Blind Image Quality Assessment via Adaptive Graph Attention

Huasheng Wang, Jiang Liu, Hongchen Tan, Jianxun Lou, Xiaochang Liu, Wei Zhou, and Hantao Liu

***Abstract*—Recent advancements in blind image quality assessment (BIQA) are primarily propelled by deep learning technologies. While leveraging transformers can effectively capture long-range dependencies and contextual details in images, the significance of local information in image quality assessment can be undervalued. To address this challenging problem, we propose a novel feature enhancement framework tailored for BIQA. Specifically, we devise an Adaptive Graph Attention (AGA) module to simultaneously augment both local and contextual information. It not only refines the post-transformer features into an adaptive graph, facilitating local information enhancement, but also exploits interactions amongst diverse feature channels. The proposed technique can better reduce redundant information introduced during feature updates compared to traditional convolution layers, streamlining the self-updating process for feature maps. Experimental results show that our proposed model outperforms state-of-the-art BIQA models in predicting the perceived quality of images. The code of the model will be made publicly available.**

***Index Terms*—Image quality assessment, no-reference, graph, convolutional neural networks, deep learning.**

I. Introduction

Image quality assessment (IQA) aims to gauge the overall quality of an image as perceived by human viewers. While obtaining an IQA measure from human subjects is the most reliable mean, it entails time-consuming and expensive conduct of psychovisual experiments. There is an urgent need to establish accurate and robust image quality assessment (IQA) models, which align with the perception and judgment of human subjects. These IQA models play a fundamental role in various perception-based vision applications [1]. In general, IQA models can be categorised into three types based on their requirement of the reference/pristine image, including full-reference IQA (FR-IQA) [2], [3], [4], reduced-reference IQA (RR-IQA) [5], and no-reference/blind IQA (BIQA) [6], [7], [8], [9]. Although both FR-IQA and RR-IQA have demonstrated remarkable performance, the availability of reference is rather limited in many practical scenarios. BIQA is hence of highly practical relevance as the models aim to evaluate image quality without relying on a reference or pristine image.

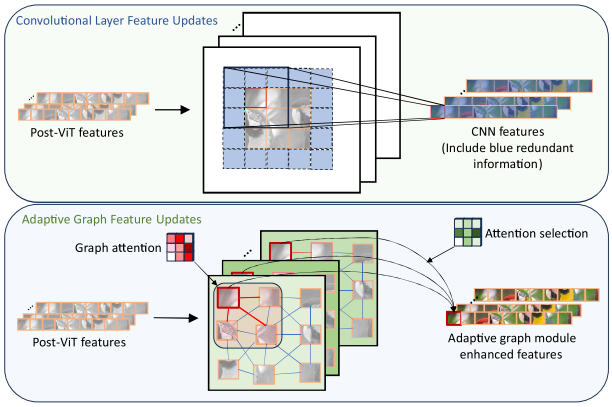


Fig. 1. Illustration of limitations of relying on convolutions to enhance local information extraction from post-ViT feature maps. The top figure shows the use of a convolutional layer for feature updates, where the involvement of zero-padding introduces redundancies into the feature representations. The bottom figure shows the proposed *Adaptive Graph Attention (AGA)* model, which refines post-transformer features into a graph structure. It uses a *graph attention block* to enhance local information and an *attention selection mechanism* to discern feature channel interactions.

In the context of deep learning, a BIQA model is derived by learning discriminative features for perceived quality through a progressively complex neural network architecture. The recent advances include the application of the vision transformer (ViT), which has the architectural advantages of capturing contextual information in images. A popular architecture design approach for BIQA models is to employ a hybrid combination of convolutional neural networks (CNNs) and transformer features to harness their respective capabilities in modelling local and global feature representations. This approach effectively applies the transformers in the IQA models, however, there remains untapped potential to further enhance the feature representation for the feature maps following the transformer module. To unleash the potential, we devise a novel Adaptive Graph Attention (AGA) module, aiming to enhance the post-transformer feature maps for a refined feature expression of local intricacies. Figure 1 illustrates the limitations of relying on convolutions to extract local information from post-ViT feature maps. Convolution operations that involve zero-padding around feature maps can detrimentally affect the update of features in the central region, introducing redundancies into the feature representations. Moreover, zero-padding might cause information loss on the edges of the feature maps. To overcome these challenges and augment the representational

The work of Huasheng Wang was supported by China Scholarship Council under Grant 202106060056. The work of Jiang Liu was supported by China Scholarship Council under Grant 202206420009. The work of Jianxun Lou was supported by China Scholarship Council under Grant 202008220129 (Corresponding author: Jiang Liu.)

Huasheng Wang, Jiang Liu, Jianxun Lou, Wei Zhou, and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF224AG Cardiff, U.K. (email: wanghs@cardiff.ac.uk; liuj137@cardiff.ac.uk; louj2@cardiff.ac.uk; zhouw26@cardiff.ac.uk; liuh35@cardiff.ac.uk)

Hongchen Tan is with College of Future Technology, Dalian University of Technology, Dalian 116024, China (e-mail: tanhongchenphd@bjut.edu.cn.)

Xiaochang Liu is with the School of Mathematics, Sun Yat-Sen University, Guangzhou 510275, China (email: liuxch68@mail2.sysu.edu.cn)

ability of post-ViT feature maps, we develop an adaptive graph attention module which consists of a graph attention block (GAB) and attention selection mechanism (ASM). This module not only prevent the introduction of unnecessary redundant information during feature updates but also comprehensively enhance local and distant interactions when updating feature maps.

More specifically, for the purpose of enabling local interactions amongst feature maps, GAB conceptualises all feature maps as a graph structure, in which each pixel within a feature map corresponds to a node, and the channel information of feature maps is represented as node features within the graph. By considering adjacent nodes as connected through edges and self-connecting each node, we derive an adjacency matrix. We assign a weight of 1 to the central node, complementing it with weighted insights from neighboring nodes during the node updating process. By doing so, the feature maps are transformed into an adaptive graph. Importantly, during the central node's update, the sum of weights assigned to other adjoining nodes equals 1, which allows further enhancement of local interactions between nodes. Then, the ASM is developed to amplify comprehensive interactions spanning across all channels within the feature structure. To achieve faster convergence and higher consistency, we propose a Patch-wise-based Hierarchical Perceptual (PHP) regression module for the inference of quality scores. PHP regression allows the IQA model to consider the relative ordering between scores on the perceptual quality scale, leading to a perceptually more relevant mechanism for an IQA model. The contributions of this work are:

- We devise a novel Adaptive Graph Attention module for deep learning-based IQA. This module not only optimise the utilisation of local features within post-transformer feature maps but also mitigates the introduction of redundant information during feature updates, introducing novel adaptations tailored to the BIQA problem. The synergy between the GAB and ASM modules allows for the selection of the most advantageous attention information for refining feature maps.
- We propose a Patch-wise-based Hierarchical Perceptual regression module to combine MSE and deep ordinal (DO) regression for inferring scores from different patches at various depths of the network. The combination of the DO loss and the PHP is contextually distinct, enhancing the regression ability specifically for the BIQA task.
- We show the substantial superiority of the proposed BIQA model over existing alternative models, through extensive experiments on many benchmark datasets.

