EyeQoE: A Novel QoE Assessment Model for 360-degree Videos Using Ocular Behaviors

HUADI ZHU, The University of Texas at Arlington
TIANHAO LI, The University of Texas at Arlington
CHAOWEI WANG, The University of Texas at Arlington
WENQIANG JIN*, Hunan University
SRINIVASAN MURALI, The University of Texas at Arlington
MINGYAN XIAO, The University of Texas at Arlington
DONGQING YE, The University of Texas at Arlington
MING LI, The University of Texas at Arlington

As virtual reality (VR) offers an unprecedented experience than any existing multimedia technologies, VR videos, or called 360-degree videos, have attracted considerable attention from academia and industry. How to quantify and model end users' perceived quality in watching 360-degree videos, or called QoE, resides the center for high-quality provisioning of these multimedia services. In this work, we present EyeQoE, a novel QoE assessment model for 360-degree videos using ocular behaviors. Unlike prior approaches, which mostly rely on objective factors, EyeQoE leverages the new ocular sensing modality to comprehensively capture both subjective and objective impact factors for QoE modeling. We propose a novel method that models eye-based cues into graphs and develop a GCN-based classifier to produce QoE assessment by extracting intrinsic features from graph-structured data. We further exploit the Siamese network to eliminate the impact from subjects and visual stimuli heterogeneity. A domain adaptation scheme named MADA is also devised to generalize our model to a vast range of unseen 360-degree videos. Extensive tests are carried out with our collected dataset. Results show that EyeQoE achieves the best prediction accuracy at 92.9%, which outperforms state-of-the-art approaches. As another contribution of this work, we have publicized our dataset on https://github.com/MobiSec-CSE-UTA/EyeQoE_Dataset.git.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**.

Additional Key Words and Phrases: QoE assessment, graph learning, eye-based cues

1 INTRODUCTION

Motivation. With the development of Virtual Reality (VR) technologies, 360-degree videos, also referred to as omnidirectional or VR videos, have seen a revolutionary rise over the last decade. As a novel type of multimedia, 360-degree videos provide an immersive and interactive watching experience by rendering spherical frames covering all directions around the viewer, attracting great interest from customers, researchers, and industry. In the meantime, these videos are mostly rendered in high resolutions to maintain fair visual quality. Given the limited

*The work was done when the author was a Ph.D. student at the University of Texas at Arlington

Authors' addresses: Huadi Zhu, huadi.zhu@mavs.uta.edu, The University of Texas at Arlington; Tianhao Li, tianhao.li@mavs.uta.edu, The University of Texas at Arlington; Chaowei Wang, chaowei.wang@mavs.uta.edu, The University of Texas at Arlington; Wenqiang Jin, wqjin@hnu.edu.cn, Hunan University; Srinivasan Murali, srinivasan.murali@mavs.uta.edu, The University of Texas at Arlington; Mingyan Xiao, mingyan.xiao@mavs.uta.edu, The University of Texas at Arlington; Dongqing Ye, dongqing.ye@mavs.uta.edu, The University of Texas at Arlington; Ming Li, ming.li@mavs.uta.edu, The University of Texas at Arlington.

2474-9567/2022/3-ART39
https://doi.org/10.1145/3517240

network bandwidth, the network and service providers have to strike a balance between resource consumption and service quality for 360-degree video streaming. Hence, it is of essential importance for them to get an in-depth understanding of the user's experience and take necessary adaptive actions in service management. As a critical evaluation indicator, quality of experience (QoE), defined by ITU-T [34] as a measure of the acceptability of an application or service perceived subjectively by end-users, has been widely adopted. In current multimedia services, user's QoE is mainly obtained by asking people to measure their perceived quality via surveys or self-reports. However, such procedures are time-consuming and may be annoying for the users.

To address this issue, extensive prior efforts have been devoted to developing QoE assessment models that map a diverse spectrum of impact factors, such as underlying network conditions and video qualities, to a QoE score given a specific multimedia service type. In this way, it avoids bothering users with questions to collect opinions and feedback. QoE assessment is automatically carried out with significantly reduced human labor efforts. Nonetheless, this topic in the context of 360-degree videos in VR environments is yet far from well investigated. One mainstream of existing approaches can be classified as *video-centric models* [14, 20, 26, 64, 66, 67, 74, 80, 81, 85]. QoE is derived by analyzing distortions of videos displayed under various video quality assessment (VQA) metrics. These models are criticized for overlooking subjective factors. More recently, [43, 44, 75] integrate human visual attention to their QoE models. Their basis is that viewers mostly focus on objects of interest in a scene. Thus, distortions on different parts should impose a nonuniform impact on QoE estimation. These works then assign weights in accordance with the viewer's visual attention in aggregating pixel-wise distortions. The above ideas are inherited from QoE modeling of conventional 2D videos and thus incapable of capturing unique characteristics of 360-degree videos. As pointed out by prior studies [30, 38, 71, 86], subject feelings, such as cybersickness, immersiveness, and fatigue, are of essential importance in determining their perceived QoE of watching 360-degree videos, in addition to the well-recognized factors such as video quality. Hence, a QoE model that effectively harnesses all the above factors is in dire need for service management of 360-degree video streaming.

Recently, ocular behaviors, such as eye gaze, fixations, saccades, pupillometry, and blinks, have emerged as a new sensing modality to measure human perceptions. For example, eye blinking rates are reported to increase as the evolvement of visual fatigue [37, 72]. Strong correlations are also observed between visual fatigue and saccade peak velocity, saccade duration, and fixation duration [82]. Eye-based sensing has extended the current multimedia applications and services with an additional perceptive dimension and opened up grand opportunities to enhance service provisioning. For instance, Tesla is starting to use the camera above the rear-view mirror in some car models to help make sure people pay attention to the road while using Autopilot [52]. In the meantime, eye trackers have been embedded into many prevalent commercial VR headsets [2–6] to assist in simulating depth of field and focus, providing a more realistic and natural visual experience. It is widely accepted that incorporating eye-tracking technology is a trend of VR headsets [18].

