AN EFFECTIVE FOVEATED 360° IMAGE ASSESSMENT BASED ON GRAPH CONVOLUTION NETWORK

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

08-02-2022 / 10-02-2022

CITATION

Truong, Thu Huong; Do Thu, Ha; Tran, Thi Thanh Huyen; Truong, Thang; Nam, Pham; Nguyen, Thanh; et al. (2022): AN EFFECTIVE FOVEATED 360° IMAGE ASSESSMENT BASED ON GRAPH CONVOLUTION NETWORK. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.19134935.v1>

DOI

[10.36227/techrxiv.19134935.v1](https://doi.org/10.36227/techrxiv.19134935.v1)

AN EFFECTIVE FOVEATED 360° IMAGE ASSESSMENT BASED ON GRAPH CONVOLUTION NETWORK

Truong Thu Huong, *Member, IEEE*, Do Thu Ha, Huyen T. T. Tran, *Member, IEEE*, Ngo Duc Viet, *Student Member, IEEE*, Bui Duy Tien, Nguyen Huu Thanh, *Member, IEEE*, Truong Cong Thang, *Senior Member, IEEE*, and Pham Ngoc Nam, *Member, IEEE*

Abstract—Virtual reality (VR) has been adopted in various fields such as entertainments, education, healthcare, and military, due to its ability to provide immersive experience to users. However, 360° images, one of the main components in VR systems, have bulky sizes, and thus require effective transmitting and rendering solutions. One of the potential solutions is to use foveated technologies that take advantage of the foveation feature of human eyes. Foveated technologies are able to significantly reduce not only data required for transmission, but also computation complexity in rendering. However, understandings on the impact of foveated 360° images on human quality perception are still limited. In this paper, we address this problem by proposing an accurate machine-learning-based quality assessment model for foveated 360° images. The proposed model is proven to outperform the three cutting edge machine-learning-based models and 25 traditional-metric-based models. It is also expected that our model helps to evaluate and improve 360° content streaming and rendering solutions to further reduce data sizes while ensuring user experience. Also, this model could be used as a building block to construct quality assessment methods for 360° videos that is reserved for our future work.

Index Terms—foveated, omnidirectional image, VR, GCN, QoE

I. INTRODUCTION

VIRTUAL reality (VR) has become a cutting-edge technology in the world since its invention in the 1950s; and its applications have been expanding and evolving over the past ten years due to the advancement on both the hardware and software technologies [1]. In contrast to traditional images (i.e., 2D images), VR images are typically recorded with a 360° camera that captures 360° space of a scene [2]. The problem is that omnidirectional contents of VR applications have huge data sizes, and thus require effective transmitting and rendering solutions.

To cope with this problem, one of the most potential solutions is to use foveated technologies that are based on

the foveation feature of human eyes. This feature refers to spatially non-uniform visual acuity due to the heterogeneous distribution of photo receptors in the retina. In foveated technologies, image areas gazed by the retina region of higher photo receptor density have higher quality levels than the outside regions. This allows significantly reducing not only the data size in transmission, but also the computation complexity in rendering.

However, foveated technologies result in spatial quality variations that may cause negative impacts on user quality perception [3]. Therefore, to build a high-quality VR service, it is of utmost importance to understand how human perceives the quality of foveated 360° images. It is expected that answering this question helps figuring out the most effective way to reduce a huge amount of data while ensuring the user experience.

There have been a number of researches that apply machine learning-based methods to successfully assess the quality of 360° images [4], [5]. However, most of these studies focus on uniform quality images. Due to the significant difference between the characteristics of uniform and foveated 360° images, existing models for uniform images can not be used effectively for foveated images. To the best of our knowledge, the quality model, called W-VPSNR, in [6] is the first and the only one dedicated to the quality assessment of foveated 360° images. In their work, the authors use a weighted sum of mean squared errors (MSEs) of image areas corresponding to different retina regions. It is worth noting that, in this model, all pixels in the same area have the same weight, and so the same impact on user quality perception.

In this paper, we aim at building a Graph Convolution Network (GCN)-based quality model that could automatically and effectively learn the contribution of each pixel to the overall perceptual quality of foveated 360° images. For this purpose, each pixel is represented by its quality value and the distance to the foveation point. To the best of our knowledge, this is the first quality model utilizing a deep learning algorithm to assess the quality of foveated 360° images. The proposed assessment method is proved to outperform the three cutting edge machine-learning-based solutions and 25 traditional-metric-based methods.

The rest of the paper is organized as follows. Section II describes state of the art of the VR image assessment methods. In section III, we elaborate our proposed model

Pham Ngoc Nam and Ngo Duc Viet are with the College of Engineering and Computer Science, College of Engineering and Computer Science, Hanoi, Vietnam.

Truong Thu Huong, Do Thu Ha, Ngo Duc Viet, Bui Duy Tien and Nguyen Huu Thanh are with the School of Electrical and Electronics Engineering, Hanoi University of Science and Technology, Hanoi, Vietnam.

Huyen T. T. Tran, Truong Cong Thang are with the Department of Computer and Information Systems, The University of Aizu, Aizuwakamatsu, Japan.

Corresponding author: Pham Ngoc Nam (e-mail: nam.pn@vinuni.edu.vn)

Manuscript received February, 2022; revised xxx, 2022.

based on Graph Convolution Network- FoVGCN - to assess the quality of foveated 360° images. Section IV shows the performance of FoVGCN in comparison with the three state-of-the-art machine-learning-based methods and 25 metric-based schemes. FoVGCN is also evaluated in some cross-validation experiments. Finally, the conclusion and future work are briefly discussed in Section V.

II. RELATED WORK

A. Foveated 360° Content Quality Assessment

In the literature, there have been a lot of studies on foveated contents (i.e., images or videos) [3], [7], [8], [9], [10], [11], [12], [6], and [13]. Among them, there are, however, only some on 360° contents [3], [10], [11], [12], and [6].

In [10], the authors proposed a framework to compare the performance of four subjective quality assessment methods: Double Stimulus Quality Comparison (DSQC), Single Stimulus Absolute Category Rating (ACR), Ascending Method (AM), and Descending Method (DM). By the analysis of Quality of Experience (QoE) scores of foveated 360° images, it was found that the DSQC method obtains the highest consistency, but requires more judgments and time to converge to the general consensus. Meanwhile, ACR was found to be the most efficient method. In [12], the authors focused on the subjective comparison between 2D and 3D foveated 360° videos in terms of users' perceptual quality. The results showed that, the perceptual quality of 2D videos was more affected by the quality of the image area corresponding to the peripheral region. Meanwhile, for 3D videos, the perceptual quality was largely impacted by the image area quality associated with the fovea region. Also, based on the results, a performance evaluation of 12 objective quality metrics were conducted. Foveated Wavelet Quality Index (FWQI) was found to be the most effective model for both 2D and 3D foveated 360° videos. In [11], the key question the authors focused on was how to spatially reduce data size without noticeable perceptual quality degradation by taking advantage of the foveation feature. In particular, a subjective quality assessment for foveated 360° images was conducted taking into account three regions of the human retina, i.e., the central vision area with one-side eccentricity $\theta \in [0^\circ, 9^\circ]$, the near peripheral area with $\theta \in (9^\circ, 30^\circ]$, and the far peripheral area. In this experiment, the image quality corresponding to each region was reduced step by step until the participants notice a perceptual difference. By utilizing encoding parameters (i.e., quantization parameters and resolutions) that had been recorded, the authors proposed a rendering solution that is indicated to be able to significantly improve rendering throughput about $10\times$ without perceptual loss, in comparison to the traditional solution of uniform quality. [3] is the first study that could quantify the impacts of different retina regions on user quality perception. In particular, the authors performed a subjective quality assessment of foveated 360° images. Through experimental results, it is quantitatively shown that image areas corresponding to the fovea and parafovea regions are extremely important for quality perception, while the impacts of the other zones are small. Besides, a performance evaluation of twenty-five

objective quality metrics were conducted. It turned out that all of them, even the fovea quality metrics, are not effective for the quality assessment of foveated 360° images.

Based on the subjective dataset in [3], the authors in [6] proposed a simple quality model called W-VPSNR. W-VPSNR is the first and only quality model for quality assessment of foveated 360° images. Instead of considering the entire image, this model predicts the quality of a small part, called viewport, that users actually observe due to the limited Field of View (FoV) of human eyes. To consider the foveation feature, a weighted sum of mean squared errors (MSEs) of image areas corresponding to different retina regions was used to predict the quality of foveated 360° images. It is worth noting that, in this model, all pixels in the same area have the same weight, and so the same impact on user quality perception.

B. Quality Models for Omnidirectional Image/Video Contents

In general, in the domain of objective Image Quality Assessment, IQA is categorized into three methods: no-reference (NR-IQA), reduced-reference (RR-IQA), and full-reference image quality assessment (FR-IQA), depending on their degree of dependence on the reference image. Full-reference approaches compare a distorted image to an entire reference image, while reduced-reference approaches just need a portion of the reference image's information. And no-reference (or blind) approaches only work with distorted images received at the client's side [14].

Many techniques have been proposed for evaluating image quality. Traditional technical image quality metrics that are leveraged commonly in FR-IQA and RR-IQA include Mean Squared Error (MSE), Frequency Mean Square Error (FMSE) [15], Universal Quality Index (UQI) [16], Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [17], Foveated Peak Signal-to-Noise Ratio (FPSNR) [8], Multi-scale SSIM (MS-SSIM) [18], Information content Weighted SSIM (IW-SSIM), Noise Quality Measure (NQM) [19], Visual Information Fidelity (VIF) [20] Visual Information Fidelity in the pixel domain (VIFp) [21], Weight Signal-to-Noise Ratio (WSNR), Feature similarity index measure (FSIM), Feature similarity measure (FSIMc) for color image, Foveal feature similarity measure (F-SSIM) [22], Perceptual Similarity (PSIM) [23], Analysis of Distortion Distribution-based (ADD-SSIM) [24], Foveal Structural Similarity [25], and Foveated Wavelet image Quality Index (FWQI) [7], Generic Statistical Information Model (GSIM), Riesz Transforms based Feature Similarity (RFSIM) [26], Information content Weighted PSNR (IW-PSNR) [27], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [28], No-reference Free Energy-Based Robust Metric (NFERM) [29], Spectral Residual based Similarity (SR-SIM) [30], and Weighted Viewport PSNR (W-VPSNR) [3]. Therefore in this paper, we will evaluate our solution's performance in terms of 25 metrics: MSE, FMSE, UQI, PSNR, FPSNR, SSIM, MS-SSIM, IW-SSIM, NQM, VIF, VIFp, WSNR, FSIM, FSIMc, F-SSIM, PSIM, ADD-SSIM, FWQI, GSIM, RFSIM, IW-PSNR, BRISQUE, NFERM, SR-SIM, and W-PSNR in order to have an insight into our proposed solution from variety of angles.

In most cases, machine learning based methods have been found to perform better than traditional metric based methods. There have been a small number of studies that applied machine learning in the domain of full-reference uniform image quality assessment, for instance [31], [32], and [33]. In [31], the authors proposed a new framework that applies a deep neural network to study the human visual sensitivity (HSV), based on distorted images, a subjective score and an objective error map (DeepQA model) or without an objective error map (DeepQA-s model) in an uniform image quality dataset. In [32], the author considered the important role of multiple viewports related to the image inside field of view (FoV). Those viewports are extracted by viewport sampling with inputs being reference images (i.e., original images at the server side) and distorted images (i.e., received images at the client side). Their proposed stereoscopic omnidirectional image quality assessment (SOIQA) model then learnt those viewport features using a deep neural network and support vector regression (SVR). Machine learning techniques have been applied efficiently in [31] and [32] to learn the characteristics of uniform immersive image quality in terms of full-reference quality approach. However, the characteristics of uniform 360° image quality are vastly different from those of foveated 360° image quality. Therefore, the research direction of assessment methods that work effectively for foveated (i.e., non-uniform) 360° image is still an open issue.

In the direction of NR-IQA methods, machine-learning-based approaches have been utilized quite commonly. Zhang *et al* [34] proposed a deep bi-linear model for non-reference image quality assessment (BIQA) to deal with synthetic distortions and authentic distortions in images. Afterward, Xu *et al* [5] developed a novel Viewport oriented Graph Convolution Network (VGCN) that concatenates a global branch based on Zhang's work [34], which predicts the global quality score by handling the synthetic and authentic distortions, and a local branch that learns the interactions among different viewports by using graph convolution network to get the overall image quality. Kim *et al* in [35], first, extracted features of distorted images to predict quality scores, then proposed a user perception guidance by using adversarial learning to enhance the prediction performance. Sun *et al* [4] introduced a multi-channel convolution neural network (i.e. MC360IQA), in which the overall quality is predicted using six simultaneous ResNet-34s that extract features from six created viewports. In [36], the authors introduced meta-learning based image quality assessment (Meta IQA) method that successfully deals with different types of distortions in image quality assessment. Machine learning approaches have been proved their effectiveness in terms of studying the characteristics of uniform distorted images to derive the image quality. However, to deal with a foveated image dataset, for example in [3], we should take into account the quality distribution in different regions as well as the priority of human attraction factors, both of which have a huge impact on the overall user experience. In our point of view, this is the main reason why all proposed NR-IQA machine-learning-based approaches have only been successful in dealing with uniform image quality assessment up to now.

Besides many studies on image quality assessment (IQA),

video quality assessment (VQA) has also been investigated by several research groups. In [37], the authors split a video into a sequence of frames by which the video quality is calculated based on the MOS of each frame. Then, they used a viewport-based convolutional neural network (V-CNN) consisting of viewport proposals to extract the viewports of the video. The overall video quality is then revised by processing viewports with the predicted saliency map. Instead of handling the entire video, Qian *et al* [38] created a video bag by grouping a specific number of video blocks, which are extracted features to predict the overall quality of the video. However, constructing a 360° video dataset requires a lot of time and resources, and questions such as “how long a video should be to be efficient for evaluating the quality”, “in what proper way we should change quality in video adaptively”, etc. are still pending. Therefore, IQA is still an effective approach to deal with the field of both image and video quality assessments.

In this work, we focus on image quality assessment of spatially foveated images for two reasons. Firstly, 360° images are still a critical topic with a wide range of applications and play an important role in evaluating immersive video quality. Secondly, almost all up-to-date IQA methods have got only modest performance so that looking for a new scheme to improve it is necessary.

Therefore, we propose a Foveated-Graph-Convolution-Network-based 360° Image Processing Method - FoVGCN - which will be compared with 25 different traditional-metric-based methods and 3 other machine-learning based models found in [5], [35], [4]. Although, FoVGCN is designed to work efficiently for foveated image quality, its performance is also cross validated with uniform datasets to evaluate its application generality over heterogeneous cases in reality.

III. DESIGN OF THE FOVEATED GRAPH CONVOLUTION NETWORK BASED 360° IMAGE PROCESSING METHOD

In this section, we will describe the detailed technical design of our proposed FoVGCN model (i.e., Foveated Graph Convolution Network based 360° Image Processing Method) for assessing the retina-related zones of omnidirectional image quality. The processing phase and the FoVGCN model as well as the training and testing implementation will be elaborated in the following subsections.

Our proposed method FoVGCN consists of two main blocks which are the preprocessor and the graph convolution network block as illustrated in Figure 1. The preprocessor plays an important role in creating an error map and an attention weight matrix which represents the quality changing in different zones of an image and the priority of human attention, respectively. In the preprocessor block, reference and distorted images are fed into the error map creator block to create an error map in the form of a graph matrix. The attention weight matrix block takes the shape of a viewport as the input to create a distribution matrix. After being processed, both the error map and the attention weight matrix are fed into the graph convolution network block to predict the overall quality score of the image.

In the next subsections, the two main blocks and parameter settings of the FoVGCN model will be described in detail.

Fig. 1: FoVGCN Model Operation

A. Preprocessor

1) *Error Map Processing*: In this section, we describe the detail of how to create an error map as the first step of the Preprocessor to generate the desired input for the GCN model in the latter phase. Intuitively, an Error Map represents spatial quality changing in a distorted image when comparing the image with a reference one. Besides, it also serves as the graph matrix input of the graph convolution network [39].

Fig. 2: Errormap Creator

According to human retina characteristics, the quality of omnidirectional images is divided into five regions (Fovea, Parafovea, Perifovea, Near periphery, Far periphery) [3]. It means that the perceived quality of each pixel in a distorted image changes from the center region to the outer regions. Our goal is to intuitively represent the changing of the quality through each zone by using the so-called Error Map. There are some pixel-wise metrics such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE) that can be used to construct the error map. However, we use the normalized log difference function (Eq.1) following [31] for better correlation with the perceived quality of viewers:

$$ErrorMap = \frac{\log(1/((I_r - I_d)^2 + \alpha))}{\log(1/\alpha)}, \quad (1)$$

where $\alpha = \varepsilon/255^2$ is a constant and $\varepsilon = 0.1$, according to [31].

The detail of the Error map Creator is shown in Figure 2. First, the reference and distorted images are converted to gray-scale images. Let I_r and I_d be the values of each pixel of the reference and distorted gray images, respectively. Then, error maps are created following Eq.1, as illustrated in Figure 3. As can be seen in Figure 4, the distorted viewport 1 has a quite good quality and distorted viewport 2 has a lower quality, while the reference viewport has the highest quality.

(a) Image 1

(b) Image 2

Fig. 3: Error Maps of two example distorted images

In the Errormap Transform block, the error map is first divided into four quadrants, then each of those is rotated and sorted as four consecutive quadrants following the rule in Figure 5.

2) *Attention Weight Matrix Processing*: The concept of the attention weight matrix is used to describe the priority of human attraction in each region of an immersive image. It can be understood that pixels in the central zones are considered to have higher weights than pixels farther away from the

Fig. 4: Comparison of two Error Maps

Fig. 5: Errormap Transform

center zone. Therefore, our attention weight matrix is created in such a way that it has high density representing high priority in the center while decreasing to lower and lower density for further zones from the center. This degradation follows two proposed distributions: Linear degradation distribution and Gaussian degradation distribution. Later in the Evaluation section, we will present the solution results based on each type of distribution to have more insights into the performance.

The detail of the Attention Weight matrix processing block is illustrated in Figure 6. First, an attention weight matrix is created in the form of a chosen degradation distribution depending on the viewport shape. Then, we leverage the first quadrant of the viewport attention weight matrix as the input of our GCN network.

Fig. 6: Attention Weight Matrix Processing

Linear degradation distribution

We propose a formula that describes the linear degradation

from the center pixel to the edge pixels in Eq. (2). The value of the matrix corresponding to each pixel is inversely proportional to the distance from that pixel to the central pixel and decreases gradually and constantly from the center to the periphery.

$$A = 1 - \frac{dis(x, x_0)}{dis_{max}} + \delta, \quad (2)$$

where

- A is the attention weight matrix of the error map, in range of $[0,1]$;
- $dis(x, x_0)$ is the distance between pixel x and center pixel x_0 ;
- dis_{max} is the maximal distance between point x and x_0 ;
- Threshold δ is applied to avoid the zero value, and is set equal to 0.0001.

This attention weight matrix idea is illustrated in Figure 7.

Fig. 7: Attention Weight matrix idea

Fig. 8: Attention weight matrix in Linear distribution

Applying Eq.(2) to calculate distributed weights in pixel $B[300, 300]$ and $C[700,700]$, we get: $A_B[300, 300] = 0.833$, $A_C[700, 700] = 0.056$. It is obviously right that B is near the center point ($A_{x_0} = 1.0001$) while C is near the corner point ($A_{corner} = 0.0001$). Eq. (2) is visualized in Figure 8.

Gaussian degradation distribution

As a second way to investigate, the Gaussian distribution is leveraged to describe the degradation of human attraction from the center to the outer region. This idea follows the concept of the human retina [3] in which user perception tends to be affected by the quality of fovea and parafovea zones from the center to the outer zone of an image. Because human eyes concentrate significantly on the fovea and parafovea zones

of the human retina which is a small region in the viewport [3]. And we need a distribution that describes a better human perspective. Therefore, we apply formula (3), following the Gaussian distribution to construct the attention weight matrix:

$$A = f(x) = e^{-\frac{1}{2}(\frac{x}{\sigma})^2}, \quad (3)$$

where σ is the standard deviation of x and A is in the range of $[0,1]$.

In fact, changing the value of σ results in the different degrees of central concentration in an image. Figure 15 illustrates the Gaussian distribution with different σ values where greater σ values correspond to larger focal regions. Therefore, in our experiment presented later in this paper, the performance of the FoVGCN model will be shown to study the impact of the sigma value σ corresponding to the human attention.

B. Graph convolution network

As the last phase of the whole FoVGCN assessment process, the Graph Convolution Network is used since it can efficiently learn graph-structured data. With GCN, an image is considered as a graph, in which each vertex represents a pixel and each edge represents the relationship between two pixels.

We design two inputs for GCN: an error map and an attention weight matrix. The error map is relevant to the changing of image spatial quality, and the attention weight matrix plays the role of a kernel in the convolution neural network (CNN).

In our work, the Graph Convolution Network is applied to extract features and study the spatial quality changes relevant to the attention weight matrix in the foveated immersive image dataset used in [3].

The error map, which is designed as a graph with a size of $(720, 720)$, is used to represent the quality changes between different pixels. This size is heuristically chosen based on the size of the attention weight matrix. Using a large weight matrix leads to a computational overload in GCN. In our design and experiment, an attention weight matrix of $(720, 720)$ is proposed in both the linear and Gaussian distribution forms, representing the priorities of human attention on focal regions.

The forward propagation of the graph convolution layer is applied using the rule provided in [39] to eliminate the vanishing and exploding gradient in back-propagation, as follows:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}) \quad (4)$$

$$\tilde{A} = A + I_N \quad (5)$$

where

- $H^{(l)}$ is the activation matrix in l^{th} layer;
- $W^{(l)}$ is the trainable weight matrix in layer l ;
- $\tilde{D}$ is the degree matrix which can be calculated by $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$;
- $\tilde{A}$ is the adjacency matrix which is the sum of attention weight matrix A and identity matrix I_N ;

(a) $\sigma = 0.1$

(b) $\sigma = 0.2$

(c) $\sigma = 0.3$

(d) $\sigma = 0.4$

Fig. 9: Attention weight matrix in Gaussian distribution

- $\sigma()$ is a non-linear activation function such as the Softmax function, ReLU function, Softplus function, etc. In our work, we use the Softplus function for more stabilization and performance to deep neural networks .

First, the adjacency matrix is symmetrically normalized by $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$, before being multiplied with the attention weight matrix $W^{(l)}$ and the output of the prior layer $H^{(l)}$. Then, the output layer is obtained after going through the activation function.

TABLE I: The configuration of the FoVGCN model

No.	Layer Name	Attention Weight Matrix Size	Input Size	Output Size
1	GCN layer 1	(360, 360)	(360, 1440)	(360, 720)
2	GCN layer 2	(360, 360)	(360, 720)	(360, 360)
3	GCN layer 3	(360, 360)	(360, 360)	(360, 180)
4	GCN layer 4	(360, 360)	(360, 180)	(360, 90)
5	GCN layer 5	(360, 360)	(360, 90)	(360, 45)
6	GCN layer 6	(360, 360)	(360, 45)	(360, 1)
7	Fully connected layer		(360, 1)	(1, 1)

C. Model Parameter Settings

To create a down-sampled image with a size of (720, 720), we use the zero-padding method to form a square matrix of (1440, 1440) in order to avoid distortion when the image is down-scaled. This square matrix is then down-sampled to the size of (720, 720) to reduce the computational cost. Next, an

Fig. 10: The details of each GCN block

error map is created from those viewports and the attention weight matrix is fed through the six blocks of the graph convolution layers which is shown in Figure 10. Each block contains a graph convolution layer, batch normalization, and a Softplus activation function as described in Eq. (6):

$$f(x) = \ln(1 + e^x) \quad (6)$$

Batch normalization is a comprehensive method for parameterizing virtually any deep neural network, and the re-parameterization significantly reduces the issue of planning updates across multiple layers. Finally, a fully-connected layer extracts the final predicted score.

In addition, following Eq.(4), a degree matrix $\tilde{D}$ is inversely calculated. Since this process takes huge computation and resources, the inputs such as attention weight matrix and error map need to be pre-processed before being fed into the FoVGCN model in order to reduce the computational complexity, thereby reducing running time. Instead of feeding the entire error map and attention weight matrix A , we process them as described in Section III-A1 and Section III-A2.

The detail of all parameters used in the FoVGCN model is presented in Table I. In the first layer of the FoVGCN model, the input size is the size of the input image.

IV. EXPERIMENTS AND RESULTS

To evaluate the performance of FoVGCN, we apply FoVGCN to the foveated dataset of Retina-related Zones Omnidirectional images, which are created to investigate the

foveation feature of human eyes. In the following sections, we will elaborate the pre-processing of those datasets and performance evaluation.

A. Dataset preparation

1) *Foveated image dataset*: Our proposed FoVGCN solution is designed to work effectively for the foveated dataset, in which the quality changes in different zones correspond to the five regions of the human retina.

FoVGCN is trained and tested with the foveated immersive image dataset of [3] that contains 16 reference and 512 distorted viewport-extracted images. These 16 reference images are retrieved from various scenes such as indoor, large conference room, containing human faces, natural landscape, etc.

However, this foveated dataset has a limited number of samples to achieve a good training performance. So, to enhance the performance and accuracy of our proposed method, we apply a data augmentation technique by flipping the viewports twice from the left to the right and from the bottom to the top, without destroying the characteristics of the foveated dataset. As the result, this technique triples the amount of data, thus helping to achieve a better training performance.

2) *Cross-validation datasets*: In order to investigate the comprehensiveness of our proposed method, two uniform image datasets - CVIQ [4] and OIQA [40] - are experimented to validate the FoVGCN model:

- The CVIQ dataset consists of 524 distorted images which were created from 16 source images. Those images are encoded by three standards: JPEG, H.264/AVC, and H.265/HEVC.
- The OIQA dataset includes 320 distorted images created from 16 reference images by four distortion types, namely JPEG compression, JPEG2000 compression, Gaussian blur, and Gaussian noise.

B. Performance evaluation

1) *Experimental settings*: To evaluate the performance of the FoVGCN model, we use common performance measures such as Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), and Root mean square error (RMSE). The experiments are conducted in various aspects to have a better understanding of the impacts of different factors.

As mentioned, we use the foveated immersive image dataset [3] that is constructed from 16 source images. It should be noted that we focus on dealing only with the viewports of full foveated immersive images in order to reflect what viewers are actually watching. More specifically, we select 2 source images among the source images. The distorted images corresponding to the selected images are used for testing while the remaining distorted images are used for training. In total, we have $\binom{16}{2} = 120$ different cases to cover all testing situations. The aforementioned performance metrics are calculated by averaging the results of those 120 cases.

Note that in our experiments, the learning model is found out to work efficiently with the learning rate set at 10^{-4}

as the model could converge after 200 epochs. The training and testing phases are executed in Google colab pro (Intel(R) Xeon(R) CPU @ 2.30GHz, Tesla P100-PCIE-16GB GPU).

In the evaluation process, firstly, we analyze the performance of our FoVGCN model on the foveated image dataset in two cases of the attention weight matrix, namely (1) the Linear degradation distribution and (2) the Gaussian degradation distribution. Secondly, we change the sigma coefficient in the Gaussian attention weight matrix in order to study the impact of σ on the performance of the FoVGCN method. Thirdly, FoVGCN is compared with 25 traditional-metric-based methods and three machine-learning-based image assessment approaches. Finally, we conduct some cross-validation experiments to investigate how FoVGCN and other fovea-quality-metrics solutions would work with the two uniform datasets.

2) *Impact of Gaussian distribution versus Linear distribution*: In this evaluation, we want to analyze how the selection of the Gaussian distribution or linear distribution for the attention weight matrix could impact the final performance of FoVGCN. This evaluation helps us to have deeper insight into what distribution should be selected for better performance of FoVGCN.

Figure 11 illustrates PLCC, SROCC, and RMSE measured for our FoVGCN method using two different attention weight matrices: (1) with Gaussian distribution, and (2) with Linear distribution. It can be obviously seen that both the two distribution attention weight matrices result in high SROCC and PLCC (i.e., over 0.85). Meanwhile, RMSE values are shown to be low, which are under 0.5 in almost 120 cases.

In more detail, in the case where the weight matrix is processed with the Gaussian distribution, the highest values of PLCC and SROCC are 0.994 and 0.991, respectively. The average values calculated for the 120 cases with the Gaussian distribution are SROCC = 0.983, PLCC = 0.967, and RMSE = 0.084. Moreover, the average values calculated for the 120 cases with the linear distribution are SROCC = 0.938, PLCC = 0.941, and RMSE = 0.056. Besides, SROCC stays in the range of [0.853, 0.982], PLCC stays in the range of [0.835, 0.982], whilst RMSE is under 0.4.

In conclusion, the FoVGCN model with the Gaussian weight matrix outperforms the linear distribution weight matrix. Moreover, it is proven that our method achieves a stable and significantly good performance in all experiments.

3) *Impact of different sigma coefficients*: As mentioned in Section III-A2, we know that any change in the sigma coefficient σ of the Gaussian distribution, which is used to construct the weight matrix, will result in different degrees of the central concentration, relating to different human attention in each zone. In order to figure out the best performance related to the σ coefficient, Figure 12 illustrates SROCC, PLCC and RMSE with different σ values. It can be seen that, in general, SROCC, PLCC, and RMSE increase as σ is decreased. When σ is less than 0.1, the accuracy values saturate while RMSE quickly increases. So, we set $\sigma=0.1$ to have a good balance among the three values SROCC, PLCC, and RMSE.

(a) PLCC

(b) SROCC

(c) RMSE

Fig. 11: SROCC, PLCC and RMSE with the Gaussian distributed and Linear distributed weight matrices

Fig. 12: The performance of FoVGCN over different sigma coefficients

C. FoVGCN versus other existing solutions

In this section, we compare FoVGCN with other existing solutions, including 25 analytical metrics and 3

(a) PLCC and SROCC

(b) RMSE

Fig. 13: Performance of 25 analytical metrics vs. FoVGCN

machine-learning-based methods. The 25 analytical metrics include MSE, FMSE, UQI, PSNR, FPSNR, SSIM, MS-SSIM, IW-SSIM, NQM, VIF, VIFp, WSNR, FSIM, FSIMc, F-SSIM, PSIM, ADD-SSIM, FWQI, GSIM, RFSIM, IW-PSNR, BRISQUE, NFERM, SR-SIM, and W-PSNR. The results of those 25 metrics are calculated by averaging the values of the 120 cases. The 3 machine-learning-based methods are DeepQA [31], MIC360IQA [4], and VGCN [5]. To have a fair comparison, all machine-learning-based methods are re-trained with the above foveated 360° image dataset, in the

same manner as the proposed FoVGCN.

Figure 13 shows that FoVGCN outperforms the 25 analytical metrics. As it can be seen, FoVGCN achieves much higher accuracy, with SROCC = 0.983 and PLCC = 0.967, while the other methods can reach to approximately 0.9 at most. In addition, FoVGCN achieves much lower RMSE (i.e., 0.084) compared to the other existing schemes.

Without loss of generality, we present specific results for one single case (i.e., case 5 using source images I1 and I6) to compare FoVGCN with 3 other machine-learning-based meth-

(b) RMSE

Fig. 14: Performance of 3 current Machine-learning-based approaches vs. FoVGCN

ods, as shown in Figure 14. It can be seen that, DeepQA [31] fails to evaluate the foveated image quality efficiently, while the VGCN and MIC360IQA models have modest performance, namely SROCC = 0.748, PLCC = 0.730, RMSE = 13.237 with MIC360IQA, and SROCC = 0.512, PLCC = 0.335, RMSE = 16.102 for VGCN. Meanwhile, FoVGCN achieves SROCC = 0.944, PLCC = 0.977, RMSE = 0.069, which are much better than the three mentioned machine-learning-based methods.

In addition, the scatter diagrams of the ground truth and the predicted MOSs of all metrics and models are shown in Figure 15. We can see that, among the analytical metrics, only the FMSE, FPSNR, and WVPSNR have reasonable relationships between the actual MOSs and predicted MOSs. This is because these metrics are specifically designed with foveation feature. As for the machine-learning-based methods, both VGCN and DeepQA provide predicted MOS values in a very narrow range. Especially, DeepQA results in very low predicted MOSs (almost zero). The scatter diagrams of FoVGCN confirm that this model (with either the Gaussian or linear degradation weight matrices) can describe exactly the trend of MOS scores.

D. Cross Validation Experiments

In the previous section, FoVGCN has been shown to be efficient in quality assessment for the foveated images. In order

TABLE II: The performance of FoVGCN over CVIQ dataset

	SROCC	PLCC	RMSE
WVPSNR	0.807	0.802	8.404
VPSNR	0.770	0.773	8.942
FVPSNR	0.804	0.797	8.512
FWPSNR	0.488	0.509	12.124
WSNR	0.670	0.671	10.440
FoVGCN	0.837	0.820	0.037

TABLE III: The performance of FoVGCN over OIQA dataset

	SROCC	PLCC	RMSE
WVPSNR	0.675	0.674	10.631
VPSNR	0.681	0.679	10.567
FVPSNR	0.670	0.670	10.684
FWPSNR	-0.477	-0.474	17.062
WSNR	0.638	0.630	11.175
FoVGCN	0.693	0.661	0.137

to see how FoVGCN will work with different types of content, we investigate the performance of FoVGCN with two other uniform image datasets: CVIQ [4] and OIQA [40]. In the cross-validation experiment, we use the foveated image dataset for training, while the CVIQ dataset and OIQA dataset are employed for testing. In this part, other foveal metrics are used for comparison.

The results are illustrated in Table II and Table III. We can see that, with the CVIQ dataset, FoVGCN achieves better performance in terms of SROCC, PLCC, and RMSE, which are 0.837, 0.820, and 0.037 respectively. With the OIQA dataset, FoVGCN achieves comparable accuracy with other metrics; however, its RMSE (0.137) is much smaller than others. That means FoVGCN is more stable than other foveal metrics.

V. CONCLUSION

In this paper, we have proposed FoVGCN as an efficient assessment model for foveated 360° images. The model uses Graph Convolutional Network to represent the complex relationships among different locations of an immersive image. It is expected that the proposed FoVGCN model will be an effective and reliable method for researchers to evaluate coding and rendering solutions of foveated image/video field. In the future work, we will employ this model to improve VR video streaming adaptation techniques to ensure good perceived quality for viewers.

ACKNOWLEDGMENT

This work was funded by Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2020.DA03.

Fig. 15: Scatter diagrams of the actual MOSs versus the predicted MOS values for all methods

REFERENCES

- [1] Wikipedia contributors, "Virtual reality — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Virtual_reality&oldid=1049963342, 2021, [Online; accessed 19-October-2021].
- [2] Chen, Meixu and Jin, Yize and Goodall, Todd and Yu, Xiangxu and Bovik, Alan Conrad, "Study of 3D Virtual Reality Picture Quality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 89–102, 2020.
- [3] H. T. T. Tran, D. V. Nguyen, N. P. Ngoc, T. H. Hoang, T. T. Huong, and T. C. Thang, "Impacts of retina-related zones on quality perception of omnidirectional image," *IEEE Access*, vol. 7, pp. 166 997–167 009, 2019.
- [4] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2020.
- [5] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1724–1737, 2021.
- [6] H. T. T. Tran, T. H. Hoang, P. N. Minh, N. P. Ngoc, and T. C. Thang, "A perception-based quality metric for omnidirectional images," in *2019 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia)*, 2019, pp. 151–152.
- [7] Z. Wang, A. C. Bovik, L. Lu, and J. L. Koulouheris, "Foveated wavelet image quality index," in *Applications of Digital Image Processing XXIV*, vol. 4472. International Society for Optics and Photonics, 2001, pp. 42–52.
- [8] S. Lee, M. Pattichis, and A. Bovik, "Foveated video quality assessment," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 129–132, 2002.
- [9] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of foveated video coding using audio-visual focus of attention," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1322–1331, 2011.
- [10] C.-F. Hsu, A. Chen, C.-H. Hsu, C.-Y. Huang, C.-L. Lei, and K.-T. Chen, "Is foveated rendering perceivable in virtual reality? exploring the efficiency and consistency of quality assessment methods," in *Proceedings of the 25th ACM international conference on multimedia*, 2017, pp. 55–63.
- [11] P. Guo, Q. Shen, Z. Ma, D. J. Brady, and Y. Wang, "Perceptual quality assessment of immersive images considering peripheral vision impact," *arXiv preprint arXiv:1802.09065*, 2018.
- [12] Y. Jin, M. Chen, T. Goodall, A. Patney, and A. C. Bovik, "Subjective and objective quality assessment of 2d and 3d foveated video compression in virtual reality," *IEEE Transactions on Image Processing*, vol. 30, pp. 5905–5919, 2021.
- [13] L. Surace, M. Wernikowski, O. Tursun, K. Myszkowski, R. Mantiuk, and P. Didyk, "Learning foveated reconstruction to preserve perceived image statistics," *arXiv preprint arXiv:2108.03499*, 2021.
- [14] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1258–1281, 2021.
- [15] J. S. Park and T. Ogunfunmi, "Dct-based image quality assessment for mobile system," in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (ICIP)*, 2014, p. 1.
- [16] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on hvs," in *Proceedings of the Second International Workshop on Video Processing and Quality Metrics*, vol. 4, 2006.
- [17] A. Horé and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369.
- [18] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thirty-Seventh Asilomar Conference on Signals, Systems Computers*, 2003, vol. 2, 2003, pp. 1398–1402 Vol.2.
- [19] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 636–650, 2000.
- [20] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1700–1705, 2013.
- [21] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [22] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
- [23] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 3903–3912, 2017.
- [24] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Transactions on Broadcasting*, vol. 62, no. 2, pp. 446–456, 2016.
- [25] H. Ha, J. Park, S. Lee, and A. C. Bovik, "Perceptually unequal packet loss protection by weighting saliency and error propagation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 9, pp. 1187–1199, 2010.
- [26] L. Zhang, L. Zhang, and X. Mou, "Rfsim: A feature based image quality assessment metric using riesz transforms," in *2010 IEEE International Conference on Image Processing*, 2010, pp. 321–324.
- [27] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.
- [28] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [29] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2015.
- [30] L. Zhang and H. Li, "Sr-sim: A fast and high performance iqa index based on spectral residual," in *2012 19th IEEE International Conference on Image Processing*, 2012, pp. 1473–1476.
- [31] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1969–1977.
- [32] J. Xu, Z. Luo, W. Zhou, W. Zhang, and Z. Chen, "Quality assessment of stereoscopic 360-degree images from multi-viewports," in *2019 Picture Coding Symposium (PCS)*, 2019, pp. 1–5.
- [33] Z. Chen, J. Xu, C. Lin, and W. Zhou, "Stereoscopic omnidirectional image quality assessment based on predictive coding theory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 103–117, 2020.
- [34] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.
- [35] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2020.
- [36] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 131–14 140.
- [37] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal cnn for 360° video quality assessment," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 169–10 178.
- [38] L. Qian, T. Pan, Y. Zheng, J. Zhang, M. Li, B. Yu, and B. Wang, "No-reference nonuniform distorted video quality assessment based on deep multiple instance learning," *IEEE MultiMedia*, vol. 28, no. 1, pp. 28–37, 2021.
- [39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
- [40] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *2018 IEEE international symposium on circuits and systems (ISCAS)*. IEEE, 2018, pp. 1–5.