Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19457-5

# Attention enabled viewport selection with graph convolution for omnidirectional visual quality assessment

Nandhini C[1] · Brindha M[1]

Received: 14 July 2023 / Revised: 15 May 2024 / Accepted: 17 May 2024

## Abstract

Omnidirectional images provide an immersive viewing experience in a Virtual Reality (VR) environment, surpassing the limitations of traditional 2D media beyond the conventional screen. This VR technology allows users to interact with visual information in an exciting and engaging manner. However, the storage and transmission requirements for 360-degree panoramic images are substantial, leading to the establishment of compression frameworks. Unfortunately, these frameworks introduce projection distortion and compression artifacts. With the rapid growth of VR applications, it becomes crucial to investigate the quality of the perceptible omnidirectional experience and evaluate the extent of visual degradation caused by compression. In this regard, viewport plays a significant role in omnidirectional image quality assessment (OIQA), as it directly affects the user's perceived quality and overall viewing experience. Extracting viewports compatible with users viewing behavior plays a crucial role in OIQA. Different users may focus on different regions, and the model's performance may be sensitive to the chosen viewport extraction strategy. Improper selection of viewports could lead to biased quality predictions. Instead of assessing the entire image, attention can be directed to areas that are more importance to the overall quality. Feature extraction is vital in OIQA as it plays a significant role in representing image content that aligns with human perception. Taking this into consideration, the proposed ATtention enabled VIewport Selection (ATVIS-OIQA) employs attention based view port selection with Vision Transformers(ViT) for feature extraction. Furthermore, the spatial relationship between the viewports is established using graph convolution, enabling intuitive prediction of the objective visual quality of omnidirectional images. The effectiveness of the proposed model is demonstrated by achieving state-of-the-art results on publicly available benchmark datasets, namely OIQA and CVIQD.

**Keywords** Omnidirectional image quality analysis · Deep learning, virtual reality · VR image quality assessment · Attention sampling · Graph convolution

✉ Brindha M
brindham@nitt.edu

Nandhini C
cn.nandhini@gmail.com

[1] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, TamilNadu 620015, India

Published online: 06 June 2024

## 1 Introduction

Nowadays, with the accelerated proliferation of VR technologies, omnidirectional videos and images have been increasingly popular and have drawn great attention. They have been employed in an abundance of application scenarios, such as TV, film, broadcasting, designing products, auto-driving, education, etc. User experience is one of the key motivations for the evolution and success of VR technologies and applications. Omnidirectional images/videos provide an immersive experience with an unlimited Field of View (FoV) in a VR environment through Head Mounted Displays (HMD). The users can interact with the visual content in an exciting way from any direction using their head movement. However, restricted by several inadequacies in graphic instruments, transfer bandwidth, and viewing devices, the content viewed by observers usually cannot quench their satisfaction. The spherical representation of the omnidirectional image and its high resolution complicates image acquisition, storage, compression, encoding, transmission, and visual display. These stages may degrade the quality by introducing white noise, blurring, compression, and projection distortions, thereby severely compromising the quality of the experience. So, quality deterioration is typical and it may cause irritation for a longer time viewing experience.

The omnidirectional spherical images are converted into 2D format using equirectangular projection (ERP) for efficient transmission and storage. During this encoding process, the bipolar regions of the sphere undergo geometrical deformations which affect the image quality. Pixel redundancy and geometric deformation additionally antagonize storing its contents. Further, the equirectangular format of these images differs substantially from the actual content viewed in the HMD. This is because in a VR environment, the viewer can see any location of the spherical image by changing the head orientation, but only a very small part of the whole content is seen at the same time. Each of these requirements is confined to lesser apparatus, high transmission bandwidth and conversion to 2D formats tends to introduce some projection distortion, compression degradation, and geometry deformation. The visual content of surrounding viewports influences the user's evaluation of every viewport in obtaining the aggregate quality score. Moreover, conventional 2-D Image Quality Assessment (IQA) models cannot be applied directly on omnidirectional images because they do not account for the characteristics of 360-degree content such as non-linear transformation like spherical image projections. Hence, 360-degree Image Quality Assessment is a prominent study that helps with effectively evaluating the quality performance of VR technologies in order to maximize the quality of realistic viewing experience.

Typical viewing process of 360-degree media contents begins with looking around the spherical interface, where the visual information of interacting multiple viewpoints is aggregated locally. At the end of the perusal, the entire scenery is reconstructed by the observer in his hallucination based on what he has seen to gain a general estimation of the quality globally. It is imperative for both local and global quality evaluation to resolve the overall quality. The proposed ATVIS-OIQA model utilizes global and local quality estimation to be more effective for Omnidirectional Image (OI) quality evaluation and better generalization performance. While CNNs dominate computer vision, transformers have demonstrated exemplary performance in various natural language processing (NLP) tasks due to their excellent ability to model sequences and long-distance dependencies. You and Korhonen [25] experiment with the employment of transformers to solve the task of 2D IQA. They achieved state-of-the-art performance on two public databases for blind IQA. Haoran and Yang et.al. [4] enhanced the vision transformer with two parallel modules and multi-lingual self-attention. The parallel-designed Dynamic Unary Convolution in Transformer (DUCT) blocks are added into the deep architecture, which enhances the computer vision tasks for classification, segmentation, and other relevant tasks. The proposed method explores an end-to-end training strategy that selects the viewport using an attention mechanism and utilizes transformer-based architectures and graph convolution to model the spatial dependencies in the omnidirectional images.

The main contribution of the proposed work are as follows:

- Develop a learning-based attention sampling model to select and extract the most important viewports in a 360° image with fewer computational steps.
- Investigate the application of transformers in OIQA to extract distinctive features from input viewport images
- Design a graph convolution network that helps extract local and global information to better model the spatial dependencies between viewports

**Implications of the proposed method** The proposed method significantly impacts almost all the broad domains of image processing tasks. This section deals with the implications of the various domains listed below.

***Image Quality Assessment (IQA):*** Proposed method introduced an advanced method for assessing the quality of the omnidirectional images. The impact of using attention and graph convolution networks significantly improves the accuracy and reliability of visual quality assessment in various image processing applications.

***Information Fusion:*** Fusing information from various viewports improves the overall visual quality. This fusion leads to better handling of distortions, thereby enhancing the robustness of the visual assessment [21].

***Impact on image processing tasks:*** The viewport selection and attention mechanism can be applied to tasks like image restoration to improve the most degraded image portions. It can also be applied to image compression while preserving quality in key areas and enhancing compression quality.

***Broaden Scope:*** Beyond omnidirectional images, proposed techniques can also influence other relevant areas like medical image fusions [20], where accurate quality assessment is crucial for diagnostics and environment perception in autonomous vehicles.

## 2 Related work

This section critically appraises the previous work published in the literature pertaining to OIQA Task. Initially, conventional FR IQA metrics had been extended to create FR OIQA metrics. On account of their coherence and statistical comfort, two very popular FR metrics - Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), were used to assess the capabilities of popular media restoration and coding technologies. Several PSNR-based FR OIQA models had been investigated. Spherical PSNR (S-PSNR) presented in [26], picked uniformly distributed points on the sphere rather than the casted panoramic image to eliminate pixel redundancy in 2-D extended omnidirectional images. Sun et al. [19] presented Weighted-to-spherically uniform PSNR (WS-PSNR) where the stretched regions are weighed and the associated map is integrated with an error map to reduce the impact of drawn out areas and establish a balanced non-uniform sampling density. Zakharchenko et al. [27] put forth the craster parabolic projection PSNR (CPP-PSNR), which computed the location-invariant PSNR on the craster parabolic projection domain promising uniform sampling density and lesser shape distortion. Xu et al. [24] predicted positional information using Non-Content-based-Perceptual PSNR (NCP-PSNR), while viewing direction was predicted by Content-based Perceptual PSNR (CP-PSNR). These metrics assigned different weights to distortions at different locations. The models relying on PSNR significantly underperformed compared to the conventional successful IQA approaches for 2D natural images to precisely predict the human perceived visual quality. This can be attributed to the discrepancy between PSNR and the HVS. Consequently, the properties of the HVS were employed to construct perception-based models using SSIM which calculates the contrast, luminescence, and structural similarities of each pixel in the spherical domain and its variants. Zhou et al. [32] proposed the Weighted-to-Spherically uniform SSIM (WS-SSIM) which combined the error and location-weighted map to ensure different weights to different distortions at different locations. Chen et al. [3] proposed the Spherical Structural Similarity Index (SSSIM) for omnidirectional video quality evaluation by computing the similarity between reference and distorted 360-degree images on the sphere. Researchers from Facebook proposed SSIM360 and 360VQM to verify the performance of 360 video pipelines on encoding and streaming [2]. SSIM360 is a result of weighing each sample SSIM on the basis of how much the sampled area is expanded when cast or projected.

However, the FR OIQA method faces significant limitations as it relies on reference images. Moreover, the aforementioned OIQA models are evaluated using 2-D image processing techniques that fail to consider the unique visual characteristics of omnidirectional images. Consequently, they struggle to achieve exceptional performance in assessing the quality of omnidirectional images.

