Perception-oriented Bidirectional Attention Network for Image Super-resolution Quality Assessment

Yixiao Li, Xiaoyuan Yang, Guanghui Yue, Jun Fu, Qiuping Jiang, *Senior Member, IEEE*, Xu Jia, Paul L. Rosin, Hantao Liu, and Wei Zhou, *Senior Member, IEEE*

***Abstract*—Many super-resolution (SR) algorithms have been proposed to increase image resolution. However, image quality assessment (IQA) metrics for comparing and evaluating different SR algorithms are limited. In this work, we propose the Perception-oriented Bidirectional Attention Network (PBAN) for image SR quality assessment, which is composed of three modules: an image encoder module, a perception-oriented bidirectional attention (PBA) module, and a quality prediction module. First, we encode the input images for feature representations. Inspired by the characteristics of the human visual system (HVS), we then construct the perception-oriented PBA module. Specifically, different from existing attention-based SR IQA methods, we conceive a Bidirectional Attention (Bi-Atten) module to bidirectionally construct visual attention to distortion, which is consistent with the generation and evaluation processes of SR images. To further guide the quality assessment towards the perception of distorted information, we propose Group Multi-scale Deformable Convolution (GMDC), enabling the proposed method to adaptively perceive distortion. Moreover, we design Sub-information Excitation Convolution (SubEC) to direct visual perception to both sub-pixel and sub-channel attention. Finally, the quality prediction module is exploited to integrate quality-aware features and regress quality scores. Extensive experiments demonstrate that our proposed PBAN outperforms state-of-the-art quality assessment methods.**

***Index Terms*—Image super-resolution, quality assessment, bidirectional attention, group multi-scale deformable convolution, sub-information excitation, human visual system.**

I. Introduction

IMAGE super-resolution (SR) aims to reconstruct high-frequency details for producing images with higher resolution that showcase detailed structures and textures compared to their low-resolution (LR) counterparts. The emergence of deep learning has significantly accelerated the development of new techniques for generating high-quality SR images. These techniques encompass a wide range of approaches, including convolutional neural network (CNN)-based approaches [1, 2], generative adversarial network (GAN)-based frameworks [3, 4], attention mechanism-based strategies [5, 6], and graph convolutional network (GCN)-based algorithms [7, 8], etc. Despite these advancements, the task of image SR remains challenging due to its highly ill-posed nature. Also, a single LR image may lead to multiple SR results that maximize the output quality — i.e. the optimal solution is not unique. Therefore, developing precise SR quality assessment metrics is of great importance.

Generally, image quality assessment (IQA) methods are categorized into full-reference (FR), reduced-reference (RR), and blind/no-reference (NR) based on the availability of reference images. When the corresponding original image is available, FR-IQA can directly compare the original and distorted images. The RR-IQA uses partial information from the reference image, while NR-IQA estimates image quality without any reference information. These methodologies aim to obtain quality predictions that closely align with subjective ratings. Among them, the peak signal-to-noise ratio (PSNR) or mean square error (MSE) is the earliest metric, which computes the pixel-level differences without consideration of the human visual system (HVS) characteristics. Afterward, the structural similarity (SSIM) index [9] was developed, and takes into account the brightness, contrast and structure perceptual elements of images, and has given rise to many variants [10, 11]. However, these methods still lack exploration of high-level semantic information. Recently, deep learning-based methods [12–17] have been actively explored, focusing on deep mining of semantic information of common IQA task. These methods are developed for common distortions (e.g. compression distortion, noise, and transmission errors), which involve degraded distortions. However, they will not be well suited to the enhanced distortions introduced by SR algorithms, such as over-sharpening, reconstruction artifacts, and false textures.

Except for common IQA, there have emerged some works for developing SR IQA methods. The recent SR FR-IQA methods [18–20] are predominantly rely on hand-crafted features. Although these methods have been proposed to tackle SR image artifacts like jagged edges, blurring, or non-existent artifacts in the original scene, their inability to analyze high-level semantic features curtails their performance. Meanwhile, there have been many attempts to apply deep learning to SR IQA, including CNNs [21–23], knowledge distillation [24], attention mechanisms [25, 26], etc., but these are designed

This work was in part supported by the National Natural Science Foundation of China under Grant 62371017. Corresponding author: X. Yang and W. Zhou.

Y. Li and X. Yang are with the School of Mathematical Sciences, Beihang University, Beijing 100191, China (e-mail: 18335310648@163.com; xiaoyuanyang@vip.163.com).

G. Yue is with the School of Biomedical Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: yueguanghui@szu.edu.cn).

J. Fu is with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: fujun@mail.ustc.edu.cn).

Q. Jiang is with the School of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jiangqiuping@nbu.edu.cn).

X. Jia is with the School of Artificial Intelligence, Dalian University of Technology, Dalian 116081, China (email: xjia@dlut.edu.cn).

P.L. Rosin, H. Liu and W. Zhou are with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom (email: rosinpl@cardiff.ac.uk; liuh35@cardiff.ac.uk; zhouw26@cardiff.ac.uk).

for either NR or RR scenarios. Therefore, there remains a significant gap in exploring deep-learning-based approaches for SR FR-IQA.

Recently, advanced SR NR-IQA methods [25, 26] integrates features across channel or spatial attention mechanisms, which are also common in other image processing tasks like image classification [27], segmentation [28], and object detection [29]. However, pooling operations in these attention mechanism may overlook essential information, thus is ineffective for assessing enhanced distortions. This suggests the potential for refining the specificity of attention mechanisms to better address the enhanced distortions of SR images. Thus, we propose a new Perception-oriented Bidirectional Attention Network (PBAN) more suitable for SR FR-IQA task.

To fully leverage reference information, we propose assessing the enhanced distortions of SR images from two perspectives: the "generate distortion" and "evaluate distortion" processes. Initially, acquiring SR images from original reference images or real-world scenarios introduces the unique enhanced distortions. Then, subjective evaluations by individuals assess such distortion by comparing SR images against the original reference images, or implicitly with hallucinated scenes in mind. Inspired by the dual aspects of SR image distortion—"generation" and "evaluation"—we propose the **Bidirectional Attention (Bi-Atten)**, which aligns SR images with their corresponding references in two directions (HR→SR and SR→HR). Unlike vanilla spatial or channel attention mechanisms commonly used in SR NR-IQA methods, Bi-Atten avoids pooling operations, but using the multiplication of feature matrices from both SR and reference sources to align the two feature spaces.

Besides, we obtain effective perception of distortion from the following two aspects, and conceive the overall Perception-oriented Bidirectional Attention (PBA). On the one hand, we upgrade deformable convolution [30, 31] to adaptively offer the perception of distorted areas. Deformable convolutions can effectively perform dense spatial prediction due to its adaptive sampling mechanism, which aligns with the spatial frequency sensitivity [32] of the HVS. Therefore, it has been a viable candidate for preliminary integration within existing IQA methods [33, 34]. Nonetheless, an inherent challenge persists in striking an optimal balance between sampling efficiency and computational complexity within these deformable convolutions. Under this consideration, we devise a novel architecture named **Grouped Multi-scale Deformable Convolution (GMDC)**. This approach effectively controls computational complexity through channel dimension grouping while employing multi-scale convolutional kernels to adeptly capture hierarchical features, which is inspired by the hierarchy [35] of the HVS.

On the other hand, our method draws inspiration from sub-pixel analysis [36] known for its assumption that there is even finer microscopic information between macroscopic physical pixels, which has not been expolred in SR FR-IQA task. Building upon this foundation, we assume that there exists finer distortion information between existing artifacts, thus we introduce the concept of **Sub-information Excitation Convolution (SubEC)**. This innovation leverages the mixed attention mechanism through the synergistic integration of sub-pixel and sub-channel information.

We will make our code publicly available at: https://github.com/Lighting-YXLI/PBAN. To sum up, our contributions are listed as follows:

1) We propose a deep learning-based SR FR-IQA metric – PBAN, with more accurate perceptual quality predictions, which is in line with the HVS.

2) We introduce Bi-Atten, which is inspired by the "generation" and "evaluation" processes of SR image distortion.

3) We develop GMDC to provide hierarchical and space sensitive perception to the distortion information.

4) We design SubEC which can improve the mining of finer microscopic degraded information through a mix of sub-channel and sub-pixel attention weights.

5) Our approach shows significant improvements in the experimental results, particularly when compared to other FR-IQA methods that have been evaluated in scenarios involving both natural and SR images. The visualization results and ablation studies further verify the significance of each component and the generalization to common IQA task.

II. Related Work

A. Image quality assessment

When original reference images are available, FR-IQA methods can be developed, which estimate image quality by comparing distorted images with the corresponding reference images. The earliest FR-IQA metrics include PSNR and MSE. Later, SSIM [9] combined attributes such as structure, brightness, and contrast to design a metric that is more in line with the HVS. Further variants like multi-scale SSIM (MS-SSIM) [11], complex wavelet SSIM (CW-SSIM) [10], quaternion structural similarity (QSSIM) [37] and gradient magnitude similarity deviation (GMSD) [38] were proposed to better align with the HVS. However, these methods are not always as effective as deep learning-based methods.

Deep learning-based FR-IQA methods have shown improved performance over traditional methods by leveraging the powerful feature extraction capabilities of deep learning models. MGCN [39] proposed a mask gated convolutional network for simultaneously evaluating image quality and identifying distortion. WaDIQaM [12] proposed joint learning of local quality and local weights. LPIPS [40] calculated the distance of features extracted from pre-trained networks between the reference and distorted images. However, these methods were not specifically designed for evaluating SR images and may not be effective in assessing SR image distortion. Recently, there have been some handcrafted FR-IQA methods specially tailored for SR images, including the structure-texture decomposition based algorithm (SIS) [18], the structural fidelity versus statistical naturalness (SFSN) method [19], as well as the deterministic and statistical fidelity (SRIF) method [20]. However, there is a lack of deep learning-based FR-IQA methods tailored for SR images.

In practical scenarios, original images with perfect quality are not always easy to obtain. Therefore, NR-IQA methods have been widely developed in the literature. For tradi-

tional NR-IQA, many methods extract handcrafted distortion-discriminative features to predict perceptual image quality, such as the natural image quality evaluator (NIQE) [41], and local binary pattern statistics method (LPSI) [42]. With the surge of deep learning, methods like MetaIQA [14], HyperIQA [15], GraphIQA [16] and LIQA [17] were proposed. Specifically, MetaIQA adopted meta learning to learn the prior knowledge shared by diversified distortion, while HyperIQA proposed a self-adaptive hyper network. GraphIQA represented each distortion as a graph and distinguished the type of distortion by learning the distortion graph. LIQA adopted a Split-and-Merge distillation strategy and proposed a new lifelong blind image quality assessment method to alleviate catastrophic forgetting of learning knowledge.

