Stereo Image Restoration via Attention-Guided Correspondence Learning

Shengping Zhang, Wei Yu, Feng Jiang, Liqiang Nie, Hongxun Yao, Qingming Huang, and Dacheng Tao

Abstract—Although stereo image restoration has been extensively studied, most existing work focuses on restoring stereo images with limited horizontal parallax due to the binocular symmetry constraint. Stereo images with unlimited parallax (e.g., large ranges and asymmetrical types) are more challenging in real-world applications and have rarely been explored so far. To restore high-quality stereo images with unlimited parallax, this paper proposes an attention-guided correspondence learning method, which learns both self- and cross-views feature correspondence guided by parallax and omnidirectional attention. To learn cross-view feature correspondence, a Selective Parallax Attention Module (SPAM) is proposed to interact with cross-view features under the guidance of parallax attention that adaptively selects receptive fields for different parallax ranges. Furthermore, to handle asymmetrical parallax, we propose a Non-local Omnidirectional Attention Module (NOAM) to learn the non-local correlation of both self- and cross-view contexts, which guides the aggregation of global contextual features. Finally, we propose an Attention-guided Correspondence Learning Restoration Network (ACLRNet) upon SPAMs and NOAMs to restore stereo images by associating the features of two views based on the learned correspondence. Extensive experiments on five benchmark datasets demonstrate the effectiveness and generalization of the proposed method on three stereo image restoration tasks including super-resolution, denoising, and compression artifact reduction.

Index Terms—Stereo Image Restoration, Stereo Correspondence, Stereo Matching

1 Introduction

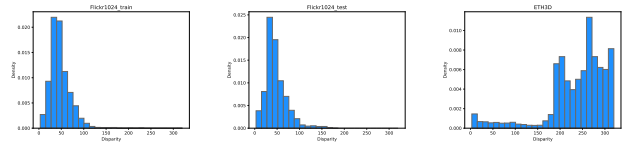
With the wide use of binocular cameras in mobile photography [1], robots [2], and autonomous vehicles [3], many stereo vision tasks have been attracting increasing attention, such as depth estimation [4] and 3D reconstruction [5]. However, owing to the physical limitations of binocular cameras [6], low resolutions and severe degradations (e.g., noise, artifacts) lead to low-quality stereo images and bring non-negligible difficulties to practical applications [7], [8]. Therefore, there is a high demand to restore stereo images for stereo vision tasks. Unlike single-view image restoration, stereo image restoration needs to take advantage of complementary cross-view information in two shifted stereo images. For example, details occluded or lost in one image may be recovered from its counterpart in the other image. In practice, the pixel shifts between two stereo images are usually from vertical or horizontal directions, which are referred to as horizontal and vertical parallax, respectively. To restore stereo images, horizontal parallax has been used as prior knowledge to establish stereo correspondence for super-resolution [8], denoising [9], and compression artifact removal [10]. Recent approaches [8], [11] utilize Convolutional Neural Networks (CNNs) with a parallax attention mechanism to capture the stereo correspondence along a horizon epipolar line to reconstruct high-resolution stereo images. To recover spatial details, some methods [7], [9], [10], [12] fuse cross-view information by using bidirectional disparity transformer modules to learn the correlation of similar features between two images.

Fig. 1: Visual result comparison between the Bicubic, SSRDEFnet [8] and our method for stereo image super-resolution on a pair of stereo images with both large parallax (208 pixels) and asymmetrical parallax (vertical offset as marked by the red line). Please zoom in for details.

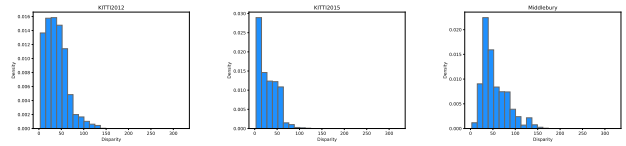
However, existing methods usually focus on restoring stereo images with limited horizontal parallax due to the binocular symmetry constraint, which fail to generalize well in real-world applications where the parallax of the captured stereo images varies significantly (e.g. large out-of-distribution or asymmetrical type) due to the installation

S. Zhang and W. Yu are with the School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, P.R. China (e-mail: s.zhang@hit.edu.cn; 20b903014@stu.hit.edu.cn).
F. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P.R. China (e-mail: fjiang@hit.edu.cn).
L. Nie is with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, P.R. China (e-mail: nieliqiang@gmail.com).
H. Yao is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, P.R. China (e-mail: h.yao@hit.edu.cn).
Q. Huang is with the University of Chinese Academy of Sciences, Beijing, China (e-mail: qmhuang@ucas.ac.cn).
D. Tao is with the UBTECH Sydney Artificial Intelligence Centre and the School of Information Technologies, Faculty of Engineering and Information Technologies, The University of Sydney, Darlington, NSW 2008, Australia (e-mail: dacheng.tao@sydney.edu.au).
Corresponding author: Liqiang Nie.

positions and types of binocular cameras [13]. As shown in Fig. 1, two stereo images have large horizontal parallax (208 pixels). In addition, the top of the chair is under the red line in the left image and above the red line in the right image. Such an asymmetrical parallax usually happens in real-world applications since the stereo cameras are not installed strictly on the horizon [6], which leads the corresponding pixels in two stereo images not to be on the same horizontal line. The state-of-the-art method [8] assumes that stereo complementary features are only derived from the horizontal direction within a limited parallax range and therefore fails to capture accurate feature correspondence and generates blurry textures.



(a) Flickr1024_train (b) Flickr1024_test (c) ETH3D



(d) KITTI2012 (e) KITTI2015 (f) Middlebury

Fig. 2: Parallax distributions on one training dataset (a) and five testing datasets (b-f).

On the other hand, existing methods are usually trained on some stereo datasets (e.g., Flickr1024 [14]) and tested on other stereo datasets (e.g., ETH3D [15], KITTI2012 [16], KITTI2015 [17], and Middlebury [18]). Here, we show the parallax distributions on one training dataset and five testing datasets in Fig. 2, from which we can see that the parallax distribution of the testing data Flickr1024_test is very similar to the training data Flickr1024_train (see Figs. 2(a) and 2(b)) since they are both from the Flickr1024 dataset. In contrast, the parallax distributions of other testing datasets are obviously different from the Flickr1024_train (see Fig. 2(a) and Figs. 2(c)-2(e)).

To address the aforementioned problems, we propose an attention-guided correspondence learning method to establish the correspondence of both self- and cross-view features, which is guided by parallax and omnidirectional attention. Specifically, to learn cross-view feature correspondence between stereo images, a Selective Parallax Attention Module (SPAM) is proposed to interact with cross-view features under the guidance of parallax attention, which is calculated by the global correlation between stereo image features. Based on the disparity shifts obtained from parallax attention, SPAM adaptively adjusts receptive fields for different parallax ranges by selecting large weights to integrate the cross-view information under large parallax. Furthermore, to fuse the corresponding features of stereo images with asymmetrical parallax, we propose a Non-local Omnidirectional Attention Module (NOAM) to learn the non-local correlation of both self- and cross-view contexts, which guides the establishment of correspondence between spatially similar features in any direction for aggregating global contextual information. Finally, to restore stereo images for different tasks, we propose a unified Attention-guided Correspondence Learning Restoration Network (ACLRNet) upon SPAMs and NOAMs to associate the features of two views based on the learned correspondence. Extensive qualitative and quantitative experiments demonstrate the effectiveness and generality of the proposed ACLRNet on three stereo image restoration tasks including super-resolution, denoising, and compression artifact removal. The proposed ACLRNet significantly outperforms other state-of-the-art approaches. The contributions of this work can be summarized as follows:

- To restore high-quality stereo images with unlimited parallax, this paper proposes an attention-guided correspondence learning method, which learns both self- and cross-views feature correspondence guided by parallax and omnidirectional attention.
- To learn the cross-view features correspondence, we propose a Selective Parallax Attention Module (SPAM) to interact with cross-view features under the guidance of parallax attention that adaptively selects receptive fields for different parallax ranges.
- To handle asymmetrical parallax, we propose a Non-local Omnidirectional Attention Module (NOAM) to learn the non-local correlation of both self- and cross-view contexts, which guides the aggregation of global contextual features.
- Extensive experiments on five benchmark datasets demonstrate the effectiveness and generalization of our ACLRNet on three stereo image restoration tasks including super-resolution, denoising, and compression artifact reduction.

2 Related Work

In this section, we review the previous work that is most related to the proposed method.

2.1 Single Image Restoration

Driven by the success of deep learning, many ingenious networks with excellent performance are designed for various single image restoration tasks, such as single image denoising, deblurring, super-resolution, compression artifact removal [19], [20], [21], [22]. Dong et al. [19] first adapt CNN layers for single image super-resolution, which outperforms traditional methods and leads the application of deep learning in image restoration. Zhang et al. [20] further design a deeper twenty-layer network to tackle image denoising and JPEG compression deblocking. Subsequently, numerous network architectures are proposed to perform restoration including residual dense connections [23], [24], progressive reconstruction [21], [25], and attention mechanisms [26], [27], [28]. Some algorithms [29], [30] exploit similarity information in images by constructing self-view correlations to improve the performance of restoration. Recently, several representative algorithms [31], [32], [33] are proposed to learn rich features by combining contextual information at multiple scales and achieving state-of-the-art performance

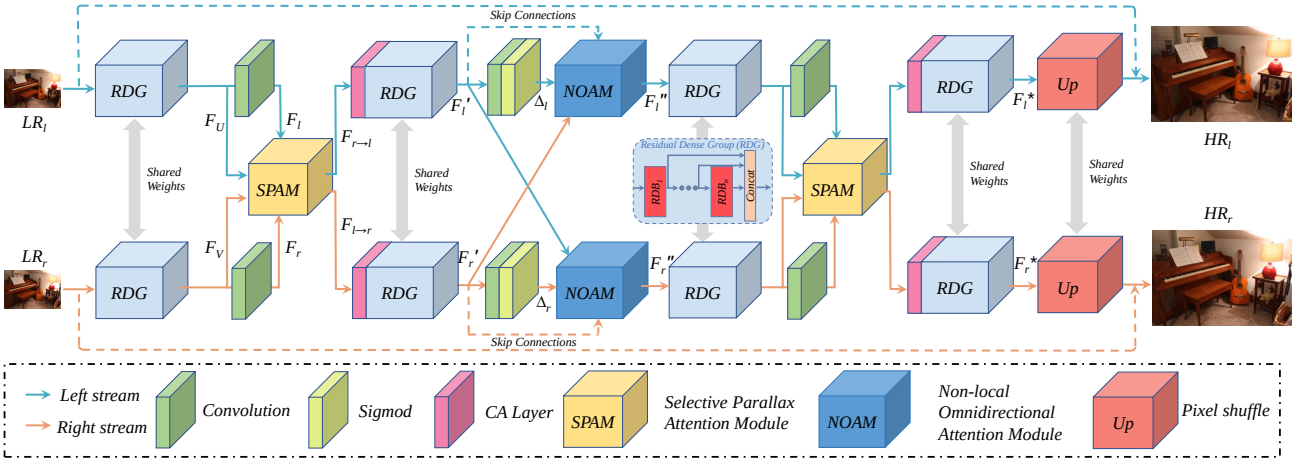


Fig. 3: The architecture of our ACLRNet for stereo image super-resolution.

on image restoration. In this paper, we propose a novel cross-view interactive correlation method that can extend any general single image restoration baseline to stereo image restoration tasks by introducing the proposed parallax and omnidirectional attention mechanisms.

