better visual perception effect. From the experimental results above, we can draw a conclusion that multi-scale has better results than single-scale because the multi-scale information processing mechanism is in line with the visual characteristics of human eyes.

TABLE V

ABLATION STUDIES ON FTM FOR 4× SR

Methods	Patch size (P×P)	HA	SA	Mask	KITTI2015 (PSNR/SSIM)	Middlebury (PSNR/SSIM)
Base	3×3	✓			25.57/0.782	28.92/0.830
Base+SA	3×3	✓	✓		25.61/0.786	29.01/0.834
Base+SA+Mask	3×3	✓	✓	✓	**25.65/0.786**	**29.04/0.834**
Base+SA+Mask	1×1	✓	✓	✓	25.62/0.786	29.01/0.833

TABLE VI

THE EFFECT OF APPLYING FTM ON MULTIPLE SCALES FOR 4× SR

Methods	KITTI2012 (PSNR/SSIM)	KITTI2015 (PSNR/SSIM)	Middlebury (PSNR/SSIM)	Flickr1024 (PSNR/SSIM)
0	26.26/0.790	25.45/0.776	28.80/0.824	23.15/0.716
1	26.40/0.796	25.56/0.783	28.92/0.830	23.29/0.725
2	26.47/0.799	25.61/0.787	29.01/0.832	23.36/0.729
3	**26.52/0.800**	**25.65/0.786**	**29.04/0.834**	**23.39/0.730**

2) Feature Transformer Module (FTM): Firstly, to investigate the advantages of FTM compared with other cross-view interaction methods, we replace the FTMs in the whole network with CONCAT, SAM [2] and CPAM [40], respectively and all the networks are trained with only SR loss for fair comparison. Therein, CONCAT means directly concatenating left view features with right view features to obtain complementary information. SAM exploits the weighted sum of all pixels along the horizontal epipolar line as the corresponding information. CPAM captures the global correspondence of complementary information for each view rather than the horizontal correspondence. From Table IV, we can find that our proposed FTM exceeds other methods by a large margin on

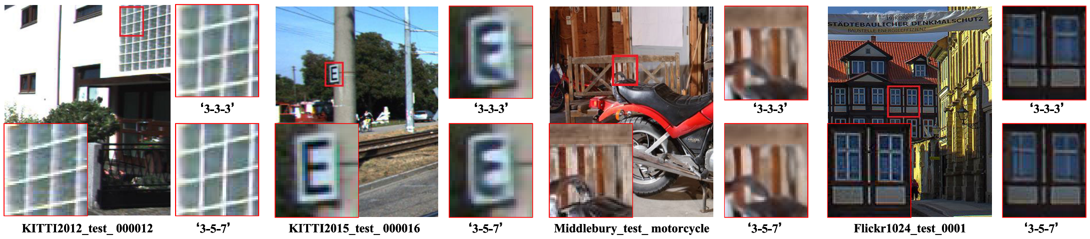


Fig. 6. Visual comparison between '3-3-3' and '3-5-7' on benchmark datasets for 4× SR.

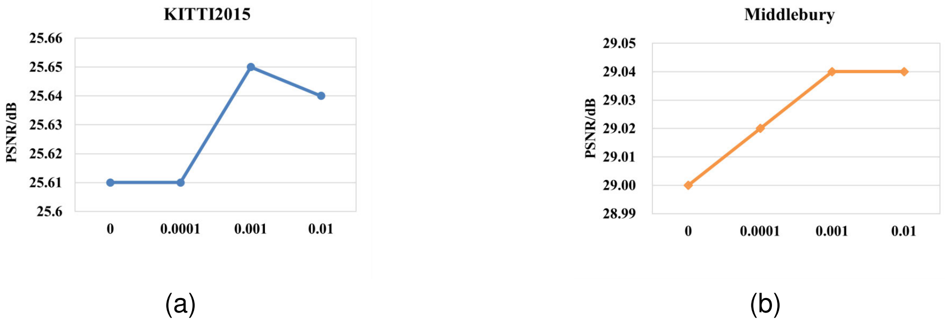


Fig. 7. Results of our proposed method with different λ on (a) KITTI2015 and (b) Middlebury dataset dataset.

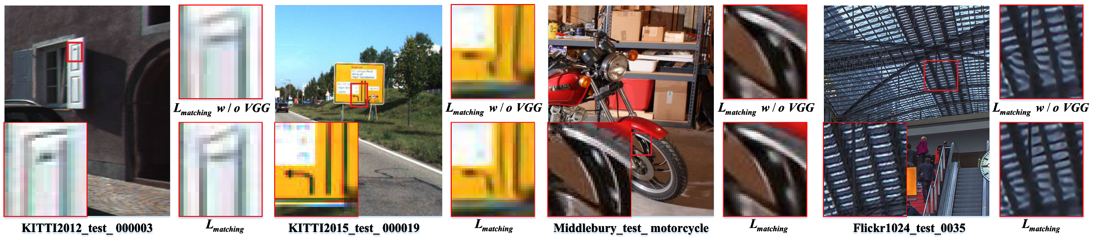


Fig. 8. Visual comparison between $L_{matching}$ and $L_{matching}$ w/o VGG.

all datasets, which illustrates that only transferring the most relevant features to the target view helps to reconstruct SR images with more accurate and proper texture information and obtain performance gain.