Existing deep learning-based SR IQA methods mainly focus on NR-IQA. For instance, DeepSRQ [21] used a two-stream CNN to extract separate structural and texture features. Considering artifacts caused by SR algorithms are sensitive to frequency information, HLSRIQA [22] utilized high-frequency and low-frequency maps of SR images. EK-SR-IQA [24] predicted SR image quality by employing a semi-supervised knowledge distillation strategy. Recently, SR NR-IQA methods have employed various attention mechanisms to further improve performance. For instance, JCSAN [25] jointly utilized channel and spatial attention to obtain more perceptually discriminative features. TADSRNet [26] constructed a triple attention mechanism to acquire more significant portions of SR images through cross-dimensionality.

Additionally, attention mechanisms are also widely applied to other tasks including image classification [27], semantic segmentation [28], and object detection [29]. However, none of them are designed for evaluating the perceptual quality of image super-resolution based on the characteristics of SR distortion, indicating that there is still room for improvement in the SR IQA task. To fill this gap, we develop a dual branch framework based on the "generation" and "evaluation" processes of SR distortion, resulting in the perception-oriented bidirectional attention network.

B. Deformable Convolution and Sub-pixel Convolution

Deformable convolution, initially introduced in Deformable ConvNets [30] and later improved in More Deformable ConvNets [31], is a technique that incorporates adaptive sizes and shapes of convolution kernels by adding offsets to sampling points. It has proven to be effective in enhancing spatially dense prediction and has found applications in various low-level vision tasks like image super-resolution [43] and video deblurring [44]. Recently, researchers have explored the use of deformable convolution in IQA to achieve performance improvements. Some notable works in this direction include RADN [33] and DQM-IQA [34]. RADN [33] utilized deformable convolutions to better leverage information from reference images, thereby increasing sensitivity to errors and misalignments in distorted images. DQM-IQA [34] employed deformable convolutions with adaptive receptive fields to extract perceptual features. However, the high computational complexity of deformable convolutions restricts the use of multi-scale convolution kernels in these methods, resulting in limitations in capturing local information.

In the field of camera imaging, sub-pixels refer to the micro-pixels that cannot be detected between macro-pixels due to limitations in the imaging capability of the camera's photosensitive element. Inspired by the concept of sub-pixels, ESPCN [36] introduced the concept of sub-pixel convolution for efficient image SR. Instead of padding or interpolation, it utilizes pixel shuffle to achieve image upsampling. Sub-pixel convolution aligns well with the human visual perception of SR images and has been widely applied in tasks such as image reconstruction [45] and pan-sharpening [46].

Inspired by the methods mentioned above, we propose Group Multi-scale Deformable Convolution (GMDC) and Sub-information Excitation Convolution (SubEC) in our proposed network for SR IQA.

III. Proposed SR IQA Method

In this section, we present the overall framework of our proposed deep learning-based SR FR-IQA method, called **Perception-oriented Bidirectional Attention Network (PBAN)**, which consists of three modules, namely the Image Encoder Module, Perception-oriented Bidirectional Attention (PBA) Module, and Quality Prediction Module. As illustrated in Fig. 1, the network takes a HR reference and SR image pair as input. The main component of our method is the PBA Module, which is the stack of PBA blocks. Each block has three key components: Group Multi-scale Deformable Convolution (GMDC), Bidirectional Attention (Bi-Atten) and Sub-information Excitation Convolution (SubEC). Specifically, we design Bi-Atten to comprehensively assess the enhanced distortions of SR images by bi-directional feature spaces alignment (HR$\rightarrow$SR and SR$\rightarrow$HR). Before calculating the attention map, we propose GMDC to provide the perception of distorted regions. After obtaining the Bi-Atten map, we propose the SubEC to further refine perception of distortion to a finer level. The design details are as follows.

A. Image Encoder Module

Given a pair of input images (i.e. SR image and the corresponding original HR reference), we first crop the images into non-overlapping patches. Following the settings of DeepSRQ [21], the patch size is set to 32×32 in our experiments. The image patches and their corresponding reference patches are then fed into a stack of a convolutional layer with 3×3 kernels, a batch normalization layer, and an activation function (i.e. ReLU) for discriminative feature extraction.

B. Perception-oriented Bidirectional Attention (PBA) Module

The feature maps of both branches are then input into the PBA Module. The feature map passes through **Bi-Atten** and **SubEC** in sequence.

1) Bidirectional Attention: Bi-Atten is an improvement upon cross-attention [47] specifically for SR FR-IQA. Cross attention is commonly used for multi-modal data (such as text and images), assuming there are inputs X_1, X_2 from

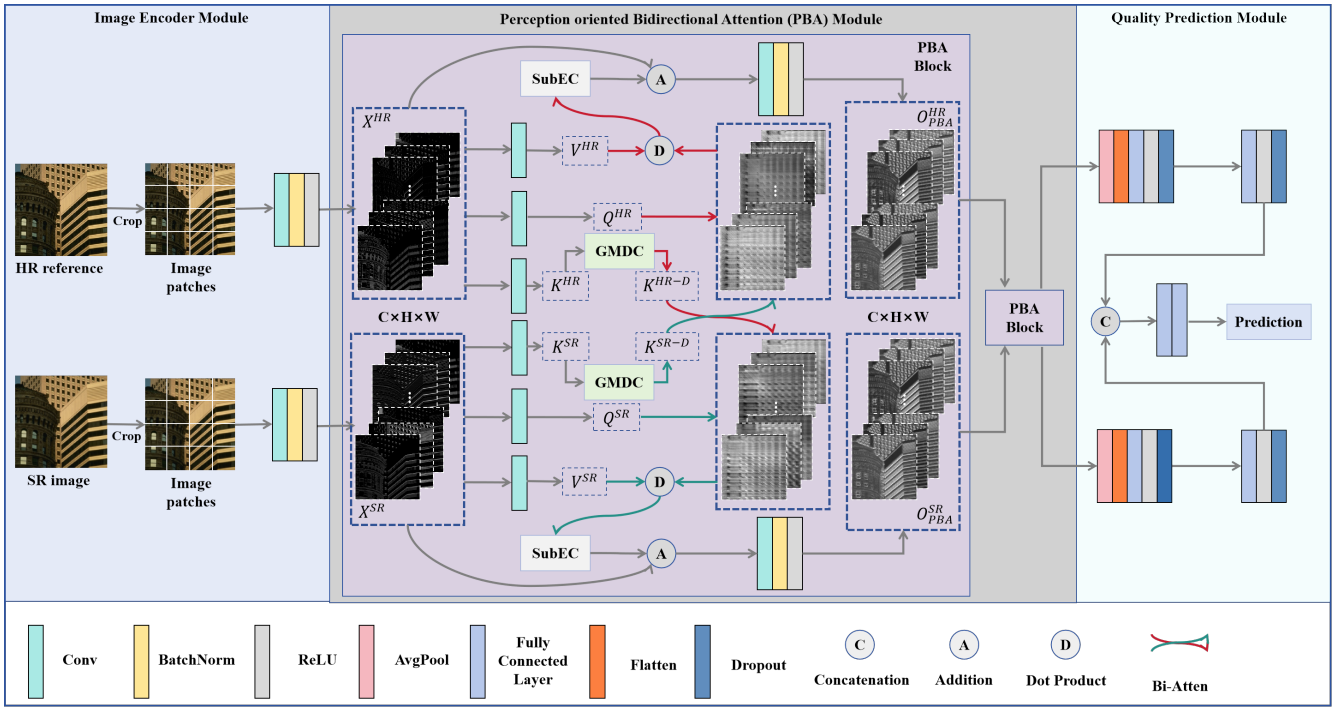

Fig. 1. The overall framework of PBAN. In the PBA Module, X^{HR}, X^{SR} are input feature maps, $O^{HR}_{PBA}, O^{SR}_{PBA}$ are output feature maps. C, H, W are the dimensions of channel, height and width of the feature map, respectively. **Bi-Atten** refers to bidirectional attention, which takes "K" from Group Multi-scale Deformable Convolution (GMDC) to compute bidirectional attention. In Bi-Atten, Q, K, V represent "query", "key", "value" of the attention mechanism. **SubEC** denotes Sub-information Excitation Convolution.

two sources. Cross attention is calculated using $Q(Query)$, $K(Key)$, and $V(Value)$ matrices to find information dependencies between cross modal data. Meanwhile, due to the inconsistency of data between different modalities and in order to ensure contextual consistency between $K\&V$, it is necessary to ensure that $K\&V$ are homologous. The cross attention first obtains through linear layers:

$$\begin{aligned} Q_1, Q_2 &= W^Q(X_1, X_2), \\ K_1, K_2 &= W^K(X_1, X_2), \\ V_1, V_2 &= W^V(X_1, X_2), \\ D_1, D_2 &= Var(Q_1 K_2^T), Var(Q_2 K_1^T), \end{aligned} \tag{1}$$

where W refers to the fully connected linear layers. D represents the variance of the dot product of Q and K. Then the cross attention is calculated as follows:

$$\begin{aligned} \text{Cross-Attention}(X_1) &= \text{Softmax}\left(\frac{Q_1 K_2^\top}{\sqrt{D_1}}\right) V_2, \\ \text{Cross-Attention}(X_2) &= \text{Softmax}\left(\frac{Q_2 K_1^\top}{\sqrt{D_2}}\right) V_1. \end{aligned} \tag{2}$$

Here, the cross attention reflects the relevance between each element from one source X_1 and elements from the other source X_2, thereby achieving effective attention between multi-model data.

To comprehensively leverage the reference information of SR images, we provide a concise overview of the baseline, illustrated in Fig. 1. Inspired by the "generate distortion" and "evaluate distortion" processes, we propose the Bi-Atten, a dual-branch framework tailored for SR FR-IQA task. For the "generate distortion" process, the upper branch uses HR reference images as input. Through Bi-Atten, the feature space of HR references is progressively aligned to that of SR images. This is achieved by multiplying the HR feature matrix with the SR feature matrix, where each row of the former is weighted by each column of the latter, resulting in an attention map with in the SR feature space. In the "evaluate distortion" process, human evaluators typically compare SR images to corresponding references or real scenes to assess distortion levels. Considering this, the lower branch takes SR images as input and incrementally aligns them with HR references using Bi-Atten, effectively capturing the degree of misalignment.