2.2 Stereo Image Restoration

To recover HR images, stereo image restoration not only utilizes self-view information but also mines similar contents between cross-views [6]. Recently, several deep learning-based methods [6], [34], [35] utilize horizontal parallax information to boost the restoration performance. For example, in the stereo image super-resolution task, Jeon et al. [6] first employ a parallax offset prior via CNNs to learn the binocular similarity correspondence for stereo image SR. However, the maximum disparity handled is fixed due to the predefined offset. To cope with parallax variations, a series of works are proposed [7], [8], [34], [36], [37] based on the parallax attention module (PAM), which explores the pixel correlation along the horizontal epipolar line to implement cross-view interaction, achieving state-of-the-art SR results. In other fields, Jiao et al. [35] optimize the denoising and stereo matching iteratively in a multi-scale manner, and then utilize the cross-scale information to further improve the denoising effect. Zhou et al. [12] refine the spatial feature of stereo images by explicitly estimating the bidirectional disparity and aligning the symmetric features to achieve stereo image deblurring. Jiang et al. [10] propose a parallax transformer network to achieve feature level matching and fusion for both views, which effectively removes compression artifacts. Yan et al. [9] capture cross-view information and hunt for more accurate disparity estimation by inserting disparity prior, which helps the stereos image restoration. So far, most existing methods focus on restoring stereo images with limited horizontal parallax. They have limited performance in unlimited parallax situations (e.g., large ranges and asymmetrical types) since the fixed horizontal receptive field restricts the flexibility of capturing arbitrary stereo correspondence and the ability to aggregate the stereo features.

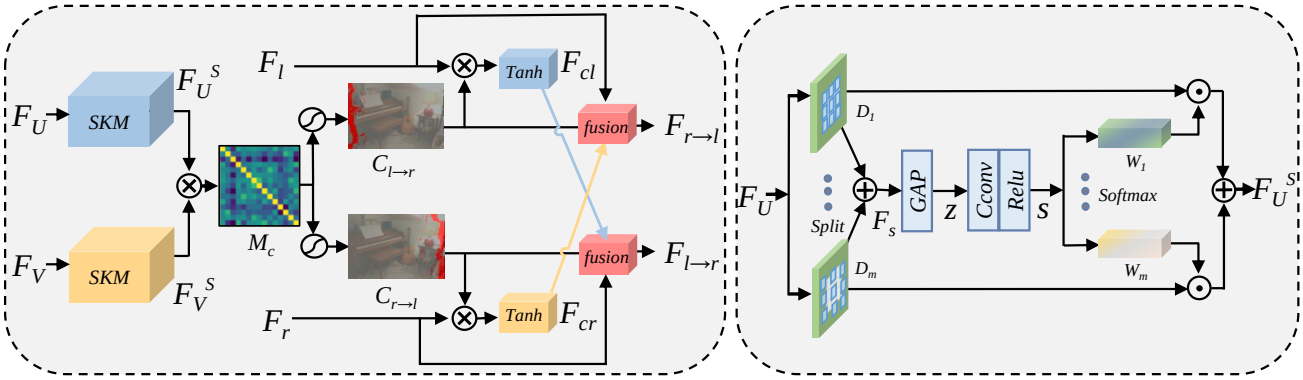
2.3 Attention Mechanisms

Generally, due to the limitation of equal treatment of multi-dimensional features by convolution, the attention mechanism biases the allocation of resources to the most challenging regions and disregards irrelevant regions, which effectively establishes long-range dependencies between features and thus benefits many computer vision tasks [38], [39]. The attention mechanism adaptively weights the features of different dimensions according to the importance of the input, which can be divided into spatial-attention [40], channel-attention [41], [42] and temporal-attention [43]. Similarly, the non-local attention network [27], [44] aggregates information from non-local regions by computing a weighted sum of features at all locations, which can be considered as a generalization of attention mechanisms. In addition, the transformer-based methods [22], [45] achieve considerable performance gains with self and cross attention to capture more accurate associations between pixels through matrix multiplication. Recently, some powerful and versatile architectures [46], [47] effectively integrate the advantages of CNN and Transformer to maximize the preservation of global and local features. Inspired by the transformer's attention mechanism, to handle stereo images with unlimited parallax, our work designs two attention mechanisms for stereo global correspondence, which can be employed as a guidance to capture the long-range dependencies between the stereo images themselves and each other.

3 Proposed Methods

3.1 Overview

To restore high-quality stereo images with unlimited parallax, this paper presents an Attention-guided Correspondence Learning Restoration Network (ACLRNet) upon the proposed Selective Parallax Attention Modules (SPAMs) and Non-local Omnidirectional Attention Modules (NOAMs). As shown in Fig. 3, we take stereo image super-resolution as an example to illustrate the architecture of our restoration network. Given a pair of low-resolution stereo images LR_l and LR_r, we first extract the hierarchical features F_U and F_V, and then concatenate all layers to generate the concatenated multi-level features F_l and F_r. To



(a) *Selective Parallax Attention Module (SPAM)* (b) *Select Kernel Module (SKM)*

Fig. 4: Illustration of Selective Parallax Attention Module (SPAM).

interact cross-view information, we feed both the hierarchical features and concatenated features to SPAM to obtain the cross-view interactive features $F_{r\to l}$ and $F_{l\to r}$, which are further combined by a Channel Attention (CA) layer and stacked Residual Dense Group (RDG) [23] to produce the combined interactive features F_l' and F_r'. To achieve self- and cross-view contextual feature aggregation, the stereo offsets Δ_l and Δ_r are further learned from the combined interactive features and then fed to NOAM to obtain the omnidirectional contextual aggregation features F_l'' and F_r''. To achieve effective cross-view interaction, we repeat the above process to extract the global contextual aggregation features F_l^* and $F *_e mphr$, which are finally fed to the pixel shuffle block and added by skip connection to reconstruct the high-resolution stereo images HR_l and HR_r. In the following, we introduce the SPAM and NOAM in detail.

3.2 Selective Parallax Attention Module

To learn cross-view feature correspondence between stereo images with large parallax ranges, we propose a selective parallax attention module (SPAM) to interact with the corresponding features between two views under the guidance of parallax attention to adaptively select receptive fields for different parallax ranges, as shown in Fig. 4 (a). Since hierarchical features with rich contextual information are beneficial to stereo correspondence learning [11], we first concatenate the multi-level features extracted from RDG to obtain the hierarchical features F_U and F_V, whose all layers are further concatenated to generate the concatenated features F_l and F_r. Moreover, to enable neurons to adaptively adjust their receptive field sizes under different parallax distributions, we use a selective kernel module (SKM) [48] to merge the hierarchical features while enriching the diversity of convolutions as shown in Fig. 4 (b). Without loss of generality, we introduce the feature fusion on the left view image. Within SKM, we perform feature separation on the hierarchical features $F_U \in \mathbb{R}^{H\times W\times C}$ by using multiple dilated convolutions with the same 3×3 kernel and four different dilation rates (2, 4, 6 and 8) to generate the multiple branch features $D_1, \cdots, D_m$ with different receptive field sizes, which are fused to facilitate the feature interaction via an element-wise summation to obtain the multi-scale features F_s

$$F_s = \sum_{i=1}^{m} D_i, D_i \in \mathbb{R}^{H\times W\times C} \tag{1}$$

Then, the channel statistics Z is obtained by using a global average pooling f_{GP} on the multi-scale features

$$Z = f_{GP}(F_s) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_s(i,j,:) \tag{2}$$

For better improving the computing efficiency, the channel statistics Z is fed to a convolution operator $\mathcal{C}$ with kernel size $1\times 1\times C\times r$ followed by a RelU layer $\mathcal{R}$ to obtain the compact feature descriptor S

$$S = \mathcal{R}(\mathcal{C}(Z)) \tag{3}$$

Further, to adaptively select different spatial scales, the i-th scale branch is assigned with the soft attention weight

$$W_i = split(softmax(reshape(Sw_s)))_i \tag{4}$$

where $w_s \in \mathbb{R}^{r\times mC}$ is the projection weight for the compact feature descriptor S. $reshape$ reshapes the input tensor to the shape of $\mathbb{R}^{1\times 1\times m\times C}$. $softmax$ computes the softmax activation along the last dimension of the input tensor. $split$ splits the input tensor into m tensors along the last dimension. The output feature F_U^s of SKM can be merged from

$$F_U^s = \sum_{i=1}^{m} W_i \odot D_i \tag{5}$$

where $\odot$ represents the broadcast element-wise multiplication along the last two dimensions to achieve weighted merging.

To further generate the cross attention map of stereo images, the merged features of the right image F_V^s are obtained by the weighted merging in SKM with the same processing as the left image. Then, the combined features of the left and right images are multiplied to generate the stereo cross-correlation map

$$M_c = F_U^s \otimes (F_V^s)^T \tag{6}$$

where $\otimes$ denotes the batch-wise matrix multiplication. To avoid inaccurate correspondence between stereo images (e.g., details lost or occluded in boundary regions), we

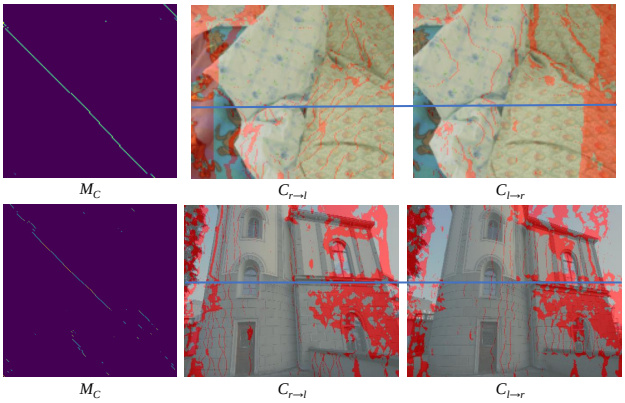


Fig. 5: Visualization of correlation and parallax attention maps. The estimated parallax range (i.e. red areas) of stereo images with two different disparity types. The first row represents symmetrical parallax and the second row represents asymmetrical parallax.

disentangle the cross-correlation map M_c by applying a softmax operation to M_c and its transpose M_c^T to generate the parallax attention maps $C_{l\to r}$ and $C_{r\to l}$ for the left and right views, respectively

$$\begin{aligned} C_{l\to r} &= softmax(M_c) \\ C_{r\to l} &= softmax(M_c^T) \end{aligned} \tag{7}$$

The visualization of the generated parallax attention maps is shown in Fig. 5, where the red regions represent the corresponding relevant features that are not found due to wrong correspondence in the textureless regions or parallax offset in the boundary occlusion regions. The first row stands for the ideal horizontal epipolar, which can be utilized to effectively establish the feature correlation between any two locations along the blue horizontal line in the cross-correlation map M_c by our SPAM. The asymmetrical parallax leads to a partial loss of feature correspondence in the cross-correlation map M_c as shown in the second row, which indicates that there is no corresponding relationship between these regions along the blue horizontal line.