Our approach. Based on these observations, we propose to leverage ocular behaviors captured by eye trackers in VR headsets to model and predict viewer's perceived QoE in watching 360-degree videos. We call the novel prediction model EyeQoE. As presented in our measurement study (Section 4), strong correlations are broadly found between eye-based cues and various impact factors of QoE for 360-degree videos, including the objective ones (e.g., video quality) and subjective ones (e.g., cybersickness, immersiveness, and fatigue). EyeQoE treats the cues as indicators of the viewer's perceived experience and aims to bridge these two. It takes the observed cues as inputs and produces a corresponding QoE score. In a holistic view, our model is superior to the state-of-the-art approaches from two aspects. First, it takes into account human feelings during QoE assessment, which are largely overlooked by prior works. Second, most prior works endeavor to exhaustively enumerate and include all impact factors in QoE modeling, which are impractical to implement in real-world scenarios. Alternatively, EyeQoE merely utilizes eye-based cues to reflect the viewer's perceived QoE as a whole. Extensive experiment results show that it outperforms representative prior works in terms of prediction accuracy.

Despite the attractive sense of exploiting ocular behaviors for 360-degree video QoE assessment, enabling it involves several non-trivial challenges. First, ocular behaviors are affected by external visual stimuli [9, 33, 60] and biologically distinct across human subjects [28, 65]. For instance, a subject's gazing patterns tend to be more static when focusing on a tree than tracking a flying bird in a scene [10, 42, 47]. As human eyes have unique physical characteristics (e.g., sizes, biophysical structures, etc.), ocular behaviors may vary among individuals even watching the same video. Thus, EyeQoE needs to cope with variations introduced by subjects and visual stimuli heterogeneity. Second, because of the intrinsic diversity of visual stimuli, i.e., video clips, the QoE assessment model, once trained over existing videos, may be hard to generalize to unseen videos. To deal with this challenge, a naive approach is to gather as many annotated training samples as possible. It means to cover videos of all kinds, which would lead to considerable overhead.

The proposed EyeQoE is inspired by some advanced techniques in deep neural networks. We first organize observed eye-based cues into a basic graph, where fixations and saccades are its nodes and edges, respectively. They are connected in chronological order. The constructed graph preserves the visual patterns of the raw data in the temporal domain through modeling the local pairwise relation between adjacent fixations and saccades. We notice that high correlations also exist among fixations associated with the same object of interest in the scene, though they may be separated in the timeline. We thus extend the basic graph by adding additional edges between nodes of high similarity to preserve the content-dependent features. To facilitate learning over graph-structured data, the core of EyeQoE adopts a graph convolution network (GCN) based classifier. GCN is a superior network to produce useful feature representations of nodes and edges from graphs. In this work, it runs over every fixation and saccade and aggregates their layer-wise representation with those of its neighbors. The useful features accumulate and propagate throughout the entire graph as the convolution evolves. The output of the GCN classifier is a QoE score of the given video clip.

To tackle the challenge of subjects and visual stimuli heterogeneity, we enhance the GCN classifier by applying a Siamese network framework with devised training sample selection strategies. The idea of the Siamese network is to employ a pair of substructures with the same GCN and weights. The selected pair of samples are passed through the two substructures separately. The distance metric between two outputs is computed and guides the updates of both substructures. The designed structure, together with the training process, allow the model to tolerate inconsistency in ocular behaviors caused by heterogeneous subjects and visual stimuli. To accommodate unseen videos, we formulate our problem as *domain adaption*. We first categorize all 360-degree videos into various types according to their *colorfulness*, *luminance*, and *motion*. Datasets associated with existing and unseen videos are treated as the source domain and the target domain, respectively. Hence, our problem involves multiple source domains. We then propose a multi-source adversarial domain adaptation (MADA) network based on the classic domain adaptation network [29] that is originally designed for single-source-domain scenarios.

The discussion of this work pertains to PC-tethered VR[1] (e.g., HTC VIVE, Oculus Rift, MS MR) and powerful standalone VR, both with the necessary computing capacity to carry out online inference and domain adaption. The QoE model is first trained offline, say, at servers or cloud, and then transferred to VR devices, while the prediction is carried out in an online manner.

We highlight our contributions of this paper as follows:

- We introduce EyeQoE, a novel QoE assessment model for 360-degree videos using eye-based cues. We then construct the cues into a graph that preserves both features in the temporal domain and content dependency.
- We develop a GCN-based classifier to facilitate learning over graphs. The classifier is then combined with a Siamese network to deal with subjects and visual stimuli heterogeneity. MADA is further proposed to easily adapt our model to unseen videos.

[1]Tethered VR means that the headset is physically connected to a computer by cables, such as HDMI and/or USB.

- We build our own dataset via a three-month data collection campaign. 50 volunteers and 5 student workers get involved. To our knowledge, it would be the first data source of annotated ocular behaviors for 360-degree video QoE assessment.
- We carry out extensive tests to evaluate EyeQoE based on our dataset. Results indicate that EyeQoE achieves the best prediction accuracy of 92.9%.