The strong contextual relationship between SR images and corresponding HR references is quite different from cross model data. **Therefore, maintaining "$K\&V$" homology is not necessary for SR FR-IQA, but rather "$Q\&V$" homology can emphasize the subtle differences (e.g. distortion) between SR images and corresponding HR references.** So the proposed Bi-Atten utilizes the visual information from one branch as "query & value" and information from the other branch as "key" to achieve attention on distortion, which is different from generic cross attention.

We use 3×3 convolutional layers to obtain the Q, K, and V matrices:

$$\begin{aligned} Q^{HR}, K^{HR}, V^{HR} &= Conv(X^{HR}), \\ Q^{SR}, K^{SR}, V^{SR} &= Conv(X^{SR}), \end{aligned} \tag{3}$$

where $X^{HR}, X^{SR} \in R^{C \times 32 \times 32}$ are SR image patch and its corresponding HR reference patch. C is the channel dimension.

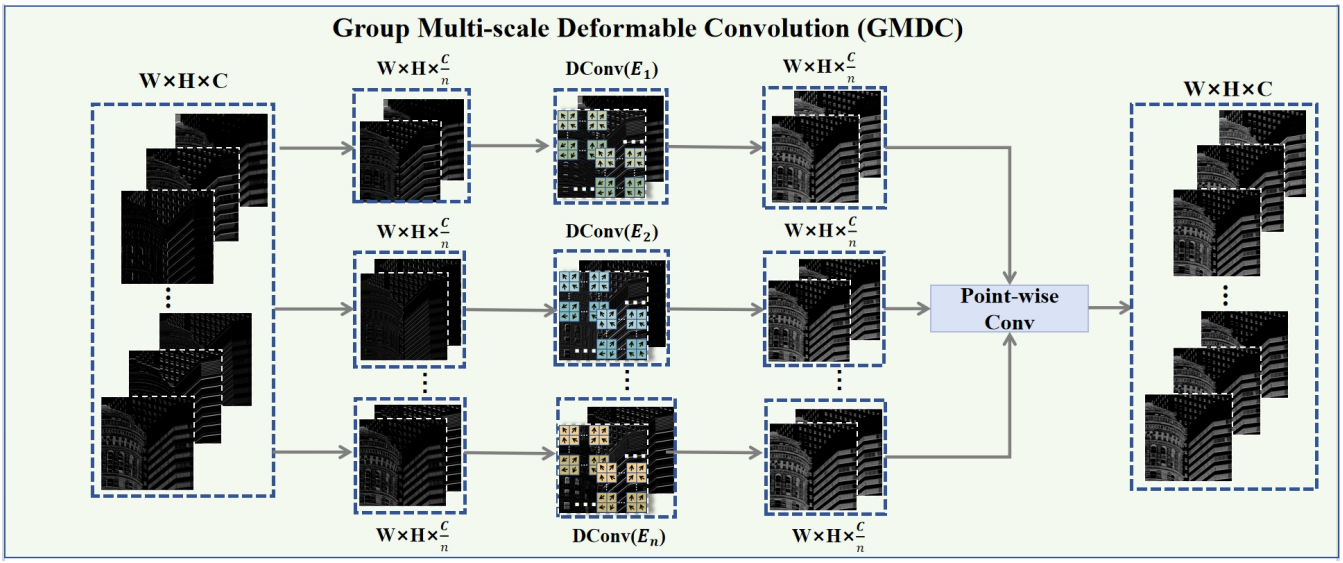

Fig. 2. The framework of **GMDC**. Given an input feature map of $(W \times H \times C)$, where C is the number of channels, H, W are the height and width. It is divided into "n" separate groups. Each group contains "$(\frac{C}{n})$" channels, and is then fed into deformable convolutions (i.e. DConv) with $E_i \times E_i, i = 1, \ldots, n$ kernel sizes. E_i can be set to multi-scale. The final Point-wise Convolution is utilized to provide the interaction between groups.

Then the two branches exchange K^{HR} and K^{SR} for information interaction. In Bi-Atten, we further propose **Group Multi-scale Deformable Convolution (GMDC)** to provide the adaptability of spatial dense predictions, thereby improving the perception of visual attention. The following is a detailed introduction to GMDC:

In standard convolution, the convolution operation can be calculated as follows:

$$F\left(p_0\right)=\sum_{p_n \in r} w\left(p_n\right) \cdot y\left(p_0+p_n\right), \tag{4}$$

where p_0 is a pixel of the output feature map y, r denotes the regular grid of its corresponding kernel, p_n enumerates the relative locations of sampling points in r. w denotes the weight for p_n. Deformable Convolution (DConv) utilized in IQA methods [33, 34] adds vertical and horizontal offsets Δp_n to each sampling point of a convolution kernel:

$$F\left(p_0\right)=\sum_{p_n \in r} w\left(p_n\right) \cdot y\left(p_0+p_n+\Delta p_n\right). \tag{5}$$

The offsets are learned from the preceding feature map y via a convolution layer. However, the addition of offsets in DConv introduces a significant number of extra parameters, making it impractical to use larger convolution kernels and limiting its learning capacity. Instead, group convolution [48] can obtain more feature maps while maintaining the same parameter quantity as standard convolution, which can provide the feature extraction capability. As shown in Fig. 2, given inputs with the size of $(W \times H \times C)$, where C is the number of channels, H, W are the height and width of the input. We divide channels into n groups (i.e. each group has $(\frac{C}{n})$ channels). Different groups undergo DConv operations independently. To capture information at multiple scales, we apply multi-sized convolution kernels with edge length $E_i, i = 1, \ldots, n$ to different groups. Take one convolutional layer as example, the original parameters for DConv is $3 \times C' \times C \times E^2$, and the parameters for the proposed GMDC is $3 \times C' \times C \times \frac{E_1^2+\cdots+E_n^2}{n}$, where the C' refers to channel of the output features, $E, E_i, i = 1, \ldots, N$ refer to kernel sizes. In our experiments, we set E to 5, n to 2, E_1 to 3, and E_2 to 7. With these settings, the parameter count of GMDC is 1.16 times that of DConv, while the feature maps obtained are twice as many as those of DConv. After going through GMDC, K^{HR} and K^{SR} are swapped and involved in the computation of the Bi-Atten map in the other branch:

$$\begin{aligned} K^{HR-D}, K^{SR-D} &= GMDC(K^{HR}, K^{SR}), \\ D^{HR} &= Var(Q^{HR}(K^{SR-D})^T), \\ D^{SR} &= Var(Q^{SR}(K^{HR-D})^T), \end{aligned} \tag{6}$$

where D^{HR} and D^{SR} are the variance of the dot product of Q^{HR}, K^{SR-D} and Q^{SR}, K^{HR-D}, respectively. Note that the proposed Bi-Atten is different from cross attention and common spatial/channel attention, which are further analyzed in SectionV-E and Section V-F, respectively. Then the Bidirectional Attention maps are calculated as:

$$\begin{aligned} &\text{Bi-Atten}^{\text{HR}}(Q^{HR}, K^{SR-D}, V^{HR}) \\ &= \text{Softmax}\left(\frac{Q^{HR}{K^{SR-D}}^{\top}}{\sqrt{D^{HR}}}\right)V^{HR}, \end{aligned} \tag{7}$$

$$\begin{aligned} &\text{Bi-Atten}^{\text{SR}}(Q^{SR}, K^{HR-D}, V^{SR}) \\ &= \text{Softmax}\left(\frac{Q^{SR}{K^{HR-D}}^{\top}}{\sqrt{D^{SR}}}\right)V^{SR}. \end{aligned} \tag{8}$$

2) Sub-information Excitation Convolution: After Bi-Atten, we introduce SubEC to further extract finer distortion information in both the spatial (pixel) and the channel dimensions of the obtained Bi-Atten feature maps.

Based on the concept of sub-pixel [36], we assume that

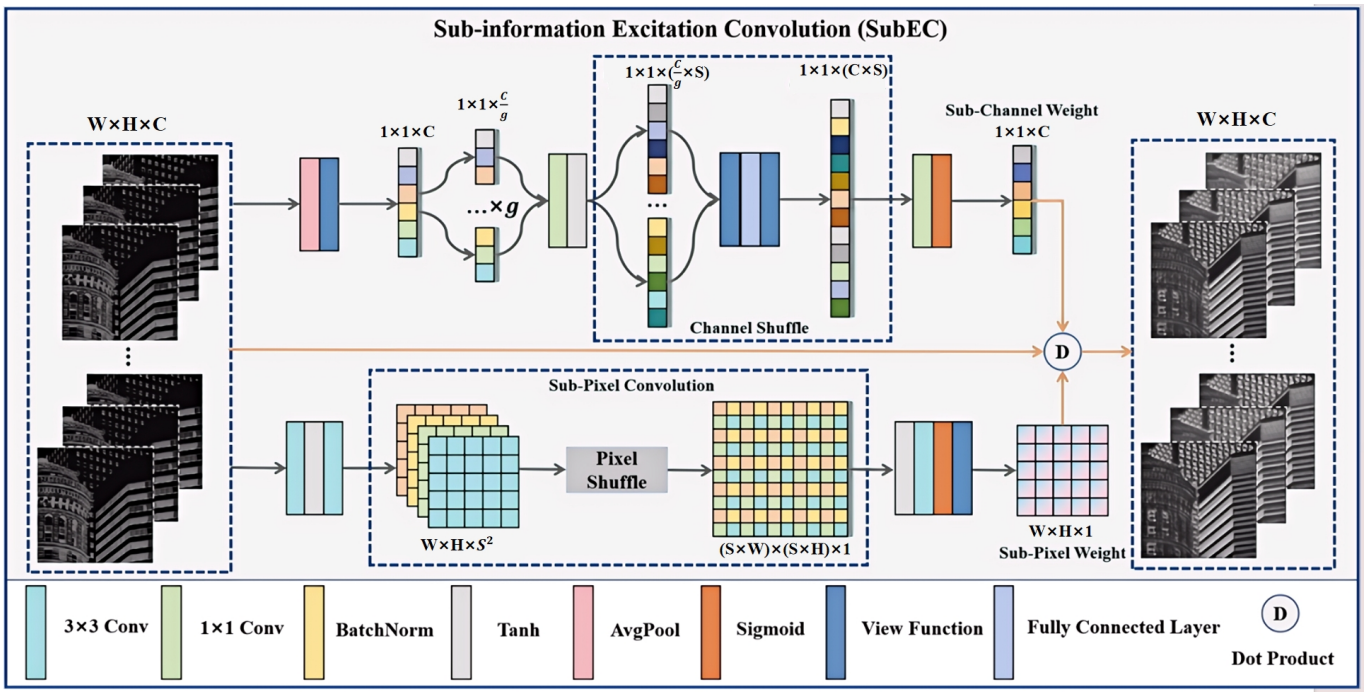

Fig. 3. The framework of **SubEC**. It is a three-branch architecture, including **Sub-Channel Weight** branch (i.e. the top branch), **Sub-Pixel Weight** branch (i.e. the bottom branch), and the identity shortcut branch (i.e. the mid one). The output feature map is obtained by multiplying the input feature map with these two weights. S is the magnification of the channel and pixel dimensions, which is set to 2 in our experiment.

micro-distorted information exists in the sub-pixels between pixels in the feature maps. Simultaneously, it also exists in sub-channels, which are assumed to be mined between the existing channels. Therefore, we mine sub-information through super-resolution of both the channel and spatial dimensions.