It is worth noting that the proposed adaptive graph attention (AGA) module can be effortlessly integrated with other transformer-based methods, enhancing the expressive capability of their features for vision tasks.

II. Related Work

A. Blind Image Quality Assessment

Blind image quality assessment (BIQA) aims to emulate the capabilities of the human visual system (HVS) in accurately assessing the perceived quality of images. Despite humans can adeptly evaluate image quality, this proves challenging for machines due to the absence of sufficient knowledge about the HVS. Conventional approaches seek to replicate the responsiveness of the HVS to diverse image signals by integrating explicit human vision models in a BIQA model. These models include the simulation of perceived structural information in natural scenes as described in Wang et al. [2], as well as the exploration of natural scene statistics (NSS) as detailed in studies by Mittal et al. [10], Moorthy et al. [11], and Gao et al. [12]. However, these approaches that are based on hand-crafted features of images face challenges in generalising to the intricacies of real-world scenarios. Especially, the IQA features are often specifically optimised for specific types of distortions, therefore they are constrained in dealing with unseen data.

Fortunately, the advancement of deep learning and convolutional neural networks (CNNs) has paved the way for approaching more reliable BIQA; and CNN-based models have remarkably surpassed their predecessors using traditional approaches, particularly when confronting real-world distortions. The improved performance is primarily attributed to the deep learning techniques that allows directly deriving discriminative features for image quality perception [13], [14], [15], [6]. Hyper-IQA [16] partitions features into low-level and high-level attributes, subsequently transforming the latter to redirect the former. However, CNNs exhibit few prominent limitations for the IQA tasks. First, their inherent difficulty in capturing non-local features and their notable locality bias hinders the models' capacity to leverage information across all regions of an image for IQA prediction. In addition, the spatial translation invariance imposed by shared convolutional kernel weights makes CNNs inadequate for handling intricate amalgamations of features. Inspired by the recent technological breakthrough in natural language processing (NLP), where Transformers [17] are invented at capturing global feature dependencies, the area of computer vision has experienced substantial advancements via the application of the Vision Transformer (ViT) [18]. You et al. [19] implement the ViT in the context of BIQA, leveraging the success of the ViT architecture. TReS [20] integrates relative ranking and self-consistency loss mechanisms to harness the abundant self-supervisory information and diminish the network's susceptibility. DOR-IQA [6] integrates the deep ordinal loss (DO-loss) function into the IQA model to enhance the accuracy of prediction.

Li et al. [8] conducted a theoretical analysis, demonstrating that embedded normalisation stabilizes the gradients of the loss function, leading to faster convergence of the IQA model.

MANIQA [7] employs multi-dimensional feature interactions and leverages spatial and channel attention mechanisms to enhance the performance of image quality prediction.

Nevertheless, how to best unleash the potential of post-ViT features remains unexplored, which is the focus of this work.

Fuzzy-QA [21] uses a fuzzy neural network to predict the opinion score distribution (OSD) of image quality. This approach aims to capture the diversity and uncertainty inherent in subjective evaluations. TempQT [22] utilizes a pre-trained Transformer model to generate an error map. It then integrates a vision Transformer branch to extract perceptual quality tokens for feature fusion with the error map, followed by regressing the fused features to the final image quality score. AFF-QA [23] introduces a model to integrate feature fusion with an attention mechanism. By extracting multilayer features and employing a hierarchical approach, the model effectively captures diverse image distortions. StairIQA [24] introduces a specialised model for in-the-wild images, tackling issues regarding feature representation and training sample diversity. It achieves this goal by employing a hierarchical feature integration method and an iterative mixed database training strategy. DEIQT [25] presents a method that efficiently generates quality-aware feature representations with reduced data requirements by incorporating a Transformer decoder to refine perceptual information and introducing a novel attention panel mechanism.

B. Feature Enhancement

In recent years, the integration of feature enhancement techniques within deep learning architectures has emerged as a pivotal approach to improve model performance and generalisation capabilities. Feature enhancement involves the incorporation of supplementary information, data transformations, or domain expertise to enhance the representational capacity of input data, thereby augmenting model robustness and efficacy. Many BIQA methods utilise a multi-task learning framework to improve the feature expression for the BIQA task. Xu et al. [26] use a single CNN architecture for image quality estimation and distortion identification simultaneously. The model aims to compress the parameters of the CNN model, therefore no interactions between these two sub-tasks are explored. By decomposing the BIQA task into two sub-tasks with different priorities, the distortion type information becomes transparent to the primary quality assessment sub-network. Ma et al. [9] concurrently optimized a pair of image and language encoders across multiple IQA datasets for tasks encompassing BIQA, scene classification, and distortion type identification. To adequately capture the interplay amongst distortion types and the distribution of samples featuring the same distortion type but varying distortion levels, GraphIQA [27] introduces the concept of Distortion Graph Representation (GDR). This approach possesses the capability to encapsulate the distinctive attributes of individual distortions and their underlying structural relationships. Consequently, GraphIQA not only creates Distortion Graph Representations (DGRs) as prior knowledge during the assessment of familiar distortions, but it also enhances the model's feature representation. By doing so, the primary sub-network can obtain more robust features to predict quality. In contrast to GraphIQA, our newly introduced adaptive graph attention (AGA) module is added as an attachment to the transformer module, addressing the decay of local information. Critically, we characterise the post-transformer feature maps as an internal graph configuration without invoking graph neural networks.

III. Method

In this section, we first describe the proposed AGA framework as illustrated in Figure 2. Then we provide detailed information regarding the three core components of the framework, including the Graph Attention Block (GAB), Attention Selection Mechanism (ASM), and Patch-wise-based Hierarchical Perceptual (PHP) regression.

A. Overall Framework

The key concept of the proposed adaptive graph attention (AGA) module as shown in Figure 2 is to augment the capabilities of vision transformers (ViT), preventing local information decay while maintaining the strong ability of capturing contextual information in images. Given a distorted image $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$, where H and W denote height and width, let f_ϕ represent the ViT with learnable parameters ϕ, $F_i \in \mathbb{R}^{b\times c_i\times H_i W_i}$ denote the features from the i^{th} layer of ViT, where $i \in \{1, 2, \cdots, 12\}$, b denote the batch size, and c_i, H_i, and W_i denote the channel size, width, and height of the i^{th} feature, respectively. To extract the information amongst different channels, we concatenate the features from 4 to 12 layers and denote the output as $\hat{F} \in \mathbb{R}^{b\times\sum_i c_i\times H_i W_i}$, where $i \in \{7, 8, 9, 10\}$. Then, we use a Transposed Attention Block (TAB) [7] to enrich the interaction between local and global regions. Compared to the traditional self-attention applied only within patches in the spatial dimension, TAB can build connections across channels, enabling the encoding of global contextual information. In implementation, the concatenated feature $\hat{F}$ is firstly transformed into three different groups of vectors, the query group ($\mathbf{Q}$), the key group ($\mathbf{K}$) and the value group ($\mathbf{V}$), which allow encoding the pixel-wise cross-channel context. Then we compute the dot product of $\mathbf{Q}$ and reshaped $\mathbf{K}$, divide each by σ, and apply a softmax function to obtain a transposed-attention map. In addition, a residual connection is added to strengthen the flow of information and improve the performance. We denote the output of TAB as $\tilde{F}$,which is computed as:

$$\tilde{F} = W_1 Attention(Q, K, V) + \hat{F}, \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sigma}) \cdot V, \tag{2}$$

where W_1 is the linear projection matrix, σ is the spatial dimension of Q, K and V.