The rest of this paper is organized as follows. Section 2 reviews prior works related to our topic. Section 3 introduces some necessary background of using ocular behaviors for QoE assessment. A measurement study that validates the feasibility of our idea is presented in Section 4. The novel graph modeling of eye-based cues is introduced in Section 5 followed by Section 6 that provides design details of EyeQoE. We evaluate EyeQoE in Section 7. A discussion over the limitations of EyeQoE is provided in Section 8. We conclude the paper in Section 9.

2 RELATED WORK

Video-centric Models. Like conventional videos, some existing QoE assessment models for 360-degree videos directly analyze the displayed videos. QoE is derived by comparing distortions of the displayed video with its original version. This kind of approach is called video quality assessment (VQA). For 360-degree videos, new VQA metrics have been investigated [14, 67, 74, 80, 81, 85]. For example, built upon peak-signal-to-noise ratio (PSNR), a commonly adopted VQA metric for traditional videos, Yu *et al.* [80] modified it into sphere PSNR (S-PSNR) by further considering the impact of the so-called *sphere-to-plane mappings*. Basically, pixels would be distorted when projected from a two-dimensional plane to spherical surface. Sun *et al.* [67] took into account the projection distortion in their VQA metric by multiplying a weight to each pixel that reflects the relation between the sphere and the plane. In the above works, the calculation of VQA metrics is in need of the reference 360-degree videos, i.e., the original version without distortion. Unfortunately, this assumption is impractical in most real-world video streaming scenarios. To overcome the limitation, QoE assessment models with no reference videos have been developed [20, 26, 64, 66]. VQA metrics are directly derived from the features of impaired videos or network parameters, e.g., bandwidth, packet loss, and latency. Nonetheless, video-centric models are criticized for overlooking viewer's perceptive feelings during QoE assessment, such as immersiveness [30, 86] and cybersickness [35, 38]. As validated through many prior works [7, 32, 50], viewer's subjective experience of watching videos does not necessarily comply with their displayed qualities in many cases.

Visual attention enhanced models. Recently, some works start introducing human factors to QoE assessment of 360-degree videos. In an immersive environment, people cannot see the whole scene from a single viewpoint. Instead, they usually look around and focus on what attracts them. Hence, distortions on different parts of the projection sphere impose a nonuniform impact on QoE. With the basis of the traditional PSNR metric, Xu *et al.* [75] assigned weights on the pixel-wise distortion in calculating the PSNR according to the distribution of the viewer's visual attention. A similar idea is adopted by VQA-OV [43]. Visual attention is generated by tracking the viewer's head and eye movements via the embedded inertial sensors and eye tracker in a VR headset. In [44], they further constructed the subject's field of view (FoV) and saliency map to guide VQA assessment. The strategy of using visual attention or saliency map to boost the video-centric QoE models has also been adopted in the context of conventional videos [25, 41, 46, 75]. As a note, all the above works are still in need of reference videos to calculate pixel-wise distortions. Although these works utilize visual information in their models, it is essentially subject's visual attention. Instead, our work exploits physiological features in viewer's ocular behaviors to infer her satisfaction in watching 360-degree videos. Therefore, our problem formulation and the corresponding inference technique are totally different.

Among the prior works, [57] is the closest to ours. It combines facial expression and gaze direction for traditional video QoE assessment. Our work differs in two main aspects. First, we target 360-degree videos in VR

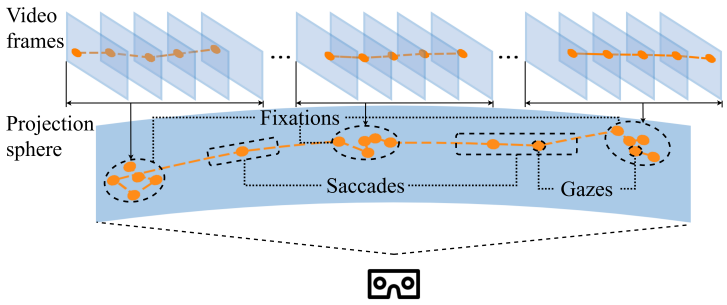


Figure 1. Relations of gazes, fixations, and saccades when a viewer is watching a 360-degree video.

environments while they are for traditional videos. Second, our work addresses critical challenges in data-driven QoE modeling, such as subjects and visual stimuli heterogeneity and adaption to unseen videos. These issues are overlooked in [57].

Some other works investigate the feasibility of leveraging human behavior related data, such as heart rate, facial expression, electrodermal activity (EDA), and electroencephalogram (EEG), to evaluate QoE on various VR applications, including assistive technique systems [61], speech and language assessment applications [36], and general-purpose applications [12, 24]. None of them is designed for 360-degree videos. Besides, to our knowledge, no existing commercial VR headset nowadays is equipped with necessary sensors to acquire these human behavior data.

3 BACKGROUND

Eye-based cues as indicators of human perceptions. A connection between the eye-based cues and human perceptions has been accepted for a decade [27, 51, 82]. Such cues include eye gaze, fixations, saccades, pupillometry, and various forms of eye opening and closure events. In neurophysiological literature, it is demonstrated that pupils are unconsciously regulated by autonomic nervous system stimulation, which is known to produce responsive output under numerous emotional states. Hess [31] reported behavior changes in subjects who view image stimuli that cause different pupil sizes; images with dilated pupils are deemed more attractive than those with constricted pupils. Eye blinks and gaze behaviors are treated as crucial indicators for visual fatigue, defined as eyestrain or asthenopia, which can be caused by both two-dimensional and stereoscopic moving images [27]. Studies show that eye blinking rates increase due to a prolonged period of time working in front of video display terminals. The exacerbated drying of the ocular surface causes subjects to blink more frequently to lubricate the surface of the cornea and conjunctiva [37, 72]. Prior works also demonstrate strong correlations in visual fatigue versus saccade peak velocity, saccade duration, and fixation duration [82]. Specifically, saccadic oculometrics, saccade peak velocity, and saccade duration significantly decrease as working time progresses, whereas the duration of medium-length fixations increases with fatigue development. All these findings motivate us to exploit eye-based cues to infer human perceived QoE toward 360-degree videos.