The proposed SubEC framework, illustrated in Fig. 3, takes an input feature map of size ($W \times H \times C$) and generates the Sub-Channel Weight and Sub-Pixel Weight through separate branches.

In the Sub-Pixel Weight branch, we employ sub-pixel convolution [36] for pixel-level super-resolution. With an upscaling factor of S, each pixel in the feature map of size ($W \times H$) is assigned weights. Consequently, the final size of Sub-Pixel Weight should be ($W \times H \times 1$). Sub-pixel convolution upsamples a single pixel into a square matrix of size ($S \times S$). To achieve this, we first reduce the channel dimension to S^2 and then use sub-pixel convolution to shuffle the pixels, as illustrated in Fig. 2. This process generates an upsampled feature map of size ($(S \times W) \times (S \times H) \times 1$). To avoid losing channel information when reducing the number of channels from C to S^2, we gradually decrease the dimension using two 3×3 convolutional layers. Finally, the upsampled feature map is downsampled to ($W \times H \times 1$) to obtain the Sub-Pixel Weight.

In the Sub-Channel Weight branch, we amplify the channel dimension to capture micro-information hidden within sub-channels. Since the final weights are applied to each input channel, the size of Sub-Channel Weight needs to be ($1 \times 1 \times C$). Given an input feature map, we first reduce the spatial dimension from ($W \times H$) to (1×1) using adaptive average pooling. Then, we perform super-resolution on the channel dimension by a factor of S. To provide computational efficiency, we employ group convolution [48] to divide the feature map into g groups, with each group having a size of ($1 \times 1 \times \frac{C}{g}$). Subsequently, a 1×1 convolution is applied to expand the channels by S times. To restore the interdependencies among channels lost during grouping, we utilize the channel shuffle operation [49], resulting in a feature map of size ($1 \times 1 \times (C \times S)$). Finally, we compress the channels using a 1×1 convolution to obtain the Sub-Channel Weight. In summary, the output feature map O_{SubEC} of SubEC can be expressed as:

$$O_{SubEC} = I_{SubEC} \times W_{Sub-Channel} \times W_{Sub-Pixel}, \quad (9)$$

where I_{SubEC} is the input of SubEC. Taking into account that the Identity Shortcut [50] has been proven to effectively alleviate model overfitting issues, we use the Shortcut as the main architecture:

$$\begin{aligned} O^{HR}_{PBA} &= O^{HR}_{SubEC} + X^{HR}, \\ O^{SR}_{PBA} &= O^{SR}_{SubEC} + X^{SR}, \end{aligned} \quad (10)$$

where $O^{HR}_{PBA}, O^{SR}_{PBA}$ are output feature maps of PBA block.

C. Quality Prediction Module

After passing through the PBA Module, feature maps are fed into the **Quality Prediction Module**, where the feature maps of both branches are flattened and concatenated:

$$\begin{aligned} O^{HR} &= ReLU(L(F(AvgPool(O^{HR}_{PBA})))), \\ O^{SR} &= ReLU(L(F(AvgPool(O^{SR}_{PBA})))), \\ O^{HR} &= Dropout(ReLU(L(Dropout(O^{HR})))), \\ O^{SR} &= Dropout(ReLU(L(Dropout(O^{SR})))), \end{aligned} \quad (11)$$

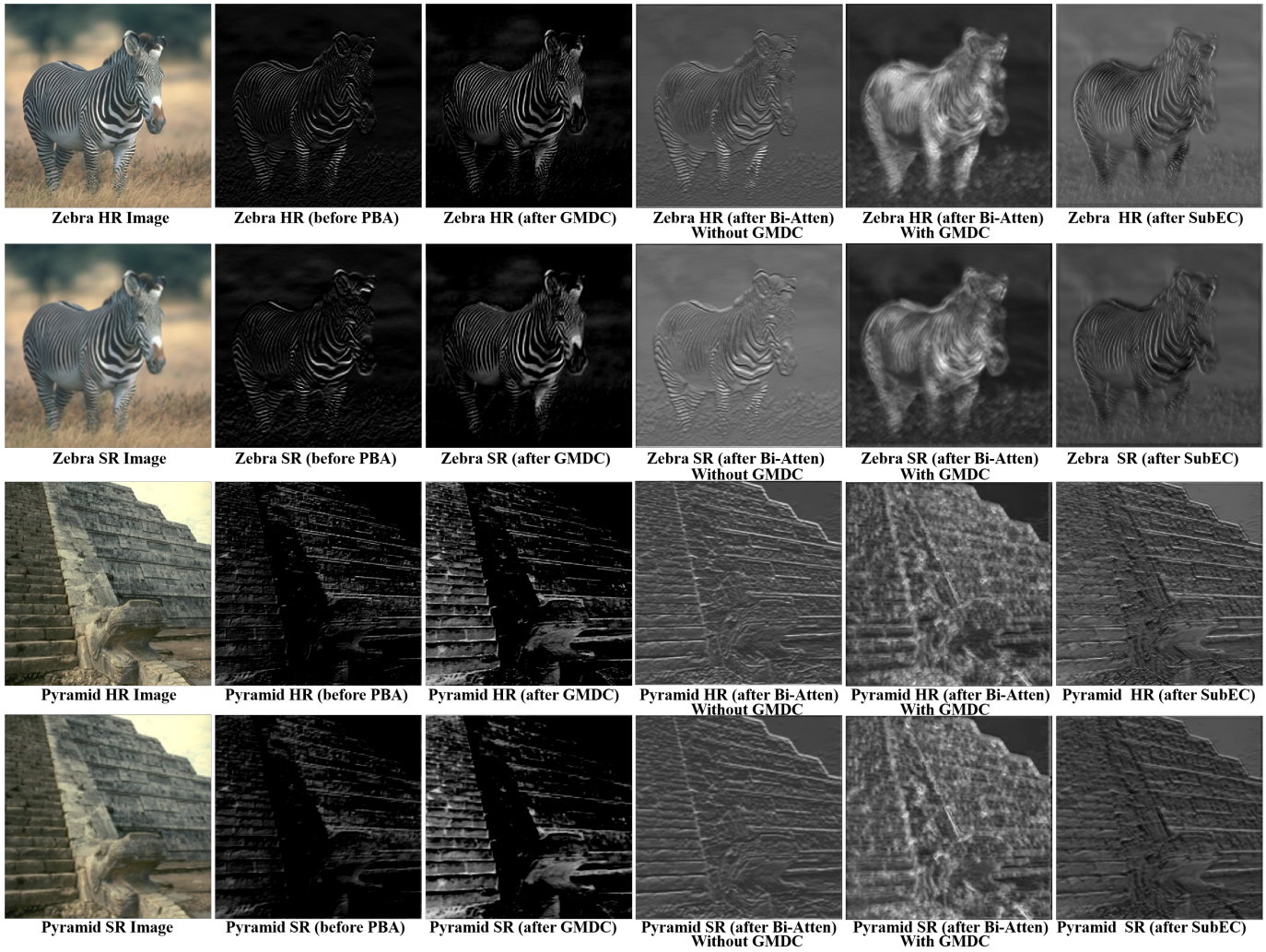

Fig. 4. Visualization comparisons of feature maps regarding the proposed PBA block. **Zebra/Pyramid HR image** and **Zebra/Pyramid SR image** are HR reference and SR image, respectively. The remaining images are feature maps before and after the first PBA block and its components in two branches.

where the Average Pooling layer is utilized to downsample O_{PBA}^{HR} and O_{PBA}^{SR} to prevent the input dimension of the fully connected layer from being too large. F is the flattened layer and L is the fully connected layer. These Linear layers gradually reduce the feature dimension from 1,024 to 1, resulting in the final prediction score.

Afterward, the perceptual quality predictions are ultimately obtained:

$$\begin{aligned} O &= Concate(O^{HR}, O^{SR}), \\ Prediction &= Linear(Linear(O)). \end{aligned} \tag{12}$$

Finally, following the protocol of DeepSRQ [21], we adopt the MSE loss to provide guidance in the network optimization process:

$$Loss = \frac{1}{B}\sum_{i=1}^{B}(Prediction_i - MOS_i)^2, \tag{13}$$

where B is the batch size of the input data.

IV. Visualization

To illustrate the proposed PBAN and its components in effectively enhancing the model's visual attention on SR image distortion, Fig. 4 shows the intermediate feature maps of the two branches. The two branches take the HR reference images and SR images of "zebra" and "pyramid" as inputs, respectively. It can be observed that the initial feature maps extracted by the image encoder module (referred to as "image before PBA") contain fewer distortion details. After passing through GMDC, the feature maps of K^{HR-D}, K^{SR-D} exhibit improved visual attention to the "zebra" and "pyramid", indicating that the multi-scale adaptive sampling of GMDC effectively provides spatially dense prediction capabilities. Then, the Bi-Atten is calculated to obtain Bi-Atten maps between two branches, resulting in feature maps that significantly provide the visual attention to areas with artifacts. Finally, after going through SubEC and outputting one PBA block, further refinement is made to the distortion details.

In order to visually illustrate the perceptual gain that GMDC brings to Bi-Atten, we visualized the feature maps of Bi-Atten

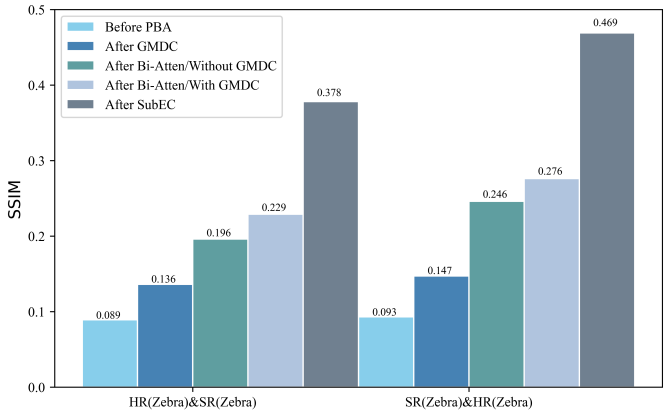

Fig. 5. The structure similarity (i.e. SSIM) of "$X_1\&X_2$", where X_1 is initial HR reference or SR image, X_2 is the **feature map** before or after one PBA block and its components (i.e. GMDC located in Bi-Atten, Bi-Atten w/o GMDC, Bi-Atten with GMDC, and SubEC). Note that the higher the SSIM, the more similar the two images are.