Secondly, to verify the effect of each component in FTM, we modify FTM with only hard attention (HA) for feature transfer as baseline, then gradually add soft attention (SA) and occlusion handling (Mask) to it. As shown in Table V, with the involvement of each component, the performance of the network is gradually improved, the PSNR value is increased from 25.57 to 25.65 on KITTI2015 dataset and increased from 28.92 to 29.04 on the Middlebury dataset. This fully explains that by paying attention to the importance of transferred features at different position and removing the features of occluded regions can make the feature transfer more effective and accurate, assisting the generation of SR images. In addition, to testify the advantage of using patches rather than pixels for relevance estimation, we replace the patch 3×3 with 1×1, it can be seen the network suffers a decrease, which demonstrates that patch-wise similarity does improve the accuracy of feature transfer.

Thirdly, to illustrate the merits of applying cross-view interaction on multiple scales, we gradually add an FTM on each branch. The results are shown in Table VI. "0" represents that there is no interaction between the left and right stream, the features from each branch are directly fused by the FB, which is equivalent to the single image super-resolution. "1" represents that we only perform cross-view interaction once before image reconstruction and the FTM is on the first branch (the first convolution layer on the branch is 3×3). "2" represents that there is an FTM on the first and second branch (the first convolution layer on the branch is 5×5) respectively. "3" represents every branch is equipped with an FTM. It's obvious that without any interaction, the network suffers an average decrease of 0.24dB in PSNR on the four datasets as compared to our proposed network. And as FTM is applied on more branches, the performance of the network gradually improves, which corroborates that by transferring features with different receptive fields at different scales, cross-view information can be fully utilized, thereby improving the quality of reconstruction.

3) Perceptual Texture Matching Loss: Firstly, to demonstrate the validity of our proposed perceptual texture matching loss and select the proper hyperparameter λ, we retrain our network with $\lambda = 0,\ 0.0001,\ 0.001$ and 0.01 respectively. The PSNR values on KITTI2015 and Middlebury for 4× SR are presented in Fig.7. We find that the highest values are obtained when λ is set to 0.001. $\lambda = 0$ refers to the network trained

TABLE VII

THE EFFECT OF VGG FEATURE EXTRACTION PART ON BENCHMARK DATASETS FOR 4× SR

Methods	KITTI2012 (PSNR/SSIM)	KITTI2015 (PSNR/SSIM)	Middlebury (PSNR/SSIM)	Flickr1024 (PSNR/SSIM)
$L_{matching}\ w/o\ VGG$	26.49/0.799	25.63/0.785	29.02/0.831	23.38/0.730
$L_{matching}$	**26.52/0.800**	**25.65/0.786**	**29.04/0.834**	**23.39/0.730**

only with SR loss. It can be observed that the performance declines if the perceptual texture matching loss is removed, which shows that according to supervise the accuracy of hard attention and soft attention maps, features can be transferred in a more effective way.

Secondly, to verify the merit of calculating the matching loss ($L_{matching}$) using textures in feature space rather than textures in pixel space, we removed the VGG feature extraction part in perceptual texture matching loss (named as $L_{matching}\ w/o\ VGG$) and retrained our network, the results are listed in Table VII and the visual comparison is also presented in Fig.8. We can observe that, with the VGG feature extraction part, the PSNR is only improved a little, but the visual quality has been greatly improved. This fully demonstrates that by calculating the matching loss using textures in feature space, our FTM can focus on the perceptually similarity features for transfer, which is conducive to obtaining more visually satisfying SR results.

V. CONCLUSION

In this paper, we propose a multi-scale visual perception based progressive feature interaction network (MS-PFINet) for stereo image SR. To be specific, based on the multi-scale information processing mechanism of HVS, we carefully design a two-stream network with multi-branch structure. With different scales on each branch, it can not only capture multi-scale intra-view information but also obtain reliable and comprehensive cross-view through progressive feature interaction for stereo image reconstruction. Moreover, to avoid the impact of irrelevant information transferred from another view, we bring up a feature transformer module (FTM) to transfer the most relevant textures in one view as the assistance of another view. Finally, for optimization, we propose a perceptual texture matching loss to additionally supervise the learning of FTM, making our network transfer the complementary features more effectively. Extensive experiments on benchmark datasets have demonstrated that the proposed MS-PFINet achieves better performance than other state-of-the-art SR methods in terms of both quantitative metrics and visual quality. However, as with other stereo image SR methods, this paper also uses bicubic interpolation to generate LR images. Therefore, it does not have good generalization ability in real-world stereo SR due to the large domain gap between real-world and such synthetic LR images. In the future, we will explore how to improve the network's ability to reconstruct real scene stereo image.