Next, the feature $\tilde{F}$ will be sent to the adaptive graph attention (AGA) module which consists of a Graph Attention Block (GAB) and an Attention Selection Mechanism (ASM) to enhance the local representational capacity of feature maps. More details of GAB and ASM are described below.

Furthermore, to collaboratively enhance spatial information, we use the Scale Swin Transformer Block (SSTB) [7], which

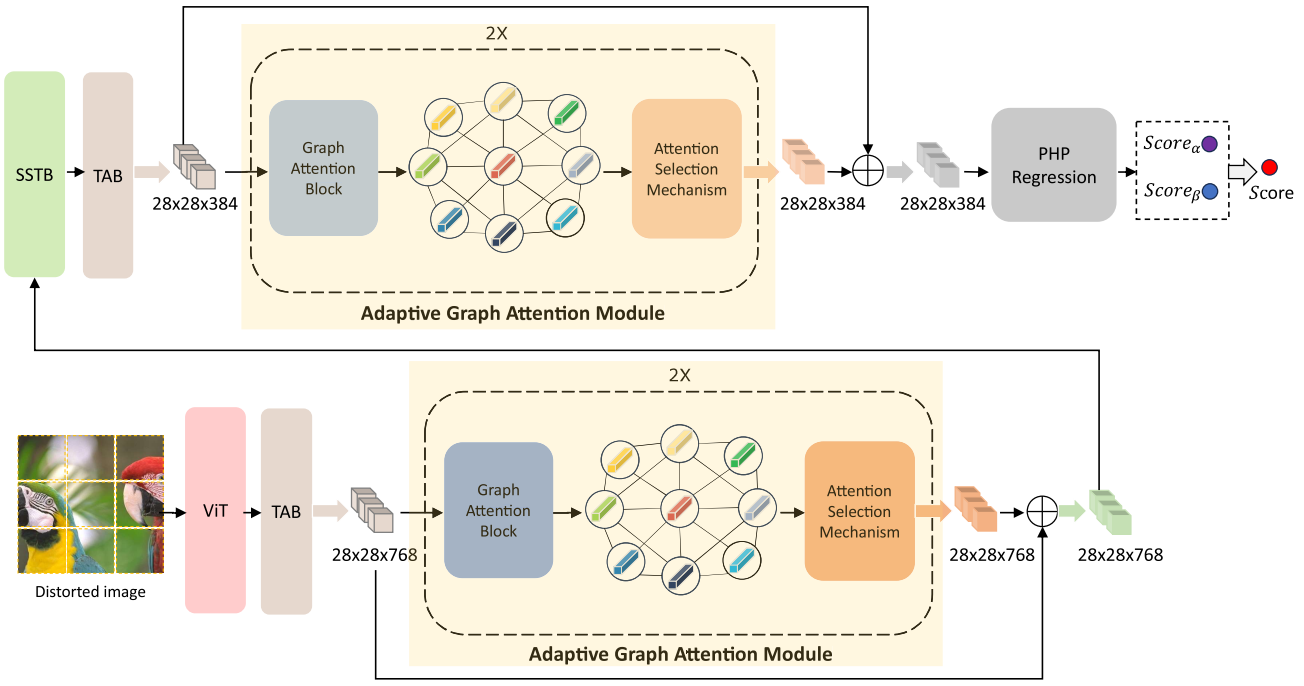


Fig. 2. Outline of our framework. A distorted image is segmented into patches of size 8×8. These patches are then input into ViT and TAB for feature extraction. The newly introduced module, comprising GAB and ASM, elaborated in Sec.III-B and Sec.III-C, is applied twice to reinforce channel attention between feature maps and node attention between the feature maps. An SSTB is employed to collaboratively enhance spatial information between two adaptive graph attention modules. Lastly, a PHPR module is introduced to compute the average of scores obtained using MSE regression and DO regression.

consists of two Swin Transformer Layers (STL), a convolutional layer and a scale factor α. Given the input feature F_{in}, the process of SSTB is defined as:

$$F_{out} = \alpha \cdot H_{CONV}(H_{STB}(F_{in})) + F_{in}, \tag{3}$$

where F_{out} is the output of SSTB, $H_{CONV}(\cdot)$ is the convolutional layer, and $H_{STB}(\cdot)$ denotes two successive Swin Transformer Blocks.

To further amplify feature expression capabilities, the adaptive graph attention (AGA) module is applied twice. Finally, the Patch-wise-based Hierarchical Perceptual (PHP) regression module is employed to achieve faster convergence and higher consistency for the prediction of quality scores. More details of PHP is described in 3.4.

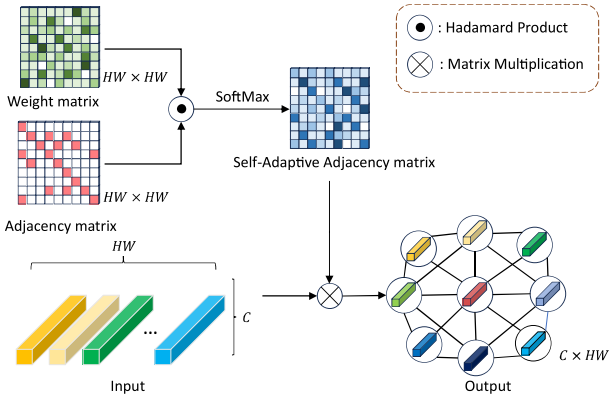


Fig. 3. Illustration of graph attention block (GAB).