(without GMDC) and found that Bi-Atten (with GMDC) pays more attention to distortion (e.g. zebra texture, brick structure).

Furthermore, the PBA Module is designed in accordance with SR image distortion "generation" and "evaluation" to obtain more comprehensive predictions that are in line with human perception. Thus during iterative learning, the feature maps of the branch with the HR reference as input should gradually approach the SR image, and vice versa for the branch with the SR image as input, which aligns with the "generation" and "evaluation" processes, respectively. We measure the level of approximation between the feature maps of "zebra" using similarity metrics such as SSIM, as shown in Fig. 5.

The feature maps of the SR image show a significant improvement in SSIM with the HR reference after receiving visual attention from the first PBA block. Similarly, the feature maps of the HR reference also exhibit a significant SSIM improvement with the SR image after the PBA block. This indicates that our method dynamically provides visual perception to distortion as the "HR" and "SR" transform into each other.

V. Experiments

In this section, we compare our method with many existing quality metrics on publicly subject-rated quality databases for SR image. Besides, ablation studies are also conducted to verify the performance of each proposed component.

A. Experimental setup and implementation

In training and testing, experiments are running on a NVIDIA DGX Station with a Tesla V100-DGXS-32GB GPU and Ubuntu18.04 LTS. All codes are implemented in PyTorch.

1) Datasets: We conduct experiments on QADS [18], CVIU [51], and Waterloo [52] databases. The QADS database contains 20 original HR references and 980 SR images created by 21 SR algorithms, including 4 interpolation-based, 11 dictionary-based, and 6 DNN-based SR models, with upsampling factors equaling 2, 3, and 4. Each SR image is associated with the MOS of 100 subjects. In the CVIU database, 1,620 SR images are produced by 9 SR approaches from 30 HR references. Six pairs of scaling factors and kernel widths are adopted, where a larger subsampling factor corresponds to a larger blur kernel width. Each image is rated by 50 subjects, and the mean of the median 40 scores is calculated for each image as the MOS. The Waterloo database involves 8 interpolation algorithms with 3 interpolation factors of 2, 4, and 8, respectively. 312 SR images are generated from 13 source images.

2) Implementation details: All databases are randomly divided into non-overlapping 80% and 20% sets, with 80% of the data used for training and the remaining 20% for testing. During training, we apply data augmentation methods to alleviate overfitting. These methods include dividing each SR image and its corresponding reference image into non-overlapping patches. Following the protocol in DeepSRQ [21], each patch is assigned the same MOS as corresponding SR image, and the final prediction of each SR image is calculated by averaging the predictions of its patches. Then using random-horizontal-flip when loading the data. We utilize 5-fold cross validation on the training set, with each fold training for 100 epochs, and ultimately tested it on the test set. We use MSE loss to measure the difference between predicted scores and MOSs. The optimizer used is stochastic gradient descent (SGD), with an initial learning rate of 0.01, momentum of 0.9, and weight decay setting to 10^{-6}.

We adopt four commonly used evaluation criteria to compare performance, including Spearman rank-order correlation coefficient (SRCC), Kendall rank-order correlation coefficient (KRCC), Pearson linear correlation coefficient (PLCC), and root mean square error (RMSE). SRCC and PLCC/RMSE are employed to assess the monotonicity and accuracy of predictions, respectively. KRCC is used to measure the ordinal association between two measured quantities. An ideal quality metric would have SRCC, KRCC, and PLCC values close to one, and RMSE close to zero. It should be noted that a five-parameter nonlinear fitting process [20] is applied to map the predicted qualities into a standardized scale of subjective quality labels before calculating PLCC and RMSE for different quality metrics.

B. Performance Comparisons

To validate the proposed method, we compare it with state-of-the-art FR-IQA, NR-IQA, and SR IQA methods. FR-IQA methods include PSNR, SSIM [9], MS-SSIM [11], CW-SSIM [10], GMSD [38], WaDIQaM [12], and LPIPS [40]. NR-IQA methods consist of NIQE [41], LPSI [42], MetaIQA [14], and HyperIQA [15]. Among them, WaDIQaM, LPIPS, MetaIQA, and HyperIQA are deep learning-based models. SR IQA methods contain SIS [18], SFSN [19], SRIF [20], DeepSRQ [21], HLSRIQA [22], EK-SR-IQA [24], JCSAN [25], and TADSRNet [26]. Besides, DeepSRQ, HLSRIQA, EK-SR-IQA, JCSAN and TADSRNet are deep learning-based methods.

The performance comparison is presented in Table I, indicating that deep learning-based methods generally outperform

TABLE I
PERFORMANCE COMPARISONS ON THE QADS [18], CVIU [51], AND WATERLOO [52] QUALITY DATABASES, WHERE THE BEST TEST PERFORMANCE VALUES OF FR AND NR ARE IN RED AND BLUE, RESPECTIVELY.

		QADS				CVIU				Waterloo			
Types	Methods	SRCC ↑	KRCC↑	PLCC ↑	RMSE↓	SRCC↑	KRCC ↑	PLCC↑	RMSE↓	SRCC↑	KRCC↑	PLCC↑	RMSE ↓
FR-IQA	PSNR	0.354	0.244	0.390	0.253	0.566	0.394	0.578	1.962	0.632	0.442	0.630	2.002
	SSIM [9]	0.529	0.369	0.533	0.233	0.629	0.443	0.650	1.828	0.613	0.431	0.621	2.022
	MS-SSIM [11]	0.717	0.530	0.724	0.190	0.805	0.601	0.811	1.405	0.825	0.623	0.837	1.411
	CW-SSIM [10]	0.326	0.228	0.379	0.254	0.759	0.541	0.754	1.579	0.863	0.666	0.906	1.094
	GMSD [38]	0.765	0.569	0.775	0.174	0.847	0.650	0.850	1.267	0.797	0.592	0.811	1.509
	WaDIQaM [12]	0.871	_	0.887	0.128	0.872	_	0.886	1.304	_	_	_	_
	LPIPS [40]	0.881	_	0.873	0.129	0.849	_	0.852	1.313	_	_	_	_
NR-IQA	NIQE [41]	0.398	0.279	0.404	0.251	0.653	0.478	0.666	1.794	0.626	0.465	0.672	1.911
	LPSI [42]	0.408	0.289	0.422	0.249	0.488	0.350	0.537	2.027	0.667	0.464	0.701	1.840
	MetaIQA [14]	0.826	_	0.790	0.178	0.720	_	0.746	1.718	_	_	_	_
	HyperIQA [15]	0.954	0.815	0.957	0.099	0.933	0.772	0.928	1.017	_	_	_	_
SR NR-IQA	DeepSRQ [21]	0.953	_	0.956	0.077	0.921	_	0.927	0.904	0.907	_	0.904	_
	HLSRIQA [22]	0.961	0.829	0.950	0.741	0.948	0.810	0.948	0.775	_	_	_	_
	EK-SR-IQA [24]	0.963	_	0.966	_	0.953	_	0.951	_	0.915	_	0.905	_
	JCSAN [25]	0.971	0.858	0.973	0.065	0.949	0.808	0.957	0.777	_	_	_	_
	TADSRNet [26]	0.972	0.862	0.974	0.067	0.952	0.812	0.959	0.797	_	_	_	_
SR FR-IQA	SIS [18]	0.913	0.740	0.914	0.112	0.869	0.686	0.897	1.061	0.878	0.677	0.891	1.169
	SFSN [19]	0.841	0.655	0.845	0.147	0.871	0.680	0.885	1.120	0.887	0.692	0.906	1.093
	SRIF [20]	0.916	0.746	0.917	0.109	0.886	0.704	0.902	1.039	0.916	0.730	0.953	0.786
	Proposed PBAN	0.986	0.923	0.987	0.044	0.978	0.872	0.981	0.397	0.965	0.848	0.979	0.587

TABLE II
ABLATION OF INDIVIDUAL PROPOSED COMPONENTS. COMPARISON OF THE TEST RESULTS OF PBAN AND VARIANTS WITHOUT EACH COMPONENT (I.E. BI-ATTEN, GMDC, AND SUBEC) ON BOTH THE QADS AND CVIU DATABASES.

Component			QADS				CVIU			
Bi-Atten	GMDC	SubEC	SRCC↑	KRCC↑	PLCC↑	RMSE ↓	SRCC↑	KRCC ↑	PLCC ↑	RMSE↓
×	×	×	0.936	0.795	0.937	0.103	0.942	0.808	0.954	0.815
✓	×	×	0.981	0.895	0.982	0.055	0.972	0.862	0.976	0.515
✓	×	✓	0.984	0.908	0.986	0.050	0.976	0.871	0.977	0.436
×	×	✓	0.961	0.849	0.962	0.074	0.966	0.845	0.969	0.641
✓	✓	✓	**0.986**	**0.923**	**0.987**	**0.044**	**0.978**	**0.872**	**0.981**	**0.397**

TABLE III
VALIDITY OF GMDC ON QADS AND CVIU DATABASES. THE MODEL WITHOUT GMDC IS TO UTILIZE STANDARD CONVOLUTION (3×3) TO REPLACE GMDC FOR EXPERIMENTATION. NOTE THAT THE SUBEC ARE INVOLVED IN ALL VARIANTS. THE FLOPS ARE CALCULATED WITH THE INPUT SIZE $3 \times 32 \times 32$.

				QADS				CVIU			
Models	Params(M)	Flops(G)	Speed(Ms)	SRCC ↑	KRCC↑	PLCC ↑	RMSE ↓	SRCC↑	KRCC↑	PLCC↑	RMSE↓
w/o GMDC	1.407	0.062	7.946	0.984	0.908	0.986	0.050	0.976	0.871	0.977	0.436
with 5×5 Dconv	1.451	0.084	9.825	0.983	0.916	0.985	0.049	0.973	0.864	0.976	0.515
with GMDC	1.458	0.088	10.785	**0.986**	**0.923**	**0.987**	**0.044**	**0.978**	**0.872**	**0.981**	**0.397**

TABLE IV
VALIDITY FOR THE NUMBER OF GROUPS IN GROUP CONVOLUTION OF GMDC. AS GMDC UTILIZES MULTI-SCALE KERNELS, THE MINIMUM NUMBER OF GROUPS IS 2. IN EXPERIMENTS WITH DIFFERENT NUMBERS OF GROUPS, HALF OF THE GROUPS HAD KERNEL SIZES SET TO 3, WHILE THE OTHER HALF HAD KERNEL SIZES SET TO 7.