REFERENCES

[1] L. Wang et al., "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12242–12251.

[2] 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, 2020.

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

[4] 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, 2022.

[5] 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.

[6] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[7] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proc. Roy. Soc. London B, Biol. Sci.*, vol. 204, no. 1156, pp. 301–328, May 1979.

[8] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[9] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[10] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[11] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. Multimedia*, vol. 21, no. 12, pp. 3106–3121, Dec. 2019.

[12] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, Oct. 2021.

[13] J. Li, Z. Pei, and T. Zeng, "From beginner to master: A survey for deep learning-based single-image super-resolution," 2021, *arXiv:2109.14335*.

[14] 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.

[15] 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. (CVPR)*, Jun. 2016, pp. 1646–1654.

[16] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

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

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

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

[20] F. Li, H. Bai, and Y. Zhao, "FilterNet: Adaptive information filtering network for accurate and fast image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1511–1523, Jun. 2020.

[21] Y. Liu, Q. Jia, X. Fan, S. Wang, S. Ma, and W. Gao, "Cross-SRN: Structure-preserving super-resolution network with cross convolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 4927–4939, Aug. 2022.

[22] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

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

[25] Y. Hu, X. Gao, J. Li, Y. Huang, and H. Wang, "Single image super-resolution via cascaded multi-scale cross network," 2018, *arXiv:1802.08808*.

[26] J. Qin, Y. Huang, and W. Wen, "Multi-scale feature fusion residual network for single image super-resolution," *Neurocomputing*, vol. 379, pp. 334–342, Feb. 2020.

[27] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, "MDCN: Multi-scale dense cross network for image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2547–2561, Jul. 2021.

[28] H. Wu et al., "Multi-grained attention networks for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 512–522, Feb. 2021.

[29] 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.*, 2018, pp. 286–301.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] 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.

[32] J. Zhang et al., "A two-stage attentive network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1020–1033, Mar. 2022.

[33] X. Zhu, K. Guo, S. Ren, B. Hu, M. Hu, and H. Fang, "Lightweight image super-resolution with expectation-maximization attention mechanism," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1273–1284, Mar. 2022.

[34] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1721–1730.

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

[36] J. Dan, Z. Qu, X. Wang, and J. Gu, "A disparity feature alignment module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 28, pp. 1285–1289, 2021.

[37] 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. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1985–1993.

[38] W. Song, S. Choi, S. Jeong, and K. Sohn, "Stereoscopic image super-resolution with stereo consistent feature," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12031–12038.

[39] 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, 2021.

[40] C. Chen, C. Qing, X. Xu, and P. Dickinson, "Cross parallax attention network for stereo image super-resolution," *IEEE Trans. Multimedia*, vol. 24, pp. 202–216, 2022.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[42] Y. Wang, L. Wang, J. Yang, W. An, and Y. Guo, "Flickr1024: A large-scale dataset for stereo image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1–6.

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

[44] 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.*, Jun. 2012, pp. 3354–3361.

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

[46] R. Timofte, R. Rothe, and L. Van Gool, "Seven ways to improve example-based single image super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1865–1873.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[48] 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. (CVPR)*, Jul. 2017, pp. 5835–5843.

Anqi Liu received the master's degree from the School of Electrical and Information Engineering, Tianjin University, where she is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. Her research interests include single/stereo image super-resolution, neural networks, and deep learning.

Sumei Li (Member, IEEE) received the Ph.D. degree from Nankai University, Tianjin, China. Since 2006, she has been an Associate Professor with the Communication Engineering Department, Tianjin University. She chaired the National 863 Project, the National Natural Science Foundation, and the Key Fund in Tianjin. Her research interests include 3D image/video transmission, processing and quality evaluation, depth/image SR reconstruction, sparse representation, neural networks, and deep learning. She is a member of the China's Neural Network Committee.

Yongli Chang received the master's degree from the School of Electrical and Information Engineering, Tianjin University, in 2019, where she is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. Her research interests include 2D/3D image quality evaluation, neural networks, deep learning, and signal sparse representation.

Yonghong Hou (Member, IEEE) received the B.Eng. degree in electronic engineering from Xidian University, Xi'an, China, in 1991, and the M.Eng. and Ph.D. degrees in communication and information systems from Tianjin University, Tianjin, China, in 2003 and 2009, respectively. He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include computer vision, artificial intelligence, and multimedia signal processing.