B. Graph Attention Block

To enhance local feature expression in feature maps, the convolution operation is often employed to exact local information. However, as mentioned above, zero-padding can potentially introduce redundant information during feature updates. Alternatively, a graph structure provides a suitable solution since there is inherent connectivity between pixels in the spatial dimension that contains rich local information. However, the conventional graph convolutional network (GCN) [27] only considers binary connection relationships between nodes, which can downplay the strength of connections. Thus, we propose Graph Attention Block (GAB) to not only convert post-transformer feature maps into a connected topology graph but also capture the importance of connections between these connected feature maps. As shown in Figure 3, given an input $X \in \mathbb{R}^{C \times HW}$, we regard each feature vector of the post-transformer feature maps as a node in a graph. The number of nodes is defined as N, where $N = HW = 768$. According to the position of each node in the spatial dimension, we build edges between adjacent nodes and the nodes themselves to obtain the adjacency matrix $A \in \mathbb{R}^{768 \times 768}$. Then, to learn the significance of the connectivity relationship adaptively, we define a trainable weight matrix as $W \in \mathbb{R}^{768 \times 768}$, which is adjusted during the training phase to optimise the performance of the model. In addition, these learned weights are not fixed in the final inference model; instead, they remain adjustable during inference, allowing the model to adapt and optimise its performance based on the

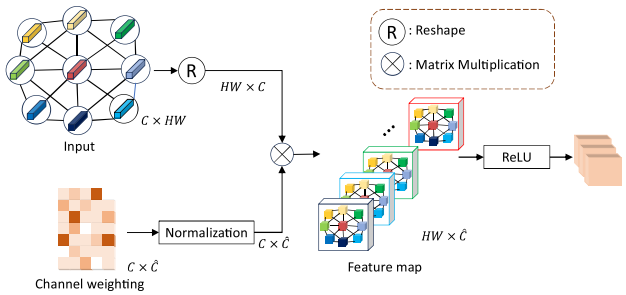


Fig. 4. Illustration of attention selection mechanism (ASM).

input data during deployment. The GAB can be calculated as:

$$GAB(A, X, W) = X \cdot SoftMax(A \odot W), \tag{4}$$

where $\odot$ denotes the element-wise Hadamard product.

C. Attention Selection Mechanism

GAB as mentioned above can establish adaptive connectivity between post-transformer feature maps and extract local features from the spatial dimension. However, it ignores the rich information between different channels. As shown in Figure 4, we propose an ASM to learn the significance of different channels, aiming to obtain enhanced representational ability. Firstly, we denote a weight matrix $W \in \mathbb{R}^{C\times\hat{C}}$ to assign importance scores to different channels. Let W_{ij} represent the probability that channel i is important relative to channel j. To ensure the probabilities sum up to 1 for each row, the W is normalised to $\hat{W} = (\hat{w}_{i,j}) = \frac{exp(w_{i,j})}{\sum_{k=1}^{C} exp(w_{k,j})}$. The ASM can be defined as:

$$F_{ASM} = ReLU(X^T \cdot \hat{W}). \tag{5}$$

D. Patch-wise Hierarchical Perceptual Regression

Each pixel in the deep feature map is extracted from various patches of the input image, and each patch has a unique impact on the perception of overall image quality. To fuse contributions from different patches, we employ a patch-wise strategy for quality prediction rather than utilising a pooling strategy to obtain a single quality score, which may ignore interactions amongst various patches. To this end, we propose a patch-wise-based hierarchical perceptual (PHP) regression module as illustrated in Figure 5 to combine MSE and ordinal regression for aggregating scores from different feature maps across various depths, which aims to augment the model's adaptability and generalisation capabilities. The PHP regression module consists of two blocks including the MSE regression block and deep ordinal (DO) regression block.

For the MSE regression block, given an input F_{MSE}, we use an SSTB to enhance the spatial information; and the embedding dimension of $SSTB_\alpha$ is set to $D_1 = 384$. Then we fed it into two linear projection branches, with one branch computing the probability score for each pixel in the feature map and the other branch calculating the attention map corresponding to each generated probability score. MSE loss between the predicted score and the ground truth score is

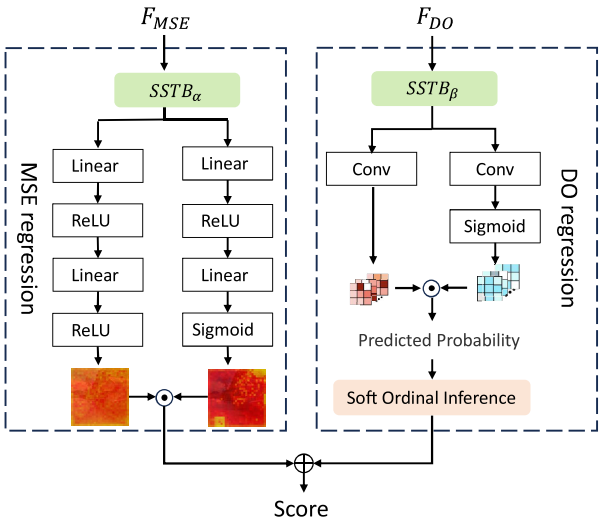


Fig. 5. Illustration of patch-wise-based hierarchical perceptual (PHP) regression module, including an MSE regression block (left) and a deep ordinal (DO) regression block (right).

utilised for the training process in our proposed method. The $Score_\alpha$ can be obtained by weighted summation of individual patch scores:

$$Score_\alpha = \frac{s \odot w}{\sum w}, \tag{6}$$

where s with dimensions $28 \times 28 \times 384$ represents the probability maps of the scores, w with dimensions $28 \times 28 \times 384$ corresponds to the corresponding attention map, and $\odot$ denotes the element-wise Hadamard product. It should be noted that the Hadamard product preserves the dimensionality of the channels.

For the deep ordinal (DO) regression block, given an input F_{DO}, we incorporate the recently proposed deep ordinal regression [6]. This method has been proven effective in boosting the baseline BIQA model's performance by enabling the model to factor in the relative ordering of quality scores on a perceptual image quality scale. In the DO regression block, the produced probabilities are transformed into quality scores using a soft ordinal inference. Similar to the MSE regression block, the embedding dimension of $SSTB_\beta$ is set to $D_2 = 384$. We feed F_{DO} into two convolution branches and calculate the $Score_\beta$ by weighted summation of individual patch scores:

$$s_{do} = \frac{\hat{s} \odot \hat{w}}{\sum \hat{w}}, \tag{7}$$

where $\hat{s}$ with dimensions $28\times28\times30$ represents the probability map of the scores, while $\hat{w}$ with dimensions $28 \times 28 \times 30$ corresponds to the associated attention map. The symbol $\odot$ denotes the element-wise Hadamard product. The predicted score's probability, denoted as $s_{do} \in \mathbb{R}^{1\times1\times30}$, is emphasized. It's crucial to note that the Hadamard product preserves the dimensionality of the channels, ensuring that the resulting s_{do} maintains a 30-dimensional confidence distribution of score probabilities. In order to get a continuous variable for image quality, denoted as $score_\beta$, we employ soft ordinal inference

to transform the predicted probabilities. The final score can be calculated by taking the element-by-element average of $score_\alpha$ and $score_\beta$.

For the loss function, we combine MSE loss and DO loss to constrain the final score, which can be defined as:

$$L_{\text{total}} = L_{\text{MSE}}(score_\alpha, score_{gt}) + \lambda \times L_{\text{DO}}(score_\beta, score_{gt}), \tag{8}$$

where $score_{gt}$ denotes the discretized ground truth labels used during the training process. λ is set to 0.5 in our experiments. An empirical approach is taken in our study to assign weights to individual losses, with the overall guiding principle being that very imbalanced loss contributions will cause the model representations to be optimised preferentially for the target with the largest individual loss, at the expense of the other targets. To remedy this, we assign different levels of importance to the loss values in their contribution to the final loss, as the losses' values use different scales (e.g., one loss takes value around 3-5, whereas the other loss can be as low as 0.1). For the overall loss, λ is utilised to provide different levels of importance to the sub-losses (i.e., L_{MSE} and L_{DO}) to balance their contribution to the final loss as the sub-losses' values use different scales as observed in our experiments. For instance, during training, if L_{MSE} takes value around 1 while L_{DO} takes value around 2, introducing λ helps balance their individual contribution to the overall loss, thus enabling more stable training of the model. The coefficient preceding L_{DO} is introduced to ensure that the values computed by all sub-loss functions during training remain within a consistent magnitude. The specifics of the deep ordinal loss are extensively explained in DOR-IQA [6].