With the fast proliferation of CNN-based models, researchers adopted deep learning based strategies to construct objective OIQA indexes. Despite expansive research attention in the realm of IQA in the present day, there is still an insufficiency in efforts to predict the objective quality of panoramic pictures especially because in most practical scenarios, reference or backing images are difficult to acquire and hence, NR/Blind IQA metrics are being researched extensively although it is challenging.

Early Blind OIQA methods leveraged patch-level features to perform quality prediction. Several modern deep quality assessment metrics regard images as an assemblage of bit-sized patches. Patch-wise metrics tend to conduct individual computations on patches only to be accumulated later to attain the final quality rating rather than handling inputs in their complete resolution. Most existing methods, however, do not focus on complex attention models and simply process each patch independently which is overcome by [11] and [9]. Li et al. [11] utilized the concept of multi-task guided prediction of saliency maps and proposed a model that predicted the weight plots of head movement and eye movement with visual attention models of DHP and SalGAN. The model was able to correlate human behaviour with the task of quality assessment. Kim et al. [9] presented a deep learning pipeline for Virtual Reality image quality assessment (DeepVR-IQA), accounting for the spherical representation of the omnidirectional content based on the concept of adversarial learning. The authors used the spherical positional information of the sampled patches to weigh individual quality ratings and predict the final quality score. The final quality rating is extracted from accumulating the local quality values with their weights. However, due to significant deformation revealed on the projection plane of the sampled images, patches failed to manifest the actual viewing information contained in the image. Hence viewport-oriented methods were developed. Viewport-based Neural Network (V-CNN) [12] proposed an approach conceptualizing images as viewports rather than patches. The authors decomposed the task of OVQA into two subsidiary tasks - to propose potential viewports based on the locations and predict the saliency of HM and EM maps. VQA score rating produced the weighted average over quality scores of selected viewports to ultimately predict the perceptual quality of omnidirectional images. MC360IQA [18] proposed a multi-channel CNN for blind OIQA. Images from 6 viewpoints are processed parallelly through 6 hyper-ResNet-34 [6] networks. The fusion of features in the image quality regressor generates the final quality score. Despite their success, they failed to consider the spatial dependence between different viewports.

Xu et al. [23] proposed a viewport-oriented graph convolutional network(VGCN) which took advantage of the spatial mutual interactions between the extracted viewports using a GCN. In addition, the authors employ the DB-CNN to find the global quality with a low-resolution ERP image as an input. A culmination of both scores predicts the final quality score. [13] apply the Local Binary Pattern (LBP) operator to encode cross-channel color information and utilize the weighted LBP technique to extract structural features. Additionally, the viewport sampling method is used to extract local NSS features, then support vector regression is utilized to predict the quality score of the observed images. [31] introduced Distortion Discrimination Assisted Multi-stream Network (DDAMN), which has the capability to assess the quality and distortion type distinguish ment task. The auxiliary distortion discrimination task enhances the learning process of OIQA. Furthermore, DDAMN introduced a data augmentation strategy that involves generating multiple sets of viewport images from a single omnidirectional image. Chen and Han et al. [1] proposed a semi-supervised learning method for dense prediction models with limited labels; a virtual category is assigned to each confusing sample to contribute to model optimization. This approach is helpful for a CNN-based segmentation and detection system with extremely limited labels. Despite the astounding performance of the CNN based methods, they suffer from major limitations. For instance, the pre-processing tasks, like viewport selection, are more computationally expensive than the actual IQA task. Hand-crafted designs of viewport sampling strategies might not hold for all types of spherical distortions. Often, viewports are processed either independently or considering only spatial interactions. Quality evaluation in all methods discussed so far ignored temporal and semantic correlations like content characteristics. Further, the CNNs exhibit a very limited, skewed tendency to model spatial dependencies. Hence, all these issues raise the concern about improving the unsatisfactory and computationally expensive OIQA performance.

## 3 Proposed methodology

The proposed pipeline of the ATVIS-OIQA model is illustrated in Fig. 1 The task of perceptual quality prediction has been bifurcated into local and global branches.

The primary branch comprises the following components:

A viewport detector A viewport descriptor A viewpoint quality aggregator It is similar to the observer perceiving the scenery using the HMD; the user first chooses the FoV, then information from multiple viewports and their dependency are aggregated to realize the entire scene. The viewport detector samples the most useful viewports using attention sampling. The viewport descriptor extracts distinctive features indicative of the perceptible information inside the input viewpoints for predicting the quality by leveraging a pre-trained vision transformer architecture. The viewport aggregator models the spatial relations between the viewports in 360-degree images by building a spatial viewport graph and aggregates visual information from multiple viewports to arrive at the local quality.

The supplementary branch consists of the global quality estimator and quality regressor. It estimates the quality of the complete distorted image without viewport selection using the vision transformer architecture. Finally, the quality regressor is used to fuse the local quality obtained from the primary branch and the global quality from the supplementary quality evaluator to determine the conclusive omnidirectional image quality of the 360-degree image.

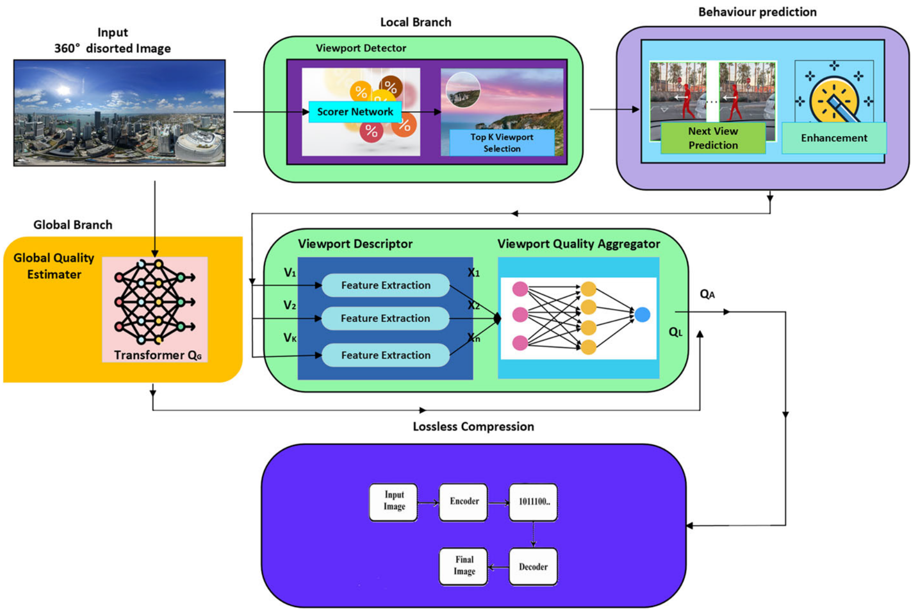

**Fig. 1** Illustration of the proposed ATVIS-OIQA Model

The following sub-sections detail each individual part in sequence.

## 3.1 Viewport detector

**Algorithm 1** Viewport Selection

```
Input:
    $x\_low$ – low resolution view of image
    $x\_high$ – high resolution view of image
Output:
    $v$ – set of top N "informative viewports"
    $attn$ – attention map computed from $x\_low$
  procedure SELECTVIEWPORTS($x\_low, x\_high$ : image)
    $attn \leftarrow attention\_network(x\_low)$
    $samples \leftarrow scorer\_network(log(attention), n\_viewports)$
    $top\_samples \leftarrow topk(samples)$
    $v \leftarrow extract\_viewports(top\_samples, x\_high)$
    return attn, v
  end procedure
```

The viewport detector aims to select the most salient viewports in accordance with structural information, which is most appealing to users following the literature [8]. Algorithm 1 briefly outlines the overall viewport selection algorithm. The viewport selector uses an attention network to select the most prominent visual features using the low-resolution image. It uses a scorer network based on attention sampling to extract the key points. The module samples locations of informative viewports from an attention distribution computed on an unclear version of the original image, thus processing only a fraction of the original image.

It then proceeds to use this attention map on the high-resolution image to sample and extract the top N viewports. This subsection first explains the generic formulation for attention in a neural network, then describes how the sampling attention significantly reduces the required computation by generating an optimal approximation, and later details the method to speed up the high-resolution image processing.

### 3.1.1 Scorer network

Attention plays a significant role in human perception. Recently, attention has been widely employed to enhance the representation of deep convolutional neural networks (DCNNs) in various computer vision tasks. The success of attention lies in its ability to discern the relevance of specific features without necessitating prior knowledge about the target task. Instead, features extracted from the internal layers of the deep neural network are analyzed to determine their significance.

The scorer network provides an effective attention-based sampling technique for selecting structurally sensitive patches. Attention sampling is used to identify the saliency region that requires further analysis. The low-resolution (LR) image is utilized in the scorer network to mitigate computation and memory bottlenecks.

By leveraging the attention model, the distorted input image $x$ is transformed into an attention map, which is subsequently employed for selecting the samples in the following stages.