To further achieve cross-view interaction, both the left and right concatenated features F_l and F_r are transformed to the other side to generate the non-corresponding features F_{cl} and F_{cr} guided by the parallax attention maps

$$\begin{aligned} F_{cl} &= C_{l\to r} \otimes F_l \\ F_{cr} &= C_{r\to l} \otimes F_r \end{aligned} \tag{8}$$

The non-corresponding features are used to correct the stereo images by feature complementation fusion to obtain the final cross-view interactive features $F_{r\to l}$ and $F_{l\to r}$

$$\begin{aligned} F_{r\to l} &= C_{l\to r} \odot F_{cr} + (1 - C_{l\to r}) \odot F_l \\ F_{l\to r} &= C_{r\to l} \odot F_{cl} + (1 - C_{r\to l}) \odot F_r \end{aligned} \tag{9}$$

where the concatenated features F_l and F_r are shifted at different ranges based on the feature correspondence of the cross-view in the horizontal direction to achieve accurate cross-view feature interaction, which corresponds to the fusion block in Fig. 4. The cross-view interactive features $F_{r\to l}$ and $F_{l\to r}$ can be adaptively filled with the concatenated features F_l and F_r to compensate for any range of parallax, thereby achieving continuous feature correspondence.

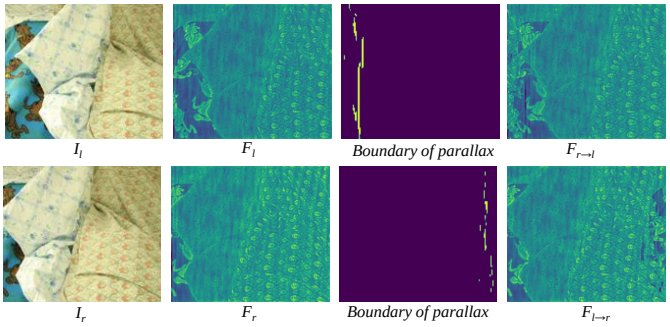


Fig. 6: Visualization of the original images, combined features, boundary of parallax, and cross-view interactive features respectively.

Fig. 6 shows the illustration of the correction process of the left and right features. It can be observed that the boundary of the parallax range can be clearly distinguished based on the parallax attention map. The cross-view interactive features are compensated to the lost features in different parallax regions, which can be filled with complementary features from the other view. Thereby, the cross-view interactive features are facilitated to establish a more comprehensive correspondence of cross-view features. In summary, SPAM predicts parallax attention and incorporates binocular complementary information in two views with various parallax ranges. Compared with existing parallax attention techniques that capture fixed correspondences [7], [8], our SPAM selectively fuses multi-scale features through the dynamic receptive fields and flexibly handles disparity variations via adaptive left-right feature correspondence.

3.3 Non-local Omnidirectional Attention Module

Since the perfect correction is difficult to achieve in real-world binocular cameras, the corresponding points can not be accurately aligned on the same horizontal line, which results in the ambiguous matching of key features and erroneous extraction of the stereo correspondence. Therefore, to handle images with asymmetrical parallax, the Non-local Omnidirectional Attention Module (NOAM) is proposed to reduce the matching ambiguity in unaligned situations, especially in occlusion and textureless regions, which can better extract global contextual information. NOAM performs non-local similar feature matching correlation between both self- and cross-views and then exploits omnidirectional attention to establish the global correlation of stereo features and guide the aggregation of global contextual features.

As shown in Fig. 7 (a), the combined interactive features F_l' and F_r' are first fed to the cross-matching block to match the cross-view dense correlation features. Then, the correlation features are fed to the self-matching block to capture their own contextual correlated features. Next, to perform accurate correlation aggregation, we utilize the right group-wise combined interactive features to achieve omnidirectional attention based on the omnidirectional offset Δ_l of the left view for obtaining the corrected right features F_{cr}'. Finally, the corrected right feature and the left combined interactive feature F_l' are utilized by the group-wise matrix

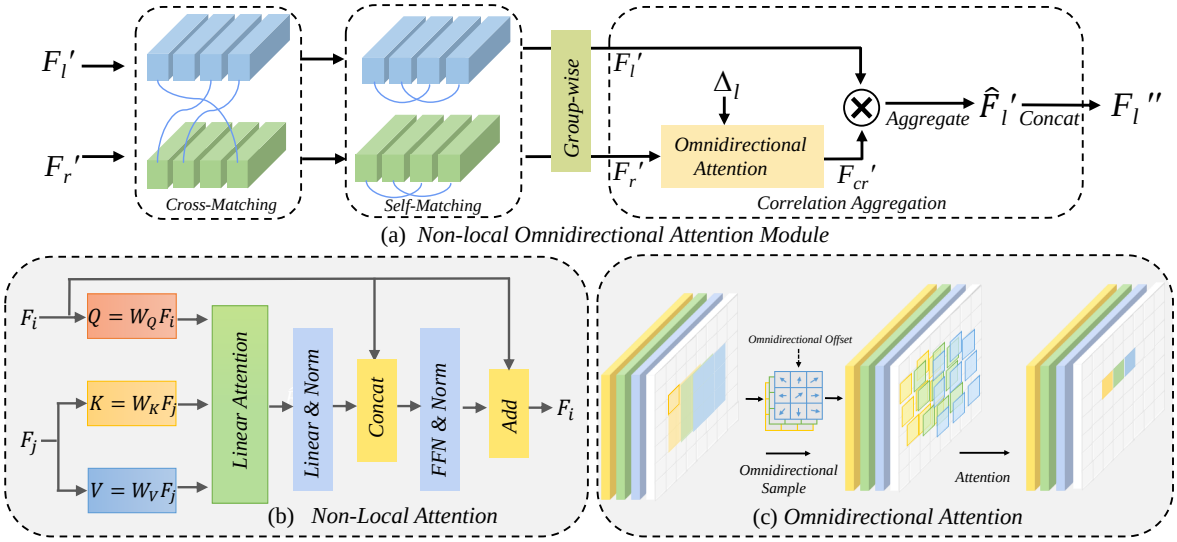


Fig. 7: Illustration of Non-local Omnidirectional Attention Module (NOAM).

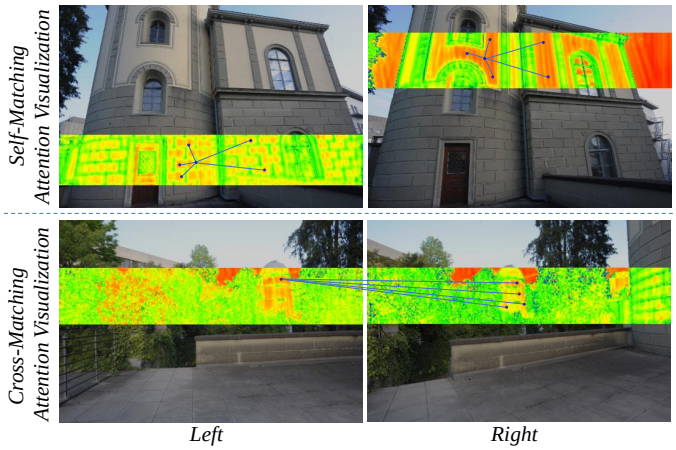


Fig. 8: Visualization of attention maps in self and cross matching. The red areas represent the higher strength of the correlation.

multiplication to achieve omnidirectional contextual feature aggregation. The outputs of all groups are averaged and concatenated to obtain the final omnidirectional contextual aggregation features $F_l^{''}$. The implementation of NOAM is based on two key elements: non-local matching and omnidirectional attention.

Non-Local Matching. To distill the self-view and cross-view similar features from stereo inputs, inspired by the transformer attention [38], we introduce the self- and cross-matching operation consisting of the non-local attention mechanism, whose structure is shown in Fig. 7 (b). For self-matching, the input features F_i and F_j are either F_l' or F_r'. For cross-matching, the input features F_i and F_j are either (F_l' and F_r') or (F_r' and F_l'), which depends on the direction of the cross-matching of two images. To achieve non-local matching, we adopt $Q = W_Q F_i$, $K = W_K F_j$, and $V = W_V F_j$ vector to represent query vector of one view, key and value vectors of another view respectively, where W_Q, W_K, and W_V are the projection matrices for one-dimensional representation, which are the input of the attention operation to obtain matching feature correlations. The matching operation based on the non-local attention [22] is denoted as

$$\text{Attention}(Q, K, V) = softmax(QK^T)V \tag{10}$$

The matching operation selects related information by measuring the similarity between two points to output the similarity scores map between the query vector and each key vector. The linear attention [49] is selected to reduce the computational consumption. The query vector Q retrieves matching information from the value vector V according to the attention weight computed from the dot product of Q and the key vector K corresponding to each value V. To extract enough correlation points, the output of the linear attention through summing of the value vectors weighted by the similarity scores, which can represent the global dense correlation of any two positions in the left and right views. The higher values indicate more relevant matching. And then the attention map is fed to a linear layer, a feed-forward network (FFN), and two Layernorm layers for optimization [38] by concatenating and adding the skip connection of the input features, resulting in the final non-local matching features F_i. By adopting the non-local attention with the global receptive field, non-local matching query relevant features between self- and cross-views to obtain self and cross non-local matching features, which can establish the global similar correspondence between point-to-point based on the dense matching strategies. The recent stereo image restoration methods often suffer from ambiguity in occlusion or textureless areas, because the local sparse feature correlation computed by CNNs fails to extract enough relevant points [4]. Thus, the dense matching is extracted from both view features based on our matching attention maps, whose visualization is shown in Fig. 8. It can be observed that the corresponding pixels are assigned bigger attention weights (i.e., red color) in attention maps, especially in repetitive patterns and textureless areas. The non-corresponding pixels are assigned smaller weights (i.e., yellow color). The visualization indicates that

we can explore the global similar correspondence features apart from the limited horizontal direction. The transform-based self- and cross-matching can not only establish the dense correspondence of similar features but also pay more attention to smooth and textureless regions to disambiguate them, which is important for the high-quality and complete restoration of the stereo images.

Omnidirectional Attention. To capture the aggregation features of global contextual information through the alignment correction in any direction, we introduce an omnidirectional attention mechanism to adopt an omnidirectional sampling strategy to adaptively align and aggregate the contextual matching features of stereo pairs. The search direction of existing methods between stereo images only lies on the horizontal epipolar line, which leads to matching ambiguity under the unrectified images. The search strategy in previous global matching sampling [50] computes the global correlation for all pixels, which can lead to misalignment and introduce additional noise interference (i.e., incorrect pixel correspondence). Inspired by the deformable convolution [51], [52], to avoid large computational and memory costs, we achieve omnidirectional attention by adopting omnidirectional sampling and correlation attention, which flexibly performs sampling from the context in the deformable window near the epipolar line to achieve omnidirectional correction of features. By adopting the right combined features F_r, the corrected right feature F_{cr} is obtained by the omnidirectional attention formula

$$F_{cr}(x, y, :) = \sum_{n=1}^{k \times k} w_n \cdot F_r(x + \Delta_{ln_x}, y + \Delta_{ln_y}, :) \tag{11}$$

where x and y denote the position coordinates of a pixel in the corrected right feature. $k \times k$ is the number of sampling directions. Δ_{ln} denotes the n-th learned omnidirectional offset of the current pixels that includes horizontal offset Δ_{ln_x} and vertical offset Δ_{ln_y}. w_n is the learned correlation attention weight that indicates the correlation strength between the offset position pixel and the current pixel.