IV. Experiments

A. Datasets and Performance Metrics

We perform a series of experiments using widely recognised IQA datasets including LIVE [28], CSIQ [29], TID2013 [30], KADID-10k [31], and PIPAL [32]. These datasets exhibit diversity in the nature of image quality perception, hence provide the opportunity to capture different characteristics of IQA models. It should be noted that these IQA databases focus on artificial/synthetic distortions, which are often used to establish a baseline evaluation of BIQA algorithms under controlled simulation of visual stimuli. In a nutshell, LIVE, CSIQ, TID2013 and KADID-10k include only traditionally distorted images via well-defined signal processing filters, while PIPAL includes a portion of distorted images generated by generative adversarial network (GAN)-based algorithms. The baseline evaluation based on artificial distortions cannot fully capture the complexity and diversity of real-world scenarios. Hence it is important to evaluate BIQA algorithms on authentic/natural distortions. To this end, we extend our experiments to the datasets of authentic distortions, including SPAQ [33], CLIVE [34], and KonIQ-10k [35]. SPAQ comprises 11,125 images captured by 66 types of mobile devices, covering a diverse range of scene categories. CLIVE consists of 1,162 images with various authentic distortions captured by different mobile devices. KonIQ-10k contains 10,073 images selected from an extensive public multimedia database, covering a broad and uniform range of distortions. Table I summarises the structure of these datasets.

TABLE I
SUMMARY OF IMAGE QUALITY ASSESSMENT (IQA) DATASETS FOR PERFORMANCE EVALUATION.

IQA Database	# Distorted Images	Distortion Type
LIVE [28]	779	artificial/synthetic
CSIQ [29]	866	artificial/synthetic
TID2013 [30]	3000	artificial/synthetic
KADID-10k [31]	10.1k	artificial/synthetic
PIPAL [32]	23.2k	artificial/synthetic
CLIVE [34]	1162	authentic/natural
KonIQ-10k [35]	10.0k	authentic/natural
SPAQ [33]	11.1k	authentic/natural

The Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SROCC) are employed for evaluating the IQA models' performance. Both PLCC and SORCC values range from 0 to 1; and a higher value indicates a better performance.

B. Implementation Details

We implement our proposed adaptive graph attention based IQA model, namley AGAIQA by PyTorch, where both training and testing are conducted on a single NVIDIA GeForce RTX4090 GPU. Firstly, the input image is randomly cropped into 224×224 pixels. Then the ViT-B/8 [18] is employed as the pre-trained model for feature extraction. In this study, we set the patch size P as 8 and the embedding dimension as 768. We concatenate the features of 4 layers $\{7, 8, 9, 10\}$ in ViT and apply TAB to enrich the global context for obtaining the post-transformer feature maps $\tilde{F} \in \mathbb{R}^{C \times H \times W}$,where $C = 768$, $H = 28$, $W = 28$.

Next, the post-transformer feature maps are fed into two successive adaptive graph attention (AGA) modules for feature augmentation. In the AGA module, the GAB and ASM are applied twice and the embedding dimensions of two ASMs are set to $\hat{C}_1 = 384$, $\hat{C}_1 = 768$, respectively. The parameters α and β are both set to 0.9. Following [7], we use SSTB to further enhance local features. The embedding dimension of SSTB is set to $D = 768$.

Following the approach taken in [36], we randomly split each dataset into 6:2:2 for training, validation and testing. During the training stage, data augmentation including random horizontal flip with a probability of 0.5 is employed. We set batch size B to 16. We use ADAM optimizer with a learning rate l of 1×10^{-5}, weight decay of 1×10^{-5}, cosine annealing learning rate T_{max} and eta_{min} of 50 and 0. The training loss is computed using the proposed L_{Total} loss function as mentioned above. During the validation and testing stage, we randomly crop each image 20 times, and use the average output of as the final score.

C. Comparison with State-of-the-art

We compare the performance of the proposed AGAIQA model against existing state-of-the-art BIQA models in terms of accuracy and generalisation. It should be noted that we

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED AGAIQA VERSUS STATE-OF-THE-ART BIQA MODEL ON FOUR STANDARD IMAGE QUALITY ASSESSMENT DATASETS. BOLD ENTRIES INDICATE THE BEST PERFORMANCE.

Method	LIVE		CSIQ		TID2013		KADID-10k	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
DIVINE [37]	0.908	0.892	0.776	0.804	0.567	0.643	0.435	0.413
BRISQUE [38]	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528
ILNIQE [39]	0.906	0.902	0.865	0.822	0.648	0.521	0.558	0.528
BIECON [40]	0.961	0.958	0.823	0.815	0.762	0.717	0.648	0.623
MEON [41]	0.955	0.951	0.864	0.852	0.824	0.808	0.691	0.604
WaDIQaM [42]	0.955	0.96	0.844	0.852	0.855	0.835	0.752	0.739
Fuzzy-QA [21]	-	-	0.948	0.933	0.852	0.844	-	-
DBCNN [43]	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851
TIQA [19]	0.965	0.949	0.838	0.825	0.858	0.846	0.855	0.85
MetaIQA [15]	0.959	0.96	0.908	0.899	0.868	0.856	0.775	0.762
P2P-BM [44]	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.84
AFF-QA [23]	0.965	0.964	0.961	0.948	0.920	0.901	0.933	0.934
HyperIQA [16]	0.966	0.962	0.942	0.923	0.858	0.84	0.845	0.852
TReS [20]	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.915
StairIQA [24]	0.970	0.966	0.941	0.920	-	-	0.875	0.867
DOR-IQA [6]	0.978	0.977	0.961	0.945	0.901	0.887	0.885	0.883
TempQT [22]	0.977	0.976	0.960	0.950	0.906	0.883	-	-
DEIQT [25]	0.982	0.980	0.963	0.946	0.908	0.892	0.887	0.889
MANIQA [7]	0.983	0.982	0.968	0.961	0.943	0.937	0.943	0.937
AGAIQA (Ours)	**0.989**	**0.988**	**0.978**	**0.973**	**0.958**	**0.951**	**0.952**	**0.947**

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED AGAIQA VERSUS STATE-OF-THE-ART BIQA MODEL ON THE PIPAL DATABASE (PUBLICLY AVAILABLE DATA). BOLD ENTRIES INDICATE THE BEST PERFORMANCE.