Let $\hat{y} = \Gamma(x;\theta)$ be the neural network parameterized by $\theta$, where $(x, y)$ is an input image and its score in the dataset is given using

$$\Gamma(x;\theta) = g(f(x;\theta);\theta) \tag{1}$$

The attention mechanism is applied to the intermediate representation of the neural network $f(x;\theta) \in R^{M\times D}$. The attention function is defined as $a(x;\theta) \in R_{+}^{M}$ such that $\sum_{i=1}^{M} a(x;\theta)_i = 1$ is used in selecting the more prominent R features of dimension D. Hence (1) can be rewritten as

$$\Gamma(x;\theta) = g(\sum_{i=1}^{M} a(x;\theta)_i f(x;\theta)_i) \tag{2}$$

the subscript $i$ is used to extract the $i$-th value from the vector.

As seen, a(.) by definition is a multinomial distribution over M pixels location in the image. Let $I$ be a random variable sampled from $a(x;\theta)$ then the attention in the neural network can be rewritten in terms of expectation of intermediate features over the attention distribution $a(.)$.

$$\begin{aligned}\Gamma(x;\theta) &= g\left(\sum_{i=1}^{M} a(x;\theta)_i f(x;\theta)_i\right)\\ &= g\left(\sum_{I\sim a(x;\theta)}^{M} [f(x;\theta)_I]\right)\end{aligned} \tag{3}$$

Now the expectation can be approximated by using the Monte Carlo estimate and thus, can avoid the computation of all the M features. Now,

$$Q = \{q_i \sim a(x;\theta) \mid i = \{1, 2, 3, \ldots N\}\}$$

is sampled, where Q is a set of N indices from the attention distribution. The attention network can be approximated as

$$\Gamma(x;\theta) \approx g\left(\frac{1}{N}\sum_{q\in Q} f(x;\theta)_q\right) \tag{4}$$

Since a neural network is used as the attention distribution, the gradient of the loss with respect to the attention function parameters should be used by sampling the set of indices of Q. The gradient that we have used in the proposed model is defined as follows

$$\frac{\delta}{\delta\theta}\frac{1}{N}\sum_{q\in Q} f(x;\theta)_q = \frac{1}{N}\sum_{q\in Q}\frac{\frac{\delta}{\delta\theta}\left[a(x;\theta)_q f(x;\theta)_q\right]}{a(x;\theta)_q} \tag{5}$$

This reduces the computation, as for the sampled indices in Q, it computes only the rows of $f(\cdot)$.

In most of the implementations that use attention in neural networks [7] it is required to compute all the features because $a(.)$ is a function of features $f(.)$. But here to avoid computing all K features in Monte Carlo estimation of (2), a low resolution image view of the original image is used. This, as a result, bargains a greater speed from sampling attention.

Now given an image $x \in R^{H\times W\times C}$ where H is the height of the image, W is the width of the image and C is the number of channels. Its corresponding view is $V(x,s) = R^{h\times w\times c}$ at scale s where h is the height of the low resolution image such that $h < H$, w is the width of low resolution image such that $w < W$ and C is the number of channels which remains unchanged. For this, the attention $\hat{A}$ is computed as follows

$$a(V(x,s);\Theta) : R^{h\times w\times C} \rightarrow R^{hw} \tag{6}$$

$$\hat{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1w_s} \\ a_{21} & a_{22} & \cdots & a_{2w_s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{h_s1} & a_{h_s2} & \cdots & a_{hw} \end{pmatrix} \tag{7}$$

where $a_{ij}$ is the probability value and the higher probability indicates the higher saliency pixel. The advantage provided by this attention block is to focus on prominent regions based on the image content and it generates the global feature representation. Directly sampling from these underlying feature maps leads to generating stronger attention masks that are more aggressive in selecting discriminating visual structures.

The attention model that generates the attention map of the low resolution image is illustrated in Fig.2

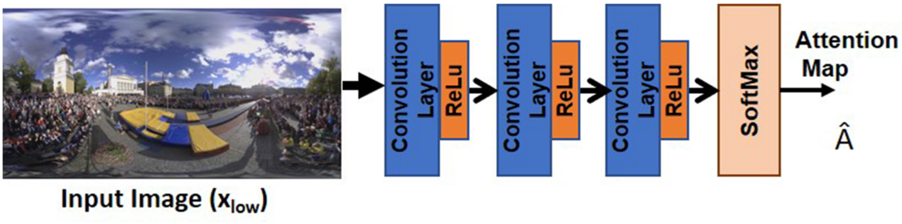

**Fig. 2** Structure of attention model to generate attention map of low resolution image

### 3.1.2 Top K viewport selection

Attention maps are used for sampling which produces optimal approximation and thereby reduces the computational complexity. As attention is acquired from multiple locations in the image, it is considered as the multinomial distribution over the discrete elements as opposed to a continuous distribution. Moreover, the attention map is generated during the forward pass, and it is not differentiable, hence the gradients cannot be backpropagated. The deterministic parts of the sampling process are attained by applying the log probabilities on the input attention $\hat{A}$, and then flattening it to reduce the computational requirements and improve numerical stability. The gumbel noise is combined with log probabilities of the gumbel softmax distribution and softmax is applied to make it differentiable and select top-k samples.

The patches are sampled from the high resolution image by defining a function P(x, i) which extracts patches centered around $i^{th}$ pixel in $V(x, s)$ from a high resolution image x. From above, only a few patches are considered from the full resolution image x and the model is defined as

$$\begin{aligned}\Psi(x;\Theta) &= g\left(\sum_{i=1}^{hw} a(V(x,s);\Theta)_i f(P(x,i);\Theta)_i\right)\\ &\approx g\left(\frac{1}{N}\sum_{q\in Q} f(P(x,q);\Theta)_q\right)\end{aligned} \tag{8}$$

The position of the patches is passed to the feature function $f(.)$. $a(.)$ is a neural network significantly smaller than feature network $f(.)$ which is called an attention network. Equation $f(.)$ is implemented by a vision transformer which is called a feature network. Finally the function $g(.)$ is a Regression Layer with $L1$ loss function and SGD as optimizer. So low and high resolution map is passed into the scorer network which calculates the attention map and selects the top N patches based on the attention called Viewports.

## 3.2 Next view prediction

*Next View Prediction* serves a critical role in enhancing blur parts before user interaction within the context of our proposed system/application. The primary objective of this mechanism is to anticipate the user's next viewpoint or perspective and pre-process the corresponding image areas to mitigate potential blurriness.

When users interact with visual content, especially in dynamic environments or virtual reality settings, there might be delays or inconsistencies in rendering high-quality images due to computational constraints or network latency. As a result, certain parts of the displayed scene may appear blurred or distorted, detracting from the overall user experience.

To address this issue, our system leverages next view prediction to anticipate the user's upcoming viewpoint based on their interactions and movement patterns. By accurately predicting the next viewpoint, we can proactively identify and prioritize the pre-processing of image regions that are likely to become the focus of attention using VGG-16 [16].

Using advanced algorithms and machine learning techniques, our system analyzes the user's previous interactions, environmental context, and motion trajectories to predict the most probable next viewpoint. Once the next viewpoint is predicted, we apply image enhancement and deblurring techniques to the corresponding image regions in real-time or pre-rendering stages.

By pre-processing the anticipated blur parts before the user interacts with them, our system ensures a smoother and more immersive user experience. Users can enjoy sharper and clearer visuals, leading to increased engagement, satisfaction, and immersion in the virtual environment or interactive application.

Overall, the integration of next view prediction to enhance blur parts before user interaction represents a significant advancement in improving the quality and realism of interactive visual experiences, particularly in applications such as virtual reality, gaming, and immersive simulations.

This module extracts effective features from the top N input viewports for quality prediction. The Vision Transformer architecture illustrated in Fig. 3 has been leveraged as Viewport Descriptor because of its success in various computer vision tasks like image recognition, classification and segmentation due to its reduced architectural complexity, scalability and training efficiency.

Figure 3 illustrates the ViT architecture. Each of top N viewports obtained for each image from the viewport detector $x \in R^{(H \times W \times C)}$, is reshaped into fixed sized patches $x_p \in R^{(N \times (P^2 \cdot C))}$, where H×W is the dimension of the omnidirectional image and C is the number of channels in the omnidirectional image, $P^2$ is the size of each image patch, and, $N = HW/P^2$. The patch embeddings are extracted from the resultant patches by flattening them and mapping them to D dimensions by a trainable linear projection with 2-D convolutional layer. Standard learnable position embeddings are added to the patch embeddings to retain positional information to get the embedding vector.

The resulting sequence of embedding vectors is fed into the encoder which contains alternating layers of Multiheaded self-attention (MSA) and Multilayer perceptron (MLP) blocks. Finally, a MLP layer at the end of transformer encoder outputs the result of viewport representation.