Database	CVIU			
GMDC (Groups)	SRCC ↑	KRCC↑	PLCC↑	RMSE ↓
Groups 2	**0.978**	**0.872**	**0.981**	**0.397**
Groups 4	0.974	0.866	0.977	0.551
Groups 8	0.972	0.864	0.976	0.541
Groups 16	0.967	0.849	0.971	0.681

TABLE V
THE COMPARISON OF COMPUTATIONAL EFFICIENCY, WITH THE BEST AND SECOND-BEST RESULTS HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY. FLOPS ARE CALCULATED USING AN INPUT TENSOR OF SHAPE $3 \times 512 \times 512$, AND INFERENCE SPEED IS AVERAGED OVER 100 RUNS. ALL EVALUATIONS ARE CONDUCTED ON THE QADS DATASET.

Model	Params/M ↓	Flops/G ↓	Speed/ms ↓	SRCC ↑	PLCC↑
WaDIQaM [12]	6.287	30.479	8.170	0.871	0.887
HyperIQA [15]	27.375	107.831	29.980	0.954	0.957
CLIP-IQA	-	61.065	21.100	0.931	0.833
TOPIQ	36.039	50.138	22.140	0.969	0.970
DeepSRQ	1.317	1.589	2.291	0.953	0.956
TADSRNet	2.119	303.076	87.568	0.972	0.974
PFIQA	38.772	127.109	18.14	0.982	0.983
PBAN	1.458	13.505	15.108	0.986	0.987

hand-crafted based methods. Among the traditional IQA approaches, FR-IQA performs better than NR-IQA, suggesting that effectively utilizing reference image information is beneficial for predicting image quality. However, traditional IQA methods overall show inferior performance compared to SR IQA methods, underscoring the importance of designing IQA metrics specifically for SR images.

In the domain of SR NR-IQA methods, attention-based approaches like JCSAN and TADSR achieve the best performance, demonstrating the effectiveness of visual attention.

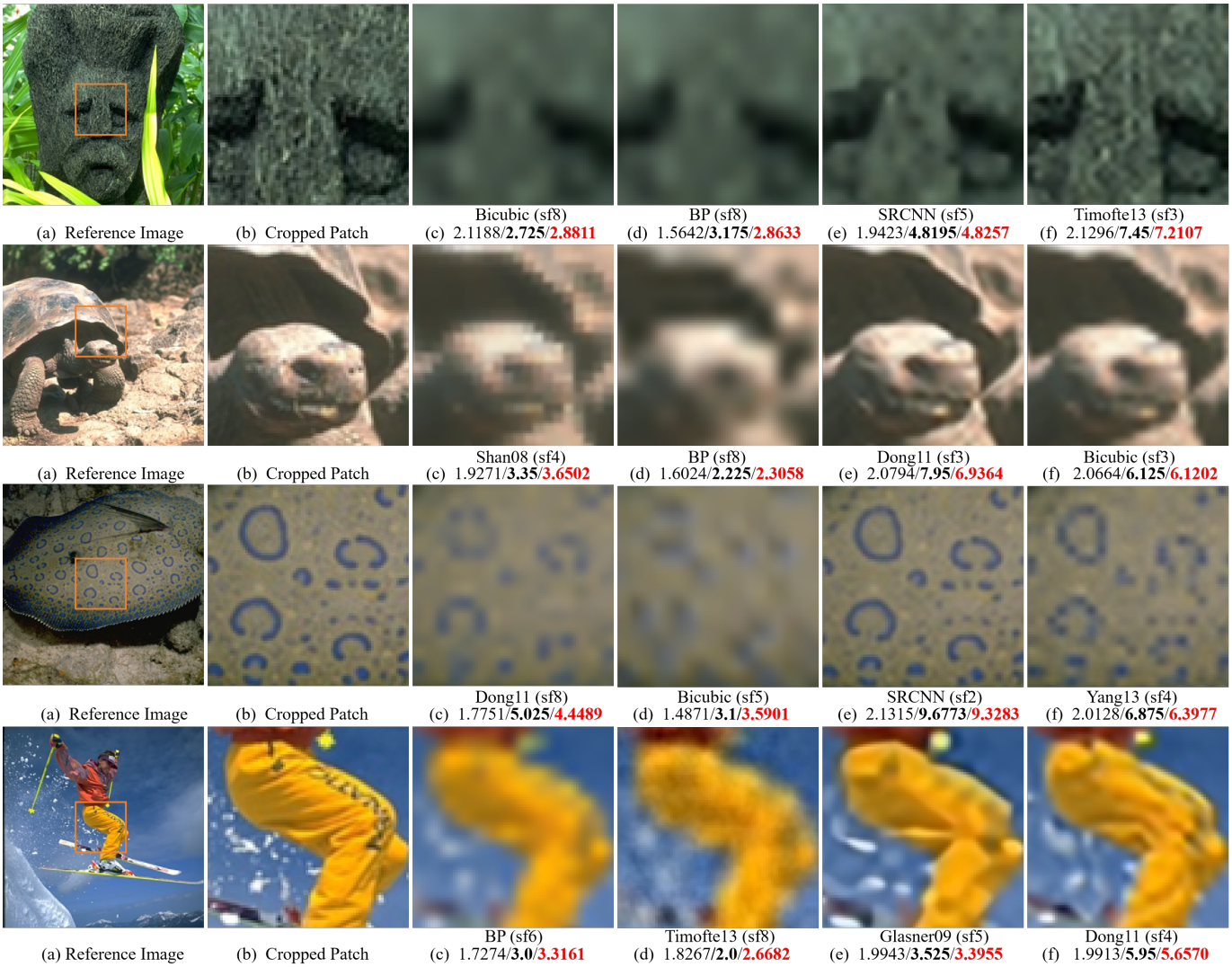

Fig. 6. Examples for predicted scores on the CVIU dataset (test): SFSN/ PBAN/ MOS. The first column has HR reference images, and column (c)-(d) are the selected SR images. All images are cropped for better visibility. In order to observe significant differences in the quality of SR images participating in the comparison, we screen images with various SR algorithms and scaling factors (i.e. sf), which are presented in detail.

On the other hand, in SR FR-IQA methods, SIS, SFSN, and SRIF are handcrafted feature-based methods. Although their design concepts are commendable, they are constrained by shallow features and fail to fully exploit the high-level semantic information present in SR images and HR references. As a deep learning-based approach, our method achieves the best performance. And there exists a considerable gap between these SR FR-IQA methods and the proposed method.

For visual comparison, we present some SR images along with the predicted quality scores by our model and another SR FR-IQA model in Fig. 6. Considering SFSN is one of the latest state-of-the-art SR FR-IQA metric and its code is easy to reproduce and validate, we take SFSN for comparison. Also, we provide MOS of these images. Fig. 6 shows that the SFSN model prefers images with relatively low quality, such as Fig. 6 (c), (d), which suffer apparent structural and textual distortion compared with their reference image. In contrast, the proposed method prefers Fig. 6 (e), (f), which is consistent with human visual perception. Besides, the low score deviation between the proposed method and MOS indicates the superiority of our model on SR IQA.

In the following sections, the performance results of each component in the PBA Module: Bi-Atten (including GMDC) and SubEC, are tested to demonstrate their importance. Furthermore, the cross-database evaluation illustrates the generalization of our method.

C. Ablation of Individual Proposed Components

In this section, we compare the effectiveness of each component in PBAN, and the results are shown in Table II. It can be seen that the complete PBAN performs the best on two datasets, while Bi-Atten has the greatest impact on performance. The model with all components removed has the worst performance, but adding only SubEC can also bring significant performance gains. In addition, the improvement effect of GMDC on Bi-Atten is also considerable.

TABLE VI
VALIDITY OF BI-ATTEN ON QADS AND CVIU DATABASES. IN ORDER TO SEPARATELY DEMONSTRATE THE ROLE OF BI-ATTEN, ALL VARIANTS DO NOT INCLUDE GMDC AND SUBEC. HR → SR AND SR → HR REPRESENT ONE-WAY "KEY" TRANSMISSION, WHILE BI-ATTEN HAS BIDIRECTIONAL "KEY" TRANSMISSION. "KEY" TRANSMISSION MEANS THE ATTENTION MAPS FOR ONE BRANCH ARE CALCULATED USING THE K MATRIX FROM THE OTHER BRANCH.

		QADS				CVIU			
Types	Models	SRCC↑	KRCC↑	PLCC↑	RMSE↓	SRCC↑	KRCC↑	PLCC↑	RMSE↓
SR FR-IQA	w/o Bi-Atten	0.936	0.795	0.937	0.103	0.942	0.808	0.954	0.815
	HR→SR	0.957	0.840	0.958	0.079	0.970	0.853	0.971	0.706
	SR→HR	0.955	0.834	0.957	0.082	0.960	0.829	0.960	0.763
	with Bi-Atten	**0.981**	**0.895**	**0.982**	**0.055**	**0.972**	**0.862**	**0.976**	**0.515**

TABLE VII
VALIDITY OF TYPES OF CROSS ATTENTION. GENERAL CROSS ATTENTION USES $K\&V$ HOMOLOGY, WHILE THE PROPOSED BI-ATTEN USES $Q\&V$ HOMOLOGY. NOTE THAT $Q\&K$ HOMOLOGY IS SELF ATTENTION, AND IS SHOWN IN TABLE VI AS W/O BI-ATTEN.

Database	QADS				CVIU			
Same Source	SRCC↑	KRCC ↑	PLCC↑	RMSE↓	SRCC ↑	KRCC↑	PLCC ↑	RMSE ↓
K&V	0.974	0.877	0.976	0.062	0.972	0.861	0.974	0.552
Q&V	**0.986**	**0.923**	**0.987**	**0.044**	**0.978**	**0.872**	**0.981**	**0.397**

TABLE VIII
VALIDITY OF SUBEC ON QADS DATABASE. BECAUSE SUBEC IS PROPOSED TO FURTHER IMPROVE BI-ATTEN, THE PERFORMANCE OF SUBEC IS COMPARED UNDER DIFFERENT ATTENTION MODES.

		Without SubEC				With SubEC			
Types	Models	SRCC ↑	KRCC↑	PLCC ↑	RMSE ↓	SRCC ↑	KRCC↑	PLCC ↑	RMSE↓
SR FR-IQA	w/o Bi-Atten	0.936	0.795	0.937	0.103	0.961	0.849	0.962	0.074
	HR→SR	0.957	0.840	0.958	0.079	0.982	0.911	0.984	0.050
	SR→HR	0.955	0.834	0.957	0.082	0.983	0.915	0.985	0.048
	with Bi-Atten	**0.981**	**0.895**	**0.982**	**0.055**	**0.984**	**0.908**	**0.986**	**0.050**

D. Validation on GMDC

1) The effectiveness of GMDC: Table III shows the validation of GMDC. It can be observed that the model with 5×5 deformable convolution (DConv) results in only a slight performance improvement or even performance degradation compared to the model without GMDC. In contrast, the use of GMDC with two groups achieves a more significant improvement in both the QADS and CVIU databases, particularly in terms of the KRCC and RMSE metrics, although the inference speed of GMDC is slightly lower. This indicates that the GMDC effectively provides the visual perception of important targets (i.e. distortion areas) through multi-level adaptive sampling.

2) The impact of channel grouping on GMDC: As group convolution [48] is employed in GMDC, we conduct validation of the number of groups. Given that the number of channels for convolutional layers is set to a multiple of 16 and the group number needs to evenly divide the channel number, we test with groups of 2, 4, 8, and 16 respectively. The results are presented in Table IV. It can be observed that GMDC with 2 groups yields optimal performance; however, as the group number doubles, there is a gradual decline in performance with more pronounced losses in terms of KRCC and RMSE. This suggests that despite an increase in the number of feature maps obtained by convolution due to more groups, there is a greater loss of inter-group information. Therefore, for other experiments within this study, we set GMDC's group number to 2.

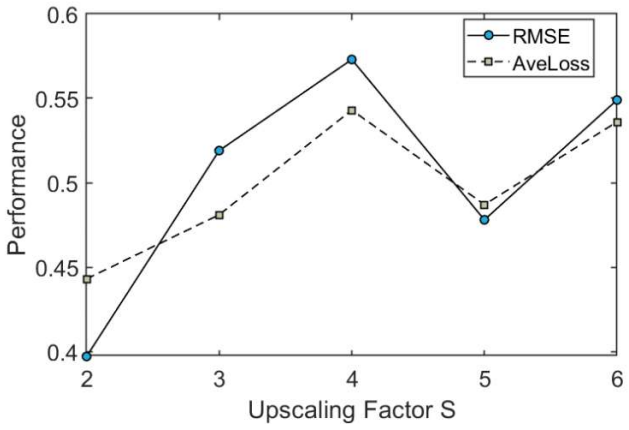

Fig. 7. The validity of the upsampling factor S on the CVIU dataset. The results include RMSE and average loss (AveLoss), where the average loss is calculated by taking the test loss of the last 30 epochs during training.

E. Validation of Bi-Atten

To validate the efficacy of Bi-Atten, we conduct a comparison on the attention interactive modes, as presented in Table VI. Specifically, we test the following scenarios: without using Bi-Atten (i.e. applying the self-attention separately to SR image and HR reference branches); transferring the "Key" information from the SR image branch to the HR reference branch while applying self-attention to the SR image branch (i.e. SR → HR); and transferring the "Key" information

TABLE IX
THE COMPARISON BETWEEN PBAN AND SPATIAL/CHANNEL ATTENTION ON CVIU DATABASE, WITH THE BEST PERFORMANCE VALUES HIGHLIGHTED IN BOLD. THE "SPATIAL ATTEN" AND "CHANNEL ATTEN" POOL THE CHANNEL AND SPATIAL DIMENSION TO 1, RESPECTIVELY. THE FLOPS IS CALCULATED WITH INPUT TENSOR SHAPE $3 \times 32 \times 32$.

Methods	Params/M	Flops/G	Speed/ms	SRCC ↑	KRCC ↑	PLCC ↑	RMSE ↓
Channel Atten	1.458	0.088	11.284	0.963	0.836	0.964	0.716
Spatial Atten	1.458	0.088	11.099	0.913	0.751	0.917	1.062
PBAN	1.458	0.088	10.785	**0.978**	**0.872**	**0.981**	**0.397**

from the HR reference branch to the SR image branch while applying the self-attention to the HR reference branch (i.e. HR $\rightarrow$ SR). Our experimental results reveal that the model without Bi-Atten exhibits the poorest performance on both databases, and even adding one-way "Key" information transmission could significantly improve the performance. Moreover, the performance improvement brought by Bi-Atten is particularly notable in KRCC and RMSE on both databases. Ultimately, the model using Bi-Atten achieves the best performance, indicating that our proposed Bi-Atten effectively provides the network's learning ability. We further compare Bi-Atten (i.e. $Q\&V$ are homology) with cross attention (i.e. $K\&V$ are homology), and the results (Table VII) show that Bi-Atten is superior in mining SR image distortion.

F. Validity on replacing Bi-Atten with spatial/channel attention

Table IX compares PBAN with Spatial and Channel Attention in terms of complexity (Params/M, FLOPs/G), inference speed (ms), and performance metrics (SRCC, KRCC, PLCC, RMSE). All methods share identical model complexity (1.458M Params, 0.088G FLOPs); however, PBAN outperforms the others, achieving the highest SRCC (0.978), KRCC (0.872), PLCC (0.981), and the lowest RMSE (0.397). While PBAN's inference speed (10.785 ms) is slightly slower, the trade-off is justified by its substantial accuracy gains, highlighting its effectiveness and efficiency. This improvement is attributed to the fact that SR images exhibit enhanced distortions, where spatial/channel attention—commonly used in common IQA with degraded distortions—may overlook critical information due to pooling operations.

G. Validation of SubEC

As group convolution [48] is also employed in SubEC (Sub-Channel Weight branch), we conduct validation of the number of groups. As mentioned earlier, we tested with group counts of 2, 4, 8, 16, and 32 respectively. However, the results show that channel grouping does not have a significant performance difference on SubEC, so we set the number of groups to 2 in the remaining experiments.

1) The effectiveness of SubEC: The SubEC is designed to effectively provide the capability of the Bi-Atten in refining attention for the perception of distorted information. Therefore, a validity of this component is conducted under different attention interactive modes, and the results are shown in Table VIII. The comparison on the QADS database reveals that adding SubEC leads to improvements in all four attention interaction modes. Performance is notably worse when SubEC is not added, highlighting the significant improvement brought by SubEC. This not only demonstrates the powerful capturing ability of SubEC for finer distorted information but also shows that SubEC can greatly compensate for the shortcomings of one-way "Key" transmission.

TABLE X
THE IMPACT OF CHANNEL SHUFFLE AND DOT-PRODUCT IN SUBEC ON CVIU DATABASE, WITH THE BEST PERFORMANCE VALUES HIGHLIGHTED IN BOLD. THE FLOPS IS CALCULATED WITH INPUT TENSOR SHAPE $3 \times 32 \times 32$.

Methods	Params/M	Flops/G	Speed/ms	SRCC ↑	KRCC ↑	PLCC ↑	RMSE ↓
w/o Channel Shuffle	1.458	0.088	10.819	0.970	0.853	0.971	0.654
Element-wise Add	1.458	0.088	11.204	0.963	0.838	0.966	0.699
PBAN	1.458	0.088	10.785	**0.978**	**0.872**	**0.981**	**0.397**

TABLE XI
VALIDITY OF NUMBER OF GROUPS IN GROUP CONVOLUTION IN SUB-CHANNEL WEIGHT BRANCH OF SUBEC.

Database	CVIU			
Sub-Channel Weight (Groups)	SRCC	KRCC	PLCC	RMSE
Groups 2	**0.978**	**0.872**	**0.981**	**0.397**
Groups 4	0.976	0.870	0.979	0.515
Groups 8	0.975	0.869	0.978	0.538
Groups 16	0.972	0.864	0.976	0.547

2) The impact of magnification factor S on SubEC: Furthermore, we delve into the impact of the upsampling factor S for both channel and spatial dimensions on experimental outcomes. Specifically, S is set to 2, 3, 4, 5, and 6 respectively; training and testing were then conducted on the CVIU dataset. Fig. 7 illustrates the variation curves for RMSE and average test loss with different values of S. It can be discerned that as S increases, there is an overall downward trend in performance. This indicates that a higher upsampling factor introduces more parameters which may actually impede the learning process of SubEC.

3) Validity on channel shuffle and dot-product of SubEC: We have incorporated ablation studies on the impact of channel shuffle and the replacement of dot-product with element-wise addition. As shown in Table X, removing the channel shuffle results in a noticeable performance drop, as the channel shuffle operation enhances the model's non-linear capability. Furthermore, replacing dot-product with element-wise addition also affects the results. Dot-product preserves the original distribution of input features, making its effects more controllable. In contrast, element-wise addition mixes the effects of channels and spatial features, changing the feature distribution and potentially compromising performance.

H. Cross-Database Evaluation

We conduct cross-database tests to validate the generalization performance of our proposed PBAN and five deep learning-based comparison methods. In practice, we train algorithms on one dataset and test them on the other. We present SRCC, KRCC, and PLCC results of the cross-dataset tests in Table XII, where the best and second-best results are labeled in red and blue, respectively. Table XII reveals that PBAN achieves desirable generalization performance on

TABLE XII
THE PERFORMANCE FOR CROSS-DATABASE EVALUATION, WHERE THE BEST AND SUBOPTIMAL PERFORMANCE VALUES ARE IN RED AND BLUE, RESPECTIVELY.

Train	CVIU			QADS		
Test	QADS			CVIU		
models	SRCC ↑	KRCC ↑	PLCC ↑	SRCC↑	KRCC↑	PLCC↑
DeepSRQ	0.723	-	0.749	0.648	-	-
HLSRIQA	0.786	-	0.773	0.785	-	0.776
EK-SR-IQA	0.645	-	-	0.682	-	-
JCSAN	0.702	0.530	0.710	0.631	0.463	0.729
TADSRNet	0.739	0.547	0.736	0.657	0.489	0.729
PBAN	0.890	0.732	0.892	0.730	0.550	0.794

TABLE XIII
THE PERFORMANCE OF COMMON NR-IQA EVALUATION, WITH THE BEST AND SUBOPTIMAL PERFORMANCE VALUES HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY. THE PROPOSED PBAN-NR UTILIZES SINGLE BRANCH TO TAKE DISTORTED IMAGES AS INPUT AND PREDICT QUALITY WITHOUT REFERENCE INFORMATION. THE FLOPS IS CALCULATED WITH INPUT TENSOR SHAPE $3 \times 512 \times 512$.

				CLIVE		KonIQ-10K	
Methods	Params/M	Flops/G	Speed/Ms	PLCC↑	SRCC↑	PLCC↑	SRCC ↑
WaDIQaM [53]	6.287	30.479	8.170	0.671	0.682	0.807	0.804
DBCNN [13]	15.311	86.219	23.688	0.869	0.869	0.884	0.875
CNNIQA [54]	0.729	1.883	71.210	0.450	0.465	0.584	0.572
HyperIQA [15]	27.375	107.831	29.980	0.882	0.859	0.917	0.906
MetaIQA [14]	-	-	-	0.802	0.835	0.856	0.887
MUSIQ [55]	27.126	17.646	14.910	0.832	0.793	0.928	0.916
Clip-IQA [56]	-	61.065	21.100	0.831	0.805	0.845	0.803
Tres [57]	152.452	500.040	251.070	0.877	0.846	0.928	0.916
TOPIQ [58]	36.039	50.138	22.140	0.884	0.870	0.939	0.926
PBAN-NR	1.598	46.632	18.219	0.899	0.886	0.924	0.916

benchmark SR-IQA databases, even when training on a small dataset (QADS) and testing on the relatively larger dataset (CVIU).

I. Validation on common IQA

Table XIII compares the performance of various No-Reference Image Quality Assessment (NR-IQA) methods on the CLIVE [59] and KonIQ-10K [60] datasets, with 80% for training and 20% for testing, evaluated by parameters (Params), computational complexity (Flops), and correlation metrics (PLCC and SRCC). The results demonstrate that the PBAN model, with the smallest parameter size (1.598M) and relatively low computational cost (46.632G), achieves the best or second-best performance across both datasets. On the CLIVE dataset, PBAN achieves the highest PLCC (0.899) and SRCC (0.886); on the KonIQ-10K dataset, PBAN outperforms other methods with PLCC 0.924 and SRCC 0.916, showcasing exceptional performance and generalization ability. Furthermore, compared to models with significantly larger parameter sizes, such as HyperIQA and TOPIQ, PBAN strikes an optimal balance between efficiency and performance. This highlights PBAN as a highly efficient and accurate NR-IQA method, particularly well-suited for scenarios with limited computational resources, offering substantial application potential.

VI. CONCLUSION

In this paper, we propose a deep learning-based SR FR-IQA method called PBAN following the inspiration from the attention characteristic, hierarchy, spatial frequency sensitivity of the HVS. Motivated by the generation and evaluation processes of SR distortion, we introduce Bi-Atten. This creates a novel attention mechanism for paying visual attention to distortion in an iterative manner. To provide the perception of distortion by utilizing the adaptive dense spatial prediction advantages of deformable convolution, we propose GMDC that balances multi-scale features and computational complexity. Inspired by sub-pixel methods in the field of camera imaging, we design SubEC to further refine the perception by focusing on mining the finer distorted information. Experimental results demonstrate that our proposed PBAN effectively provides visual perception to SR distortion and surpasses existing state-of-the-art quality assessment methods.

In the future, we plan to construct a large-scale SR quality dataset containing the latest SR models (e.g., GAN-based methods and diffusion model-based methods) to enhance the generalization of SR IQA methods. Moreover, we will design specific SR quality metrics for videos and develop quality-driven SR algorithms.

REFERENCES

[1] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *ICCV*, 2017.

[2] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018.

[3] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *ECCV*, 2019.

[4] Z. Liu, Z. Li, X. Wu, Z. Liu, and W. Chen, "DSRGAN: Detail prior-assisted perceptual single image super-resolution via generative adversarial networks," *IEEE TCSVT*, vol. 32, no. 11, pp. 7418–7431, 2022.

[5] R. Chen and Y. Zhang, "Learning dynamic generative attention for single image super-resolution," *IEEE TCSVT*, vol. 32, no. 12, pp. 8368–8382, 2022.

[6] J. Zhang, C. Long, Y. Wang, H. Piao, H. Mei, X. Yang, and B. Yin, "A two-stage attentive network for single image super-resolution," *IEEE TCSVT*, vol. 32, no. 3, pp. 1020–1033, 2022.

[7] Y. Yang and Y. Qi, "Image super-resolution via channel attention and spatial graph convolutional network," *Pattern Recognition*, vol. 112, p. 107798, 2021.

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

[9] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[10] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE TIP*, vol. 18, no. 11, pp. 2385–2401, 2009.

[11] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC*, 2003.

[12] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE TIP*, vol. 27, no. 1, pp. 206–219, 2018.

[13] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE TCSVT*, vol. 30, no. 1, pp. 36–47, 2020.

[14] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *CVPR*, 2020.

[15] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *CVPR*, 2020.

[16] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE TMM*, vol. 25, pp. 2912–2925, 2023.

[17] J. Liu, W. Zhou, X. Li, J. Xu, and Z. Chen, "LIQA: Lifelong blind image quality assessment," *IEEE TMM*, vol. 25, pp. 5358–5373, 2023.

[18] F. Zhou, R. Yao, B. Liu, and G. Qiu, "Visual quality assessment for super-resolved images: Database and method," *IEEE TIP*, vol. 28, no. 7, pp. 3528–3541, 2019.

[19] W. Zhou, Z. Wang, and Z. Chen, "Image super-resolution quality assessment: Structural fidelity versus statistical naturalness," in *QoMEX*, 2021.

[20] W. Zhou and Z. Wang, "Quality assessment of image super-resolution: Balancing deterministic and statistical fidelity," in *ACM Multimedia*, 2022.

[21] W. Zhou, Q. Jiang, Y. Wang, Z. Chen, and W. Li, "Blind quality assessment for image superresolution using deep two-stream convolutional networks," *Information Sciences*, vol. 528, pp. 205–218, 2020.

[22] Z. Zhang, W. Sun, X. Min, W. Zhu, T. Wang, W. Lu, and G. Zhai, "A no-reference deep learning quality assessment method for super-resolution images based on frequency maps," in *ISCAS*, 2022.

[23] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, "Learning-based quality assessment for image super-resolution," *IEEE TMM*, vol. 24, pp. 3570–3581, 2021.

[24] H. Zhang, S. Su, Y. Zhu, J. Sun, and Y. Zhang, "Boosting no-reference super-resolution image quality assessment with knowledge distillation and extension," in *ICASSP*, 2023.

[25] T. Zhang, K. Zhang, C. Xiao, Z. Xiong, and J. Lu, "Joint channel-spatial attention network for super-resolution image quality assessment," *Appl. Intell*, vol. 52, pp. 17 118–17 132, 2022.

[26] X. Quan, K. Zhang, H. Li, D. Fan, Y. Hu, and J. Chen, "TADSRNet: A triple-attention dual-scale residual network for super-resolution image quality assessment," *Appl. Intell*, vol. 53, p. 26708–26724, 2023.

[27] Y. Shuai, Q. Yuan, and S. Zhao, "A spatial-channel attention-based convolutional neural network for remote sensing image classification," in *IGARSS*, 2022, pp. 3628–3631.

[28] Y. Fu, J. Liu, and J. Shi, "TSCA-Net: Transformer based spatial-channel attention segmentation network for medical images," *Computers in Biology and Medicine*, vol. 170, p. 107938, 2024.

[29] D. Wan, R. Lu, S. Shen, T. Xu, X. Lang, and Z. Ren, "Mixed local channel attention for object detection," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106442, 2023.

[30] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," in *ICCV*, 2017, pp. 764–773.

[31] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More Deformable, Better Results," in *CVPR*, 2019, pp. 9308–9316.

[32] W. Kirsch and W. Kunde, "Human perception of spatial frequency varies with stimulus orientation and location in the visual field," *Scientific Reports*, vol. 13, no. 1, OCT 17 2023.

[33] S. Shi, Q. Bai, M. Cao, W. Xia, J. Wang, Y. Chen, and Y. Yang, "Region-adaptive deformable network for image quality assessment," in *CVPRW*, 2021, pp. 324–333.

[34] Z. Shi, Z. Wang, F. Kong, R. Li, and T. Luo, "Dual-quality map based no reference image quality assessment using deformable convolution," *Digital Signal Processing*, vol. 123, p. 103398, 2022.

[35] S. Hochstein and M. Ahissar, "View from the top: Hierarchies and reverse hierarchies in the visual system," *Neuron*, vol. 36, no. 5, pp. 791–804, 2002.

[36] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *CVPR*, 2016, pp. 1874–1883.

[37] A. Kolaman and O. Yadid-Pecht, "Quaternion structural similarity: A new quality index for color images," *IEEE TIP*, vol. 21, no. 4, pp. 1526–1536, 2012.

[38] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE TIP*, vol. 23, no. 2, pp. 684–695, 2014.

[39] C. Huang, T. Jiang, and M. Jiang, "Encoding distortions for multi-task full-reference image quality assessment," in *ICME*, 2019.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[41] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[42] Q. Wu, Z. Wang, and H. Li, "A highly efficient method for blind image quality assessment," in *ICIP*, 2015.

[43] G. Li, L. Qiu, H. Zhang, F. Xie, and Z. Jiang, "Multi-frame super-resolution with raw images via modified deformable convolution," in *ICASSP*, 2022, pp. 2155–2159.

[44] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *CVPRW*, 2019, pp. 1954–1963.

[45] L. Yu, X. Zhang, and Y. Chu, "Super-resolution reconstruction algorithm for infrared image with double regular items based on sub-pixel convolution," *Applied Sciences*, vol. 10, no. 3, 2020.

[46] J. Qu, J. Yin, Y. Jiang, W. Huang, and Q. Chen, "A sub-pixel convolution-based improved bidirectional feature pyramid network for pansharpening," *Remote Sensing Letters*, vol. 14, no. 1, pp. 91–101, 2023.

[47] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *ICCV*, 2021, pp. 347–356.

[48] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017, pp. 5987–5995.

[49] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018, pp. 6848–6856.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[51] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *CVIU*, vol. 158, pp. 1–16, 2017.

[52] H. Yeganeh, M. Rostami, and Z. Wang, "Objective quality assessment of interpolated natural images," *IEEE TIP*, vol. 24, no. 11, pp. 4651–4663, 2015.

[53] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE TIP*, vol. 27, no. 1, pp. 206–219, 2018.

[54] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE CVPR*, 2014, pp. 1733–1740.

[55] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *IEEE ICCV*, 2021, pp. 5148–5157.

[56] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.

[57] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 3989–3999.

[58] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE TIP*, 2024.
[59] D. Ghadiyaram and A. C. Bovik, "Live in the wild image quality challenge database," *Online: http://live. ece. utexas. edu/research/ChallengeDB/index. html[Mar, 2017]*, 2015.
[60] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE TIP*, vol. 29, pp. 4041–4056, 2020.