Method	Validation		Test	
	PLCC	SROCC	PLCC	SROCC
BRISQUE [38]	0.015	0.059	0.087	0.097
NIQE [10]	0.005	0.115	0.030	0.112
PI [45]	0.079	0.133	0.123	0.153
MA [46]	0.129	0.131	0.173	0.224
SSIM [47]	0.332	0.386	0.377	0.407
FSIM [48]	0.473	0.575	0.528	0.610
LPIPS-Alex [49]	0.581	0.616	0.584	0.592
DBCNN [43]	0.643	0.631	0.635	0.628
MetaIQA [15]	0.651	0.642	0.647	0.638
HyperIQA [16]	0.679	0.662	0.671	0.657
TReS [20]	0.685	0.677	0.677	0.661
DOR-IQA [6]	0.687	0.679	0.683	0.673
StairIQA [24]	0.692	0.688	0.685	0.674
DEIQT [25]	0.695	0.690	0.685	0.676
MANIQA [7]	0.721	0.713	0.704	**0.740**
AGAIQA (Ours)	**0.748**	**0.737**	**0.735**	0.729

TABLE IV
COMPARISON OF STATE-OF-THE-ART NR-IQA ALGORITHMS ON CLIVE, KONIQ-10K, AND SPAQ DATASETS, SORTED BY PLCC ON CLIVE WITH THE HIGHEST SCORE AT THE BOTTOM.

Method	CLIVE		KonIQ-10k		SPAQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
TIQA [19]	0.600	0.590	0.600	0.590	-	-
BRISQUE [38]	0.623	0.619	0.687	0.678	0.806	0.802
MetaIQA [15]	0.700	0.690	0.700	0.690	-	-
PieAPP [4]	0.785	0.782	0.822	0.820	-	-
DBCNN [43]	0.800	0.790	0.800	0.790	0.915	0.911
AFF-QA [23]	0.838	0.828	0.932	0.919	-	-
HyperIQA [16]	0.849	0.846	0.846	0.845	0.919	0.916
Fuzzy-QA [21]	0.864	0.829	0.921	0.901	-	-
TempQT [22]	0.886	0.870	0.920	0.903	-	-
DEIQT [25]	0.894	0.875	0.934	0.921	0.923	0.919
TReS [20]	0.906	0.902	0.883	0.877	0.909	0.905
MANIQA [7]	0.907	0.902	0.932	0.931	0.916	0.911
DOR-IQA [6]	0.911	0.908	0.901	0.889	0.921	0.916
StairIQA [24]	0.917	0.892	0.936	0.921	0.927	0.924
AGAIQA(Ours)	**0.919**	**0.918**	**0.939**	**0.936**	**0.933**	**0.929**

replicate the experimental protocol as described in [7] to conduct the comparative study, ensuring consistent results are reported to facilitate IQA model comparisons.

Table II illustrates the performance comparison of our proposed AGAIQA model versus other state-of-the-art BIQA models across four standard datasets including LIVE, CSIQ, TID2013, and KADID-10k. It can be seen from the reported PLCC and SROCC values that that AGAIQA achieves superior performance, as evidenced by consistently producing the highest PLCC and SROCC values (highlighted in bold) across all IQA datasets. The results reported on the lab-based datasets (i.e., LIVE, CSIQ and TID2013) demonstrate the high perceptual relevance of the our AGAIQA model, outperforming the other state-of-the-art models. The results reported on the crowdsourcing-based dataset (i.e., KADID-10k) show the superiority of the our AGAIQA model in generalising to real-world IQA scenarios and complexities.

Given that the PIPAL dataset is curated to encompass images distorted by various techniques, including the GAN-based algorithms. This database introduces challenges in terms of assessing novel distortions that exhibit distinctive characteristics compared to those encompassed by the standard datasets as shown in Table I. This unique attribute necessitates a dedicated assessment for IQA models. To this end, we perform a comparative study on the PIPAL database; and the results are listed in Table III. For the sake of fairness for IQA comparison, we follow the same experimental setup and training strategy as described in [7]. Note, we use the publicly available data of the PIPAL dataset for reporting on results of IQA models. It can be seen that our AGAIQA demonstrates considerable efficacy when evaluating challenging distortions e.g., from GAN-based algorithms. The PIPAL is so far the largest and most diverse IQA database of artificial/synthetic

TABLE V
RESULTS OF STATISTICAL SIGNIFICANCE TESTING FOR MODEL PERFORMANCE ON ARTIFICIAL/SYNTHETIC DATASETS. M-DE-T-S-DO REPRESENTS STATE-OF-THE-ART MODELS INCLUDING MANIQA, DEIQT, TRES, STAIRIQA, AND DOIQA. "1" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT (P<0.05 AT THE 95% CONFIDENCE LEVEL). "0" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS NOT STATISTICALLY SIGNIFICANT.

	LIVE	TID2013	CSIQ	KADID-10k	PIPAL
	M-DE-T-S-DO	M-DE-T-S-DO	M-DE-T-S-DO	M-DE-T-S-DO	M-DE-T-S-DO
AGAIQA (ours)	1-1-1-1-1	1-1-1-1-1	1-1-1-1-1	1-1-1-1-1	1-1-1-1-1

TABLE VI
RESULTS OF STATISTICAL SIGNIFICANCE TESTING FOR MODEL PERFORMANCE ON AUTHENTIC/NATURAL DATASETS. M-DE-T-S-DO REPRESENTS THE STATE-OF-THE-ART MODELS INCLUDING MANIQA, DEIQT, TRES, STAIRIQA, AND DOIQA. "1" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT (P<0.05 AT THE 95% CONFIDENCE LEVEL). "0" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS NOT STATISTICALLY SIGNIFICANT.

	SPAQ	KonIQ-10k	CLIVE
	M-DE-T-S-DO	M-DE-T-S-DO	M-DE-T-S-DO
AGAIQA (ours)	1-1-1	1-1-1	1-1-1

distortions, and the reported results reveal the robustness of AGAIQA in handing unseen data.

In addition, we conduct a comparative experiment of BIQA models on three IQA datasets of authentic/natural distortions, namely CLIVE, KonIQ-10k, and SPAQ, to validate the effectiveness of our proposed AGAIQA model in more demanding real-world scenarios. The experimental results of Table IV demonstrate that **AGAIQA** remains the top-performing IQA model in handling authentically distorted images.

To critically evaluate the generalisation ability of the proposed AGAIQA model, we conduct a cross-dataset validation experiment. This entails training an BIQA model on the PIPAL dataset and subsequently subjecting it to testing on the LIVE and TID2013 datasets, without engaging in any form of fine-tuning or parameter adaptation. By doing so, the bad performance of an BIQA cannot be masked. We select top-performing deep learning-based NR-IQA models that have made the implementation code transparently available in our study, including DBCNN, TReS, MetaIQA, HyperIQA, StairIQA, DEIQT and MANIQA. The experimental results in terms of the PLCC and SROCC values are shown on Table VIII. It is evident that the AGAIQA significantly outperforms the top-ranked BIQA models. Especially, there is an average increase of 3% in the performance when comparing the AGAIQA to MANIQA. These findings demonstrate the good generalisation capability of our proposed AGAIQA model.