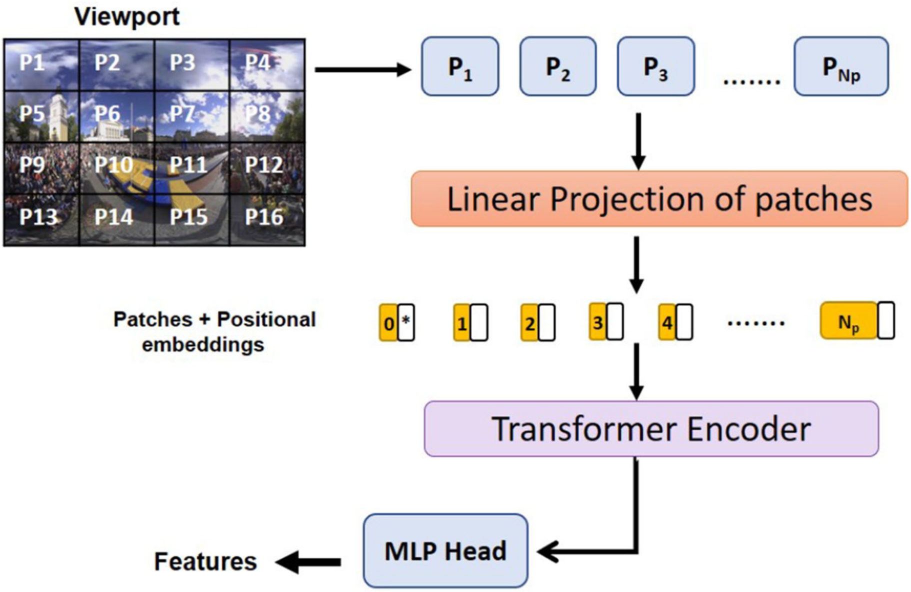

**Fig. 3** ViT model

### 3.3 Viewport quality aggregator

The process of the viewport quality aggretator is shown in Fig. 4. The viewport features of different viewports is given to the graph convolution network, then average pooling is used to estimate the local quality score.

A spatial viewport graph is constructed to create a mutual reliance between different viewports. This is done by taking the N selected viewports as graph nodes, and connecting pairs of these viewpoints with different edges. The procedure is elaborated as shown in Fig. 5 based on [23], as the relation between two viewports A and B is connected only when the medial point of viewpoint B is in the Field of View (FoV) of viewpoint A, the viewport A and C are separated as the medial point of viewpoint C is not in the FoV.

The feature vector for each of the $N$ viewports acquired from the viewport descriptor is denoted by $X$. The feature representation of the $N$ viewports is represented as $X = x_1, x_2, ..., x_N$. Each pair of viewpoints shares a correspondence, imitating the relationship as shown in (9).

$$A\left(x_{i'} x_j\right) = \begin{cases} 1 & \text{if AngularDist}\left[(\Phi_{i'}, \theta_i), (\Phi_{j'}, \theta_j)\right] \\ & \quad \leq 45^\circ, \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where, A denotes the affinity matrix, $AngularDist(.)$ computes the angular distance between two viewpoints i and j on the 3D sphere, $(i, i)$ and $(j, j)$ represent the longitudes and the latitudes of viewports i and j. Assuming the viewport size to be angular 90 degrees, the angular distance threshold in the above equation is set to angular 45 degrees. Then, normalization is implemented as follows:

$$\widehat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{10}$$

where, A shows the spatial viewport graph (un-directed) after normalization represented as an adjecency matrix, D represents the diagonal matrix and $D_{ij} = \sum_i A_{ij}$. Each layer in the graph convolution network propagates using the soft plus activation function as follows:

$$H^{(l+1)} = \sigma\left(BN_{\gamma,\beta}\left(\widehat{A} H^{(l)} W^{(l)}\right)\right) \tag{11}$$

$$\sigma(x) = \log(1 + e^x) \cdot \text{BN}\gamma, \beta(.) \tag{12}$$

Where BN is the batch normalization and $\gamma$, $\beta$ are the trainable parameters. The trainable weight matrix for every layer $l$ is denoted by $W^{(l)}$, and the after-activation matrix in layer $l$

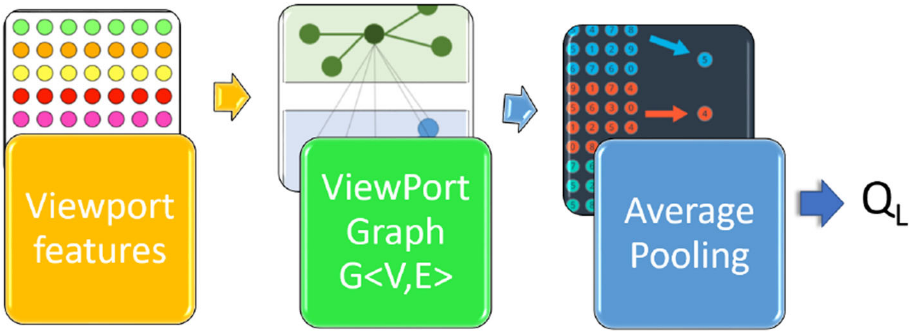

**Fig. 4** Viewport quality aggregator

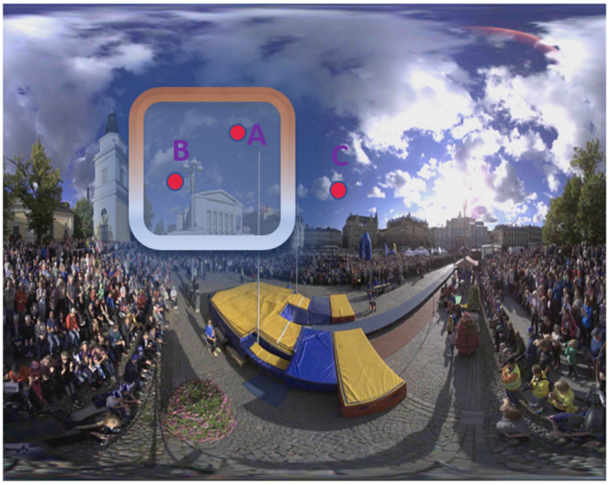

**Fig. 5** Visual examples of spatial relations

is represented by $H^{(l)}$. $H^{(0)} = X$. The features of each viewpoint represented as a node can be changed through interactions between different nodes in graph convolutions. Eight layers of GCN are adopted, and each layer's specific dimensions of feature vectors are 768, 384, 192, 96, 48, 24, 12, 6, and 1, respectively. In the end, a single max-pooling layer cumulates the features from each viewpoint and acquires the local quality $Q_L$. The entire process of quality aggregation training is outlined in Algorithm 2.

**Algorithm 2** Viewpoint Quality Aggregation

```
Input:
    $X = \{x_1, x_2, ..., x_N\}$ $where\ X\ is\ the\ feature\ representation\ of\ N\ viewports$
Output:    $Q_L - Local\ Quality\ Aggregate$
  while $k \leq batch\_size$ do
    if AngularDist$\left[\left(\Phi_{i'}, \theta_i\right), \left(\Phi_{j'}\theta_j\right)\right] \leq 45°$ then
      $A\left(x_{i'}x_j\right) \leftarrow 1$
    else
      $A\left(x_{i'}x_j\right) \leftarrow 0$
    end if
    $D_{ij} \leftarrow \sum_i A_{ij}$
    $\widehat{A} \leftarrow D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$                                    ▷ Normalize A
    $H^{(0)} \leftarrow X$
    while $l \leq 8$ do                                                                   ▷ GCN
      $H^{(l+1)} \leftarrow \text{Softplus}\left(BN_{\gamma,\beta}\left(\widehat{A}H^{(l)}W^{(l)}\right)\right)$
      $W^{(l)} \leftarrow W^{(l)} - \alpha \times \frac{dE^{(l)}}{dW^{(l)}}$
    end while
    $k \leftarrow k + 1$
  end while
  $Q_L \leftarrow maxPool(H^{(8)})$
```

### 3.4 Global quality estimator

Viewpoints are selected and need to be reconstructed when the observers view the 360○ complete scenery. This is illustrated in the local quality branch of ATVIS OIQA model. The global branch of the entire omni-directional image is directly taken in its ERP format as an input and the quality is estimated directly. The vision transformer model inscribed in the section 3.2 extracts the effective features of the whole image. Lastly, a linear layer is applied to get the global quality of the omnidirectional image QG.

**Algorithm 3** Final Quality Estimation

```
Input:
    $Q_L - Local\ Quality,\ I - InputImage$
Output:
    $Q - Final\ Perceptual\ Quality$

 while $k \le batch\_size$ do
     $features \leftarrow ViT(I_k)$
     $Q_G \leftarrow features^T \times W + bias$
     $Q_A \leftarrow Regressor(Q_L, Q_G)$
 end while
```

The Local quality Aggregate QL from the local branch and the Global quality QG from the global quality estimator in the global branch are weighted automatically by utilizing a linear layer, and the final output is the predicted perceptual image quality score QA .

### 3.5 Loseless compression

We researched on efficiency of lossless compression techniques, particularly in the context of low-end devices. These compression methods offer significant advantages in terms of data reduction and computational efficiency, making them particularly suitable for resource-constrained environments.

Low-end devices, such as smartphones, IoT devices, and embedded systems, often have limited storage capacity and processing power. As a result, it is essential to minimize the size of data without sacrificing quality to ensure optimal performance and user experience.

Lossless compression techniques, shown in Fig. 6 unlike their lossy counterparts, enable data to be compressed without any loss of information. This preservation of data integrity is crucial for applications where data accuracy and consistency are paramount.