Gazes, fixations, and saccades. Saccades are rapid stepwise movements of both eyes in the same direction that typically last 10-100 ms, depending on the distance covered [23]. They are used to shift the gaze to another location. In contrast to saccades, fixations are relatively focused, low-velocity eye movements with a typical duration of 100-400 ms and are used to stabilize the retina over a stationary object of interest. A visual gaze is the instantaneous visual point landing on the stimulus. A fixation consists of multiple time-series gazes concentrated around the same viewpoint. As shown in Figure 1, as a subject watches a 360-degree video, her fixations move

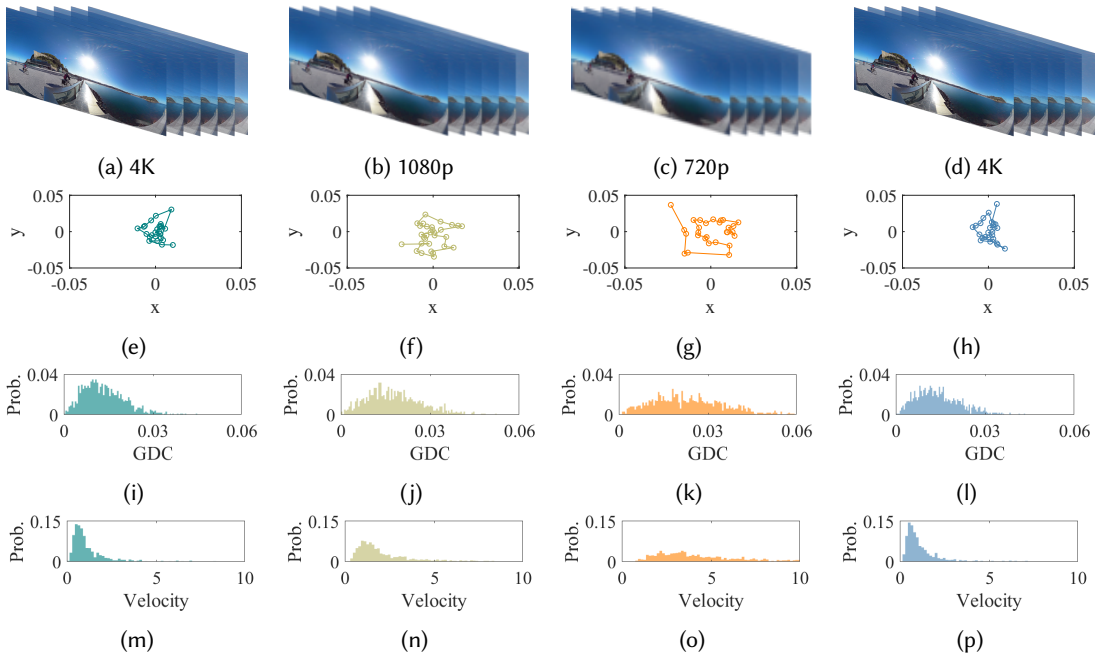


Figure 2. (a)-(d): 360-degree videos in resolutions of 4K, 1080p, 720p, and 4K (same subject in a second trial). (e)-(h): Normalized coordinates of gazes in one fixation. (i)-(l): Distribution of GDC. (m)-(p): Distribution of gaze velocity.

over the projection sphere in accordance with the object of interest. Each fixation is associated with a series of frames that typically display a similar scene, in which the location of objects of interest is basically unchanged.

4 MEASUREMENT STUDY

While the correlation between eye-based cues and human perceptions is well recognized, whether the former can serve as an indicator for 360-degree video QoE is unclear. Our measurement study intends to answer this question by carrying out extensive experiments. A total number of 10 subjects are invited to watch 360-degree videos of different qualities via the HTC Vive headset. Each video is of 25 seconds duration. Subjects' ocular behaviors are captured by a Pupil Labs eye tracker that is integrated into the headset. We then examine how they are influenced by various well- recognized impact factors of 360-degree video QoE, including video quality, cybersickness, immersiveness, and fatigue.

Observation 1: Eye-based cues are impacted by video quality. Figure 2 exhibits the impact of video resolutions to ocular behaviors. Figure 2e-2g show coordinates of time-series gazes from one fixation with the image resolution of 4K, 1080p, and 720p, respectively. In these figures, the origin is the fixation center and the x-/y-coordinate of each gaze is its horizontal/vertical distance to the center. For fair comparison, we extract the fixations on the same object across the three videos. We find that gazes are more focused when the video is in a higher resolution. This phenomenon is further validated through Figure 2i-2l where the probabilistic distribution of gaze distance-to-center (GDC) is displayed. GDC mainly concentrates on the lower end of the x-axis, mostly lower than 0.03 for 4K videos. It becomes scattered as the resolution decreases. We have a similar observation over the gaze velocity in Figure 2m-2p; eye movements within a fixation tend to slow down when watching a high-quality video, whereas they become faster as the quality is degraded.