It should be noted that NOAM can be considered as a novel correlation mechanism for complex stereo correspondence situations, which enables pixel-level view alignment and captures the omnidirectional feature interaction in global contexts. For non-textured or repetitive textured regions, NOAM makes the matching more robust to establish dense feature correspondence due to the non-local receptive field to restore fine structure details, which are verified in the qualitative experiment results (see the smooth cloth and repeating pattern in Fig. 12). In addition, the omnidirectional attention strategy replaces the conventional biased prediction with the more accurate context attention search, which achieves the robust alignment to aggregate the contextual features due to the omnidirectional dense correspondence. The global context aggregation information provided by self- and cross-view images is useful for the quality improvement of stereo image restoration.

3.4 A Unified Framework for Stereo Image Restoration

To achieve stereo restoration with various degradations for different tasks, we build a unified Attention-guided Correspondence Learning Restoration network (ACLRNet) to achieve high-quality image restoration, which cooperates with the above two attention modules to exploit additional complementary information through reliable correspondence learning of binocular similar features. Specifically, as shown in Fig. 3, ACLRNet is essentially an end-to-end network with the proposed SPAM and NOAM based on the residual dense groups (RDGs) stacked by two residual dense blocks (RDBs) [23], which can obtain high-resolution stereo images HR_l and HR_r from a pair of low-resolution stereo input images LR_l and LR_r. Since the hierarchical features are demonstrated to be effective for stereo tasks and the weight sharing strategy facilitates the fusion of features between the stereo image [7], [8], we first apply the weight-sharing residual dense group (RDG) to extract multi-level hierarchical features

$$\begin{aligned} F_U &= f_{RDG}(LR_l) \\ F_V &= f_{RDG}(LR_r) \end{aligned} \tag{12}$$

where the obtained hierarchical features F_U and F_V are concatenated for maximum utilization and then integrated by a 1×1 convolution to generate the concatenated features F_l and F_r. To obtain the cross-view interactive image features $F_{r \to l}$ and $F_{l \to r}$, SPAM utilizes parallax attention to capture cross-view correspondence based on the concatenated features

$$F_{r \to l}, F_{l \to r} = f_{spam}(F_U, F_V, F_l, F_r) \tag{13}$$

Next, the RDGs and Channel Attention (CA) layer are stacked to extract the combined interactive features F_l' and F_r'

$$\begin{aligned} F_l' &= f_{CA}(f_{RDG}(F_{r \to l})) \\ F_r' &= f_{CA}(f_{RDG}(F_{l \to r})) \end{aligned} \tag{14}$$

After capturing the horizontal directional correspondence of the left and right views, their vertical counterparts are often ignored in stereo images due to the epipolar constraint. To achieve omnidirectional aggregation of self- and cross-views contextual features, the omnidirectional offsets $\Delta_l and \Delta_r$ of left and right views are first learned from the combined interactive features F_l' and F_r'

$$\begin{aligned} \Delta_l &= f_{sigmod}(f_{conv}(F_l')) \\ \Delta_r &= f_{sigmod}(f_{conv}(F_r')) \end{aligned} \tag{15}$$

Then, to obtain accurate omnidirectional contextual aggregation features F_l'' and F_r'', the left and right offsets are fed to the NOAM based on the combined interactive features F_l' and F_r' respectively

$$\begin{aligned} F_l'' &= f_{NOAM}(F_l', \Delta_l, F_r') \\ F_r'' &= f_{NOAM}(F_r', \Delta_r, F_l') \end{aligned} \tag{16}$$

Finally, we repeat the above process to extract the global contextual aggregation features F_l^* and F_r^* to achieve effectively cross-view interaction. The high-quality restoration results HR_l and HR_r are reconstructed by using the generic upsampling operation based on the extracted stereo accurate features and the skip connection

$$\begin{aligned} HR_l &= f_{UP}(F_l^*) + f_{Bic}(LR_l) \\ HR_r &= f_{UP}(F_r^*) + f_{Bic}(LR_r) \end{aligned} \tag{17}$$

where f_{UP} is the pixel shuffle upsampling operation for stereo image super-resolution, f_{Bic} represents the Bicubic

TABLE 1: Comparison of PSNR, SSIM on the ETH3D dataset with asymmetric parallax for stereo image super-resolution, denoising, and compression artifact removal. The performance of ACLRNet surpasses other algorithms in the majority of the table entries. The red numbers indicate the best results.

Method	Scale	Bicubic	RDN [23]	RCAN [26]	iPASSR [7]	SSRDE-FNet [8]	Ours
Super-Resolution	×2	35.47/0.9542	40.06/0.9736	40.09/0.9740	39.60/0.9734	-/-	40.28/0.9750
	×4	30.96/0.8793	33.85/0.9263	34.10/0.9298	33.85/0.9264	33.95/0.9276	34.15/0.9316
Method	Noise	Noisy	DnCNN [20]	CBDNet [53]	FFDNet [54]	iPASSR [7]	Ours
Denoising	10	28.22/0.4918	41.16/0.9649	36.43/0.9146	41.58/0.9675	44.46/0.9796	44.71/0.9808
	30	18.95/0.1387	35.56/0.9332	26.26/0.5529	35.76/0.9369	41.18/0.9667	41.46/ 0.9692
Method	QF	JPEG	DCSC [55]	QGCN [56]	iPASSR [7]	PTNet [10]	Ours
Compression Artifact Removal	10	34.87/0.9056	37.04/0.9394	37.42/ 0.9421	37.31/0.9403	37.50/0.9416	38.21/0.9461
	30	39.89/0.9542	41.95/0.9671	42.34/0.9689	42.22/0.9682	42.36/0.9691	42.79/ 0.9704

interpolation for skip connection. For stereo image denoising and compression artifact removal tasks, the input and output dimensions are the same, a convolution layer is used to replace the pixel shuffle operation and without interpolation to reconstruct high-quality stereo images.

Different from the existing methods, our network attempts to make full use of the most relevant features in both self- and cross-view for image restoration regardless of disparity range or type, which is achieved by flexible selective kernel modeling and robust non-local omnidirectional attention. Note that, the two attentions are learned in an unsupervised manner to explore the stereo correspondence, and the image restoration is learned in a supervised manner. Our approach not only mines the similar features of self- and cross-views to construct global associations but also can be seamlessly applied to different restoration tasks under any disparity type, which is more suitable for real-world applications. Since our method aims to achieve several stereo image restoration tasks, we only set the Mean Square Error (MSE) loss constraint for all tasks to optimize the proposed restoration network

$$\mathcal{L}_{\mathcal{HR}} = \|HR_l - GT_l\|_2^2 + \|HR_r - GT_r\|_2^2 \qquad (18)$$

where HR_l and HR_r represent the left and right images restored by the algorithm, respectively. GT_l and GT_r represent the corresponding ground truth images.

4 Experiments

In this section, we first describe the datasets, metrics, and implementation details of our experiments. Next, we conduct qualitative and quantitative comparisons with other state-of-the-art methods for stereo image super-resolution, denoising, and compression artifact removal tasks on five stereo datasets with different disparity distributions. Finally, we report the results of ablation experiments.

4.1 Datasets and Implementation Details

We quantitatively and qualitatively evaluate our ACLRNet on five popular public stereo datasets against state-of-the-art restoration algorithms [8], [9], [10], [37]. To verify the generalization of our ACLRNet, we only adopt 800 pair images from Flickr1024 [14] as the training images. These images

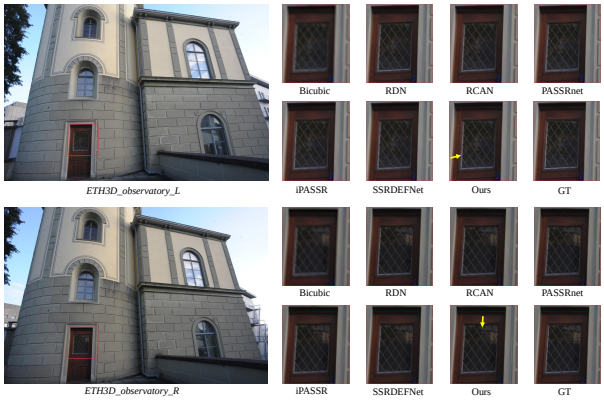


Fig. 9: Qualitative results (×4) of stereo image super-resolution task on ETH3D [15] with asymmetric epipolar lines

are augmented by means of random flips horizontally and vertically, rotation, and cropped into patches of size 30×90 with a stride of 20 as training patches. To compare the performance of stereo image restoration methods, we follow the existing work to choose 5 pair images from Middlebury [18], 20 pair images from KITTI 2012 [16], 20 pair images from KITTI 2015 [17], 112 pair images from Flickr1024 [14] as the testing dataset with symmetrical parallax. Furthermore, to verify the effectiveness of our algorithm in stereo images with asymmetrical parallax, we select 6 pair images from the ETH3D [15] dataset with different views as the testing dataset and perform ×2 bicubic downsampling to facilitate the testing. To produce training and testing data pairs, for the super-resolution task, we further bicubic downsampling all images with the scale factors of 2 and 4 to generate low-resolution images. For the denoising task, we adopt white Gaussian noise with 10 and 30 levels [10] to generate noisy images. For the compression artifacts removal task, we convert clean images to low-quality images with 10 and 30 quality factors by the normal JPEG compression.

Following previous work, we compute the popular metrics including Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), and Peak Signal to Noise Ratio Block (PSNR-B) metrics using the Y channel (in YCrCb space) to quantitatively evaluate the compared algorithms. To comprehensively evaluate the performance of restoration,

TABLE 2: Quantitative results achieved by different stereo image super-resolution methods on four stereo datasets with symmetric parallax. #Params, #FLOPs, and Times represent efficiency metrics. PSNR/SSIM values achieved on both the left images (i.e., Left) and a pair of stereo images (i.e., (Left + Right) / 2). The red numbers indicate the best results.