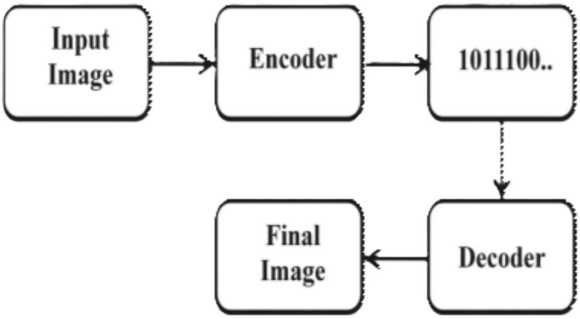

**Fig. 6** Loseless compression

In our investigation, we focus on the utilization of lossless compression techniques, such as HEVC (High Efficiency Video Coding) and JPEG formats with consistency improvements. HEVC, in particular, offers superior compression efficiency compared to previous standards, allowing for significant reductions in file size while maintaining high-quality video content.

Additionally, we explore methods to enhance the consistency and reliability of lossless compression techniques, ensuring consistent performance across different devices and platforms. By optimizing compression algorithms and reducing computational overhead, we aim to improve the efficiency and effectiveness of data compression for low-end devices.

Our research highlights the importance of lossless compression techniques in enabling efficient data storage and transmission on low-end devices. By leveraging these techniques, we can overcome the limitations of resource-constrained environments and enhance the overall user experience on such devices.

## 4 Experiment and result analysis

This section introduces the details of the dataset, performance metrics used to evaluate the model and perormance analysis. Then, the proposed ATVIS-OIQA model is compared against existing competitive IQA metrics on the publicly available OIQA and CVIQD dataset.

### 4.1 Dataset description

The efficacy of the proposed method is evaluated on two benchmark omnidirectional image quality assessment public datasets OIQA [5] and CVIQD [17] dataset.

- **OIQA**: This dataset contains 336 omnidirectional images formed from 16 reference images, which is degraded using JPEG, JPEG2000, Gaussian blur, and Gaussian noise. The MOS values are in the range [1,10].
- **CVIQD**: This database is the largest 360-degree image quality assessment database composed of1 6 uncompressed reference images and 528 lossy compressed images. It considers three compression distortiontypes, i.e., H.265/HEVC,H.264/AVC, and JPEG. TheMOS constituting the target label is normalized to the range [0,100].

During implementation, following the literature in [10], the database is split for training and evaluation, and ten viewpoint images are cropped from each distorted 360-degree image. Figure 6 shows a sample from the dataset with images in decreasing order of quality, in terms of MOS values, from top-left to bottom-right.

### 4.2 Evaluation metrics

Performance analysis is done on the CVIQD [17] and OIQA [5] datasets by the adoption of three standard measures - Root Mean Square Deviation (RMSE), Spearman's Rank Order Correlation Coefficient (SROCC), and Pearson's Linear Correlation Coefficient (PLCC). The metrics have been chosen such that SROCC estimates the monotonicity in prediction, while RMSE and PLCC assess the prediction accuracy. Since SROCC and PLCC assess the likeness or closeness of the target and forecast values, higher values indicate better performance. Conversely, RMSE measures the extent of distance between target and predicted values, indicating that the smaller the value, the higher the accuracy (Fig. 7).

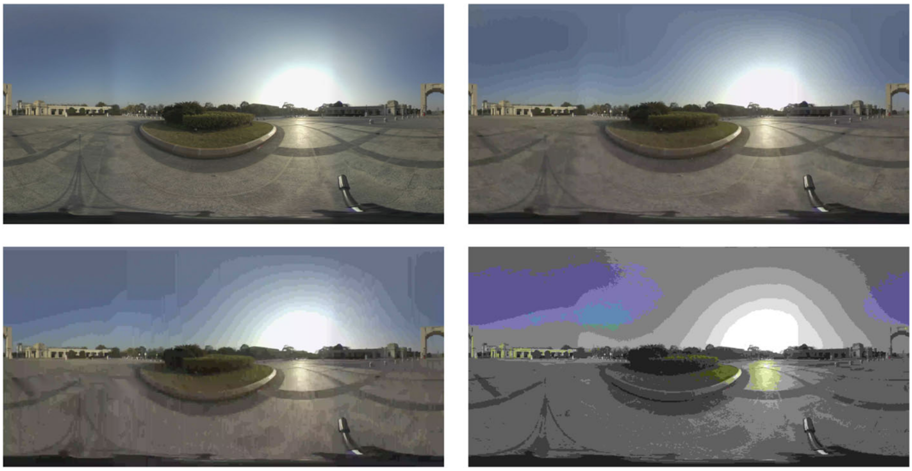

**Fig. 7** Sample from CVIQD dataset

PLCC is formulated as follows:

$$r = \frac{\sum \left(x_{gi} - \bar{x_g}\right)\left(x_{pi} - \bar{x_p}\right)}{\sqrt{\sum \left(x_{gi} - \bar{x_g}\right)^2 \sum \left(x_{pi} - \bar{x_p}\right)^2}} \tag{13}$$

where, $\boldsymbol{r}$ is the correlation coefficient; $x_{gi}$ represents the ground truth MOS scores; $\bar{x_g}$ represents the mean of the ground truth scores; $x_{pi}$ represents predicted MOS and $\bar{x_p}$ represents mean of the predicted scores.

SROCC is given in the equation as below

$$r_s = \rho_{\mathrm{R}(X_g),\mathrm{R}(X_p)} = \frac{\operatorname{cov}(\mathrm{R}(X_g), \mathrm{R}(X_p))}{\sigma_{\mathrm{R}(X_g)}\sigma_{\mathrm{R}(X_p)}} \tag{14}$$

where, $\rho$ represents usual Pearson correlation coefficient, but applied to the rank variables, $\operatorname{cov}(\mathrm{R}(X_g), \mathrm{R}(X_p))$ represents covariance of the rank variables and $\sigma_{\mathrm{R}(X_g)}$ and $\sigma_{\mathrm{R}(X_p)}$ represents standard deviations of the rank variables.

RMSE is given as below

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_{gi} - x_{pi}\right)^2}{N}} \tag{15}$$

where, RMSE represents root-mean-square deviation between $x_{gi}$ actual quality scores and $x_{pi}$ estimated quality scores.

### 4.3 Performance evaluation

This section compares the proposed pipeline for OIQA against several existing state-of-the-art FR and Blind IQA metrics. Additionally, this section investigates the effectiveness of the strategies adopted in the ATVIS-OIQA model and how each strategy impacts performance during quality prediction tasks. Table 1 shows the performance for the proposed ATVIS-OIQA model on OIQA [5] and CVIQD [17] dataset. Fifteen representative IQA methods are chosen for performance comparison. The competitive approaches include five FR

**Table 1** Performance comparison on OIQA and CVIQD dataset

| | | OIQA | | | CVIQD | | |
|---|---|---|---|---|---|---|---|
| | Methods | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE |
| FR 2DIQA | PSNR | 0.5812 | 0.5226 | 1.7005 | 0.7008 | 0.6239 | 9.9599 |
| | SSIM [28] | 0.8718 | 0.8588 | 1.0238 | 0.9002 | 0.8842 | 6.0793 |
| | MS-SSIM [18] | 0.7710 | 0.7379 | 1.3308 | 0.8521 | 0.8222 | 7.3072 |
| | FSIM [24] | 0.9014 | 0.8938 | 0.9047 | 0.9340 | 0.9152 | 4.9864 |
| | DeepQA [7] | 0.9044 | 0.8973 | 0.8914 | 0.9375 | 0.9292 | 4.8574 |
| NR 2DIQA | BRISQUE [13] | 0.8424 | 0.8331 | 1.1261 | 0.8376 | 0.8180 | 7.6271 |
| | BMPRI [12] | 0.6503 | 0.6238 | 1.5874 | 0.7919 | 0.7470 | 8.5258 |
| | DB-CNN [25] | 0.8852 | 0.8653 | 0.9717 | 0.9356 | 0.9308 | 4.9311 |
| FR 360IQA | SP-PSNR [22] | 0.5997 | 0.5399 | 1.6721 | 0.7083 | 0.6449 | 9.8564 |
| | WS_PSNR [17] | 0.5819 | 0.5263 | 1.6994 | 0.6729 | 0.6107 | 10.3283 |
| | CPP-PSNR [23] | 0.5683 | 0.5149 | 1.7193 | 0.6871 | 0.6265 | 10.1448 |
| NR 360IQA | MC360 [16] | 0.9267 | 0.9139 | 0.7854 | 0.9429 | 0.9428 | 4.6506 |
| | VGCN [19] | 0.9584 | 0.9584 | 0.5967 | 0.9597 | 0.9539 | 3.9220 |
| | MFILGN [26] | 0.9695 | 0.9614 | 0.5146 | 0.9751 | 0.9670 | 3.1036 |
| | SCSOIQA [11] | 0.9746 | 0.9673 | 3.2582 | 0.9714 | 0.9669 | 0.5063 |
| | ATVIS | **0.9751** | 0.9654 | 0.6132 | **0.9762** | 0.9668 | 3.0171 |

Bold entries indicate the highest accuracy of the proposed model

2DIQA metrics, i.e., PSNR, SSIM [32], MS-SSIM [22], FSIM [28], and DeepQA [9]; three learning-based NR 2DIQA metrics, i.e., BRISQUE [15], BMPRI [14], and DB-CNN [29]; three FR 360IQA metrics, i.e., S-PSNR [26], WS-PSNR [19], and CPP-PSNR [27]; four viewport-oriented NR 360IQA metrics, i.e., MC360IQA [18], VGCN [23], MFILGN [30] and SCSOIQA [13]. Compared to MC360IQA, MFILGN, and SCSOIQA, VGCN considers interactions between viewports.

PSNR and its variants are substandard to SSIM-based IQA metrics. The PSNR-based metrics reflect only the pixel-level distortion, while SSIM-based ones measure the structural distortion related to HVS. From Table 1, it can be observed that even though the FR DIQA models based on HVS characteristics like SSIM, MS-SSIM, and FSIM that use structural features, fail to achieve satisfactory performance as they primarily focus on low-level features and they are not correlated with human perception. Similarly, the NR-2DIQA models did not take the OIs' properties into account, so they cannot apply to evaluate OIs' quality effectively.

NR 360IQA viewpoint-oriented metrics exhibit discernible supremacy over 2D Full Reference and No Reference IQA metrics. The characteristics of panoramic images, like viewpoint data or spheric depiction, are not accounted for 2D IQA metrics, e.g., sphere representation, viewport information. This confirms the gap between 2DIQA and 360IQA, and points out the importance of viewport-level information for 360IQA. MC360IQA observes a significant enhancement in prediction accuracy because it replaced the conventional ERP format of omnidirectional images with six viewport images.

VGCN significantly outperforms MC360IQA in terms of PLCC, SROCC, and RMSE because it considers keypoint interactions and the regression of local and global forecasts. This shows that it is essential to consider mutual dependency between viewports in OIQA. The proposed method performs better than the state-of-the-art approaches on OIQA and CVIQD datasets. This is mainly owing to two aspects; The ATVIS model utilizes an attention sampling method for viewport selection. The second aspect involves modeling the interactions between viewports through graphs and feature extraction using transformers. The performance comparison results prove that the proposed model employs more representative features than VGCN and can be effectively applied to evaluate OIs' quality with limited data.

### 4.4 Distortion type specific performance comparison

The performance of each distortion type is evaluated on the CVIQD dataset, and the results are shown in Table 2 with top results denoted in bold. From Table 2, the PLCC of the proposed model outperforms other state-of-the-art models in terms of JPEG and HEVC artifacts. The development of image compression technology poses more challenges to IQA. JPEG introduces tonal disruption and blockiness that is less discernible when compared to products of newer encoders. AVC/HEVC presents a more limited quality range than JPEG, making quality prediction extremely difficult.

Table 3 shows the performance of ATVIS-OIQA on various distortion types on OIQA dataset with best performing results marked in bold. According to Table 3, the SROCC IQA metric demonstrates a significant decrease in potential from BLUR to WN to JP2K and further to JPEG. The ATVIS-OIQA model outperforms the state-of-the-art models for JPEG distortions with the highest PLCC and lowest RMSE. ATVIS-OIQA exhibits exciting power in evaluating compressed 360-degree images, especially the ones encoded by JPEG. For JPEG compression, the proposed ATVIS-OIQA delivers the best performance, which is mainly due to the attention-based viewport selection and the spatial relation between the viewports that helps in capturing block distortions.

### 4.5 Visualization of prediction performance

In order to know the effectiveness of the proposed scheme, the actual MOS value is plotted against the predicted scores for the two datasets as shown in Fig. 8 On the top row of the figure, the ATVIS model's overall performance is depicted for the CVIQD2018 database (a) and the OIQA database (b). The performance under different distortions is shown on the bottom row for the CVIQD2018 database (c) and the OIQA database (d). A more focused scatter plot exhibits the promising performance of the ATVIS-OIQA model. The ATVIS-OIQA model outperforms in the CVIQD dataset, and there is a very few deviations for JPEG compared to AVC, and HEVC has more deviations. The scatter plot shows a high correlation between MOS and predicted MOS for BLUR distortion on the OIQA dataset. There are a few more deviations for JPEG and JP2K distortions. The figure clearly demonstrates that the ATVIS method consistently and accurately assesses the quality of omnidirectional images.

### 4.6 Analysis of sampling strategies in viewport detector

The viewport detector module carries out the process of viewport selection to choose the most influential viewports in an omnidirectional image, taking into account the high sensitivity of observers to constructional properties in accordance with the HVS. The proposed method adopts attention sampling to model effective viewport sampling. To validate the efficacy of the proposed sampling strategy in choosing the most beneficial viewports for quality prediction to achieve best performance, two other commonly used sampling strategies have been adopted

**Table 2** Performance comparison for different distortion types on CVIQD dataset

| | JPEG | | | AVC | | | HEVC | | |
|---|---|---|---|---|---|---|---|---|---|
| | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE |
| PSNR | 0.8682 | 0.6982 | 8.0429 | 0.6141 | 0.5802 | 10.5520 | 0.5982 | 0.5762 | 9.4697 |
| SSIM [28] | 0.9822 | 0.9582 | 3.0468 | 0.9303 | 0.9174 | 4.9029 | 0.9436 | 0.9452 | 3.9097 |
| MS-SSIM [18] | 0.9636 | 0.9047 | 4.3355 | 0.7960 | 0.7650 | 8.0924 | 0.8072 | 0.8011 | 6.9693 |
| FSIM [24] | 0.9839 | 0.9639 | 2.8928 | 0.9534 | 0.9439 | 4.0327 | 0.9617 | 0.9532 | 3.2385 |
| DeepQA [7] | 0.9526 | 0.9001 | 4.9290 | 0.9477 | 0.9375 | 4.2683 | 0.9221 | 0.9288 | 4.5694 |
| BRISQUE [13] | 0.9464 | 0.9031 | 5.2442 | 0.7745 | 0.7714 | 8.4573 | 0.7548 | 0.7644 | 7.7455 |
| BMPRI [12] | 0.9874 | 0.9562 | 2.5597 | 0.7161 | 0.6731 | 9.3318 | 0.6154 | 0.6715 | 9.3071 |
| DB-CNN [25] | 0.9779 | 0.9576 | 3.3862 | 0.9564 | 0.9545 | 3.9063 | 0.8646 | 0.8693 | 5.9335 |
| S-PSNR [22] | 0.8661 | 0.7172 | 8.1008 | 0.6307 | 0.6039 | 10.3760 | 0.6514 | 0.6150 | 8.9585 |
| WS-PSNR [17] | 0.8572 | 0.6848 | 8.3465 | 0.5702 | 0.5521 | 10.9841 | 0.5884 | 0.5642 | 9.5473 |
| CPP-PSNR [23] | 0.8585 | 0.7059 | 8.3109 | 0.6137 | 0.5872 | 10.5616 | 0.6160 | 0.5689 | 9.3009 |
| MC360IQA [16] | 0.9698 | 0.9693 | 3.9517 | 0.9487 | 0.9569 | 4.2281 | 0.8976 | 0.9104 | 5.2557 |
| VGCN [19] | 0.9857 | 0.9666 | 2.7310 | 0.9684 | 0.9622 | 3.3328 | 0.9367 | 0.9422 | 4.1329 |
| SCSOIQA [11] | 0.9817 | 0.9580 | – | 0.9697 | 0.9614 | – | 0.9580 | 0.9524 | – |
| MFILGN [26] | 0.9862 | 0.9591 | 2.7904 | 0.9785 | 0.9683 | 2.4998 | 0.9581 | 0.9485 | 3.3950 |
| ATVIS-OIQA | **0.9867** | **0.9766** | 2.6040 | 0.9637 | 0.9439 | 3.5568 | **0.9590** | 0.9321 | 3.4174 |

Bold entries indicate the highest accuracy of the proposed model

**Table 3** Performance comparison for different distortion types on OIQA dataset

| | JPEG | | | JP2K | | | WN | | | BLUR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE |
| PSNR | 0.6941 | 0.7060 | 1.6141 | 0.8632 | 0.7821 | 1.1316 | 0.9547 | 0.9500 | 0.5370 | 0.9282 | 0.7417 | 0.8299 |
| SSIM [28] | 0.9077 | 0.9008 | 0.9406 | 0.9783 | 0.9679 | 0.4643 | 0.8828 | 0.8607 | 0.8474 | 0.9926 | 0.9777 | 0.2358 |
| MS-SSIM [18] | 0.9102 | 0.8937 | 0.9288 | 0.9492 | 0.9250 | 0.7052 | 0.9691 | 0.9571 | 0.4452 | 0.9251 | 0.8990 | 0.7374 |
| FSIM [24] | 0.8938 | 0.8490 | 1.0057 | 0.9699 | 0.9643 | 0.5454 | 0.9170 | 0.8893 | 0.7197 | 0.9914 | 0.9902 | 0.2544 |
| DeepQA [7] | 0.8301 | 0.8150 | 1.2506 | 0.9905 | 0.9893 | 0.3082 | 0.9709 | 0.9857 | 0.4317 | 0.9623 | 0.9473 | 0.5283 |
| BRISQUE [13] | 0.9160 | 0.9392 | 0.8992 | 0.7397 | 0.6750 | 1.5082 | 0.9818 | 0.9750 | 0.3427 | 0.8663 | 0.8508 | 0.9697 |
| BMPRI [12] | 0.9361 | 0.8954 | 0.7886 | 0.8322 | 0.8214 | 1.2428 | 0.9673 | 0.9821 | 0.4572 | 0.5199 | 0.3807 | 1.6584 |
| DB-CNN [25] | 0.8413 | 0.7346 | 1.2118 | 0.9755 | 0.9607 | 0.4935 | 0.9772 | 0.9786 | 0.3832 | 0.9536 | 0.8865 | 0.5875 |
| S-PSNR [22] | 0.6911 | 0.6148 | 1.6205 | 0.9205 | 0.7250 | 0.8757 | 0.9503 | 0.9357 | 0.5620 | 0.8282 | 0.7525 | 0.0910 |
| WS-PSNR [17] | 0.7133 | 0.6792 | 1.5713 | 0.9344 | 0.7500 | 0.9128 | 0.9626 | 0.9500 | 0.4890 | 0.8190 | 0.7668 | 1.1172 |
| CPP-PSNR [23] | 0.6153 | 0.5362 | 1.7693 | 0.8971 | 0.7250 | 0.9904 | 0.9276 | 0.9143 | 0.6739 | 0.7969 | 0.7185 | 1.7280 |
| MC360IQA [16] | 0.9450 | 0.9008 | 0.7272 | 0.9165 | 0.9036 | 0.8966 | 0.9718 | 0.9464 | 0.4251 | 0.9526 | 0.9580 | 0.5907 |
| VGCN [19] | 0.9540 | 0.9294 | 0.6720 | 0.9771 | 0.9464 | 0.4772 | 0.9811 | 0.9750 | 0.3493 | 0.9852 | 0.9651 | 0.3327 |
| ATVIS-OIQA | **0.9546** | 0.9286 | **0.6241** | 0.9778 | 0.9465 | 0.6568 | 0.9802 | 0.9631 | 1.4174 | 0.9818 | 0.9682 | 0.4174 |

Bold entries indicate the highest accuracy of the proposed model

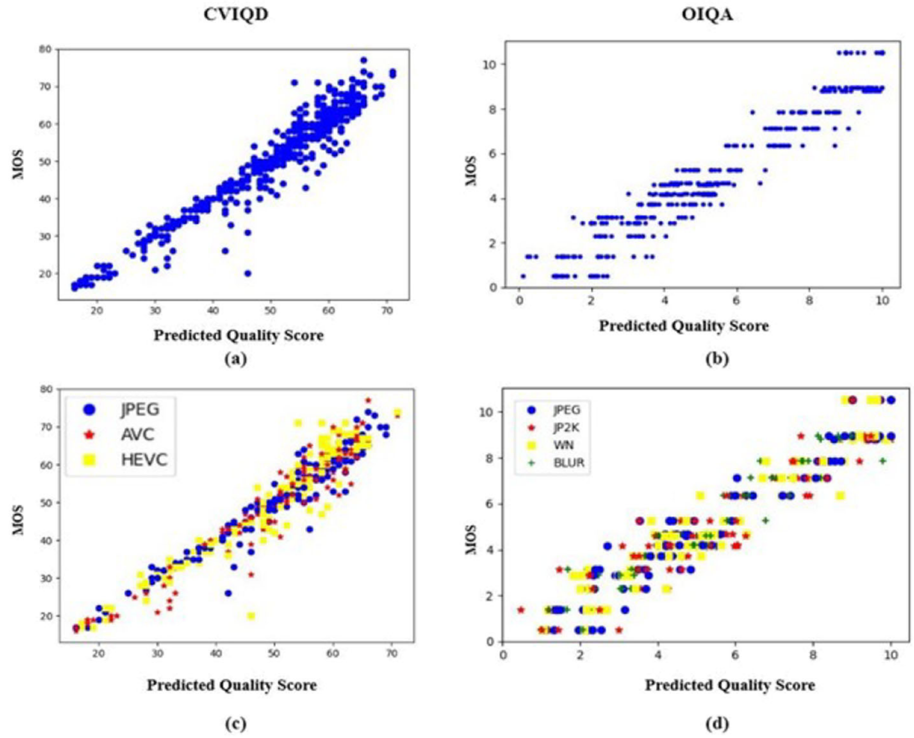

**Fig. 8** Predicted MOS versus actual MOS (a) & (c) CVIQD and (b) & (d) OIQA

specifically uniform sampling and random sampling. Uniform sampling is a strategy where viewports are sampled uniformly at fixed intervals of latitude. Random sampling selects viewports at random during training and evaluation.

The VGCN method for Panaromic Image Quality Assessment appearing in the literature [23] adopted a selective sampling strategy. The method adopts Speeded Up Robust Features (SURF) local feature detector to select salient keypoints and generate a kepypoint map using structural information in accordance with HVS. Further, a heatmap is generated using a 2D Gaussian filter. Finally N viewpoints are selected based on angular distance on the sphere. Table 4 illustrates the performance assessment of various sampling strategies in the Viewport Detector signifying the how the proposed attention sampling is beneficial to predicting quality ratings on CVIQD Database.

**Table 4** Performance of various sampling strategies in viewport detector

| Dataset | OIQA | | | CVIQD | | |
|---|---|---|---|---|---|---|
| Sampling Strategy | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE |
| Random | 0.9341 | 0.9322 | 0.5241 | 0.9443 | 0.9310 | 4.6941 |
| Uniform | 0.9532 | 0.9501 | 0.5811 | 0.9484 | 0.9320 | 4.5215 |
| VGCN [19] | 0.9588 | 0.9532 | 0.5611 | 0.9597 | 0.9539 | 3.9220 |
| ATVIS | **0.9752** | **0.9642** | **0.6132** | **0.9762** | **0.9663** | **3.0171** |

Bold entries indicate the highest accuracy of the proposed model

**Table 5** Effect of various feature extractors in viewport descriptor

| Dataset | OIQA | | | CVIQD | | |
|---|---|---|---|---|---|---|
| Feature Extractors | PLCC | SROCC | RMSE | PLCC | SROCC | RMSE |
| VGG-16 [14] | 0.9499 | 0.9463 | 0.7812 | 0.9504 | 0.9363 | 4.4356 |
| ResNet-18 [4] | 0.9600 | 0.9592 | 0.6611 | 0.9597 | 0.9539 | 3.9220 |
| Transformer [21] | **0.9752** | **0.9642** | **0.6132** | **0.9762** | **0.9663** | **3.0171** |

Bold entries indicate the highest accuracy of the proposed model

### 4.7 Analysis of construction of viewport descriptor

The novelty of this proposed framework for 360-degree IQA lies in the application of Vision Transformers to solve the task of quality assessment. The chosen feature extractor is highly influential on the performance of the Viewport Descriptor module in the overall quality prediction task. Table 5 illustrates the influence of various Viewport Descriptor architectures on effective feature extraction on CVIQD Dataset over all distortion types by comparing their performance under the same conditions. The proposed Viewport Descriptor is competitive with other architectures owing to the exceptional ability of transformers in capturing both positional and content characteristics in images because of positional encodings.

### 4.8 Cross database evaluation

To evaluate the generalization performance, the model trained on one database is used to test on remaining databases as shown in Table 6 The CVIQD and OIQA

Datasets contain only JPEG as the common distortion type. CVIQD contains only distortions related to compression artifacts, and OIQA contains Gaussian noise and blur in addition to compression artifacts. Following the approach of MC360IQA [18], the model trained on the CVIQD dataset is evaluated on JPEG and JP2K compression from the OIQA dataset, and the results are shown in Table 6.

It is observed from Table 6 that the cross-dataset evaluation on JPEG compression is superior, as it is a common distortion type, and the proposed model can effectively capture the known distortions. The ATVIS-OIQA model outperforms the other state-of-the-art models in generalization ability. However, the model finds it hard to generalize on other noise-related distortions; hence the overall generalization capability is significantly low due to unknown distortions.

In conclusion, the proposed ATVIS-OIQA model exhibits a strong generalization capability.

**Table 6** Cross database comparison

| Distortions | JPEG | | JP2K | | ALL | |
|---|---|---|---|---|---|---|
| Method | PLCC | SROCC | PLCC | SROCC | PLCC | SROCC |
| MC360IQA [16] | 0.8898 | 0.8412 | 0.6211 | 0.6221 | 0.7443 | 0.6981 |
| MFILGN [26] | 0.9027 | 0.8889 | 0.7107 | 0.6781 | 0.7885 | 0.7589 |
| ATVIS-OIQA | **0.9042** | **0.8910** | **0.7202** | **0.6820** | **0.7892** | **0.7583** |

Bold entries indicate the highest accuracy of the proposed model

**Table 7** Performance comparison with different number of viewports

| Dataset | OIQA | | CVIQD | |
|---|---|---|---|---|
| No. of Viewports | PLCC | SROCC | PLCC | SROCC |
| 5 | 0.9413 | 0.9264 | 0.9571 | 0.9317 |
| 10 | 0.9751 | 0.9654 | 0.9762 | 0.9668 |
| 15 | 0.9756 | 0.9659 | 0.9764 | 0.9670 |
| 20 | 0.9758 | 0.9662 | 0.9764 | 0.9671 |

### 4.9 Effect of number of viewports

The minimal number of viewports used to predict the omnidirectional image quality is one of the significant parameters during testing. Insufficient viewports will result in incorrect quality estimation due to the inability to extract features from the entire image. In contrast, excessive viewports will significantly increase the processing requirements. As a result, the proposed work investigated the number of viewports to be selected that will improve the performance without increasing the computation. From Table 7, it is observed that the SROCC and PLCC values with more than ten viewports do not change considerably. So, the proposed method chooses ten viewports for training.

### 4.10 Runtime analysis

In Fig. 9 DeepQA and DB-CNN uses the FR 2DIQA learning method which basically process the 2-D images, hence the average runtime of these two methods are less than other SOTA methods. CPP-PSNR uses FR 360IQA learning method whose average inference time is 1.2172 sec per frame. MC360 uses NR 360IQA learning method whose average runtime is 1.101 sec per frame which is less than VGCN, MFILGN, SCSOIQA. VGCN uses NR 360IQA learning method whose average runtime is 1.48 sec which is higher than all the other methods due to number of training parameters. MFILGN uses NR 360IQA learning method whose average runtime is 1.125 sec which is second best in the NR 360IQA methods. SCSOIQA also uses NR 360IQA learning method whose average runtime is 1.2401 sec. Our proposed ATVIS algorithm outperforms due to compatability of algorithm to work with low computing devices with loss less compression technique, ATVIS uses NR 360IQA learning method with average runtime is 0.60121 sec.

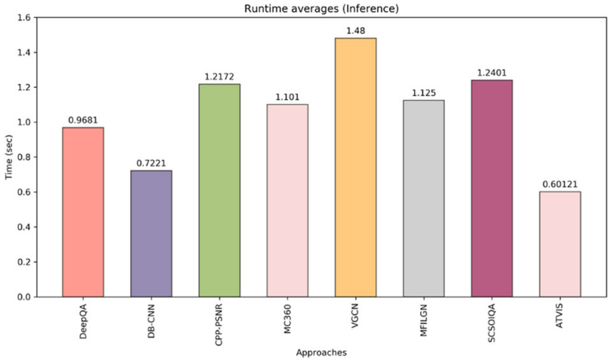

**Fig. 9** Runtime analysis of different viewport descriptor

## 5 Conclusion

This chapter proposed an application of image quality assessment on omnidirectional images for VR applications. The proposed method utilizes attention-based viewport selection with graph convolution for building a spatial viewport graph and vision transformers for feature extraction. It contains a local branch to capture fine-grained details and a global branch to capture real-world distortions.

The local branch is built upon a scorer network which uses the attention map to select the top-k viewports, and graph convolution is employed to model the relation between the viewports. The global branch module fuses the features trained using the vision transformers, which helps in extracting the global features. Finally, the local and global quality features regress to the quality score.

Extensive experiments have shown the effectiveness of the proposed ATVIS-OIQA on the two publicly available benchmark datasets. The future extension of the ATVIS-OIQA model could be to focus on Omnidirectional Video Quality Assessment for capturing complex spatial-temporal interactions between viewports in videos.

**Acknowledgements** This publication is an outcome of the R&D work undertaken in the project under TiHAN faculty fellowship of NMICPS technology innovation hub on autonomous navigation foundation being implemented by department of science & technology national mission on interdisciplinary cyber-physical systems (DST NMICPS) at IIT Hyderabad.

**Data Availability Statement** The codes and network models generated during the current study are available from the corresponding author upon reasonable request.

## Declarations

**Conflict of Interest Statement** The authors declare that they have no conflict of interest to disclose. All the authors have participated in writing this paper, and the work is original and is not published elsewhere. This manuscript has not been submitted to, nor is it under review at, another journal or other publishing venue.

## References

1. Chen C, Han J, Debattista K (2024) Virtual category learning: A semi-supervised learning method for dense prediction with extremely limited labels. IEEE Trans Pattern Anal Mach Intell pp 1–17
2. Chen S (2018) Quality assessment of 360 video view sessions. https://engineering.fb.com/2018/03/09/video-engineering/quality-assessment-of-360-video-view-sessions/
3. Chen S, Zhang Y, Li Y, Chen Z, Wang Z (2018) Spherical structural similarity index for objective omnidirectional video quality assessment. In: 2018 IEEE international conference on multimedia and expo (ICME), pp 1–6. IEEE
4. Duan H, Long Y, Wang S, Zhang H, Willcocks CG, Shao L (2023) Dynamic unary convolution in transformers. IEEE Trans Pattern Anal Mach Intell 45(11):12747–12759
5. Duan H, Zhai G, Min X, Zhu Y, Fang Y, Yang X (2018) Perceptual quality assessment of omnidirectional images. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), pp 1–5
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778
7. Ilse M, Tomczak J, Welling M (2018) Attention-based deep multiple instance learning. In: International conference on machine learning, pages 2127–2136. PMLR
8. Katharopoulos A, Fleuret F (2019) Processing megapixel images with deep attention-sampling models. In: International Conference on Machine Learning, pages 3282–3291. PMLR
9. Kim HG, Lim HT, Ro YM (2019) Deep virtual reality image quality assessment with human perception guider for omnidirectional image. IEEE Trans Circ Syst Video Tech 30(4):917–928

10. Krasula L, Fliegel K, Callet PL, Klíma M (2016) On the accuracy of objective image and video quality models: New methodology for performance evaluation. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp 1–6. IEEE
11. Li C, Xu M, Du X, Wang Z (2018) Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In: Proceedings of the 26th ACM international conference on Multimedia, pp 932–940
12. Li C, Xu M, Jiang L, Zhang S, Tao X (2019) Viewport proposal cnn for 360° video quality assessment. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 10169–10178. IEEE
13. Liu Y, Yin X, Yue G, Zheng Z, Jiang J, He Q, Li X (2023) Blind omnidirectional image quality assessment with representative features and viewport oriented statistical features. J Vis Commun Image Represent 91:103770
14. Min X, Zhai G, Gu K, Liu Y, Yang X (2018) Blind image quality estimation via distortion aggravation. IEEE Trans Broadcast 64(2):508–517
15. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
16. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition
17. Sun W, Gu K, Ma S, Zhu W, Liu N, Zhai G (2018) A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison. In: 2018 IEEE 20th international workshop on multimedia signal processing (MMSP), pp 1–6. IEEE
18. Sun W, Min X, Zhai G, Gu K, Duan H, Ma S (2019) Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment. IEEE J Sel Top Signal Process 14(1):64–77
19. Sun Y, Lu A, Yu L (2017) Weighted-to-spherically-uniform quality evaluation for omnidirectional video. IEEE Signal Process Lett 24(9):1408–1412
20. Wang Z, Li X, Duan H, Su Y, Zhang X, Guan X (2021) Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform. Expert Syst Appl 171:114574
21. Wang Z, Li X, Duan H, Zhang X (2022) A self-supervised residual feature learning model for multifocus image fusion. IEEE Trans Image Process 31:4527–4542
22. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, vol 2, pp 1398–1402. Ieee
23. Xu J, Zhou W, Chen Z (2020) Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. IEEE Trans Circ Syst Video Tech 31(5):1724–1737
24. Xu M, Li C, Chen Z, Wang Z, Guan Z (2018) Assessing visual quality of omnidirectional videos. IEEE Trans Circuits Syst Video Technol 29(12):3516–3530
25. You J, Korhonen J (2021) Transformer for image quality assessment. In: 2021 IEEE international conference on image processing (ICIP), pp 1389–1393. IEEE
26. Yu M, Lakshman H, Girod B (2015) A framework to evaluate omnidirectional video coding schemes. In: 2015 IEEE international symposium on mixed and augmented reality, pp 31–36. IEEE
27. Zakharchenko V, Choi KP, Park JH (2016) Quality metric for spherical panoramic video. In: Optics and Photonics for Information Processing X, vol 9970, pp 57–65. SPIE
28. Zhang L, Zhang L, Mou X, Zhang D (2011) Fsim: A feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386
29. Zhang W, Ma K, Yan J, Deng D, Wang Z (2018) Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans Circ Syst Video Tech 30(1):36–47
30. Zhou W, Xu J, Jiang Q, Chen Z (2021) No-reference quality assessment for 360-degree images by analysis of multi-frequency information and local-global naturalness
31. Zhou Y, Sun Y, Li L, Gu K, Fang Y (2021) Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. IEEE Trans Circ Syst Video Tech 32(4):1767–1777
32. Zhou Y, Yu M, Ma H, Shao H, Jiang G (2018) Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video. In: 2018 14th IEEE International Conference on Signal Processing (ICSP), pp 54–57. IEEE

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.