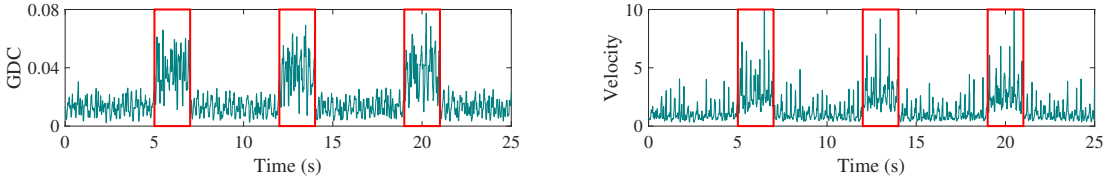


Figure 3. (a) The impact of video stalling on viewer's GDC. (b) The impact of video stalling on viewer's gaze velocity.

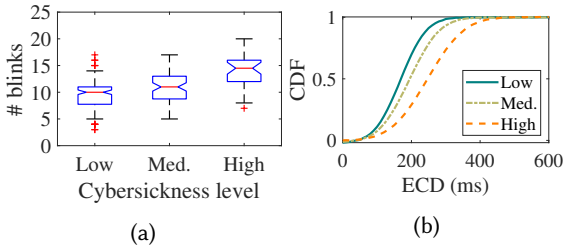


Figure 4. Impact of cybersickness. (a) Number of blinks observed in 25 seconds (sample video duration) under various cybersickness levels. (b) The CDF of ECD under various cybersickness levels.

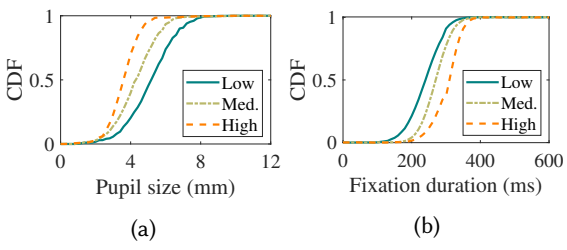


Figure 5. Impact of fatigue. (a) The CDF of pupil sizes. (b) The CDF of fixation duration.

Apart from the spatial distortion, we also explore the impact of the video's temporal distortion with stalling events in the video. Figure 3a shows the GDC of each observed gaze as time proceeds. There are three surges in GDC at the 5th, 12th, and 19th second, which are exactly time instances of the stalling events. It implies that visual attention becomes less focused as stalling occurs. As indicated in Figure 3b, gaze velocity also experiences significant increases as the video freezes.

Observation 2: Eye-based cues are impacted by subjective factors. As verified in prior studies [30, 38, 71, 86], aside from the video quality, 360-degree video QoE is also influenced by subjective factors, namely cybersickness, fatigue, and immersiveness. Cybersickness, or motion sickness, refers to the subject's feeling of sickness, dizziness, nausea, etc., caused by, for example, the physical device, the VR environment, video contents, and the subject's physical status. Fatigue describes the subject's tiresome and is mainly impacted by the time duration of watching videos. Immersiveness reflects the subject's perception of being physically present in the VR environment. In the measurement study, subjects are asked to rate their feelings towards cybersickness, fatigue, and immersiveness on a 3-point scale indicating low, medium, and high, respectively, after watching each video.

A correlation is observed between cybersickness and the subject's blink events. Figure 4a shows the number of blinks that a subject performs in watching a 360-degree video of 25 seconds under three cybersickness levels. Subjects tend to exhibit a higher blink rate when experiencing cybersickness. It may be due to more intense eye-strain symptoms, which leads to higher frequent blinks. Meanwhile, the eye closure duration (ECD) in each blink increases with higher perceived cybersickness, as presented in Figure 4b. Our measurement study also reveals viewer's oculomotor and pupillary behaviors as potential indicators of her fatigue. As shown in Figure 5, higher perceived fatigue is associated with shrunk pupil sizes and longer fixation durations. A similar finding in contexts other than VR is reported in prior studies [48, 83].

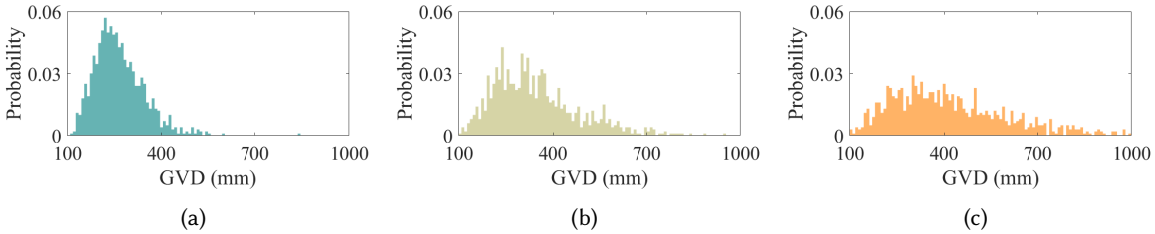


Figure 6. Impact of immersiveness on viewer's GVD. (a) Low. (b) Medium. (c) High.

Lastly, we present the impact of immersiveness. Figure 6 shows the probabilistic distribution of gaze vergence distance (GVD) subject to various immersiveness levels. Specifically, GVD is defined as the distance between the viewer's eyes and the focused object on display. For low immersiveness, the GVD is more concentrated, while it becomes scattered under higher perceived immersion. It indicates that a viewer's visual attention follows objects of interest that may cover a wide range on the sphere under good immersiveness; it tends to stay in the center of the scene as the perception becomes less satisfactory.