D. Statistical Significance Testing

To verify whether the observed differences in the model performance results (i.e., as shown in Table II, III and IV) are statistically significant, we perform hypothesis testing using the statistical methods as prescribed in [50]. By doing so, we can better interpret the meaningfulness of the performance gain achieved by our proposed AGAIQA model in comparison to other models. To conduct realistic and efficient testing, we select the top-performing deep learning-based IQA models that have their implementation code transparently available in our study. More specifically, this is to demonstrate that our proposed AGAIQA model is statistically significantly better than the IQA models that exhibit comparative performance. The selected models are MANIQA, DEIQT, TReS, StairIQA, and DOIQA. The statistical significance test is based on the test set (20% of the entire data) of each of the IQA datasets used in our experiments, i.e., LIVE, TID2013, CSIQ, KADID-10k, PIPAL, SPAQ, KonIQ-10k, and CLIVE. For instance, on the test set of the TID2013 dataset, each model generates 600 data points for performance residuals between the ground truth and predicted quality scores. The comparison of performance between two IQA models is based on their produced residuals (i.e., 600 data points each). When both residual samples under consideration are normally distributed, we perform an independent samples t-test. In the case of non-normality, a non-parametric version, i.e., Mann-Whitney U test, analogous to an independent samples t-test, is conducted. The results of the statistical evaluation are presented in Table V and Table VI. The tables indicates that our proposed AGAIQA model is statistically significantly (P<0.05 at the 95% confidence level) better than any of the other top-performing IQA models in predicting perceived image quality.

TABLE VII
ABLATION STUDY TO VERIFY THE CONTRIBUTION OF INDIVIDUAL KEY COMPONENTS (I.E., GAB, ASM AND PHP) PROPOSED IN OUR MODEL TOWARDS THE OVERALL PERFORMANCE IMPROVEMENT. THE KADID-10K DATASET IS USED.

#	GAB	ASM	PHP	PLCC	SROCC
1				0.939	0.939
2	✓			0.947	0.943
3	✓	✓		0.946	0.942
4			✓	0.945	0.940
5	✓	✓	✓	**0.952**	**0.947**

E. Ablation Study

To verify the contribution of individual key components (i.e., GAB, ASM and PHP) proposed in our model towards the overall performance improvement, we carry out a series of ablation studies on the KADID-10k dataset, as detailed in Table VIII. The reason of using KADID-10k for this particular study is that this dataset is the largest and most challenging standard dataset as shown in Table I, hence the evaluation is

TABLE VIII
RESULTS OF CROSS-DATASET VALIDATION. MODELS ARE TRAINED ON PIPAL AND TESTED ON LIVE AND TID2013.

Train on		PIPAL			
Test on		LIVE		TID2013	
		PLCC	SROCC	PLCC	SROCC
Method	DBCNN [43]	0.638	0.625	0.518	0.511
	TReS [20]	0.646	0.658	0.523	0.545
	MetaIQA [15]	0.712	0.704	0.622	0.589
	HyperIQA [16]	0.756	0.739	0.637	0.601
	StairIQA [24]	0.773	0.758	0.661	0.645
	DEIQT [25]	0.798	0.776	0.682	0.659
	MANIQA [7]	0.839	0.825	0.712	0.626
	AGAIQA (Ours)	**0.869**	**0.856**	**0.734**	**0.653**

critical. Table VII lists the results of the ablation study, include #1 being the baseline model where none of the aforementioned components is employed. Models #2, #3 and #4 are model variants; and each include only a single component, i.e., GAB, ASM or PHP. Model #5 represents the proposed AGAIQA model that integrates all three key components. Note that the ASM module cannot function independently and is inherently dependent on the GAB module for effective attention selection, as it operates based on the graph's connectivity. Thus, in #3, we conducted ablation study to evaluate the ASM component with GAB+ASM alone, excluding PHP. The results show that all AGAIQA variants give better performance than the baseline model, demonstrating the contribution of the proposed components. Also, the final AGAIQA model provides the best results, suggesting the importance of combing all three components in a BIQA model.

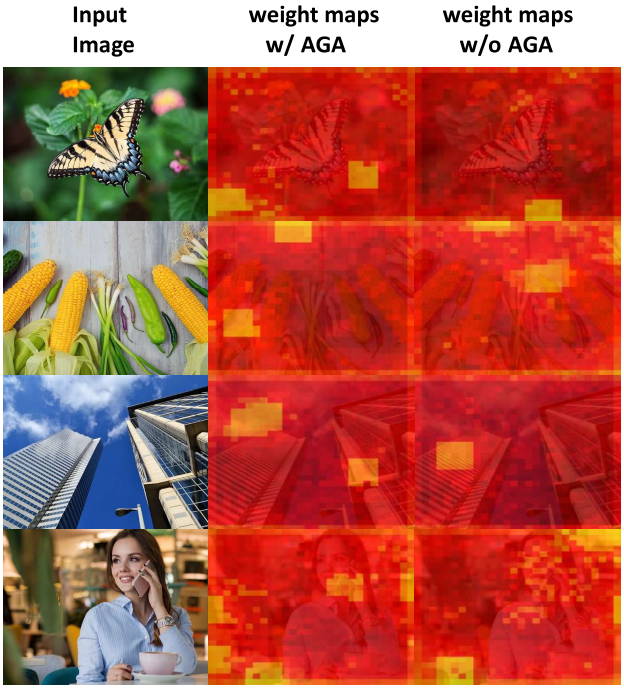


Fig. 6. Visualization of the weight map for the weighting branch with the test dataset of KADID-10k. "w/ AGA" indicates the use of the adaptive graph attention module, whereas "w/o AGA" signifies the absence of the adaptive graph attention module.

TABLE IX
COMPARISION OF MODEL TRAINING TIME AND PARAMETER COUNT.

Model	Training Time (per epoch)	Parameter Count
DEIQT	25 mins	40.51M
DOR-IQA	31 mins	107.32M
StairIQA	32 mins	122.47M
MANIQA	34 mins	135.62M
AGAIQA (ours)	36 mins	148.77M

To provide further insights into the contribution of GAB and ASM components (i.e., the AGA module) in boosting local representational capacity of post-ViT feature maps, we visualise the intermediate weighting maps as produced after the AGA module and situated in the PHP module. The weighting map is already illustrated in Figure 5 at the end of the left branch of the MSE regression block. Figure 6 shows some examples of the weighting maps with and without using the AGA module. It can be seen that with the application of the AGA module, the model becomes adept at focusing on more perceptually meaningful regions for the IQA task, enhancing its performance in predicting perceived image quality.

V. DISCUSSION

In many real-world applications, it is crucial to understand the practical implications and performance characteristics of IQA methods in relation to runtime efficiency and computational complexity. We provide below measurements of runtime, encompassing both time consumption and parameter count, for our proposed method in comparison to some state-of-the-art (SOTA) methods (note this selection is limited to the top-performing deep learning-based NR-IQA models that have made the implementation code transparently available in our study) on the KADID-10k dataset, as depicted in Table IX. The results reveal that while our model's runtime performance and parameter count are comparable to MANIQA, StairIQA and DOR-IQA models and that our model is more complex than the DEIQT model, the proposed AGAIQA exhibits superior prediction accuracy compared to its counterparts. However, it should be noted that a comprehensive analysis of computational complexity is nontrivial, and it often involves additional factors such as algorithmic efficiency and resource requirements. Furthermore, in many practical scenarios, there exists a delicate balance between model's complexity and its prediction accuracy, necessitating careful considerations of trade-offs for specific circumstances.