Method	#Params (M)	#FLOPs (G)	Times (ms)	Scale	Left			(Left + Right) /2			
					KITTI 2012	KITTI 2015	Middlebury	KITTI 2012	KITTI 2015	Middlebury	Flickr1024
bicubic	-	-	-	×2	28.44/0.8808	27.81/0.8814	30.46/0.8979	28.51/0.8842	28.61/0.8973	30.60/0.8990	24.94/0.8186
				×4	24.52/0.7310	23.79/0.7072	26.27/0.7553	24.58/0.7372	24.38/0.7340	26.40/0.7572	21.82/0.6293
EDSR [57]	38.90	407.30	123.94	×2	30.83/0.9199	29.94/0.9231	34.84/0.9489	30.96/0.9228	30.73/0.9335	34.95/0.9492	28.66/0.9087
				×4	26.26/0.7954	25.38/0.7811	29.15/0.8383	26.35/0.8015	26.04/0.8039	29.23/0.8397	23.46/0.7285
RDN [23]	22.00	451.93	88.81	×2	30.81/0.9197	29.91/0.9224	34.85/0.9488	30.94/0.9227	30.70/0.9330	34.94/0.9491	28.64/0.9084
				×4	26.23/0.7952	25.37/0.7813	29.15/0.8387	26.32/0.8014	26.04/0.8043	29.27/0.8404	23.47/0.7295
RCAN [26]	15.36	313.32	114.47	×2	30.88/0.9202	29.97/0.9231	34.80/0.9482	31.02/0.9232	30.77/0.9336	34.90/0.9486	28.63/0.9082
				×4	26.36/0.7968	25.53/0.7836	29.20/0.8381	26.44/0.8029	26.22/0.8068	29.30/0.8397	23.48/0.7286
StereoSR [6]	1.15	23.54	7.68	×2	29.42/0.9040	28.53/0.9038	33.15/0.9343	29.51/0.9073	29.33/0.9168	33.23/0.9348	25.96/0.8599
				×4	24.49/0.7502	23.67/0.7273	27.70/0.8036	24.53/0.7555	24.21/0.7511	27.64/0.8022	21.70/0.6460
PASSRnet [11]	1.42	25.49	15.49	×2	30.68/0.9159	29.81/0.9191	34.13/0.9421	30.81/0.9190	30.60/0.9300	34.23/0.9422	28.38/0.9038
				×4	26.26/0.7919	25.41/0.7772	28.61/0.8232	26.34/0.7981	26.08/0.8002	28.72/0.8236	23.31/0.7195
iPASSR [7]	1.43	29.60	17.36	×2	30.97/0.9210	30.01/0.9234	34.41/0.9454	31.11/0.9240	30.81/0.9340	34.51/0.9454	28.60/0.9097
				×4	26.47/0.7993	25.61/0.7850	29.07/0.8363	26.56/0.8053	26.32/0.8084	29.16/0.8367	23.44/0.7287
SSRDE-FNet [8]	2.26	266.54	123.31	×2	31.08/0.9224	30.10/0.9245	35.02/0.9508	31.23/0.9254	30.90/0.9352	35.09/0.9511	28.85/0.9132
				×4	26.61/0.8028	25.74/0.7884	29.29/0.8407	26.70/0.8082	26.43/0.8118	29.38/0.8411	23.59/0.7352
Ours	1.24	25.13	12.39	×2	31.20/0.9236	30.23/0.9262	35.17/0.9513	31.33/0.9266	31.02/0.9363	35.25/0.9512	29.25/0.9183
				×4	26.75/0.8071	25.85/0.7936	29.54/0.8465	26.84/0.8128	26.59/0.8168	29.61/0.8467	23.84/0.7475

we also report the average metrics on stereo image pairs (i.e., (Left + Right)/2). Different from the denoising and compression artifacts removal tasks, we crop the borders (64 pixels) to calculate the metric value for the super-resolution task for a fair comparison. All models are implemented by Pytorch and trained on an Nvidia GTX 3090 GPU for 80 epochs with a batch size of 12 and are optimized using the Adam [58] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to 2×10^{-4} and reduced to half after every 30 epochs. For each restoration task, the compared methods can be grouped into single image restoration methods and stereo image restoration methods. Note that the compared methods are retrained with the same training data as our method on the same task.

4.2 Stereo Image Super-Resolution

To validate our ACLRNet on the super-resolution task, three state-of-the-art single image super-resolution methods (EDSR [57], RDN [23] and RCAN [26]) and four stereo image super-resolution methods (StereoSR [6], PASSRnet [34], iPASSR [7] and SSRDE-FNet [8]) are selected for quantitatively and qualitatively comparison at two scales.

Quantitative results. In Table 2, we show the quantitative comparison of performance and efficiency in stereo testing datasets with symmetrical parallax, from which we can see that ACLRNet achieves the best performance on the four datasets with two scale factors ($\times 2$ and $\times 4$). It can be seen that our method has comparable efficiency while achieving the best performance. Specifically, our ACLRNet outperforms the best single image SR method (RCAN [26]) by an average of 0.4 dB in terms of PSNR on the Flickr1024 [14] dataset for $\times 2$ SR, demonstrating the effectiveness of our stereo restoration framework. Since more reliable stereo correspondence can be captured by our attention mechanism, the proposed method obtains PSNR higher than the best stereo method (SSRDE-FNet [8]) by an average of 0.3 dB.

In addition, results on stereo image pairs with asymmetrical parallax are shown in Table 1. Our method achieves the optimum performance in PSNR by 0.2dB and 0.1dB higher than the best stereo SR method (SSRDE-Fnet) and single SR method (RCAN) respectively. Due to the absence of symmetric features in the horizontal epipolar, current stereo methods cannot fully extract interactive information, which makes their performance inferior to single-image methods. Our correspondence learning method can effectively capture more accurate omnidirectional correspondences under any parallax type, which is beneficial to the overall image restoration performance.

Qualitative results. Fig. 10 illustrates the qualitative results on two different real-world scenarios. From the zoomed-in areas, it can be observed that the single-image image SR methods cannot reliably recover the lost details and have obvious artifacts. In contrast, our ACLRNet produces finer details and fewer artifacts based on binocular correspondence information, e.g. the tiles are independent of each other, and the wings with clear structure. We further test our ACLRNet with other single and stereo SR methods under asymmetrical parallax types and the results are shown in Fig. 9. It can be observed that the window bar in the results of ACLRNet is more complete and clearer in detail than other SR methods Compared to stereo image SR methods, ACLRNet reconstructs a more complete structure and preserves the desired spatial details.

4.3 Stereo Image Denoising

The proposed ACLRNet and other state-of-the-art algorithms, including DnCNN [20], FFDNet [54], CBDNet [53], DASSR [9], IPASSR [7] are compared at 10 and 30 noise

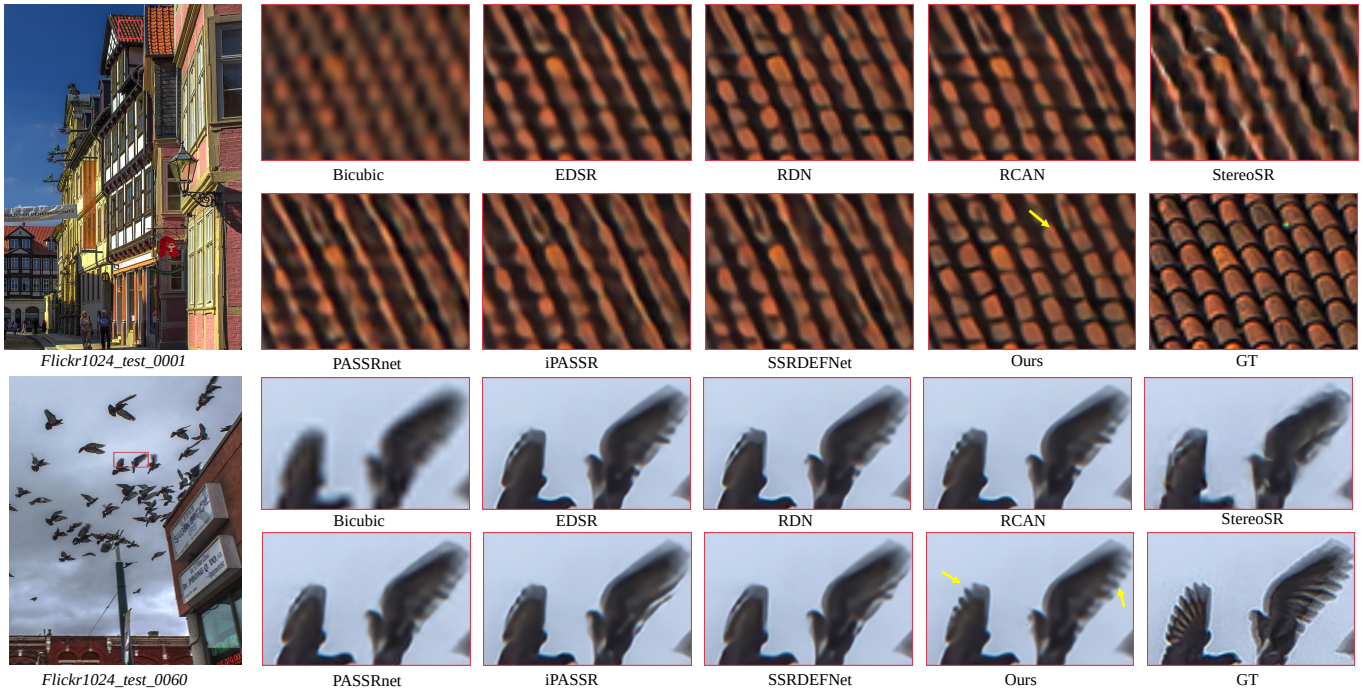


Fig. 10: Stereo image super-resolution of 0001 and 0060 from Flickr1024 [14], for scale factor 4.

TABLE 3: Quantitative results achieved by different stereo image denoising methods on four stereo datasets. #Params, #FLOPs, and Times represent efficiency metrics. PSNR/SSIM values achieved on both the left images (i.e., Left) and a pair of stereo images (i.e., (Left + Right) / 2). The red numbers indicate the best results.

Method	#Params (M)	#FLOPs (G)	Times (ms)	Noise	Left			(Left + Right) / 2			
					KITTI 2012	KITTI 2015	Middlebury	KITTI 2012	KITTI 2015	Middlebury	Flickr1024
Noisy	-	-	-	10	28.38/0.6986	28.41/ 0.7325	28.24/0.6879	28.38/0.6979	28.40/0.7217	28.24/0.6882	28.52/0.7422
				30	19.37/0.3350	19.38/0.3690	19.11/0.3185	19.38/0.3349	19.36/0.3563	19.11/0.3181	19.38/0.4182
DnCNN [20]	0.67	13.68	3.38	10	30.92/0.8899	28.65/0.8617	35.39/0.9396	31.38/0.8991	29.25/0.8737	35.37/0.9398	31.22/0.8915
				30	28.01/0.8177	26.66/0.7883	30.55/0.8524	28.22/0.8250	27.12/0.8035	30.55/0.8530	27.07/0.7858
FFDNet [54]	0.49	10.12	3.44	10	33.35/0.9164	32.09/0.9053	36.79/0.9490	33.61/0.9205	32.56/0.9108	36.78/0.9494	33.99/0.9164
				30	28.24/0.8176	27.32/0.7929	30.89/0.8533	28.39/0.8236	27.73/0.8073	30.90/0.8541	27.77/0.7909
CBDNet [53]	4.37	18.90	8.56	10	32.37/0.8696	28.43/0.8327	32.37/0.8696	32.34/0.8695	28.88/0.8589	32.34/0.8695	30.95/0.8880
				30	23.69/0.5477	22.77/0.5593	23.69/0.5478	23.67/0.5467	22.91/0.5602	23.67/0.5467	23.43/0.6361
iPASSR [7]	1.36	27.76	17.21	10	36.58/0.9603	35.48/0.9569	39.18/0.9732	36.58/0.9599	35.51/0.9534	39.19/0.9733	37.13/0.9732
				30	32.61/ 0.9266	31.32/0.9177	34.80/ 0.9408	32.61/0.9264	31.36/0.9118	34.80/0.9408	32.68/0.9436
DASSR [9]	5.96	241.26	106.38	10	-/-	-/-	-/-	-/-	-/-	39.12/0.9850	-/-
				30	-/-	-/-	-/-	-/-	-/-	34.73/0.9640	-/-
Ours	1.18	23.34	13.26	10	36.91/0.9623	35.90/0.9595	39.68/0.9759	36.90/0.9620	35.91/0.9563	39.68/0.9761	37.58/0.9759
				30	32.89/0.9296	31.65/0.9203	35.43/0.9478	32.93/0.9299	31.69/0.9152	35.44/0.9479	33.05/0.9475

levels. Among them, DnCNN [20], FFDNet [54] and CBDNet [53] are single image denoising methods, DASSR [9] is currently the only stereo denoising method, and IPASSR [7] is trained as a baseline for fair experimental comparisons.

