Parallax-based second-order mixed attention for stereo image super-resolution

Chenyang Duan | Nanfeng Xiao

School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

Correspondence
Chenyang Duan, School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China.
Email: chenyang.duan@outlook.com

Funding information
Basic and Applied Basic Research of Guangdong Province, Grant/Award Number: 2015A030300810; Public Research and Capacity Building of Guangdong Province, Grant/Award Number: 2014B010104001

1 | INTRODUCTION

Image super-resolution (SR) aims at reconstructing a visually high-resolution (HR) image from its degraded low-resolution (LR) counterparts, which is a fundamental problem in low-level vision tasks. However, this is typically an ill-posed problem since infinite HR images can map to any LR input. In order to address this problem, there are plenty of convolutional neural network (CNN)-based SR methods that have been proposed [1–7], which construct a well-designed network to learn the mapping between LR and HR images. Recently, with the rapid development of the stereoscopic 3D industry, stereoscopic image processing techniques have received widespread attention in computer vision fields. Therefore, more stereo SR methods, which use multiple LR images such as stereo image pairs as input, have been advanced to obtain enhanced performance [8–12].

Finding local correspondences between stereo image pairs is central to stereo matching [13–16]. The recent methods [17–20] utilise 3D or 4D cost volumes in their convolutional neural network to measure the dissimilarity. Although these methods can effectively improve disparity estimation, they are not suitable for stereo image SR. This is because 4D cost volumes-based methods [17, 18] take up unbearable memory and computing resources. Although 3D cost volumes-based methods [19, 20] improve the computation efficiency by reducing cost volumes from 4D to 3D, both 3D and 4D cost volume techniques preset a maximum disparity value and cannot handle stereo image pairs with disparities larger than this value.

Making full use of the cross-view information in stereo image pairs is extremely challenging for stereo image SR. Recent methods have adopted different algorithms to capture the correspondence between stereo images. Specifically, Bhavsar and Rajagopalan [8, 9] and Park et al. [10] used stereo block matching to obtain pixel correspondences. Then the correspondences were used to register the stereo input image pair. These methods try to use the per-pixel disparity from stereo matching, but they only use the discrete estimation of parallax. Jeon et al. proposed StereoSR [11] to directly learn the end-to-end mapping from an...
LR image to an HR image, which avoids using the discrete disparity from stereo matching. In [11], the cross-view information is obtained from the image stack which is generated by concatenating the left image with 64 draft images obtained by shifting the right image by different pixels. Recently, Wang et al. proposed a parallax-attention stereo super-resolution network (PASSRnet) [12] to integrate the information from two stereo images to achieve state-of-the-art SR performance. In [12], the cross-view information is obtained from a parallax attention mechanism. However, the inherent feature correlations of the intermediate layers are ignored in their methods, thus reducing the representation ability of their network.

In this work, we propose a parallax-based second-order mixed attention stereo SR network (PSMASSRnet) for more powerful feature correlation and feature expression. In particular, we propose a parallax-based second-order mixed attention module (PSMAM) for better feature correlation, which not only exploits the second-order statistics of features instead of the first-order ones but also combines the second-order channel features and spatial features. Such PSMAM can adaptively learn feature interdependencies and make our network focus on more informative features. Moreover, in order to extract rich high-frequency details for PSMAM, we propose an efficient feature extraction module named dense cross- atrous spatial pyramid pooling (ASPP) module, which can effectively explore the local and multi-scale features with different dilation rates to extract more discriminative features with fewer parameters and less execution time. Finally, we adopt a dual regression scheme [21] to improve the SR performance by employing an additional constraint on the LR data to reduce the space of the possible mapping from LR to HR images. Comparative experiments demonstrate that compared with state-of-the-art single image SR and stereo image SR methods, our method can recover more image details and achieve superior visual quality.

In summary, the main contributions of this work are listed as follows:

We propose a novel PSMASSRnet for better handling stereo image SR. This model can effectively use cross-view information while maintaining the superiority of intra-view information exploitation.

We propose a dense cross-ASPP module for feature extraction, which can make full use of hierarchical features to effectively detect local and multi-scale features with different dilation rates to achieve competitive performance with fewer parameters and less execution time.

We propose a parallax-based second-order mixed attention mechanism, which considers both second-order channel features and spatial features to focus on more informative features and adaptively learn feature interdependencies.

2 | RELATED WORK

Because the quality of recovered SR images has a seriously influence on the accuracy of high-level computer vision tasks such as object detection [22, 23], image segmentation [24, 25], and image classification [26, 27], stereo image SR has become more and more popular in recent years.

2.1 | Super-resolution

2.1.1 | Single image super-resolution

Single image SR methods, which are widely adopted by researchers, use a single LR image to recover a HR image. Before the recent explosive development of deep learning techniques, a large number of traditional algorithms have been proposed, such as sparse-representation-based algorithms [28, 29], patch-based algorithms [30–32] and statistics-based algorithms [33, 34]. In recent years, the great potential of deep learning techniques in the field of computer vision has attracted more and more researchers to adopt CNN-based algorithms to solve the single image SR problem. Dong et al. [67] first proposed a super-resolution convolutional neural network (SRCNN) network which only uses three convolutional layers but achieves performance that far exceeds traditional algorithms. Zhang et al. [35] proposed a deep and wide residual network which adopts multiple revised residual blocks. Shi et al. [36] proposed a sub-pixel upsampling algorithm to reduce computational complexity and improve efficiency. Recently, Guo et al. [37] proposed dual-view attention networks which adopt global aware attention and local aware attention for a HR feature space and LR feature space, respectively. In order to allow the low-level layers to utilise the information passed from the high-level layers, Zhang et al. [5] designed a feedback network image super-resolution feedback network and improved the SR performance. Tian et al. [38] proposed a coarse-to-fine convolution neural network to collect more complementary contextual information and achieved better results.

2.1.2 | Stereo image super-resolution

Since the information contained in a single image is extremely limited, stereo image SR algorithms can solve this problem and improve the SR performance. Jeon et al. [11] proposed the StereoSR to directly learn an end-to-end mapping from an LR image to an HR image, which avoids using discrete disparity from stereo matching. Recently, Wang et al. [12] proposed a PASSRnet to integrate the information from two stereo images to achieve state-of-the-art SR performance. Song et al. [39] designed an efficient network to enforce the stereo-consistency constraint; they combined the self-attention mechanism with the parallax-based attention mechanism to make full use of the consistency of left and right-view textures. Yan et al. [40] proposed a unified stereo image restoration framework to effectively explore the inherent pixel correspondence between stereo images. Ying et al. [41] proposed a generic stereo attention module (SAM) and used multiple SAMs to extend any single image SR network to stereo images. Mei et al. [42] proposed a cross-
scale non-local (CS-NL) attention module and integrated it into a recurrent neural network to explore the cross-scale feature correlation of images. Jiang et al. [43] proposed a hierarchical dense recursive network to improve reconstruction performance and efficiency. However, their methods only explore the first-order statistics of features which hinders the representational ability of the network. Our module can use the second-order statistics of features to learn better feature correlation.

2.2 Attention mechanism

It is well-known that the human visual attention mechanism can quickly scan a global image to obtain the target area that needs to be focused on. In accordance with this, lots of researchers have introduced attention processing into their methods to improve the performance of CNNs for various vision tasks, such as face recognition [44, 45], object detection [46, 47] and image classification [25, 48].

Xu et al. [49] employed the attention mechanism to quickly detect important features of sparse data for image caption, which is the first time the attention mechanism has been introduced in computer vision. Wang et al. [50] proposed a residual attention network for image classification, which is built by stacking multiple attention modules which generate attention-aware features. Wang et al. [12] introduced a PASSRnet, which adopts a parallax-attention module (PAM) to capture the long-range dependency in stereo images, to improve the SR performance of stereo images. However, the attention mechanisms used by these methods only focus on the spatial characteristic of image features. Duan et al. [51] proposed a parallax-based spatial and channel attention module to effectively combine spatial and channel-wise characteristics to address the stereo correspondence problem. Wang et al. [52] presented a symmetric bidirectional parallax attention module which can effectively deal with the occluded regions. Chen et al. [53] proposed a cross parallax attention module (CPAM) to explore the global correspondence of additional information for each image. Moreover, such a CPAM can solve the problem of different types of epipolar lines and the problem of large disparity. Wang et al. [54] presented a generic parallax attention mechanism which combines epipolar constraints with the attention mechanism to capture stereo correspondence with large disparity variations.

However, most of the existing attention mechanisms only detect the first-order statistics of features and ignore the feature statistics higher than first-order, which hinders the discriminative ability of the network. Recently, Dai et al. [55] proposed a second-order attention network for SR. In [55], a novel second-order channel attention (SOCA) mechanism is proposed to adaptively learn feature interdependencies. Inspired by the SOCA, we propose a PSMAM to obtain all the similar features along the epipolar between the stereo image pair by exploring second-order channel features and spatial features.

3 METHODS

3.1 Network architecture

In this work, we propose a PSMASSRnet. As shown in Figure 1, PSMASSRnet can be divided into four stages: feature extraction stage, feature fusion stage, deep fusion stage and reconstruction stage. In the feature extraction stage, we use a convolution and a residual block for shallow feature extraction, N dense cross ASPP blocks for deep feature extraction and a hierarchical feature distillation block (HFDB) which is proposed by [56] for hierarchical features’ distillation. Note that all the parameters are shared by the stereo images in the feature extraction stage. In the feature fusion stage, we introduce a PSMAM to learn stereo correspondence and integrate the features from a stereo image pair. In the deep fusion stage, we use M residual blocks to deeply fuse the information from stereo image pairs. In the reconstruction stage, we employ the sub-pixel layer to reconstruct HR images.

Define \( L_{\text{left}}^{\text{input}} \) and \( L_{\text{left}}^{\text{SR}} \) as the input left image and the output of PSMASSRnet. \( L_{\text{left}}^{\text{input}}, L_{\text{left}}^{\text{output}} \) and \( L_{\text{left}}^{\text{mix}} \) are the input of the first dense cross ASPP block, the output of the last dense cross ASPP block, and the input of PSMAM, respectively. We first use a 3 \( \times \) 3 convolutional layer and a residual block to upgrade the LR image to a high-dimensional image and obtain its shallow features.

\[
L_{\text{left}}^{\text{input}} = F_{\text{rel}}\left(C_{3 \times 3}\left(L_{\text{left}}^{\text{LR}}\right)\right)
\]

where \( C_{3 \times 3} \) and \( F_{\text{rel}}(...) \) stand for the corresponding convolution layer and residual block, respectively. Then, \( L_{\text{left}}^{\text{input}} \) is fed to the dense cross ASPP block for local feature extraction.

\[
L_{\text{left}}^{\text{output}} = F_N\left(L_{\text{left}}^{\text{input}}\right)
\]

where \( F_N(...) \) denotes the \( N \) dense cross ASPP blocks. Similar to [56], we also use the HFDB for hierarchical feature fusion and distillation.

\[
L_{\text{left}}^{\text{HFDB}} = F_{\text{HFDB}}\left([L_{\text{left}}^{\text{input}}, L_{\text{left}}^{\text{output}}, \ldots, L_{\text{left}}^{\text{mix}}]\right)
\]

where \( F_{\text{HFDB}}(...) \) and \( L_{\text{left}}^{n} \) represent the HFDB and the output of the \( n \)-th dense cross ASPP block, respectively. And \([\ldots]\) denotes the concatenation operation. After that, we combine the extracted shallow features \( L_{\text{left}}^{\text{input}}, \) the deep features \( L_{\text{left}}^{\text{output}}, \) and the distilled hierarchical features \( L_{\text{left}}^{\text{HFDB}} \) as one input of PSMAM.

\[
L_{\text{left}}^{\text{mix}} = L_{\text{left}}^{\text{input}} + L_{\text{left}}^{\text{output}} + L_{\text{left}}^{\text{HFDB}}
\]
Similarly, we use the right image as the initial input and repeat the above processes to obtain $L_{\text{mix}}$ and $L_{\text{right}}$ as another input of PSMAM

$$L_{\text{PSMAM}} = F_{\text{PSMAM}}(L_{\text{mix}}, L_{\text{right}})$$ \hspace{1cm} (5)

where $F_{\text{PSMAM}}(\cdot)$ denotes the PSMAM. Finally, the integrated features $L_{\text{PSMAM}}$ are used to obtain the left SR image $I_{\text{left}}^\text{SR}$:

$$I_{\text{left}}^\text{SR} = F_{\text{sub}}(F_{\text{RG}}(L_{\text{PSMAM}}))$$ \hspace{1cm} (6)

where $F_{\text{RG}}(\cdot)$ represents the residual group which consists of $M$ residual blocks and $F_{\text{sub}}(\cdot)$ denotes the sub-pixel layer.

### 3.2 Dense cross ASPP block

Feature extraction is the most crucial operation in image restoration. The ASPP module is proposed by [57] to detect multi-scale information. However, the local features are hard to transfer to other layers since the ASPP module does not make the most use of the image features of the previous layers. Therefore, we propose a dense cross ASPP block to extract multi-scale features as well as make full use of the features of the previous layers.

As shown in Figure 2, the dense cross ASPP block consists of three dense networks. In order to better introduce the structure of the dense cross ASPP block, we remove the residual learning and break down the structure of the model in Figure 3. As shown in Figure 3, the module can be separated into three parts, namely DenseNet-Top, DenseNet-Middle and DenseNet-Bottom. Each subnet uses the same network structure with different dilation rates. And each subnet uses a feed-forward way to connect each layer to every other layer (black lines in Figure 3), which can effectively utilize the rich features of the previous layers. Take DenseNet-Top as an example. We can obtain the output $L_{\text{out}}^T$ by

$$T_{22} = C_{d=1}(T_{11})$$ \hspace{1cm} (7)

$$T_{33} = C_{d=1}^{2}(C_{1 \times 1}^{0}([T_{12}, T_{22}]))$$ \hspace{1cm} (8)

$$L_{\text{out}}^T = C_{1 \times 1}^{1}([T_{13}, T_{23}, T_{33}])$$ \hspace{1cm} (9)

where $C_{d=1}^{n}$ represents the $n$–depth $3 \times 3$ convolutional layer with the dilation rate of 1, which aims at feature extraction, and $C_{1 \times 1}^{0}$ denotes the $n$–depth $1 \times 1$ convolutional layer, which is used for feature fusion. In DenseNet-Middle and DenseNet-Bottom, we only need to replace the dilation rate of all $3 \times 3$ convolutional layers from 1 to 4 and 8

$$M_{22} = C_{d=4}^{1}(M_{11}), B_{22} = C_{d=8}^{1}(B_{11})$$ \hspace{1cm} (10)
where $C_{d=4}^3$ represents the $n$–depth $3 \times 3$ convolutional layer with the dilation rate of four and $C_{d=8}^3$ represents the $n$–depth $3 \times 3$ convolutional layer with the dilation rate of eight.

In order to transfer and fuse the multi-scale image features, we adopt skip connection to associate the three subnets (red lines in Figure 3). For example, $R_{14}$ and $R_{18}$ transmit the rich features extracted by DenseNet-Top to DenseNetMiddle and DenseNet-Bottom, respectively. Finally, a $1 \times 1$ convolutional layer is adopted at the end of the
module to fuse the multi-scale features and reduce the dimensions. Therefore, the final operations of the module can be defined as

\[ T_{22} = C_{d=1}^1(T_{11}), \quad M_{22} = C_{d=4}(M_{11}), \]
\[ B_{22} = C_{d=8}(B_{11}) \]

(15)

\[ T_{33} = C_{d=1}^2(C_{1\times 1}([T_{12}, T_{22}, R_{41}, R_{84}])) \]

(16)

\[ M_{33} = C_{d=4}^2(C_{1\times 1}([M_{12}, M_{22}, R_{14}, R_{48}])) \]

(17)

\[ B_{33} = C_{d=8}^2(C_{1\times 1}([B_{12}, B_{22}, R_{18}, R_{48}])) \]

(18)

\[ L_{out} = C_{1\times 1}^1([T_{23}, T_{33}, M_{23}, M_{33}, B_{23}, B_{33}, L_{in}]) \]

(19)

where \( L_{in} = T_{11} = M_{11} = B_{11} = T_{12} = M_{12} = B_{12} = T_{13} = M_{13} = B_{13}, \)
\( T_{22} = T_{23} = R_{14} = R_{18}, \)
\( M_{22} = M_{23} = R_{41} = R_{48}, \)
\( B_{22} = B_{23} = R_{81} = R_{84}, \)
which denotes the same features passed to different convolutional layers. \( L_{in} \) and \( L_{out} \) stand for the input and output of our module, respectively.

3.3 Parallax-based second-order mixed attention module

Most existing methods only utilise the first-order statistics of features in their attention mechanism, which hinders the discriminative ability of their model. Moreover, in a recent article [58], it was proven that the second-order statistics of features can extract more discriminative features than the first-order ones. Therefore, combined with the PAM introduced by [12], we propose a PSMAM, which simultaneously explores the second-order channel features and the spatial features. As shown in Figure 4, The PSMAM has two inputs, feature map \( L \) which is extracted by the left image and feature map \( R \) which is extracted by the right image. Take the feature map \( L \) as an example. In order to detect second-order information, we replace the traditional global average pooling with global covariance pooling in our model. The covariance matrix \( \Sigma \) can be obtained by

\[ \Sigma = XX^T \]

(20)

where \( X \) denotes the feature matrix with \( W \times H \) features of C-dimension, which is obtained by reshaping the feature map \( L \). And \( I = \frac{1}{WH} (I - \frac{1}{WH} 1) \), where \( I \) and \( 1 \) represent the \( WH \times WH \) identity matrix and matrix of all ones, respectively.

Because the obtained \( \Sigma \) is a symmetric positive semi-definite matrix, it can be decomposed into

\[ \Sigma = U\Lambda U^T \]

(21)

where \( U \) and \( \Lambda \) stand for orthogonal matrix and diagonal matrix with eigenvalues, respectively. Then the covariance normalisation can be expressed as

\[ \hat{Y} = \Sigma^a = U \Lambda^a U^T \]

(22)

where \( a = (1/2) \) in this work, since [58] has proven that \( a = (1/2) \) can obtain more discriminative representations. However, the GPU platform does not support eigenvalue decomposition well, which leads to low training efficiency. Therefore, as explored in [58], we also adopt the Newton-Shulz iteration [59] to accelerate the calculation of the covariance normalisation.

Let \( \hat{Y} = [\hat{y}_1, \ldots, \hat{y}_C] \), which has \( C \) feature maps with a size of \( C \). Then we apply global covariance pooling for each channel to obtain the channel-wise statistics \( V \in R^{C \times 1} \). The \( ctb \) dimension of \( V \) is calculated as

\[ V_c = F_{GCP}(y^c) = \frac{1}{C} \sum_{i=1}^{C} y^c(i) \]

(23)
where $F_{GCCP}()$ represents the global covariance pooling function. Then a simple sigmoid function is applied as a gating mechanism to exploit abundant feature interdependencies.

$$A = \delta(C_U R(C_D V))$$ (24)

where $\delta(.)$ and $R(.)$ refer to the softmax function and the Leaky ReLU function, respectively. $C_D$ represents the channel-downscaling convolution layer which can set the channel dimension of features to $C/r$ and $C_U$ denotes the channel upsampling convolution layer which can set the channel dimension of features to $C$. Then the channel-wise statistics $A$ is used to rescale the initial feature map $L$.

$$L_c = L \odot A$$ (25)

where $\odot$ represents the element-wise product operation.

Then a residual block and a convolution layer are employed to obtain the query feature map $L_c$.

$$L_c = C_{3 \times 3}(F_{re}(L_c))$$ (26)

where $F_{re}()$ and $C_{3 \times 3}()$ denote the residual block and the $3 \times 3$ convolution layer, respectively.

Similarly, we replace the initial input $L$ with $R$ and repeat the above process to produce the query feature map $R_c$. $L_c$ and $R_c$ are employed to produce the parallax attention map $M_{R \rightarrow L}$.

$$M_{R \rightarrow L} = \delta(L_c \cdot R_c)$$ (27)

where $\cdot$ refers to the matrix multiplication. The numerical value of $M_{R \rightarrow L}$ at position $(i, k, j)$ stands for the influence of the pixel at position $(i, j)$ in the right image on the pixel at position $(i, k)$ in the left image. Once the $M_{R \rightarrow L}$ is ready, $L$ and $R$ are swapped to obtain the $M_{L \rightarrow R}$ which is employed to generate the valid mask $V_{L \rightarrow R}$ to deal with the occluded regions [12].

$$V_{L \rightarrow R}(i, j) = \begin{cases} 1, \\ 0, \end{cases} \text{if} \sum_{k \in W} |M_{L \rightarrow R}(i, k, j)| > \tau$$ (28)

where $W$ represents the width of the image, $\tau$ is a threshold (0.1 in this work), which is utilised to determine whether each pixel is an occluded one. Note that the valid mask is only applied during the training process. Then, a weighted sum of features at all possible disparities $Q$ is obtained by

$$Q = C_{3 \times 3}(R) \cdot M_{R \rightarrow L}$$ (29)

Finally, $Q$, $V_{L \rightarrow R}$ and $L$ are stacked and fed to a $1 \times 1$ convolution layer for feature fusion.

$$O = C_{1 \times 1}([Q, V_{L \rightarrow R}, L])$$ (30)

where $C_{1 \times 1}$ denotes the convolution layer.

### 3.4 | Losses

In this work, in addition to employing the four losses (SR loss, photometric loss, smoothness loss and cycle loss) proposed by [12], we also introduce a regression loss to reduce the space of the possible mapping function from LR to HR images. The total loss function formula is

$$L = L_{SR} + \lambda (L_{PM} + L_S + L_C) + \gamma L_R$$ (31)

where $\lambda$ and $\gamma$ are set to 0.005 and 0.1, respectively, in our experiment. $L_{SR}$ denotes the mean square error loss.

$$L_{SR} = \|I_{LR}^{SR} - I_{LR}^{HR}\|^2_2$$ (32)

where $I_{LR}^{SR}$ represents the output of our network and $I_{LR}^{HR}$ represents the HR ground truth of the left image. $L_{PM}$ denotes the photometric loss.

$$L_{PM} = \sum_{p \in V_{LR \rightarrow right}} \| I_{LR}^{left}(p) - (M_{right \rightarrow left} \cdot I_{right}(p)) \|_1 + \sum_{p \in V_{right \rightarrow left}} \| I_{right}(p) - (M_{left \rightarrow right} \cdot I_{left}(p)) \|_1$$ (33)

where $p$ refers to a pixel with a valid mask value. $L_S$ denotes the smoothness loss.

$$L_S = \sum_{u} \sum_{i,j,k} \left( \| M(i,j,k) - M(i+1,j,k) \|_1 + \| M(i,j,k) - M(i,j+1,k+1) \|_1 \right)$$ (34)

where $M \in \{ M_{left \rightarrow right}, M_{right \rightarrow left} \}$, $L_C$ denotes the cycle loss.

$$L_C = \sum_{p \in V_{LR \rightarrow right}} \| M_{left \rightarrow right \rightarrow left}(p) - I(p) \|_1 + \sum_{p \in V_{right \rightarrow left}} \| M_{right \rightarrow left \rightarrow right}(p) - I(p) \|_1$$ (35)

where, $I$ is a stack of $H$ identity matrices and $M_{L \rightarrow R \rightarrow L}$ and $M_{R \rightarrow L \rightarrow R}$ can be produced by

$$M_{L \rightarrow R \rightarrow L} = M_{R \rightarrow L} \cdot M_{L \rightarrow R}$$ (36)

$$M_{R \rightarrow L \rightarrow R} = M_{L \rightarrow R} \cdot M_{R \rightarrow L}$$ (37)

$L_R$ denotes the regression loss.

$$L_R = \| F_{sub}(I_{LR}^{HR}) - I_{LR}^{HR} \|^2_2$$ (38)

where $F_{sub}()$ represents the subsampled function. The regression can not only reduce the space of the possible mapping function from LR to HR images but also deal with
real-world data. When the paired LR-HR data is not available in real-world applications, we can only use regression loss to train our model.

4 | EXPERIMENTS

4.1 | Dataset and evaluation metrics

For training, we followed [12] to employ the Flickr1024 dataset [60] and downsampled 60 Middlebury [61–63] images by a factor of 2 as training data for our experiments. For the test, we employed five images from the Middlebury dataset, 100 images from the KITTI2012 dataset [64] and 100 images from the KITTI2015 dataset [65] as our benchmark datasets. We evaluated the SR performance using two metrics, structural similarity index and peak signal-to-noise ratio (PSNR), where high means better.

4.2 | Implementation details

In the training stage, we first employed bicubic interpolation to subsampled HR images to generate LR images, and then these LR images were be divided into 30 × 90 patches with a stride of 20. Meanwhile, we also cropped their corresponding patches in HR images, which we used as the HR ground truth. In order to include the most disparities in our training dataset, the horizontal patch size was raised to 90. To further expand the data, these patches were flipped horizontally and vertically randomly.

Our network has been implemented with the Pytorch framework on an NVIDIA GTX 1080TI GPU. The module was optimised by the Adam [66] algorithm with β1 = 0.9, β2 = 0.999 and a batch size of 32. The learning rate was initialised as 2 × 10^-4 and then reduced to half every 30 epochs. The training process was stopped at 100 epochs since more epochs lead to an overfitting problem.

4.3 | Ablation study

In this section, we apply ablation experiments to demonstrate the effectiveness of a few major designs in our network.

4.3.1 | Dense cross ASPP block

Dense cross ASPP block is used in our network to detect multi-scale information. To test the effectiveness of dense cross ASPP block and investigate the impact of the hyper-parameter N (number of dense cross ASPP blocks), we use different numbers of dense cross ASPP blocks in our network. Note that when the number is 0 it means we remove all the dense cross ASPP blocks in our network, and to ensure fairness, we use several residual blocks to replace them.

4.3.2 | Parallax-based second-order mixed attention module

PSMAM is introduced to adaptively learn feature interdependencies. To verify the effectiveness of PSMAM, we remove it from our network and directly stack the feature map $L_{\text{left}}^{\text{mix}}$ and $L_{\text{right}}^{\text{mix}}$. Experiment result shows that the PSNR value declined from 25.12 to 24.95 dB. That is because PSMAM can effectively utilise the second-order statistics of features to extract more discriminative features.

4.3.3 | Deep fusion block

Deep fusion block is used to deeply fuse the information from stereo image pairs which is composed of M residual blocks. In order to investigate the impact of the hyper-parameter M, we use different numbers of residual blocks in our network. As shown in Table 2, the best performance can be achieved by using four residual blocks in our network. Therefore, we actually set $M = 4$ in our network.
4.3.4 Regression loss

Regression loss is introduced to reduce the space of the possible mapping function from LR to HR images. We conduct an ablation experiment to investigate the effect of the hyper-parameter $\gamma$ in Equation (31). From the comparative results shown in Table 3, we can observe that if we remove the regression loss from the total loss ($\gamma = 0$), the PSNR value will drop to 25.03 dB. Moreover, when we gradually increase $\gamma$ from 0.001 to 10, the regression loss becomes more and more important. It can provide powerful supervision to improve the SR performance at the beginning, and reach the peak at $\gamma = 0.1$, and then as $\gamma$ increases, it overwhelms other losses and hinders the final performance. Therefore, we set $\gamma = 0.1$ in practice.

4.3.5 PSMAM versus PAM

PAM only utilises the first-order statistics of features which hinders the discriminative ability of the network. PSMAM can simultaneously explore the second-order channel features and the spatial features. To verify the effectiveness of our PSMAM in stereo correspondence generation, we introduce a variant by replacing PSMAM with PAM in our network. Experimental results show that the PSNR value reduced from 25.12 to 25.09 dB if we replaced PSMAM with PAM in our network. The reason is that our PSMAM can use the second-order statistics of features to learn better feature correlation.

4.4 Comparison with state-of-the-art methods

We perform qualitative and quantitative comparisons on three benchmark datasets to evaluate the performance of our PSMASSRnet. For comparison, we choose a number of state-of-the-art single-image SR algorithms (SRCNN [67], MemNet [68], Very Deep Convolutional Network for Image Super-Resolution [66], Deeply-recursive Convolutional Network [69], Deep Recursive Residual Network [70]) and the recent stereo image SR algorithm (PASSR [12], SRResNet + SAM [41], Cross Parallax Attention Stereo Super-resolution Network [53]).

Quantitative results for scale factors $\times 2$ and $\times 4$ are listed in Table 4. It can be seen that our PSMASSRnet achieves better

| Dataset          | KITTI 2012 (100 images) | KITTI 2015 (100 images) | Middleburty (5 images) |
|------------------|-------------------------|-------------------------|------------------------|
| Scale | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Single images SR |                  |                  |                  |        |                  |                  |        |        |
| Bicubic          | 28.36 | 0.859 | 27.20 | 0.813 | 29.58 | 0.918 | 25.71 |
| SRCNN [67]       | 29.94 | 0.904 | 28.55 | 0.910 | 31.82 | 0.945 | 27.32 |
| VDSR [68]        | 30.37 | 0.908 | 28.76 | 0.912 | 32.43 | 0.950 | 27.53 |
| MemNet [66]      | 30.51 | 0.910 | 28.85 | 0.914 | 32.74 | 0.951 | 27.81 |
| DRCN [69]        | 30.39 | 0.909 | 28.79 | 0.912 | 32.55 | 0.951 | 27.59 |
| DRRN [70]        | 30.40 | 0.909 | 28.77 | 0.913 | 32.80 | 0.952 | 27.60 |
| Stereo image SR  |                  |                  |                  |        |                  |                  |        |        |
| PASSR [34]       | 30.84 | 0.919 | 29.59 | 0.925 | 34.08 | 0.960 | 28.74 |
| SRResNet + SAM [41] | --   | 0.800 | 25.10 | 0.790 | --   | 0.876 | 28.95 |
| CPASSRnet [53]   | 30.93 | 0.920 | 29.70 | 0.928 | 34.49 | 0.963 | 28.90 |
| Ours             | 30.98 | 0.922 | 29.74 | 0.930 | 34.56 | 0.966 | 29.01 |

Abbreviations: CPASSRnet, cross parallax attention stereo super-resolution network; DRCN, deeply-recursive convolutional network; DRRN, deep recursive residual network; PASSR, parallax-attention stereo super-resolution network; PSMASSRnet, parallax-based second-order mixed attention stereo SR network; PSNR, peak signal-to-noise ratio; SR, super-resolution; SRCNN, super-resolution convolutional neural network; SSIM, structural similarity index; VDSR, Very Deep Convolutional Network for Image Super-resolution.
performance than all the single image SR algorithms. Since all the stereo image SR methods used stereo images to obtain more useful information, they achieved very similar performance. However, our network achieves better performance than other stereo image SR methods because our network can make full use of second-order statistics of features to extract more discriminative features.

We also show two sets of qualitative results in Figure 5. It can be seen from zoom-in areas that the proposed PSMASSRnet can recover more image details and fewer artefacts compared with other methods. Take the output images of PASSR as examples, the burglar bars for windows in ‘img_129’ suffer from an obvious blurring artefact and the railings of the highway in ‘img_192’ suffer from serious deformation. However, the output images of our network have reliable details and are closer to the ground-truth images.

5 | CONCLUSIONS

In this work, we propose a novel PSMASSRnet to integrate the cross-view information from a stereo image pair for SR. Specifically, our PSMAM can use second-order statistics of features to focus on more informative features and adaptively learn feature interdependencies. The comprehensive experimental results have proved that the proposed PSMASSRnet could provide comparative or better performance in comparison with the current state-of-the-art single image SR algorithms and stereo image SR algorithms.

**FIGURE 5** Qualitative results for 4× SR achieved on ‘img_129’ from the KITTI2012 test dataset and ‘img_192’ from the KITTI2015 test dataset. Red colour presents the best performance, and blue colour indicates the second best performance. PASSRnet, parallax-attention stereo SR network; PSMASSRnet, parallax-based second-order mixed attention stereo SR network; SR, super-resolution
ACKNOWLEDGEMENTS
This work was supported by the Public Research and Capacity Building of Guangdong Province under Grant (No. 2014B010104001) and the Basic and Applied Basic Research of Guangdong Province under Grant (No. 2015A03030 8018).

CONFLICT OF INTEREST
We declare that we have no conflict of interest.

ORCID
Chenyang Duan https://orcid.org/0000-0003-4647-3631

REFERENCES
1. Zuo, Y., et al.: Multi-scale frequency reconstruction for guided depth map super-resolution via deep residual network. IEEE Trans. Circ. Syst. Video Technol. 30(2), 297–306. (2019)
2. Zhang, Y., et al.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision, pp. 286–301. ECCV (2018)
3. Han, W., et al.: Image super-resolution via dual-state recurrent networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1654–1663. (2018)
4. Hui, Z., Wang, X., Gao, X.: Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3867–3876. (2019)
5. Guo, Y., et al.: Hierarchical neural architecture search for single image super-resolution. arXiv preprint arXiv:2003.04619 (2020)
6. Xu, Y.S., et al.: Unified dynamic convolutional network for super-resolution with variational degradations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12496–12505. (2020)
7. Bhavsar, A.V., Rajagopalan, A.N.: Resolution enhancement for binocular stereo. In: Proceedings of the 2008 19th International Conference on Pattern recognition, pp. 1–4. (2008)
8. Bhavsar, A.V., Rajagopalan, A.N.: Resolution enhancement in multi-image stereo. IEEE Trans. Pattern Anal. Mach. Intell. 32(9), 1721–1728. (2010)
9. Park, H., Lee, K.M., Lee, S.U.: Combining multi-view stereo and super-resolution in a unified framework. In: Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–4. IEEE (2012)
10. Jeon, D.S., et al.: Enhancing the spatial resolution of stereo images using a parallax prior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1721–1730. (2018)
11. Wang, L., et al.: Learning parallax attention for stereo image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12250–12259. (2019)
12. Humenberger, M., Engkelke, T., Kubinger, W.: A census-based stereo vision algorithm using modified semi-global matching and plane fitting to improve matching quality. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 77–84. IEEE (2010)
13. Mei, X., et al.: On building an accurate stereo matching system on graphics hardware. In: Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 467–474. IEEE (2011)
14. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4334361 (2015)
15. Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1592–1599. (2015)
16. Kendall, A., et al.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6675. (2017)
17. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5418. (2018)
18. Shaked, A., Wolf, L.: Improved stereo matching with constant highway networks and reflective confidence learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4641–4650. (2017)
19. Jie, Z., et al.: Left-right comparative recurrent model for stereo matching. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, pp. 3838–3846. (2018)
20. Guo, Y., et al.: Closed-loop matters: dual resolution networks for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5407–5416. (2020)
21. Sommer, L., Schuchert, T., Beyeler, J.: Comprehensive analysis of deep learning-based vehicle detection in aerial images. IEEE Trans. Circ. Syst. Video Technol. 29(9), 2733–2747. (2018)
22. Li, H., et al.: Learning from noisy anchors for one-stage object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10588–10597. (2020)
23. Wang, H., et al.: Hierarchical image segmentation ensemble for objectness in RGB-D images. IEEE Trans. Circ. Syst. Video Technol. 29(1), 93–103. (2017)
24. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. (2018)
25. Russell, M., et al.: Feature-based image patch classification for moving shadow detection. IEEE Trans. Circ. Syst. Video Technol. 29(9), 2652–2666. (2017)
26. Han, N., et al.: Double Relaxed regression for image classification. IEEE Trans. Circ. Syst. Video Technol. 30(2), 307–319. (2019)
27. Yang, J., et al.: Image super-resolution via sparse representation. IEEE Trans. Image Process. 19(11), 2861–2873. (2010)
28. Yang, J., et al.: Image super-resolution as sparse representation of raw image patches. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
29. Glasner, B., Bagom, S., Irani, M.: Super-resolution from a single image. In: Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, pp. 349–356. IEEE (2009)
30. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbour embedding. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 (CVPR 2004), vol. 1, pp. I–IEEE (2004)
31. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Comput. Graph. Appl. 22(2), 56–65. (2002)
32. Xiong, Z., Sun, X., Wu, F.: Robust web image/video super-resolution. IEEE Trans. Image Process. 19(8), 2017–2028. (2010)
33. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. IEEE Trans. Pattern Anal. Mach. Intell. 32(6), 1127–1133. (2010)
34. Lim, B., et al.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144. (2017)
35. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883. (2016)
36. Guo, Y., et al.: Dual-view attention networks for single image super-resolution. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2728–2736. (2020)
37. Tian, C., et al.: Coarse-to-fine CNN for image super-resolution. IEEE Trans. Multimed. 23, 1489–1502. (2020)
39. Song, W., et al.: Stereoscopic image super-resolution with stereo consistent feature. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(07), pp. 12031–12038. (2020)
40. Yan, B., et al.: Disparity-aware domain adaptation in stereo image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13179–13187. (2020)
41. Ying, X., et al.: A stereo attention module for stereo image super-resolution. IEEE Signal Process. Lett. 27, 496–500. (2020)
42. Mei, Y., et al.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5690–5699. (2020)
43. Jiang, K., et al.: Hierarchical dense recursive network for image super-resolution. Pattern Recogn. 107, 107475 (2020)
44. Huang, Y., et al.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5901–5910. (2020)
45. Wang, Q., et al.: Hierarchical pyramid diverse attention networks for face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8326–8335. (2020)
46. Fan, Q., et al.: Few-shot object detection with attention-RPN and multi-relation detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4013–4022. (2020)
47. Piao, Y., et al.: Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7254–7263. (2019)
48. Wang, X., et al.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803. (2018)
49. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning, pp. 2048–2057. (2015)
50. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164. (2017)
51. Duan, C., Xiao, N.: Parallax-based spatial and channel attention for stereo image super-resolution. IEEE Access. 7, 183672183679 (2019)
52. Wang, Y., et al.: Symmetric parallax attention for stereo image super-resolution. arXiv preprint arXiv:2011.03802 (2020)
53. Chen, C., et al.: Cross parallax attention network for stereo image super-resolution. IEEE Trans. Multimed (2021)
54. Wang, L., et al.: Parallax attention for unsupervised stereo correspondence learning. IEEE Trans. Pattern Anal. Mach. Intell (2020)
55. Dai, T., et al.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11065–11074. (2019)
56. Li, J., et al.: MDCN: multi-scale dense cross network for image super-resolution. IEEE Trans. Circ. Syst. Video Technol (2020)
57. Chen, L.C., et al.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848. (2017)
58. Li, P., et al.: Is second-order information helpful for large-scale visual recognition? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2070–2078. (2017)
59. Li, P., et al.: Towards faster training of global covariance pooling networks by iterative matrix square root normalisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 947–955. (2018)
60. Wang, Y., et al.: Flickr1024: a dataset for stereo image super-resolution. arXiv preprint arXiv:1903.06332 (2019)
61. Hirschmuller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
62. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
63. Scharstein, D., et al.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Proceedings of the German Conference on Pattern Recognition, pp. 31–42. Springer, Cham (2014)
64. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The kitti vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3353–3361. IEEE (2012)
65. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3061–3070. (2015)
66. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 16461654 (2016)
67. Dong, C., et al.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 38(2), 295–307. (2015)
68. Tai, Y., et al.: Memnet: a persistent memory network for image restoration. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4539–4547. (2017)
69. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 16371645 (2016)
70. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3147–3155. (2017)

How to cite this article: Duan, C., Xiao, N.: Parallax-based second-order mixed attention for stereo image super-resolution. IET Comput. Vis. 1–12 (2021). https://doi.org/10.1049/cvi2.12063