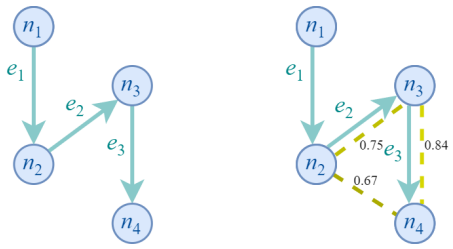
Observation 3: Eye-based patterns are consistent in multiple trials. In the measurement study, we play the same video of the same quality a couple of times to the same subjects and analyze changes in their ocular behaviors. Two trials, as indicated in the first and fourth column of Figure 2, are randomly selected. It is observed that ocular patterns, including but not limited to, the spatial distribution of gazes, GDC, and gaze velocity, are quite similar to each other. This observation implies that our QoE assessment model, once well trained on existing eye-based cues, can be reused over time.

Summary. Our measurement study lays the necessary foundation for the idea of leveraging eye-based cues to infer the subject's QoE in watching 360-degree videos. The findings are encouraging. First of all, we verify the hypothesis that there are strong correlations between viewers' eye-based cues and their perceived experience in watching 360-degree videos. Second, eye-based cues can effectively reflect both objective (e.g., video quality) and subjective (e.g., cybersickness, immersiveness, and fatigue) impact factors of perceived QoE in VR. This property can be achieved neither by the existing video-centric models [14, 26, 67, 74, 80, 81, 85] nor the visual attention enhanced models [43, 44, 75]. Nonetheless, how to perform an accurate QoE assessment based on collected ocular cues is a non-trivial task, which is also the focus of Section 5 and 6 next.

5 MODELING OCULAR BEHAVIORS INTO GRAPHS

The "node-edge" structure of subject's ocular behavior data shown in Figure 1 motivates us to transform them into graphs. In the following, we first introduce a basic version that only captures the temporal structure of eye-based cues, followed by a comprehensive version that further explores content dependencies out of the cues.

A basic version. Consider a time series of gazes captured by a VR headset. They form N fixations ($\boldsymbol{N}$) and thus $N-1$ saccades ($\boldsymbol{E}$). The corresponding basic graph is of N nodes and $N-1$ edges. We denote the graph as $G = \{\boldsymbol{N}, \boldsymbol{E}\}$, where $\boldsymbol{N} = \{n_1, ..., n_N\}$ and $\boldsymbol{E} = \{e_1, ..., e_{N-1}\}$. Each saccade $e_k \in \boldsymbol{E}$ links two fixations $n_k,\ n_{k+1} \in \boldsymbol{N}$. Saccades are directional as they present the chronological order from a fixation to its successor in the temporal domain. As depicted in Figure 1, inside each fixation, there are many gazes. Typically, these gazes reflect the subject's visual attention to the same object of interest in the same scene. Their time-stamped coordinates then serve as part of attributes of the fixation (i.e., node). Additionally, correlations are observed between the subject's pupillary and oculomotor behaviors and perceived video quality as elaborated in Section 4. Hence, time-series pupil sizes and time instances of eyelids open/close events (i.e., blink onsets/offsets) are also treated as part of



(a) (b)

A	n_1	e_1	n_2	e_2	n_3	e_3	n_4
n_1	-	1	0	0	0	0	0
e_1	0	-	1	0	0	0	0
n_2	0	0	-	1	0.75	0	0.67
e_2	0	0	0	-	1	0	0
n_3	0	0	0.75	0	-	1	0.84
e_3	0	0	0	0	0	-	1
n_4	0	0	0.67	0	0.84	0	-

(c)

Figure 7. (a) An illustration of a basic graph, where circles and arrows denote fixations and saccades, respectively. (b) An illustration of a comprehensive graph. Dashed lines represent newly added edges. (c) The adjacency matrix corresponding to the comprehensive graph.

fixation's attributes. For a saccade, its attributes are similar to those of fixations, including coordinates of gazes of that saccade and time-series pupillary and oculomotor features described above.

A comprehensive version. The basic graph only captures local pairwise relationships between the temporally adjacent fixations and saccades; a fixation (saccade) is connected to two adjacent saccades (fixations). Relationships in other domains remain unexplored. In practice, two fixations, even not directly connected by a saccade, may share high similarities in their attributes. We find in our measurement studies that these similar fixations are typically associated with the same object in a video. For example, ocular behaviors when a viewer focusing on a tree are distinct from tracking a flying bird [10, 42, 47]. Thus, we develop a comprehensive graph that preserves both the temporal and content-dependent information in the collected raw data. The comprehensive version creates additional edges between fixations of high similarities on the basic graph. As shown in Figure 7b, 3 new edges (indicated by bidirectional dashed lines) are added. Note that new edges do not have any attribute.

Now the remaining question is how to determine the "similarity" of two given fixations. In this work, we employ the *cosine similarity*, a common measure of similarity between two non-zero vectors. Specifically, the similarity score between two fixations n_i and n_j is calculated as $\theta(n_i, n_j) = (n_i \cdot n_j)/(||n_i||\,||n_j||)$. For expression simplicity, here we use the node index n_i to represent its attribute vector. Given a pre-defined threshold θ_0, an edge is added between n_i and n_j if $\theta(n_i, n_j) > \theta_0$. $\theta(n_i, n_j)$ is then treated as the weight of the new edge. The comprehensive graph is thus a weighted graph. It is possible that attributes of fixations and saccades are of unequal size. To facilitate the learning graphs with unequal attribute size, we employ an *encoding process* that transforms arbitrary-length attributes into fixed-length vectors before passing them into the learning model [68].

6 EYEQOE

In the following, we first present a basic QoE assessment model that learns from the graph-structured ocular behaviors. We realize that the intrinsic heterogeneity of human visual behaviors and the impact of diverse video contents introduce variations to the learning process. In addition, the assessment model, trained on existing video samples, may not be readily applicable to new unseen videos. Thus, the basic model is further extended to deal with these issues.

6.1 A Basic GCN-based QoE Assessment Model

We propose to use GCN neural networks to solve our learning-on-graph problem. GCN is capable of extracting the representation of non-Euclidean graphs using a "convolutional" (neighbor-weight-sharing) kernel [84]. Like other neural networks, a GCN model consists of several layers of neurons; in each layer, higher-level features

are extracted from the input and passed onto the next layer. A GCN model can be designed to classify nodes, subgraphs, or even entire graphs. Aside from GCN, graph neural network (GNN) [21, 45, 62] is another feasible model in handling non-Euclidean characteristics of the complex structure of graphs. We pick the former over the latter due to its efficiency in running backpropagation over time.

Construction of adjacency matrix. We formulate our problem as a graph classification problem, where the classifier takes the comprehensive graph (generated in Section 5) as the input and outputs a QoE score on the scale of 1-5. The input consists of an *attribute matrix* and an *adjacency matrix*. Specifically, an attribute matrix is denoted as $X \in \mathbb{R}^{(2N-1)\times D}$, where $2N-1$ comes from N fixations and $N-1$ saccades, and D is the dimension of their attributes after encoding. Each row is the encoded attributes from a fixation/saccade. An adjacency matrix is denoted as $A \in \mathbb{R}^{(2N-1)\times(2N-1)}$, where each row and column corresponds to a fixation or a saccade. The entries of the matrix indicate whether pairs of elements are adjacent or not in the graph. Take Figure 7 as an illustration. Since n_1 is linked to e_1, then $A_{1,2} = 1$. On the other hand, $A_{2,1} = 0$ as e_1 is a directional edge. Assume $\theta_0 = 0.5$. For two fixations n_2 and n_4, their corresponding matrix entries are given by their similarity score: $A_{3,7} = A_{7,3} = \theta(n_2, n_4) = 0.67$ as $\theta(n_2, n_4) > \theta_0$. Denote by v_i a node or an edge, the instantiation rule of the adjacency matrix is summarized as

$$A_{i,j} \in A = \begin{cases} 1 & \text{if } v_j \text{ is the successor of } v_i \text{ in the basic graph,} \\ \theta(v_i, v_j) & \text{if } v_i, v_j \in \boldsymbol{N} \text{ and } \theta(v_i, v_j) > \theta_0, \\ 0 & \text{otherwise.} \end{cases}$$

GCN-based model. Our GCN classifier consists of four convolutional layers followed by a max pooling layer [40]; each layer in this classifier can be written as a non-linear function

$$H^{l+1} = f(H^l, A)$$

where $H^l \in \mathbb{R}^{(2N-1)\times D}$ is the matrix of activations in the lth layer with $H^0 = X$. The model is specified by the $f(\cdot,\cdot)$ function of each layer. We adopt the propagation rule introduced in [40]

$$f(H^l, A) = \rho(\hat{\Delta}^{-1}\hat{A}H^l W^l) \tag{1}$$

where $\hat{A} = A+I$ with I being the identity matrix. $\hat{\Delta}$ is the diagonal node dimension matrix of $\hat{A}$, and $W^l \in \mathbb{R}^{(2N-1)\times D}$ is the weight matrix for the l-th layer. ρ is an activation function, e.g., a ReLU $\rho(x) = \max(0, x)$.

The "convolution" operation in Equation (1) is designed in a way such that a "one-hop" filter runs over every fixation and saccade and aggregates its layer-wise representation with those of its neighbors. Specifically, for each fixation, the filter adds to it the representations of all other fixations, weighted by their similarity scores, and the representation of its neighboring saccade. For each saccade, since it only has one predecessor fixation as its neighbor, it is only updated by taking the representation of that fixation. Then, the aggregated representation is normalized by dividing with the dimension of the representations. One can incorporate higher-order neighborhoods information by stacking multiple GCN layers. Then features are aggregated and propagated iteratively along with the graph. In the final step, the output from the last layer is passed through a max-pooling layer to generate the classification result z as the estimated QoE score for the given 360-degree video.

We adopt the mean squared error as the loss function:

$$\mathcal{L}_G = \frac{1}{N}\sum_{i=0}^{N}(y_i - z_i)^2 \tag{2}$$

where y_i and z_i denote the ground-truth label and the model prediction of the ith sample, respectively, and N stands for the number of training samples.

Table 1. Training sample selection rule. ✓ means the pair is selected and ✗ means the pair is not.

		Same subject	Diff. subjects
Same label	Same video	✗	✓
	Diff. videos	✓	✓
Diff. labels	Same video	✓	✗
	Diff. videos	✗	✗

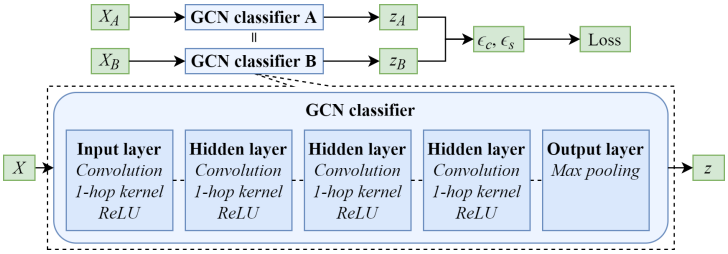


Figure 8. Top: the architecture of the Siamese network. Bottom: the GCN classifier model.

6.2 Dealing with Subjects and Visual Stimuli Heterogeneity

In practice, the training dataset, i.e., labeled eye-based cues, is obtained from a group of subjects for watching various 360-degree videos. In addition to objective and subjective impact factors of perceived QoE (as discussed in Section 4), the subjects and visual stimuli heterogeneity also affects ocular behaviors. As a result, it introduces an additional dimension of uncertainty to the learning process.