VI. CONCLUSION

In this paper, we have presented the AGAIQA, a novel model for blind image quality assessment (BIQA). The key innovation of this model is to incorporate our proposed Adaptive Graph Attention (AGA) module with Transformers. The AGA module can successfully enhance the post-transformer feature representations by leveraging the local information through a graph-based structure, while preventing the introduction of redundant information during feature updates in learning. The proposed AGAIQA also benefits from the integration

of a patch-wise hierarchical perceptual (PHP) regression to infer image quality scores aligned with the human judgements. Experimental results demonstrate the superiority of AGAIQA over state-of-the-art methods cross popular image quality assessment datasets. Critically, our AGAIQA model exhibits robust generalisability. We conclude that by harnessing transformers for extracting contextual information and AGA structure to enhance the local feature expression can significantly improve image quality assessment.

REFERENCES

[1] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63:1–52, 2020.

[2] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[3] Soomin Seo, Sehwan Ki, and Munchurl Kim. A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions. *IEEE TCSVT*, 31(7):2602–2616, 2020.

[4] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *CVPR*, pages 1808–1817, 2018.

[5] Abdul Rehman and Zhou Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE TIP*, 21(8):3378–3389, 2012.

[6] Huasheng Wang, Yulin Tu, Xiaochang Liu, Hongchen Tan, and Hantao Liu. Deep ordinal regression framework for no-reference image quality assessment. *IEEE Signal Processing Letters*, 2023.

[7] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022.

[8] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International conference on multimedia*, pages 789–797, 2020.

[9] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, pages 14071–14081, 2023.

[10] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[11] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, 2010.

[12] Xinbo Gao, Fei Gao, Dacheng Tao, and Xuelong Li. Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning. *IEEE TNNLS*, 24(12), 2013.

[13] Mingdeng Cao, Yanbo Fan, Yong Zhang, Jue Wang, and Yujiu Yang. Vdtr: Video deblurring with transformer. *IEEE TCSVT*, 33(1):160–171, 2022.

[14] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In *CVPR*, pages 732–741, 2018.

[15] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *CVPR*, pages 14143–14152, 2020.

[16] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *CVPR*, pages 3667–3676, 2020.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[19] Junyong You and Jari Korhonen. Transformer for image quality assessment. In *IEEE International Conference on Image Processing*, pages 1389–1393, 2021.

[20] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *WACV*, pages 1220–1230, 2022.

[21] Yixuan Gao, Xiongkuo Min, Yucheng Zhu, Xiao-Ping Zhang, and Guangtao Zhai. Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] Jinsong Shi, Pan Gao, and Aljosa Smolic. Blind image quality assessment via transformer predicted error map and perceptual quality token. *IEEE Transactions on Multimedia*, 26:4641–4651, 2024.

[23] Mingliang Zhou, Shujun Lang, Taiping Zhang, Xingran Liao, Zhaowei Shang, Tao Xiang, and Bin Fang. Attentional feature fusion for end-to-end blind image quality assessment. *IEEE Transactions on Broadcasting*, 69(1):144–152, 2022.

[24] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[25] Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2091–2100, 2023.

[26] Long Xu, Jia Li, Weisi Lin, Yongbing Zhang, Lin Ma, Yuming Fang, and Yihua Yan. Multi-task rank learning for image quality assessment. *IEEE TCSVT*, 27(9):1833–1843, 2016.

[27] Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. GraphIQA: Learning distortion graph representations for blind image quality assessment. *IEEE TMM*, 25:2912–2925, 2023.

[28] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE TIP*, 15(11):3440–3451, 2006.

[29] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010.

[30] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

[31] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–3, 2019.

[32] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. PIPAL: a large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, pages 633–651. Springer, 2020.

[33] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020.

[34] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015.

[35] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

[36] Shanshan Lao, Yuan Gong, Shuwei Shi, Sidi Yang, Tianhe Wu, Jiahao Wang, Weihao Xia, and Yujiu Yang. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *CVPR*, pages 1140–1149, 2022.

[37] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE TIP*, 21(8):3339–3352, 2012.

[38] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 723–727, 2011.

[39] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE TIP*, 24(8):2579–2591, 2015.

[40] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2016.

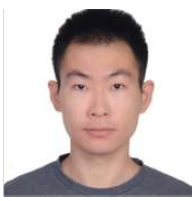
[41] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE TIP*, 27(3):1202–1213, 2017.

[42] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference

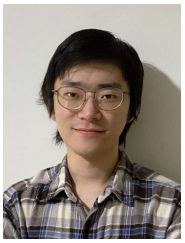
and full-reference image quality assessment. *IEEE TIP*, 27(1):206–219, 2017.
[43] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE TCSVT*, 30(1):36–47, 2018.
[44] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *CVPR*, pages 3575–3585, 2020.
[45] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCV Workshops*, pages 0–0, 2018.
[46] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *CVIU*, 158:1–16, 2017.
[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.
[48] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE TIP*, 20(8):2378–2386, 2011.
[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.
[50] Wei Zhang, Ali Borji, Zhou Wang, Patrick Le Callet, and Hantao Liu. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE transactions on neural networks and learning systems*, 27(6):1266–1278, 2015.

Huasheng Wang received the B.Eng. from Xiangtan University, in 2018 and the M.S. degree from Dalian University of Technology in 2021. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK. His interests are Saliency Prediction, Image Quality Assessment and Depth Estimation.

Jiang Liu received the B.Eng. and M.S. degrees from China University of Mining and Technology, Xuzhou, China, in 2019 and 2022 respectively. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK. His interests are Action Quality Assessment and Image Quality Assessment.



Hongchen Tan is a Teacher of College of Future Technology at Dalian University of Technology. He received his Ph.D degree in computational mathematics from the Dalian University of Technology, Dalian, China, in 2021. His research interests is Computer Vision.



Jianxun Lou received the B.Eng. from Central South University, Changsha, China, in 2018 and the M.S. degree from Cardiff University, Cardiff, UK, in 2020. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK.

Xiaochang Liu is currently pursuing the bachelor's degree within School of Mathematics at Sun Yat-sen University, China. Her research interests include mathematical modelling and data analytics.

Wei Zhou is an Assistant Professor at Cardiff University, United Kingdom. Dr Zhou was a Post-doctoral Fellow at University of Waterloo, Canada. Wei received the Ph.D. degree from the University of Science and Technology of China in 2021, joint with the University of Waterloo from 2019 to 2021. Dr Zhou was a visiting scholar at National Institute of Informatics, Japan, a research assistant with Intel, and a research intern at Microsoft Research and Alibaba Cloud. Wei is now an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems. Wei's research interests span multimedia computing, perceptual image processing, and computational vision.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently a Professor at the School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom.