IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 August 2022, accepted 4 September 2022, date of publication 6 September 2022, date of current version 22 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3204766

RESEARCH ARTICLE

An Effective Foveated 360° Image Assessment Based on Graph Convolution Network

TRUONG THU HUONG[1], (Member, IEEE), DO THU HA[1,2], HUYEN T. T. TRAN[3], (Member, IEEE), NGO DUC VIET[1,4], BUI DUY TIEN[1], NGUYEN HUU THANH[1], (Member, IEEE), TRUONG CONG THANG[3], (Senior Member, IEEE), AND PHAM NGOC NAM[4], (Member, IEEE)

[1]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 100000, Vietnam
[2]International Research Institute for Artificial Intelligence and Data Science, Dong A University, Da Nang 50000, Vietnam
[3]Department of Computer Science and Engineering, University of Aizu, Aizuwakamatsu 965-0006, Japan
[4]College of Engineering and Computer Science, VinUniversity, Hanoi 100000, Vietnam

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

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

ABSTRACT Virtual reality (VR) has been adopted in various fields such as entertainment, education, healthcare, and the military, due to its ability to provide an 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 the human eyes. Foveated technologies can significantly reduce the data required for transmission and computation complexity in rendering. However, understanding the impact of foveated 360° images on human quality perception is still limited. This paper addresses the above problems 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, which apply deep learning techniques and 25 traditional-metric-based models (or analytical-function-based-models), which utilize analytical functions. 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 are reserved for our future work. The source code is available at https://github.com/telagment/FoVGCN.

INDEX TERMS Foveated image, omnidirectional image, virtual reality, graph convolution network, quality of experience.

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 of 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 the 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 [3].

To cope with this problem, one of the most potential solutions is to use foveated technologies, that are based on the foveation feature of the human eyes. This feature refers to spatially foveated visual acuity due to the heterogeneous distribution of photoreceptors 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 areas. This allows significantly reducing not only the data

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

size in transmission but also the computation complexity in rendering [3], [4].

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 figure out the most effective way to reduce a huge amount of data while ensuring the user experience.

There have been several types of research, that apply machine learning-based methods to successfully assess the quality of 360° images [5], [6]. However, most of these studies are just designed to deal with uniform-quality images, but not with foveated-quality images in which quality is changed from the center to the periphery. Due to the significant difference between the characteristic 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 [7] 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 have the same impact on user quality perception.

In this work, we design a Graph Convolution Network (GCN)-based quality model, that could automatically and effectively learn the contribution of each pixel to the perceptual quality of foveated 360° images. Within this scope, 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 or analytical-function-based-models over both the main case study with foveated data and the cross-validation study with uniform data.

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 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 (i.e. DeepQA [8], MIC360IQA [5], and VGCN [6]) and 25 metric-based schemes (i.e. MSE, FMSE [9], UQI [10], PSNR [11], FPSNR [12], SSIM [11], MS-SSIM [13], IW-SSIM, NQM [14], VIF [15], VIFp [16], WSNR, FSIM, FSIMc, F-SSIM [17], PSIM [18], ADD-SSIM [19], FWQI [20], GSIM, RFSIM [21], IW-PSNR [22], BRISQUE [23], NFERM [24], SR-SIM [25], and W-VPSNR [7]). FoVGCN is also evaluated in some cross-validation experiments on uniform content datasets. Finally, the conclusion and future work are briefly discussed in Section V.

II. RELATED WORK

Quality of Experience (QoE) has long been investigated for different content types [26].

A. FOVEATED 360° CONTENT SUBJECTIVE QUALITY ASSESSMENT

In the literature, there have been a lot of studies on foveated contents (i.e., images or videos) [3], [4], [7], [12], [20], [27], [28], [29], [30]. Among them, there are, however, only some on 360° contents [3], [4], [7], [28], [29]..

In [28], 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 consensus. Meanwhile, ACR was found to be the most efficient method. In [29], 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 was conducted. Foveated Wavelet Quality Index (FWQI) was found to be the most effective model for both 2D and 3D foveated 360° videos.

In [4], 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 $\theta \in [0°, 9°]$, the near peripheral area with $\theta \in (9°, 30°]$, 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 by about 10× without perceptual loss, in comparison to the traditional solution of uniform quality. Reference [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 was 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 [7] proposed a simple quality model called W-VPSNR. W-VPSNR is the first and only objective 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 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 have 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 [31].

1) TRADITIONAL-METRIC-BASED METHODS (OR ANALYTICAL-FUNCTION-BASED-MODELS)

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) [9], Universal Quality Index (UQI) [10], Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [11], Foveated Peak Signal-to-Noise Ratio(FPSNR) [12], Multi-scale SSIM (MS-SSIM) [13], Information content Weighted SSIM (IW-SSIM), Noise Quality Measure (NQM) [14], Visual Information Fidelity (VIF) [15], Visual Information Fidelity in the pixel domain (VIFp) [16], Weight Signal-to-Noise Ratio (WSNR), Feature similarity index measure (FSIM), Feature similarity measure (FSIMc) for color image, Foveal feature similarity measure (F-SSIM) [17], Perceptual Similarity (PSIM) [18], Analysis of Distortion Distribution-based (ADD-SSIM) [19], Foveal Structural Similarity [32], and Foveated Wavelet image Quality Index (FWQI) [20], Generic Statistical Information Model (GSIM), Riesz Transforms based Feature Similarity (RFSIM) [21], Information content Weighted PSNR (IW-PSNR) [22], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [23], No-reference Free Energy-Based Robust Metric (NFERM) [24], Spectral Residual based Similarity (SR-SIM) [25], and Weighted Viewport PSNR (W-VPSNR) [7]. Therefore in this paper, we will evaluate our solution's performance in terms of 25 metrics: MSE, FMSE [9], UQI [10], PSNR [11], FPSNR [12], SSIM [11], MS-SSIM [13], IW-SSIM, NQM [14], VIF [15], VIFp [16], WSNR, FSIM, FSIMc, F-SSIM [17], PSIM [18], ADD-SSIM [19], FWQI [20], GSIM, RFSIM [21], IW-PSNR [22], BRISQUE [23], NFERM [24], SR-SIM [25], and W-VPSNR [7] in order to have an insight into our proposed solution from variety of angles.

2) MACHINE LEARNING-BASED METHODS

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, [8], [33], and [34]. In [8], 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 a uniform image quality dataset. In [33], the author considered the important role of multiple viewports related to the image inside the field of view (FoV). Those viewports are extracted by viewport sampling with inputs being reference images (i.e., original images on the server-side) and distorted images (i.e., received images on the client-side). Their proposed stereoscopic omnidirectional image quality assessment (SOIQA) model then learned those viewport features using a deep neural network and support vector regression (SVR). Machine learning techniques have been applied efficiently in [8] and [33] to learn the characteristics of uniform immersive image quality in terms of the 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 360° images is still an open issue.

In the direction of NR-IQA methods, machine-learning-based approaches have been utilized quite commonly. Zhang *et al.* [35] 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.* [6] developed a novel Viewport oriented Graph Convolution Network (VGCN), that concatenates a global branch based on Zhang's work [35], 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.* [36], 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.* [5] 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 [37], the authors introduced meta-learning based image quality assessment

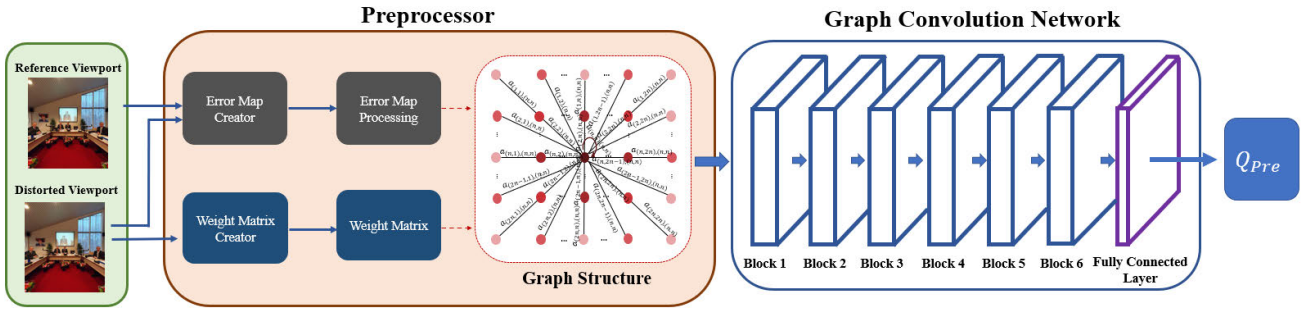


FIGURE 1. FoVGCN Model Operation based on the full-reference image quality assessment (FR-IQA) approach which leverages the information of both reference and distorted images. First, the graph structure data is constructed based on Error map and attention weight matrix. Then, the convolution graph neural network interprets the graph data to predict the final quality assessment score of viewports.

(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.

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], [6], and [36]. 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 (FoVGCN)

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 zone quality of omnidirectional images.

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 represent the spatial quality changing in different zones of an image and the priority of human attention, respectively. After being pre-processed, both the error map and the attention weight matrix construct a graph structure, that is the input of the graph convolution layer. The graph structure is then fed into the convolution graph neural network block to predict the overall quality score of the image. In the next subsections, the two main blocks of the FoVGCN model and its parameter settings will be described in detail.

A. PREPROCESOR

In the preprocessor block, reference and distorted images are fed into the error map creator block to create an error map. Error map is implied as a graph in which each vertex (node) represents a pixel and is connected via an edge to a foveation node (or foveation pixel), as Figure 2 describes. A foveation node is defined as the center point in the virtual viewport [3]. Therefore, each non-foveation pixel has only one neighbor (foveation node), and the information stored at each node is calculated based on an error map E.

The attention weight matrix block takes the shape of a viewport as the input to create a distance-based distribution matrix which has the same size as the viewport. The attention weight $a_{(i,j),(n,n)}$ measures the connective strength based on distance relation between arbitrary $E_{i,j}$ and foveation node $E_{n,n}$, $i, j \in (1, 2, \ldots, 2n)$. Consequently, error map and attention weight matrix construct the graph structure, that is the input of the graph convolution layer as shown in Figure 2.

1) ERROR MAP PROCESSING

In this section, we describe details 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.

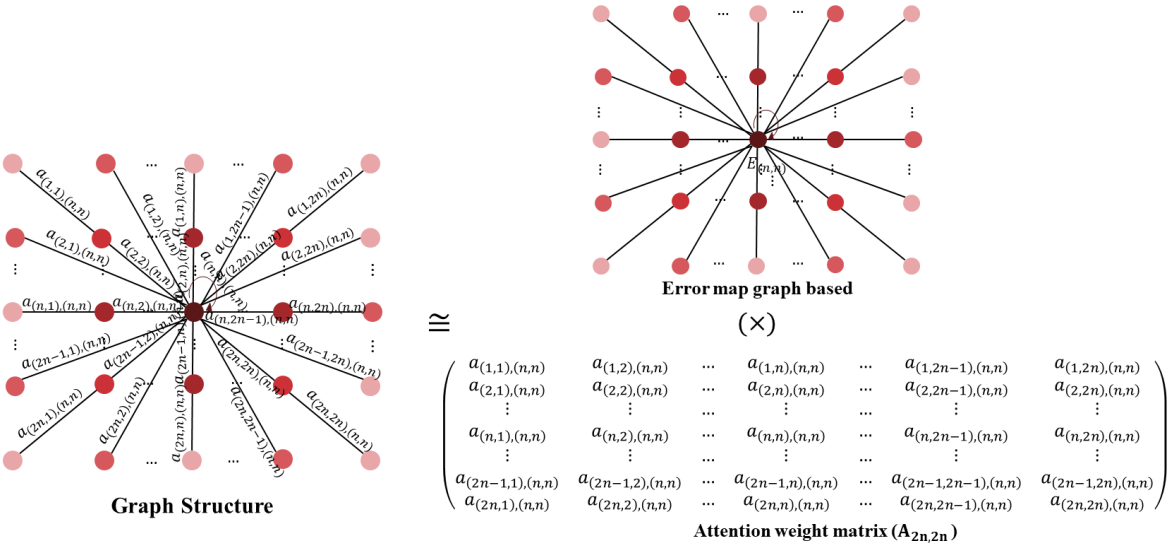


FIGURE 2. Graph structure representation.

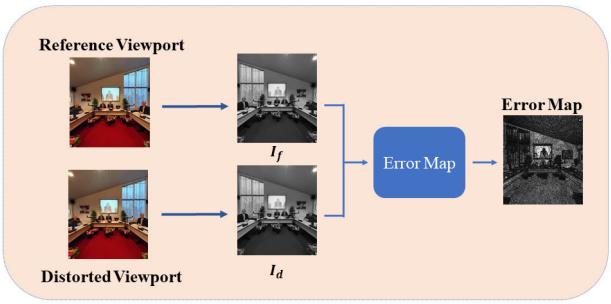


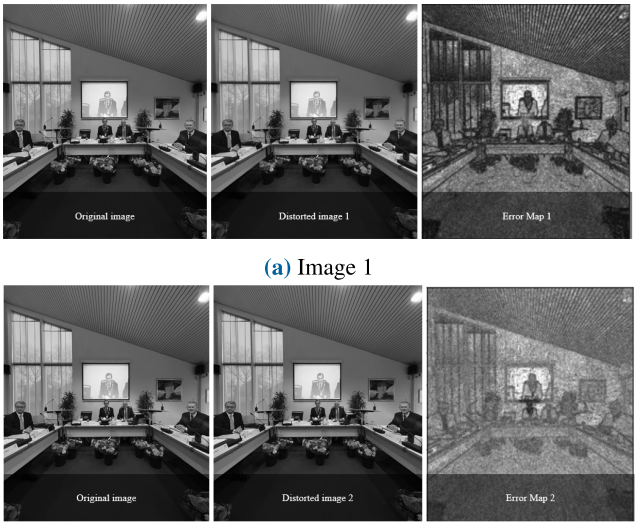
FIGURE 3. Error map Creator.

According to human retina characteristics, the quality of omnidirectional images is typically adjusted based on five human retina zones of Fovea, Parafovea, Perifovea, Near periphery, Far periphery [3]. It means that the quality of each pixel in a distorted image usually changes from the center region inside to the outside. Our goal is to intuitively imply the changing of the quality through each zone by using the so-called *Error Map* (E). 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 [8] for better correlation with the perceived quality of viewers:

$$E = \frac{\log\left(1/((I_r - I_d)^2 + \alpha)\right)}{\log(1/\alpha),} \tag{1}$$

where $\alpha = \varepsilon/255^2$ is a constant and $\varepsilon = 0.1$.

The detail of the Error map Creator is shown in Figure 3. 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 can be seen in



(a) Image 1

(b) Image 2

FIGURE 4. The visual image of error maps of two foveated images. The error map is created by the original (or reference) image and the corresponding distorted image.

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.

In the Errormap Transform block, the error map is first divided into four quadrants for adapting with Attention weight matrix transformation, then each of those is rotated and sorted as four consecutive quadrants following the rule in Figure 5. The division of the error map is to help reduce the computational complexity of the proposed model towards real-time quality assessment. As a result of this process, the attention weight matrix size is decreased by a factor of 4, thereby significantly reducing the running time of the whole model.

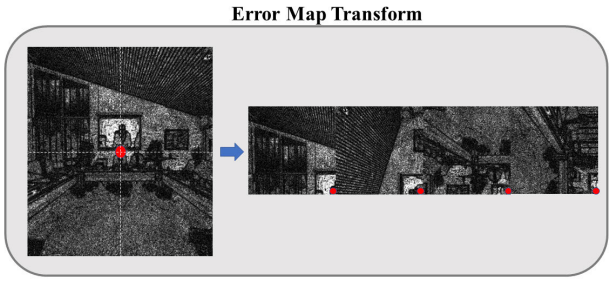


FIGURE 5. Errormap transform.

2) ATTENTION WEIGHT MATRIX PROCESSING

The concept of attention weight matrix is used to describe the priority of human attraction in each zone of an immersive image. It can be understood that pixels (nodes) in the central zones are considered to have the higher weight than pixels (nodes) farther away from the center zone. Therefore, our attention weight matrix is created in such a way that it has high density representing high priority in the middle while decreasing to lower and lower density for further areas from the middle. Then, we define attention the weight matrix as a $2n \times 2n$ symmetric matrix. The elements of the matrix indicate the edge information between arbitrary nodes to the central node of the graph structure, as described in Figure 2. The set of $a_{(i,j),(n,n)}$ entries follow 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 a broader insight 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 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.

Linear degradation distribution In our design, we propose a formula presented in Eq. (2), that describes the linear degradation from the center node to the edge nodes. The value of the matrix corresponding to each node is inversely proportional to the distance from node $E_{i,j}$ to the central node $E_{n,n}$ and decreases gradually and constantly from the center to the periphery. The elements of the attention matrix are calculated following Equ.2, and its distribution is visualized in Figure 7.

$$A_{i,j} = a_{(i,j),(n,n)} = 1 - \frac{dis(E_{i,j}, E_{n,n})}{dis_{max}} + \delta, \tag{2}$$

where

- $a_{(i,j),(n,n)}$ or $A_{i,j}$ is the element of the attention weight matrix, in range of $[\delta, 1 + \delta]$;
- $dis(E_{i,j}, E_{n,n}) = \sqrt{(i-n)^2 + (j-n)^2}$ is the distance between pixel $E_{i,j}$ and center pixel $E_{n,n}$;
- dis_{max} is the maximal distance between two nodes $E_{i,j}$ and $E_{n,n}$);
- Threshold δ is applied to avoid the zero value, and is set equal to 0.0001.

Gaussian degradation distribution As mentioned in [4], the density function of cones presented in [38] can be approximated by a Gaussian distribution. Inspired by this observation, the Gaussian distribution is leveraged to represent the perception process of human eyes. This idea follows the concept of the human retina [3] which is user perception tends to be affected by the quality of fovea and parafovea zones from the center to the outside of an image. The reason is that human eyes concentrate significantly on the fovea and parafovea zones of the human retina - a small region in the viewport [3]. Therefore, 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:

$$A_{i,j} = f(x) = e^{-\frac{1}{2}(\frac{x}{\sigma})^2}, \tag{3}$$

where

- σ: standard deviation of x;
- $A_{i,j}$ 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 8 visualizes the Gaussian distribution with different σ values where bigger 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 applied since it can efficiently learn the graph-structured data and successfully study the characteristic of foveated 360° images. In GCN, we construct the graph structure in which a node represents a pixel and an edge represents the distance-based correlation between an arbitrary node to a foveation node. The graph structure is built on the error map and attention weight matrix. Significantly, the GCN is designed by using the attention weight matrix instead of an adjacency matrix for a purpose of describing the humane eyes attention in practice. The attention weight matrix satisfies real, symmetric, positive semidefinite properties, that are adapted to the mathematical requirement in Eq. 4 and 5.

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

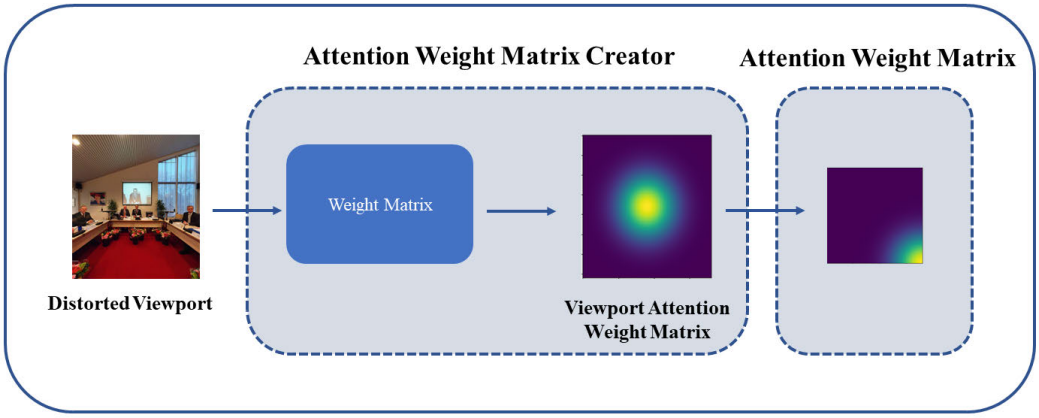


FIGURE 6. Attention Weight matrix processing.

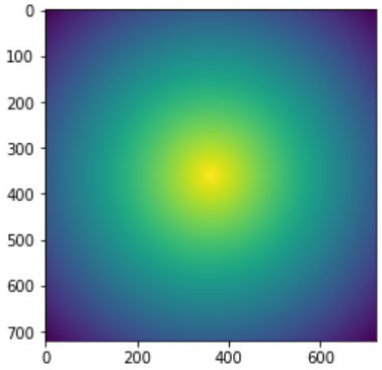
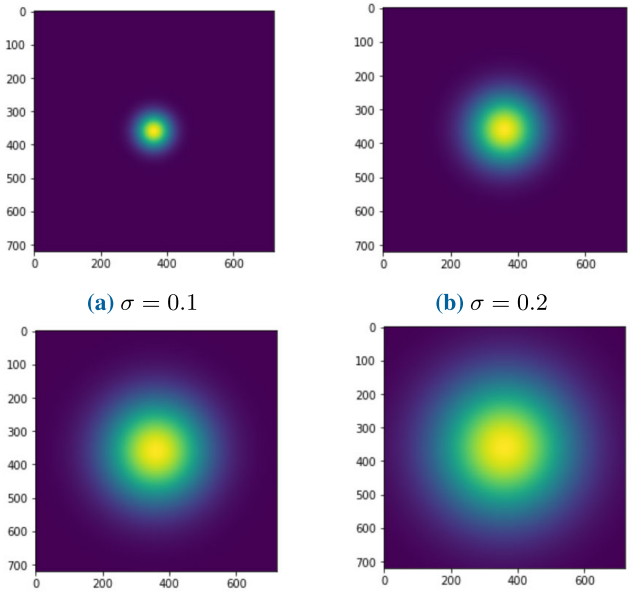


FIGURE 7. Attention weight matrix in linear distribution.



(a) $\sigma = 0.1$ **(b)** $\sigma = 0.2$ **(c)** $\sigma = 0.3$ **(d)** $\sigma = 0.4$

FIGURE 8. Attention weight matrix gaussian distribution.

(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:

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

$$\tilde{A} = A + I_N \tag{5}$$

where

- $H^{(l)}$: the activation matrix in l^{th} layer;
- $W^{(l)}$: the learnable model parameters in layer l
- $\tilde{D}$: the degree matrix which can be calculated by $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$;
- A: attention weight matrix and identity matrix I_N;
- $\sigma()$: a non-linear activation function such as the Softmax function, ReLU function, Softplus function, etc. In our work, we use the Softplus function for better stabilization and performance to deep neural networks.

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

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.

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 error map is created from those viewports and the attention weight matrix is fed through the six blocks of the graph

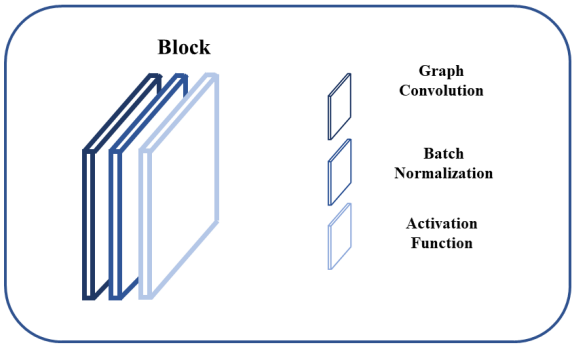


FIGURE 9. The detail of each GCN block.

TABLE 1. 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)

convolution layers which is shown in Figure 9. Each block contains a graph convolution layer, batch normalization, and a Softplus activation function as described in Eq. (6):

$$f(x) = \ln(1 + e^x) \tag{6}$$

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.

The details of all parameters used in the FoVGCN model are presented in Table 1.

IV. EXPERIMENTS AND RESULTS

To evaluate the performance of FoVGCN, we use three open datasets, one of which is a foveated image dataset, and the other two are uniform image datasets. In the following sections, we will firstly describe these datasets. Then, the experimental settings and results of FoVGCN and existing solutions are presented.

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, and natural landscape. To create the distorted images, Gaussian filters were employed with a fixed filter size of 50 and four different standard deviations. Specifically, the distortion of images was conducted based on five regions of the human retina and two basic scenarios of spatial quality changes: the quality gradually decreases or increases from the center to the periphery. For each scenario, four different quality levels were generated corresponding to four different standard deviations σ. Due to the fact that blurring in the center zones is easier to be perceived than in the peripheral zones, the values of σ are 2, 4, 8, and 12 for the first scenario and 1, 2, 4, and 6 for the second scenario. To prevent boundaries between the low and high-quality zones from irritating viewers, a linear function was used to smooth transition belts between two adjacent zones. Please refer to [3] for more details about the process of creating the distorted images.

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

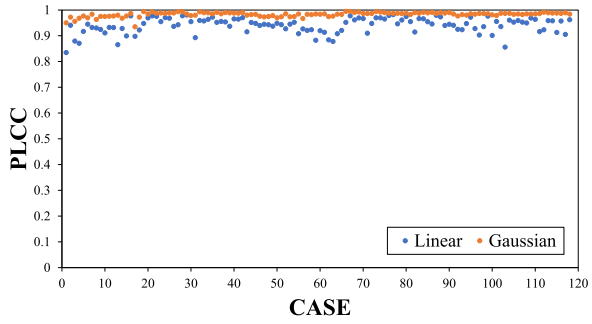
In our work, FoVGCN is specifically designed to cope with foveated-quality images. That inclusively means that it may not work well for uniform-quality images in comparison with the other existing solutions, which are designed for this uniform type. However, to investigate the effectiveness of FoVGCN for uniform images, we also evaluate the performance of FoVGCN on two other uniform image datasets - CVIQ [5] and OIQA [40] - which 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 distorted 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.

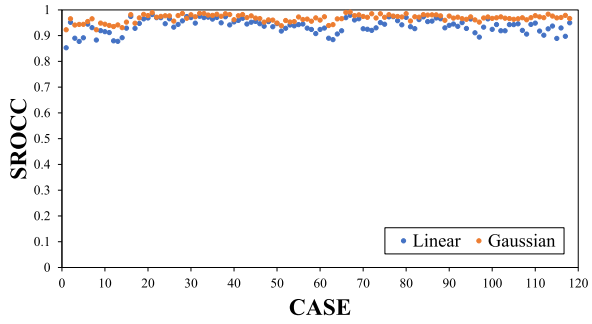
B. PERFORMANCE EVALUATION

1) EXPERIMENTAL SETTINGS

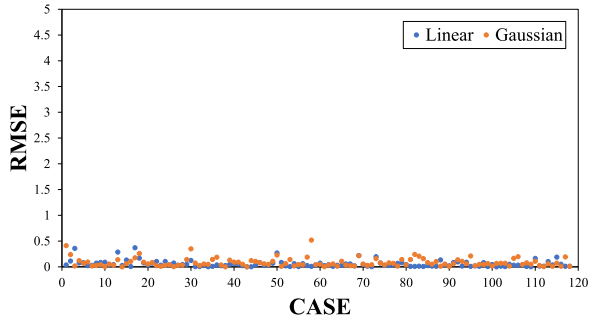
To evaluate the performance of the FoVGCN model, we use common performance measures such as Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), and Root mean square error (RMSE). In the literature, RMSE, PLCC, and SROCC are commonly considered as standard metrics to evaluate the accuracy of quality models [41], [42], [43]. Specifically, they are utilized to measure the difference, the linear and non-linear correlations between subjective quality values and objective



(a) PLCC



(b) SROCC



(c) RMSE

FIGURE 10. SROCC, PLCC and RMSE with the gaussian distributed and linear distributed weight matrices.

quality values predicted from a model, respectively [44]. The experiments are conducted in various aspects to have a better understanding of the impacts of different factors.

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. As aforementioned, we use the foveated immersive image dataset [3], that is constructed from 16 source distorted images. More specifically, we need to select 2 source distorted images for testing from the 16 source distorted images. So it means the remaining 14 distorted images are used for training.

Since we have to cover all cases of choosing any 2 distorted images from a set of 16 source images, it leads to the mathematics combination problem of choosing 2 from 16 subjects. Therefore, in total, there are totally $\binom{16}{2} = 120$ possible testing sets. 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 *sigma* 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 10 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 of the two distribution attention weight matrices result in high SROCC and PLCC (i.e., over 0.85 for both SROCC and PLCC). 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 *sigma* of the Gaussian distribution, which is used to construct the weight matrix, will result in

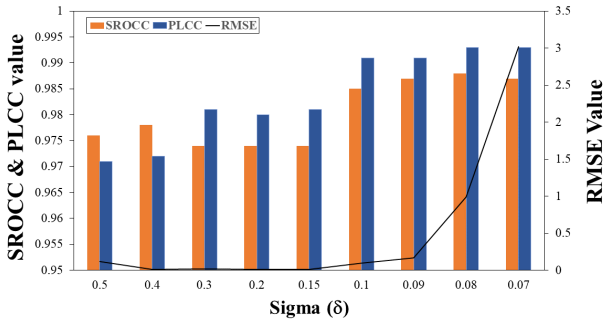


FIGURE 11. The performance of FoVGCN over different sigma coefficients.

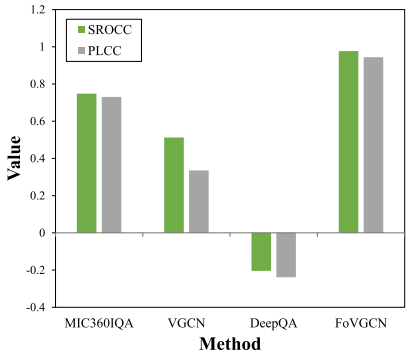
different degrees of the central concentration, relating to different human attention in each zone. In order to figure out the best performance related to the σ coefficient, Figure 11 illustrates SROCC, PLCC and RMSE with different σ values. It can be seen that, in general, SROCC, PLCC, and RMSE increase as σ is decreased. When σ is less than 0.1, the accuracy values saturate while RMSE quickly increases. So, we set σ=0.1 to have a good balance among the three values SROCC, PLCC, and RMSE.

C. FoVGCN VERSUS OTHER EXISTING SOLUTIONS

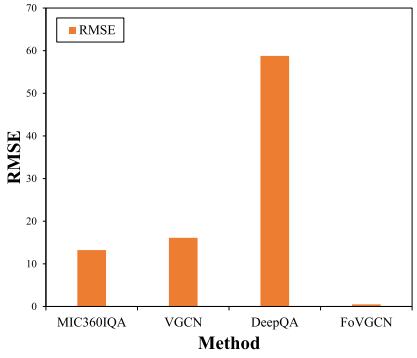
In this section, we compare FoVGCN with other existing solutions, including 25 analytical metrics and 3 machine-learning-based methods. The 25 analytical metrics include MSE, FMSE, UQI, PSNR, FPSNR, SSIM, MS-SSIM, IW-SSIM, NQM, VIF, VIFp, WSNR, FSIM, FSIMc, F-SSIM, PSIM, ADD-SSIM, FWQI, GSIM, RFSIM, IW-PSNR, BRISQUE, NFERM, SR-SIM, and W-VPSNR. The results of those 25 metrics are calculated by averaging the values of the 120 cases. The 3 machine-learning-based methods are DeepQA [8], MIC360IQA [5], and VGCN [6]. 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 methods, as shown in Figure 12. It can be seen that, DeepQA [8] 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. The overall



(a) PLCC and SROCC



(b) RMSE

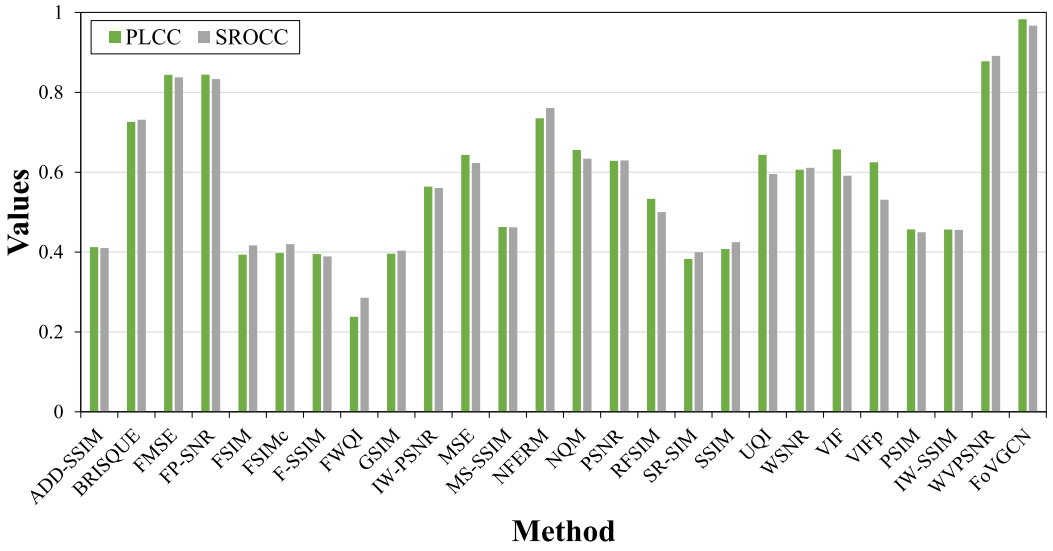
FIGURE 12. Performance of 3 current machine-learning-based approaches vs. FoVGCN.

performance comparison between FoVGCN versus other existing solutions over the foveated dataset is also summarized in Table 4.

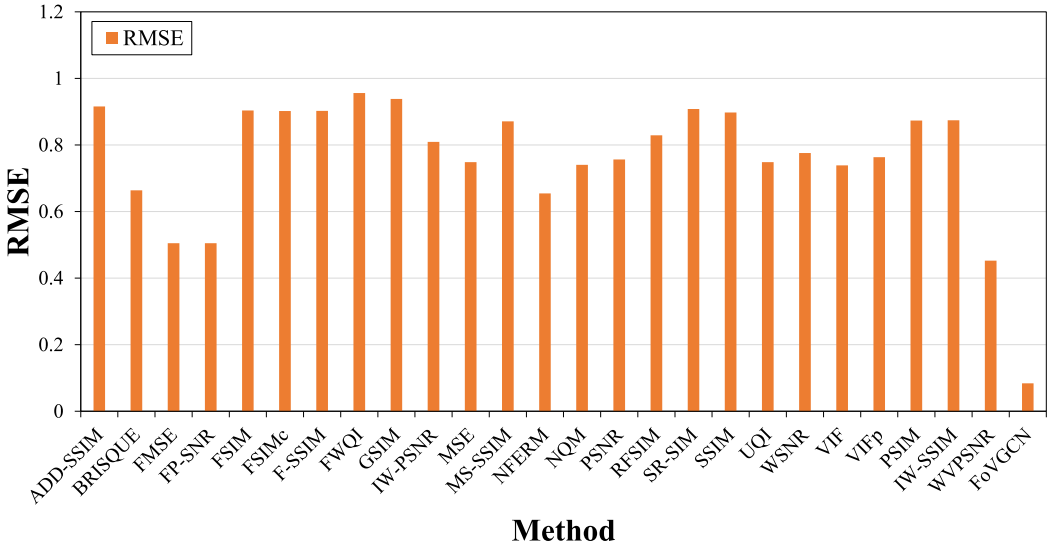
In addition, the scatter diagrams of the ground truth and the predicted MOSs of all metrics and models are shown in Figure 14. In this figure, the horizontal axis presents the MOS score, and the vertical axis shows the predicted MOS score, which is the quality image score predicted by each different model/approach. The trend of those diagrams is expected to be the shape of Identity Function Graph indicating the relationship between the predicted MOS score and real MOS score. MOS stands for Mean Opinion Score, which is a numerical measure of the human-judged overall quality of experience (QoE), normally for voice and video sessions, ranked on a scale from 1 (bad) to 5 (excellent). The definition of QoE and MOS can be found in [26].

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.



(a) PLCC and SROCC



(b) RMSE

FIGURE 13. Performance of 25 analytical metrics vs. FoVGCN.

TABLE 2. Performance of FoVGCN over the 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	**0.920**	**0.925**	**0.614**

TABLE 3. Performance of FoVGCN over the 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	**0.781**	**0.815**	**0.285**

D. CROSS VALIDATION EXPERIMENTS

In the previous section, FoVGCN has been shown to be efficient in quality assessment for the foveated images. In order to see how FoVGCN will work with different types of content, we investigate the performance of FoVGCN with two other uniform image datasets of CVIQ [5] 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.

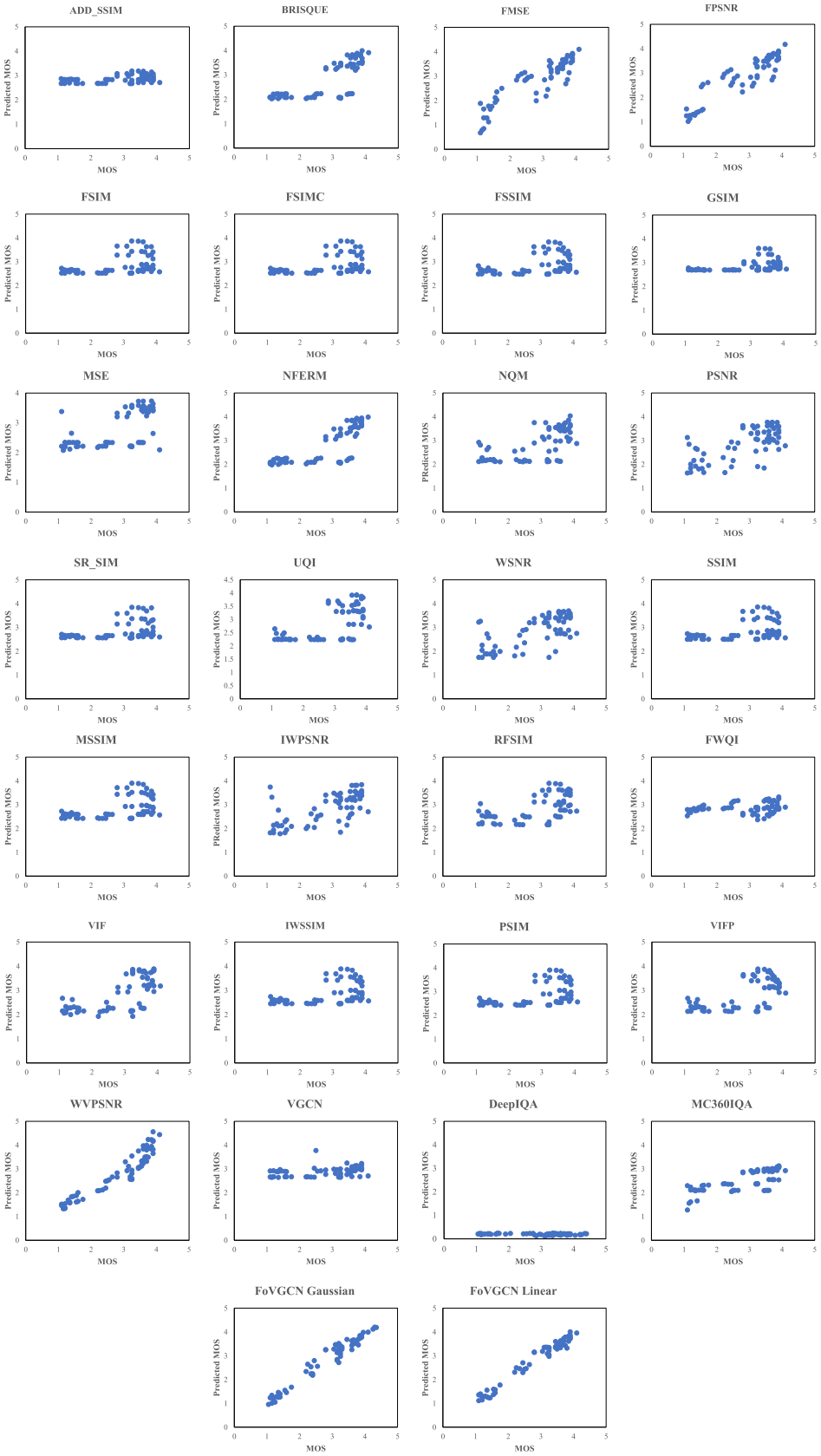


FIGURE 14. Scatter diagrams of the actual MOSs versus the predicted MOS values for all methods.

TABLE 4. FoVGCN vs. other solutions over 3 different datasets.

Method		OIQA dataset			CVIQA dataset			Foveated dataset		
		SROCC	PLCC	RMSE	SROCC	PLCC	RMSE	SROCC	PLCC	RMSE
Machine-learning-based model	MIC360IQA	0.9139	0.9267	0.7854	0.9428	0.9429	4.6506	0.7482	0.7302	13.2373
	VGCN	0.9584	0.9584	0.5967	0.9651	0.9639	3.6573	0.5124	0.3354	16.1019
	DeepQA	0.8973	0.9044	0.8914	0.9292	0.9375	4.8574	-0.2044	-0.2385	58.7569
	FoVGCN Gaussian with $\sigma = 0.5$	0.7571	0.7870	1.0224	0.8599	0.8567	3.8729	0.9760	0.9710	0.1190
	FoVGCN Gaussian with $\sigma = 0.4$	0.7733	0.7803	0.9155	0.8326	0.8618	4.3105	0.9780	0.9720	0.0080
	FoVGCN Gaussian with $\sigma = 0.3$	0.7794	0.8132	0.5374	0.8929	0.9084	3.0532	0.9740	0.9810	0.0190
	FoVGCN Gaussian with $\sigma = 0.2$	0.7986	0.8199	0.0187	0.8994	0.9113	1.6876	0.9740	0.9800	0.0090
	FoVGCN Gaussian with $\sigma = 0.15$	0.7900	0.8173	0.1517	0.9097	0.9185	1.1508	0.9740	0.9810	0.0090
	FoVGCN Gaussian with $\sigma = 0.1$	0.7813	0.8147	0.2848	0.9201	0.9256	0.6140	0.9850	0.9910	0.0940
	FoVGCN Gaussian with $\sigma = 0.09$	0.8441	0.8551	0.9270	0.9240	0.9295	0.5087	0.9870	0.9910	0.1650
	FoVGCN Gaussian with $\sigma = 0.08$	0.8461	0.8569	0.9087	0.9304	0.9335	0.0947	0.9880	0.9930	0.9930
	FoVGCN Linear	0.7676	0.7985	1.4096	0.8558	0.8536	2.7726	0.9381	0.9319	0.0967
25 Traditional-metric-based models (or analytical-function-based-models)	ADD-SSIM	0.8411	0.7040	10.2171	0.8447	0.7642	9.0835	0.4125	0.4102	0.9157
	BRISQUE	0.5032	0.5098	12.3762	0.4350	0.4258	12.7421	0.7259	0.7312	0.6637
	FMSE	0.6698	0.3846	13.2795	0.8041	0.6656	10.5096	0.8440	0.8376	0.5051
	FPSNR	0.6698	0.6696	10.6840	0.8041	0.7967	8.5119	0.8442	0.8330	0.5048
	FSIM	0.8687	0.7470	9.4768	0.8798	0.8250	7.8113	0.3935	0.4169	0.9039
	FSIMC	0.8768	0.7584	9.2908	0.8836	0.8284	7.7422	0.3976	0.4199	0.9021
	FSSIM	0.8377	0.8029	8.5762	0.8560	0.8373	7.6995	0.3950	0.3893	0.9027
	FWQI	0.7654	0.7722	9.1402	0.8055	0.8233	7.9929	0.2381	0.2858	0.9563
	GSIM	0.8663	0.7673	9.2261	0.8665	0.8129	8.2025	0.3961	0.4036	0.9383
	IWPSNR	0.7887	0.7810	8.9848	0.8225	0.8215	8.0302	0.5641	0.5605	0.8094
	MSE	0.6813	0.4365	12.9429	0.7702	0.7841	8.7409	0.6433	0.6232	0.7483
	MS-SSIM	0.8163	0.8407	7.9526	0.8442	0.8065	8.3251	0.4629	0.4621	0.8710
	NFERM	0.0162	0.0095	14.3850	0.2251	0.2673	13.5703	0.7351	0.7606	0.6544
	NQM	0.8121	0.7846	8.9183	0.8201	0.8314	7.8251	0.6555	0.6339	0.7404
	PSNR	0.6813	0.6785	10.5674	0.7702	0.7725	8.9419	0.6285	0.6296	0.7564
	RFSIM	0.7822	0.7800	9.0023	0.7937	0.7933	8.5736	0.5335	0.5003	0.8294
	SR-SIM	0.8470	0.8218	8.1223	0.8760	0.7884	8.5027	0.3830	0.3996	0.9084
	SSIM	0.8247	0.8312	7.9979	0.8382	0.7961	8.5227	0.4077	0.4248	0.8976
	UQI	0.7921	0.7892	8.8349	0.7689	0.7906	8.6226	0.6435	0.5955	0.7485
	WSNR	0.7336	0.7221	9.9512	0.7312	0.7293	9.6359	0.6064	0.6111	0.7760
	VIF	0.8001	0.8023	8.5862	0.8096	0.7859	8.7082	0.6571	0.5912	0.7390
	VIFP	0.7979	0.8000	8.6320	0.8382	0.8312	7.8293	0.6247	0.5311	0.7636
	PSIM	0.8774	0.7822	8.8797	0.8940	0.8393	7.5411	0.4568	0.4498	0.8736
	IWSSIM	0.8248	0.8010	8.4870	0.8602	0.8425	7.5854	0.4564	0.4554	0.8744
	WVPSNR	0.6750	0.6740	10.6310	0.8070	0.8020	8.4040	0.8780	0.8912	0.4526

The results are illustrated in Table 2 and Table 3. We can see that, with the CVIQ dataset, FoVGCN achieves better performance in terms of SROCC, PLCC, and RMSE, which are 0.920, 0.925, and 0.614 respectively. With the OIQA dataset, FoVGCN achieves comparable accuracy with other metrics; however, its RMSE (0.285) is much smaller than others. That means FoVGCN is more stable than other foveal metrics.

E. DISCUSSIONS

To get the overview of the performance of FoVGCN vs. other existing solutions over different datasets, we summarize all performances in Table 4. The above results show that the proposed model FoVGCN provides the best performance compared to reference methods. Also, FoVGCN is effective not only with foveated images but also with uniform-quality images. We believe that constructing a graph structure that is composed of an errormap and attention weight matrix allows the model to efficiently interpret the characteristics of data structure with spatial quality changes. It is the main reason our proposed model achieved high performance. Testing our model with three different datasets of various scenarios (i.e., uniform-quality and foveated images) also help avoid bias and guarantee that the model can work well in general.

The use of foveation feature in quality models is crucial to effectively deal with foveated images. As seen in Fig. 13, the performances of the analytical foveation-based models like FMSE, FPSNR, WVPSNR are quite good (over 0.8 for both PLCC and SROCC). Meanwhile, the three reference deep-learning-based models, namely DeepQA, MIC360IQA, and VGCN, have lower (or very low) performances (see Fig. 12).

Note that, though these deep-learning-based models are already retrained using the same foveated image dataset as the proposed FoVGCN model, their low performances imply that the deep-learning architectures of these models still cannot capture the characteristics of foveated images.

Currently, the study in this paper still has some limitations.

- First, the proposed model is just focused on image contents. It was not evaluated with video contents due to the lack of foveated video datasets.
- Second, the resolution of foveated images in this study is fixed. This is also because of the available dataset does not provide images of different resolutions.

In the future, we will carry out subjective tests to obtain more foveated content datasets, which cover different cases of resolutions, headsets, and content types (i.e. images and videos). The FoVGCN model will be extended and evaluated using these future datasets. Field studies using foveated quality models in the context of VR video streaming will be also implemented.

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.

REFERENCES

[1] Wikipedia Contributors. (2021). *Virtual Reality—Wikipedia, the Free Encyclopedia*. Accessed: Oct. 19, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Virtual_reality&oldid=1049963342

[2] M. Chen, Y. Jin, T. Goodall, X. Yu, and A. C. Bovik, "Study of 3D virtual reality picture quality," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 89–102, Jan. 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. 166997–167009, 2019.

[4] P. Guo, Q. Shen, Z. Ma, D. J. Brady, and Y. Wang, "Perceptual quality assessment of immersive images considering peripheral vision impact," 2018, *arXiv:1802.09065*.

[5] 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 J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 64–77, Jan. 2020.

[6] J. Xu, W. Zhou, and Z. Chen, "Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1724–1737, May 2021.

[7] 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 *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, Jun. 2019, pp. 151–152.

[8] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1969–1977.

[9] J. S. Park and T. Ogunfunmi, "DCT-based image quality assessment for mobile system," in *Proc. Int. Conf. Image Process., Comput. Vis., Pattern Recognit. (IPCV)*, 2014, p. 1.

[10] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. 2nd Int. Workshop Video Process. Qual. Metrics*, vol. 4, 2006, pp. 1–4.

[11] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.

[12] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, pp. 129–132, Mar. 2002.

[13] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[14] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[15] J. Wu, W. Lin, G. Shi, and A. Liu, "Reduced-reference image quality assessment with visual information fidelity," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1700–1705, Nov. 2013.

[16] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[17] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[18] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 3903–3912, May 2017.

[19] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Trans. Broadcast.*, vol. 62, no. 2, pp. 446–456, Jun. 2016.

[20] Z. Wang, A. C. Bovik, L. Lu, and J. L. Kouloheris, "Foveated wavelet image quality index," in *Applications of Digital Image Processing XXIV*, vol. 4472. Bellingham, WA, USA: SPIE, 2001, pp. 42–52.

[21] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 321–324.

[22] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[24] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[25] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1473–1476.

[26] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, and M.-C. Larabi, "Qualinet white paper on definitions of quality of experience," in *Proc. Qualinet Meeting*, Novi Sad, Serbia, Mar. 2013.

[27] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation of foveated video coding using audio-visual focus of attention," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1322–1331, Nov. 2011.

[28] 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 *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 55–63.

[29] 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 Trans. Image Process.*, vol. 30, pp. 5905–5919, 2021.

[30] L. Surace, M. Wernikowski, C. Tursun, K. Myszkowski, R. Mantiuk, and P. Didyk, "Learning foveated reconstruction to preserve perceived image statistics," 2021, *arXiv:2108.03499*.

[31] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1258–1281, Apr. 2021.

[32] H. Ha, J. Park, S. Lee, and A. C. Bovik, "Perceptually unequal packet loss protection by weighting saliency and error propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 9, pp. 1187–1199, Sep. 2010.

[33] J. Xu, Z. Luo, W. Zhou, W. Zhang, and Z. Chen, "Quality assessment of stereoscopic 360-degree images from multi-viewports," in *Proc. Picture Coding Symp. (PCS)*, Nov. 2019, pp. 1–5.

[34] Z. Chen, J. Xu, C. Lin, and W. Zhou, "Stereoscopic omnidirectional image quality assessment based on predictive coding theory," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 103–117, Jan. 2020.

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

[36] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 917–928, Apr. 2020.

[37] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14131–14140.

[38] C. W. Tyler, "Analysis of human receptor density," in *Basic and Clinical Applications of Vision Science*. Cham, Switzerland: Springer, 1997, pp. 63–71.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[40] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2018, pp. 1–5.

[41] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A tutorial on video quality assessment," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 1126–1165, 2nd Quart., 2014.

[42] M. Xu, C. Li, S. Zhang, and P. L. Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 5–26, Jan. 2020.

[43] *Tutorial: Objective Perceptual Assessment of Video Quality: Full Reference Television*, document, International Telecommunication Union, 2004. [Online]. Available: https://www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf

[44] H. T. T. Tran, D. V. Nguyen, N. P. Ngoc, and T. C. Thang, "Overall quality prediction for HTTP adaptive streaming using LSTM network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3212–3226, Aug. 2021.