Quantitative results. Table 3 shows the quantitative results of the denoising algorithm's performance and efficiency on four stereo datasets with symmetrical parallax. It can be found that ACLRNet achieves the best performance at all scenarios and noise levels, which demonstrate that our ACLRNet is more advantageous in generalization. The PSNR and SSIM gains of our method over the current best results indicate the advantages of our framework. For example, PSNR achieved by ACLRNet is higher than the best single method (FFDNet [54]) by an average of 3.1 dB on 10 noise level. Moreover, although our method is not more efficient than the single image compression artifact removal method FFDNet, our method achieves an improvement of 0.71 dB over the state-of-the-art stereo image compression artifact removal method DASSR [9] but with only about 10% computational cost and 20% parameters. In addition, comparison results on denoising images with asymmetrical parallax are shown in Table 1. Our method achieves the optimum performance in mean PSNR by 0.25dB and 4.4dB higher than the stereo denoising baseline (iPASSR) and single denoising method (FFDNet) respectively. This fully demonstrates the effectiveness and advancement of the proposed ACLRNet.

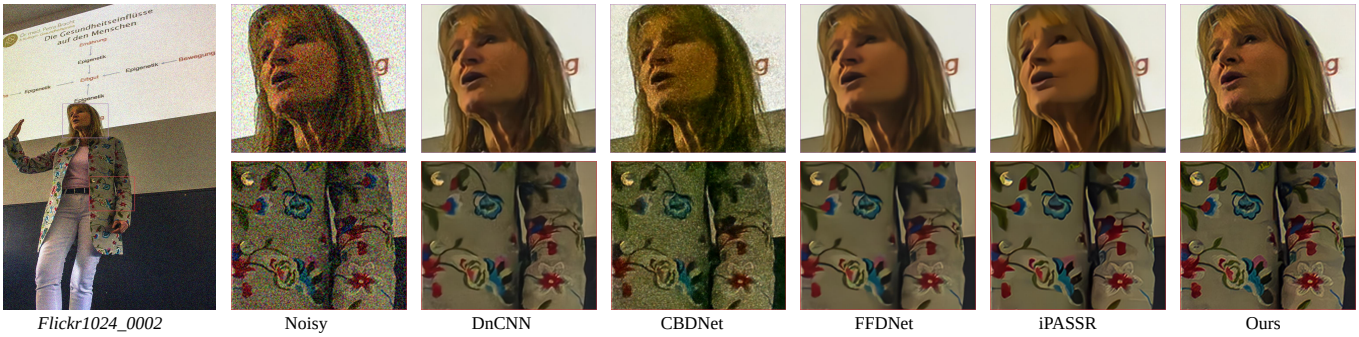


Fig. 11: Stereo image denoising of 0002 from Flickr1024 [14], for noise level 30.

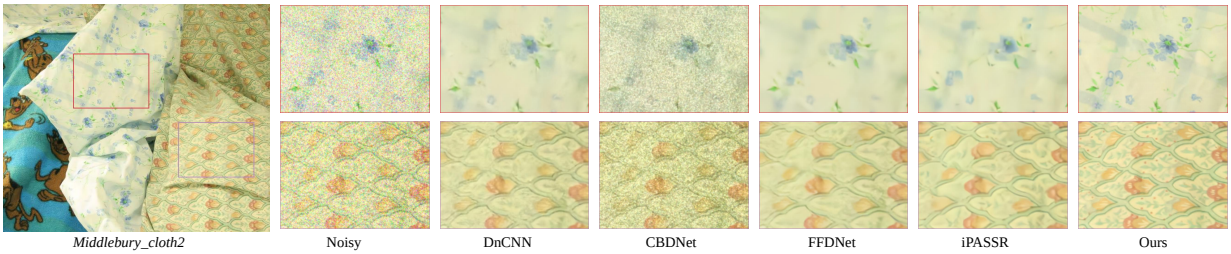


Fig. 12: Stereo image denoising of cloth2 from Middlebury [18], for noise level 30.

TABLE 4: Quantitative results achieved by different stereo image compression artifact removal methods on four stereo datasets. #Params, #FLOPs, and Times represent efficiency metrics. PSNR/SSIM/PSNR-B values achieved on both the left images (i.e., Left) and a pair of stereo images (i.e., (Left + Right) / 2). The red numbers indicate the best results.

Method	#Params (M)	#FLOPs (G)	Times (ms)	QF	Left			(Left + Right) /2			
					KITTI 2012	KITTI 2015	Middlebury	KITTI 2012	KITTI 2015	Middlebury	Flickr1024
JPEG	-	-	-	10	29.27/0.8292/26.46	29.31/0.8230/26.22	29.65/0.8114/27.09	29.12/0.8267/26.33	29.72/0.8314/26.57	29.62/0.8105/27.02	26.00/0.7860/23.74
				30	33.07/0.9170/30.27	33.54/0.9220/30.23	33.40/0.9110/30.86	32.85/0.9149/30.08	34.13/0.9279/30.73	33.38/0.9112/30.76	29.43/0.8933/27.15
DnCNN [20]	0.67	13.68	3.26	10	30.82/0.8665/30.53	30.90/0.8615/30.53	31.38/0.8529/31.22	30.64/0.8641/30.33	31.37/0.8708/31.04	31.32/0.8518/31.14	27.41/0.8223/27.03
				30	34.65/0.9347/34.02	35.13/0.9401/34.20	35.16/0.9304/34.70	34.40/0.9327/33.72	35.76/0.9455/34.93	35.15/0.9310/34.57	31.09/0.9166/30.40
DCSC [55]	0.09	2.26	1.92	10	30.99/0.8711/30.65	31.06/0.8665/30.60	31.57/0.8582/31.38	30.81/0.8687/30.44	31.54/0.8740/31.12	31.53/0.8572/31.25	27.57/0.8279/27.16
				30	34.80/0.9362/34.18	35.26/0.9415/34.39	35.35/0.9325/34.95	34.55/0.9343/33.89	35.90/0.9469/35.12	35.35/0.9331/34.79	31.26/0.9185/30.49
QGCN [56]	1.43	7.34	4.32	10	31.20/0.8759/30.95	31.31/0.8714/30.96	31.85/0.8643/31.73	31.00/0.8732/30.75	31.82/0.8807/31.50	31.74/0.8624/31.48	27.74/0.8345/27.44
				30	34.97/0.9388/34.46	35.46/0.9436/34.63	35.54/0.9361/35.23	34.71/0.9374/34.18	36.13/0.9490/35.46	35.57/0.9368/35.07	31.59/0.9240/30.86
iPASSR [7]	1.37	27.81	17.32	10	30.93/0.8713/30.49	30.98/0.8664/30.42	31.63/0.8602/31.36	30.76/0.8689/30.30	31.45/0.8757/30.93	31.59/0.8592/31.23	27.92/0.8353/27.37
				30	34.90/0.9379/34.24	35.36/0.9431/34.46	35.50/0.9348/35.12	34.66/0.9360/33.96	36.01/0.9485/35.19	35.51/0.9355/34.96	31.57/0.9226/30.74
PTNet [10]	1.38	29.46	17.85	10	31.43/0.8786/31.05	31.42/0.8730/30.92	32.05/0.8676/31.88	31.23/0.8761/30.83	31.97/0.8831/31.52	32.03/0.8672/31.75	28.07/0.8397/27.55
				30	35.18/0.9404/34.48	35.57/0.9449/34.58	35.85/0.9400/35.40	34.92/0.9384/34.18	36.28/0.9507/35.39	35.88/0.9409/35.25	31.83/0.9259/30.90
Ours	1.18	23.29	12.86	10	31.55/0.8823/31.19	31.62/0.8774/31.16	32.07/0.8687/31.90	31.35/0.8799/30.97	32.19/0.8878/31.77	32.18/0.8720/31.93	28.18/0.8436/27.66
				30	35.20/0.9407/34.59	35.66/0.9456/34.87	36.01/0.9401/35.67	34.96/0.9389/34.30	36.36/0.9510/35.64	36.03/0.9410/35.53	31.92/0.9264/31.03

Qualitative results. Compared with other methods, our ACLRNet removes noise effectively and produces cleaner results as shown in Figs. 11 and 12. We can see from the "Middlebury_cloth2" that our method not only reconstructs the untextured areas more clearly but also has richer details. The main reason is that ACLRNet makes full use of additional similarity information in stereo images to restore more complete structures and high-fidelity textures.

4.4 Stereo Image Compression Artifact Removal

In this section, we compare our ACLRNet to stereo image compression artifact removal method PTNet [10] and retrain the iPASSR [7] as a baseline. We also compare several single image compression artifact removal methods, including DnCNN [20], DCSC [55] and QGCN [56].

Quantitative results. The quantitative results of performance and efficiency on four test sets with JPEG quality factors (QF) 10 and 30 are shown in Table 4. As can be seen, the proposed ACLRNet obtains the highest PSNR on all test sets for all JPEG compression quality while achieving comparable efficiency. Although the PTNet and IPASSR exploit the binocular information, they do not consider the vertical direction and non-local similarity features, which makes them 0.2 dB lower than our method on four datasets in mean PSNR. Specifically, compared with the best stereo image compression artifact removal methods, our ACLRNet achieves a 0.25 dB improvement in terms of PSNR on the Middlebury [18] dataset at 30 quality factor. Experiments on ETH3D [15] with asymmetrical parallax lines are shown in Table 1. Our method outperforms the best stereo compression artifact removal method (PTNet)

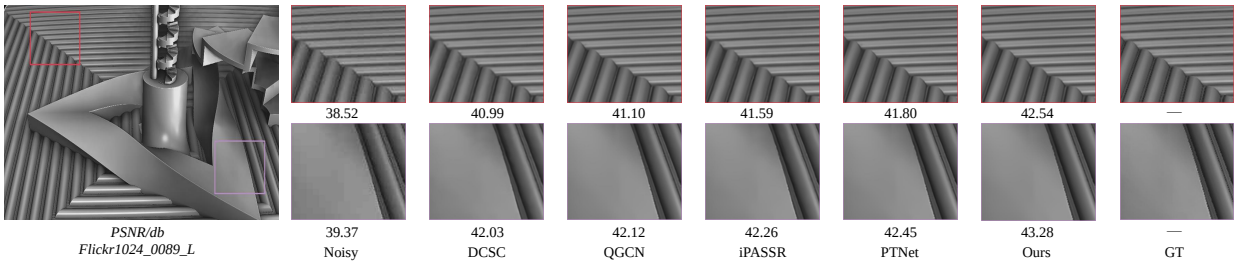


Fig. 13: Stereo image compression artifact removal of image 0089 from Flickr1024 [14], for noise level 30.

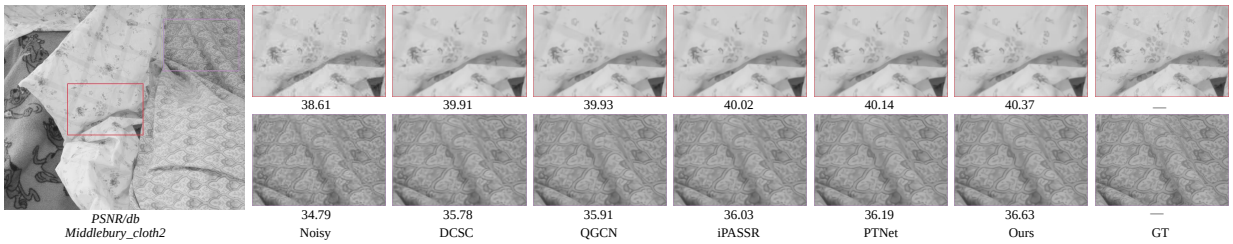


Fig. 14: Stereo image compression artifact removal of image cloth2 from Middlebury [18], for noise level 30.

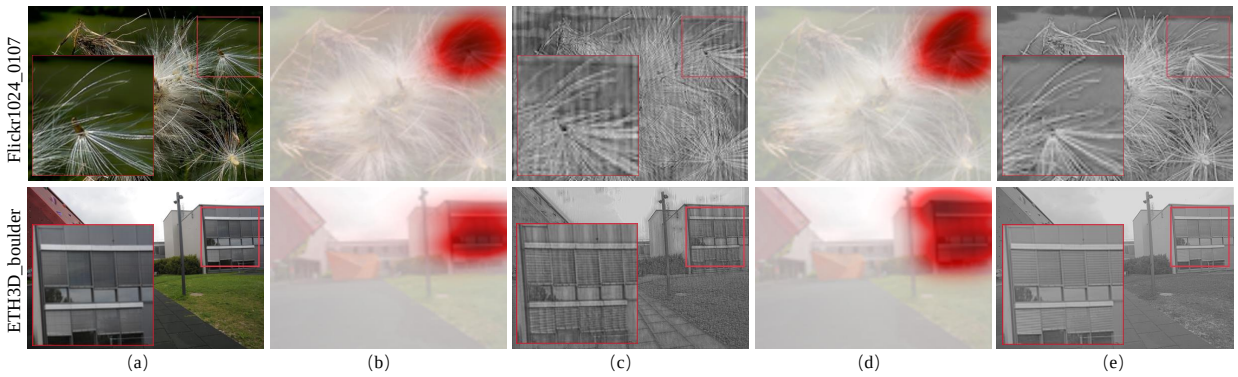


Fig. 15: Visualization comparison of receptive fields and extracted features by different kernel strategies (a) Input images with different parallax, (b) The receptive fields of the standard fixed kernel, (c) Features extracted by the standard fixed kernel, (d) The receptive fields of our selective kernel, (e) Features extracted by our selective kernel.

and single compression artifact removal method (QGCN) in mean PSNR score of 0.57 dB and 0.62 dB respectively.

Qualitative results. The qualitative comparison results are shown in Figs. 13 and 14 from which we can see that ACLRNet achieves more effective deblocking and artifacts removal. Since the visual results of compression restoration are difficult to distinguish with the naked eye, we add the PSNR metric for a fair comparison that our ACLRNet achieves an average improvement of 0.56 dB. Compared with other methods, our ACLRNet can remove compression artifacts and recover smooth planes or complex textures more effectively. Because the accurate cross-view interaction and pixel-level dense matching alignment improve the performance of feature fusion and facilitate compression reconstruction, which proves the superiority of our method.

4.5 Ablation Study

To verify the effectiveness of the proposed ACLRNet method, we conduct ablation experiments to study the impact of each module on the performance. As shown in Table 5, we derive four variants (S1, S2, S3, and S4) corresponding to different combinations of the proposed modules. We compare the variants with the baseline method [23] and baseline+biPAM [7] for stereo image super-resolution at $\times 4$ scale on the Middlebury dataset. It is worth noting that the baseline only stacks the residual dense group (RDG) module and does not use the stereo cross-view interaction mechanism.

TABLE 5: The impact of individual components in the proposed ACLRNet.

Method	SPAM	NOAM		PSNR/SSIM
		Non-Local	Omni-Atten	
Baseline [23]	-	-	-	29.16/0.8367
Baseline+biPAM [7]	-	-	-	29.27/0.8404
S1	✓	×	×	29.47/0.8441
S2	×	✓	×	29.45/0.8434
S3	×	✓	✓	29.52/0.8454
S4	✓	✓	✓	29.61/0.8467

TABLE 6: Ablation experiment results of SPAM achieved by stereo image super-resolution at $\times 4$ scale on the Middlebury [18] dataset with symmetric parallax.

SPAM	SKM	PAM	#Params (M)	#FLOPs (G)	PSNR/SSIM
Conv + SAM [34]	×	×	1.52	28.40	29.17/0.8370
Conv + biPAM [7]	×	×	1.52	27.99	29.27/0.8406
SKM + biPAM [7]	✓	×	1.45	25.79	29.46/0.8443
Conv + PAM	×	✓	1.31	27.33	29.43/0.8436
SKM + PAM (Ours)	✓	✓	1.24	25.13	29.61/0.8467

TABLE 7: Ablation experiment results of NOAM achieved by stereo image super-resolution at $\times 4$ scale on the ETH3D [15] dataset with asymmetric parallax.

NOAM		PSNR/SSIM
Non-local	Omni-Atten	
×	×	33.73/0.9215
×	✓	33.91/0.9278
✓	×	34.08/0.9296
✓	✓	34.15/0.9316

Selective Parallax Attention Module. The results of S1 show that the PSNR values of 29.16/29.27 dB from baseline and the combined previous parallax attention [7] are increased to 29.47 dB by our SPAM. According to the result, we can draw the conclusion that without the parallax attention module, the lack of stereo complementary information leads to incomplete recovery. Without the selective kernel module, the fixed receptive field prevents our network from capturing accurate correspondence on stereo images with different parallaxes.

Furthermore, to further verify the effectiveness of our SPAM, we decouple SPAM to obtain a Selective Kernel Module (SKM) and a Parallax Attention Module (PAM) for detailed analysis in Table 6. Specifically, we conduct a series of comparative experiments in terms of performance and efficiency to further elaborate the advantages and functions of the SKM and PAM, where 'Conv' denotes the standard convolution with the fixed kernel [23], 'SAM' and 'biPAM' denote the stereo parallax attention mechanism [34] and bi-directional parallax attention mechanism [7], respectively. By comparing 'SKM + PAM' and 'Conv + PAM', we can see that our SKM with the selected optimal kernels improves performance over the standard convolution by 0.18 dB while reducing the parameters and computational cost by 0.07M and 2.2 GFLOPs, respectively. In addition, we visualize the receptive fields and extracted features of the standard convolution with fixed kernels and the SKM with selective kernels for different parallax ranges. As shown in Fig. 15 (b) and (d), our selective kernel adaptively focuses on contextual regions of correlated texture details (e.g., tidbits and railings) and dynamically increases the receptive field size as the parallax increases, while the standard convolution only focuses on the center regions and its receptive field size is fixed. In Fig. 15 (c) and (e), the features extracted by our SKM preserve more accurate textures and finer details, which is beneficial for stereo correspondence learning to improve the generalization of the model. By comparing 'Conv + biPAM' and 'Conv + PAM' in Table 6, we find that our PAM improves performance over biPAM by 0.16 dB while reducing the parameters and computational cost by 0.21M and 0.66 GFLOPs, respectively. It can be attributed to the proposed parallax attention that learns the global cross-view correlation and therefore performs well in practical scenarios with unlimited asymmetric parallax. According to these experiment results, we can see that the selective parallax-attention module effectively employs context information from arbitrary parallax ranges with dynamic receptive fields to achieve better performance.

Non-local Omnidirectional Attention Module. Our NOAM aggregates similar features from self and cross non-local regions in stereo image pairs through non-local matching and omnidirectional attention operation. To demonstrate the effectiveness of the two operations separately, we add them in turn to our baseline network. The results of S2 show that the network increases by 0.29 dB (from 29.45 to 29.16) in terms of PSNR, which indicates the contribution of accurate matching associations of non-local contextual features to image restoration. From S3, the omnidirectional attention generates favorable results by alignment correlation operation, and the PSNR is improved from 29.45 to 29.52. As discussed in the previous sections, the two attention modules cooperate with each other to further enhance the utilization of complementary information. In S4, the results of our ACLRNet clearly demonstrate that the global contextual features from horizontal and vertical directions are fully extracted and aggregated, which helps to improve the performance of stereo restoration

Furthermore, to verify the ability of NOAM to handle asymmetric parallax, we conduct additional experiments on ETH3D [15] dataset in Table 7. Here, we perform ablation analysis on both the non-local matching (Non-local) and omnidirectional attention (Omin-Atten) components of NOAM by using stereo images with asymmetrical parallaxs from the ETH3D dataset. As we can see from Table 7, Non-local and Omin-Atten improve the performance by 0.35 dB and 0.18 dB in terms of PSNR. NOAM improves the performance by 0.42 dB and 0.0101 in terms of PSNR and SSIM, respectively. It should be noted that the performance gain (0.42 dB) on asymmetrical parallax is larger than symmetrical parallax (0.14 dB). According to the experiment results, we can draw that NOAM aggregates the features of self and cross views and achieves dense correlation, improving the stereo restoration effect. Experiments clearly demonstrate the excellent performance of non-local attention and omnidirectional attention components in our NOAM.

Model Size vs. Restoration Performance. It is known that increasing the height and width of the network with higher overhead computation costs can improve model performance, but how to balance and coordinate them is critical [59]. To validate the impact of the number of SPAM and NOAM, we study the comparison experiments performed for the super-resolution task with $\times 4$ scale factor on Middlebury dataset in Table 8. Increasing the number of modules leads to better scores, which indicates that the exchange of correspondence information between stereo images is beneficial to the integration of similar features. However, we

TABLE 8: The impact of the number of SPAM and NOAM modules in the proposed ACLRNet.

SPAM \ NOAM	0	1	2
0	29.16/0.8367	-/-	-/-
1	29.32/0.8412	29.39/0.8382	-/-
2	29.47/0.8441	29.61/0.8467	29.18/0.8339

TABLE 9: Performance comparison of the ACLRNet with different model sizes.

	PSNR	SSIM	PSNR-B	#Parms (M)	#FLOPs (G)
ACLRNet-T	32.18	0.8720	31.93	1.18	23.29
ACLRNet-S	32.27	0.8741	31.99	3.30	56.78
ACLRNet-M	32.42	0.8772	32.17	7.85	150.06
ACLRNet-L	32.49	0.8783	32.27	22.86	457.42

can note that increasing the number of NOAMs from 1 to 2 results in a decrease in the evaluation score since the excess global feature correspondence introduces additional noise. Therefore, ACLRNet with one NOAM and two SPAMs is chosen for all tasks.

Furthermore, to validate the effect of the model parameter and calculations on the restoration performance, we train multiple networks with different widths and heights, i.e., ACLRNet with tiny, small, medium and large (T, S, M, L) models, the RDG hyperparameters of these model variants are ($C = 24$, $N = 2$), ($C = 32$, $N = 4$), ($C = 32$, $N = 8$), ($C = 64$, $N = 8$), respectively, where C and N are the number of output channels and layers of the block, which denote the width and depth of the model. As described in Tables 4 and 9, for stereo image compression artifact removal with 10 noise level, our network outperforms the best method [10] in PSNR by 0.15 dB with only the smallest tiny model ACLRNet-T, which does not benefit from adding more convolution layers into the architecture and has fewer calculations. It can be noted that bigger networks with larger widths and depths tend to achieve better performance, but the PSNR gain saturates quickly after reaching 32.42 dB. Although scaling up any dimension of network width or depth improves the performance of restoration, the performance gain decreases with the increment of model size. Therefore, for a fair comparison and balance of parameters and performance, ACLRNet-T is chosen as the backbone network for three restoration tasks. The comparison results of efficiency and performance are shown in Tables 2, 3, and 4, respectively. It can be seen that our method has comparable efficiency while achieving the best performance. We show the efficiency and performance comparison between our method and four state-of-the-art stereo super-resolution methods in Fig. 16, from which we can see that our method achieves a good trade-off between efficiency and performance.

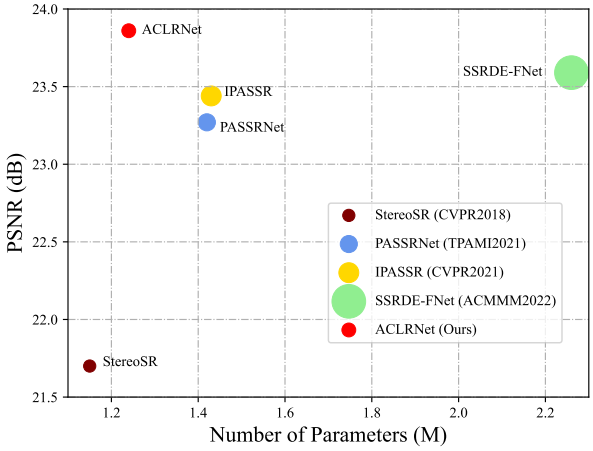


Fig. 16: Comparisons in terms of computational cost, number of parameters, and performance between the proposed method and state-of-the-art stereo super-resolution methods, including StereoSR [6], PASSRNet [11], IPASSR [7] and SSRDE-FNet [8]. The radius of the circle represents the computational cost (FLOPs) of the compared method.

5 Conclusion

To restore real-world stereo images with unlimited parallax (e.g., large ranges and asymmetrical types), this paper proposes a generic Attention-guided Correspondence Learning Restoration Network (ACLRNet) to establish both self- and cross-view feature correspondence, which is guided by the selective parallax and non-local omnidirectional attention. To learn cross-view correspondence, our Selective Parallax Attention Module (SPAM) adaptively interacts with stereo features from images with various parallax distributions, which flexibly selects appropriate receptive fields to estimate disparity offsets in different ranges. In addition, to handle asymmetrical parallax, we propose a Non-local Omnidirectional Attention Module (NOAM) to learn the omnidirectional dense correspondences between self and cross non-local features, which achieves sufficient global contextual feature aggregation. Finally, the proposed ACLRNet is built upon SPAMs and NOAMs to restore the stereo images with different degradation by associating the features of two views based on the learned stereo correspondence. Extensive experiments on five benchmark datasets show that our ACLRNet outperforms current methods over different stereo image restoration tasks and achieves state-of-the-art performance. In the future, we will extend our method to other restoration tasks such as stereo image deblurring, dehazing, and deraining for more real-world applications.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Nos. 62272134 and 62236003), in part by the National Key R&D Program of China (No. 2021ZD0110901), in part by the Taishan Scholars Program of Shandong Province (No. tsqn201812106), in part by the Shenzhen Colleges and Universities Stable Support Program (No. GXWD20220817144428005).

References

[1] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime

stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation*, pages 5893–5900. IEEE, 2019. 1

[2] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4338–4364, 2020. 1

[3] Boyu Gao, Haoxiang Lang, and Jing Ren. Stereo visual slam for autonomous vehicles: A review. In *2020 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1316–1322. IEEE, 2020. 1

[4] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022. 1, 6

[5] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1

[6] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1721–1730, 2018. 1, 2, 3, 9, 14

[7] Yingqian Wang, Xinyi Ying, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Symmetric parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–775, 2021. 1, 3, 5, 7, 8, 9, 10, 11, 12, 13, 14

[8] Qinyan Dai, Juncheng Li, Qiaosi Yi, Faming Fang, and Guixu Zhang. Feedback network for mutually boosted stereo image super-resolution and disparity estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1985–1993, 2021. 1, 2, 3, 5, 7, 8, 9, 14

[9] Bo Yan, Chenxi Ma, Bahetiyaer Bare, Weimin Tan, and Steven CH Hoi. Disparity-aware domain adaptation in stereo image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13179–13187, 2020. 1, 3, 8, 9, 10

[10] Xuhao Jiang, Weimin Tan, Ri Cheng, Shili Zhou, and Bo Yan. Learning parallax transformer network for stereo image jpeg artifacts removal. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6072–6082, 2022. 1, 3, 8, 11, 14

[11] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019. 1, 4, 9, 14

[12] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, and Jimmy S Ren. Davanet: Stereo deblurring with view aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10996–11005, 2019. 1, 3

[13] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4384–4393, 2019. 2

[14] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 8, 9, 10, 11, 12

[15] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3260–3269, 2017. 2, 8, 11, 13

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving the kitti vision benchmark suite. In *2012 IEEE conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 2, 8

[17] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 2, 8

[18] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014. 2, 8, 11, 12, 13

[19] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2015. 2

[20] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2, 8, 9, 10, 11

[21] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 2

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2, 3, 6

[23] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020. 2, 4, 7, 8, 9, 12, 13

[24] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 0–0, 2018. 2

[25] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 624–632, 2017. 2

[26] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, pages 286–301, 2018. 2, 8, 9

[27] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2, 3

[28] J. Chen, Z. Yang, T. N. Chan, H. Li, J. Hou, and L. P. Chau. Attention-guided progressive neural texture fusion for high dynamic range image restoration. 2021. 2

[29] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038. IEEE, 1999. 2

[30] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on Computer Vision and Pattern Recognition*, volume 2, pages 60–65. Ieee, 2005. 2

[31] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 2

[32] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1934–1948, 2022. 2

[33] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

[34] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 9, 13

[35] Jianbo Jiao, Qingxiong Yang, Shengfeng He, Shuhang Gu, Lei Zhang, and Rynson WH Lau. Joint image denoising and disparity estimation via stereo structure pca and noise-tolerant cost. *International Journal of Computer Vision*, 124(2):204–222, 2017. 3

[36] Xinyi Ying, Yingqian Wang, Longguang Wang, Weidong Sheng, Wei An, and Yulan Guo. A stereo attention module for stereo image super-resolution. *IEEE Signal Processing Letters*, 27:496–500, 2020. 3

[37] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, Liangyu Chen, Xiaojie Chu, Wenqing Yu, Kai Jin, et al. Ntire 2022 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 906–919, 2022. 3, 8

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 3, 6

[39] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, pages 1–38, 2022. 3

[40] Tonglai Liu, Ronghai Luo, Longqin Xu, Dachun Feng, Liang Cao, Shuangyin Liu, and Jianjun Guo. Spatial channel attention for deep convolutional neural networks. *Mathematics*, 10(10):1750, 2022. 3

[41] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 783–792, 2021. 3

[42] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5659–5667, 2017. 3

[43] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. Stat: Spatial-temporal attention mechanism for video captioning. *IEEE Transactions on Multimedia*, 22(1):229–241, 2019. 3

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

[45] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6197–6206, 2021. 3

[46] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 3

[47] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021. 3

[48] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 510–519, 2019. 4

[49] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 6

[50] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2007. 7

[51] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 973–981, 2021. 7

[52] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 7

[53] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1712–1722, 2019. 8, 9, 10

[54] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 8, 9, 10

[55] Xueyang Fu, Zheng-Jun Zha, Feng Wu, Xinghao Ding, and John Paisley. Jpeg artifacts reduction via deep convolutional sparse coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2501–2510, 2019. 8, 11

[56] Jianwei Li, Yongtao Wang, Haihua Xie, and Kai-Kuang Ma. Learning a single model with a wide range of quality factors for jpeg image artifacts removal. *IEEE Transactions on Image Processing*, 29:8842–8854, 2020. 8, 11

[57] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–144, 2017. 9

[58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9

[59] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 13

Shengping Zhang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 2013. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology at Weihai. He had been a post-doctoral research associate with Brown University and with Hong Kong Baptist University, and a visiting student researcher with University of California at Berkeley. He has authored or co-authored over 50 research publications in refereed journals and conferences. His research interests include deep learning and its applications in computer vision.

Wei Yu received the B.S. and M.S. degrees from China University of Petroleum (East China), China, in 2017 and 2020, respectively, and is currently working toward the Ph.D. degree in computer science and technology at Harbin Institute of Technology, China. His research interests include low-level vision, image super-resolution, and restoration.

Feng Jiang received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2001, 2003, and 2008, respectively. He is currently a Professor with the Department of Computer Science, HIT, and a Visiting Scholar with the School of Electrical Engineering, Princeton University. His research interests include computer vision, pattern recognition, and image and video processing.

Liqiang Nie is currently the dean with the Department of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). He received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University and National University of Singapore (NUS), respectively. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 100 papers and 4 books, received more than 15,000 Google Scholar citations. He is an AE of IEEE TKDE, IEEE TMM, IEEE TCSVT, ACM ToMM, and Information Science. Meanwhile, he is the regular area chair of ACM MM, NeurIPS, IJCAI and AAAI. He is a member of ICME steering committee. He has received many awards, like ACM MM and SIGIR best paper honorable mention in 2019, SIGMM rising star in 2020, TR35 China 2020, DAMO Academy Young Fellow in 2020, SIGIR best student paper in 2021, and ACM MM best paper in 2022.

Hongxun Yao received the B.S. and M.S. degrees in computer science from the Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and in 1990, respectively, and received Ph.D. degree in computer science from Harbin Institute of Technology in 2003. Currently, she is a professor with School of Computer Science and Technology, Harbin Institute of Technology. Her research interests include computer vision, pattern recognition, multimedia computing, human-computer interaction technology. She has 6 books and over 200 scientific papers published, and won both the honor title of the new century excellent talent in China and enjoys special government allowances expert in Heilongjiang Province, China.

Qingming Huang (Fellow, IEEE) is a professor in University of Chinese Academy of Sciences (CAS) and an adjunct professor in both Harbin Institute of Technology (HIT) and Institute of Computing Technology of CAS. He received B.S. and Ph.D. degrees in 1988 and 1994 respectively, both from HIT, China. His research areas include multimedia video analysis, image processing, computer vision and pattern recognition. He has published more than 300 academic papers on top journals and conferences. He is the associate editor of Acta Automatica Sinica and Fellow of IEEE. He has served as program chair, track chair and technical program committee member for top conferences such as ACM Multimedia, CVPR, and ICCV.

Dacheng Tao (Fellow, IEEE) is the Inaugural Director of JD Explore Academy and a Senior Vice President of JD.com. He is also a chief scientist of the digital science institute in the University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science, and his research is detailed in one monograph and over 200 publications in prestigious journals and conferences. He is a fellow of the Australian Academy of Science, AAAS, ACM and IEEE.