Compared with video quality, QoE should be much less relevant to the video content. It means that two videos are expected to produce similar QoE scores given the same quality and other subjective impact factors (e.g., cybersickness, fatigue, immersiveness, etc.), regardless of the contents displayed. In the meantime, video contents highly affect ocular behaviors, the features considered by EyeQoE for QoE assessment. For example, eyes move faster when watching high-motion scenes than the stationary ones. As one of our contributions, this work aims to eliminate the impact of video contents to QoE assessment, as called *visual stimuli heterogeneity*. To alleviate impacts from both subjects and visual stimuli heterogeneity, we modify the basic GCN-based QoE assessment model by applying the Siamese network [17]. Its idea is to employ a pair of substructures with the same architecture and weights. It passes a pair of input data through the two substructures separately, computes the distance metric between the outputs, and updates both substructures simultaneously.

The modified model is shown in Figure 8. It is composed of two identical GCN classifiers introduced in Section 6.1. X_A and X_B stand for the pair of training samples for the two classifiers, respectively. Sample pairs are carefully selected following a scheme as outlined in Table 1. Each pair of samples is classified into one of the four categories based on their subjects, video contents, and labels. If their labels are the same, we select the pairs from different subjects and/or video contents. In this way, the model can learn to tolerate differences in ocular behaviors caused by heterogeneous subjects and video contents, i.e., visual stimuli. In contrast, if their labels are different, we select the pairs from the same subjects and video contents; the model then learns to distinguish samples of similar patterns associated with different labels (i.e., QoE scores). The selected sample pairs are passed through the two twin models separately. We then calculate the distance between two outputs. The loss function of the Siamese

network is defined as

$$\mathcal{L}_S = \sum_{i=1}^{N} \left(\alpha \left(\eta\epsilon_c^2 + (1-\eta)(4-\epsilon_c)^2\right) + (1-\alpha)\left(\eta\epsilon_s^2 + (1-\eta)(4-\epsilon_s)^2\right)\right) \tag{3}$$

where N is the number of sample pairs. ϵ_c and $\epsilon_s \in [0, 4]$ denote the distances between model outputs of a sample pair concerning visual stimuli and subjects, respectively. η is a binary value indicating whether labels of a sample pair are the same ($\eta = 1$) or different ($\eta = 0$). α is a factor to balance the weight between ϵ_c and ϵ_s.

Combining (2) and (3), the final loss function is expressed as

$$\mathcal{L} = \mathcal{L}_G + \mathcal{L}_S + \lambda \cdot ||w||_2^2 \tag{4}$$

where $\lambda \cdot ||w||_2^2$ serves as a regularization term. In the training process, the final loss $\mathcal{L}$ is fed back into the network to update the weights.

6.3 Dealing with Unseen Videos

As discussed, the characteristics of the video scenery being displayed also impact the viewer's ocular behaviors. Hence, the QoE assessment model, trained over existing video clips, may not be readily scalable to an even broader set of unseen videos, especially of different characteristics. A conventional approach is to gather as many annotated samples as possible to train the model. In our case, it requires covering videos of all kinds, which would incur prohibitively expensive overhead in data collection. Alternatively, we propose to employ *domain adaptation* [58]. Under this framework, existing videos and new videos are treated as the *source domain* and the *target domain*, respectively. The domain adaption technique aims to fine-tune parameters of models trained in the source domain to adapt to new circumstances in the target domain. While this technique has been widely adopted in the context of computer vision [19], sentiment analysis [55, 69], and action recognition [16, 49], whether it is effective in 360-degree video QoE assessment is unexplored yet.

Video type categorization. To facilitate the employment of domain adaption, we first categorize all 360-degree videos in various types[2] according to their *colorfulness*, *luminance*, and *motion*. Existing methods are available to obtain the above information by inspecting *I-frames* and *P-frames* in videos [1, 54, 63]. As these computations do not involve any sophisticated operations, they can be accomplished within dozens of milliseconds in a computer with moderate settings. Assume that the entire video space is divided into κ types. κ plays an important role in the performance of EyeQoE. We will examine its value selection in Section 7.3.

Domain Adaptation Each video type is treated as a domain. Assume that the training videos cover K ($K < \kappa$) domains $\mathcal{D}_S = \{\mathcal{D}_S^1, ..., \mathcal{D}_S^K\}$. The target domain that an unseen video falls into is denoted as $\mathcal{D}_T$. We propose a multi-source adversarial domain adaptation (MADA) network. It is inspired by the classic domain adaptation network introduced in [29] but further extended to scenarios of multiple source domains as in this work. As a note, the classic domain adaptation network is originally designed to deal with single-source-domain scenarios and thus not readily applicable here.

The architecture of MADA is illustrated in Figure 9. To fine-tune the trained GCN classifier, MADA takes as inputs the samples from a specific target domain $\mathcal{D}_T$ and a set of samples from each source domain $\mathcal{D}_S^k$ ($k \in [1, K]$). MADA is constructed based on the GCN classifier with four main modules: feature extractor, label predictor, domain predictor, and loss scaler. The feature extractor, together with the label predictor, assemble the same components of the GCN classifier introduced above. Specifically, the feature extractor is comprised of the first four convolutional layers of GCN. The label predictor is simply the output layer, i.e., the max-pooling layer (Figure 8). Given any graph presentation X and A, the above two modules generate a prediction label z. The

[2]In this work, we assume that each 360-degree video clip is of one type without significant scene changes. For longer videos in multiple scenes, they can be divided into multiple segments, each in one scene. We then apply our model to each segment sequentially. The final QoE can be calculated as the aggregated QoEs of all segments.

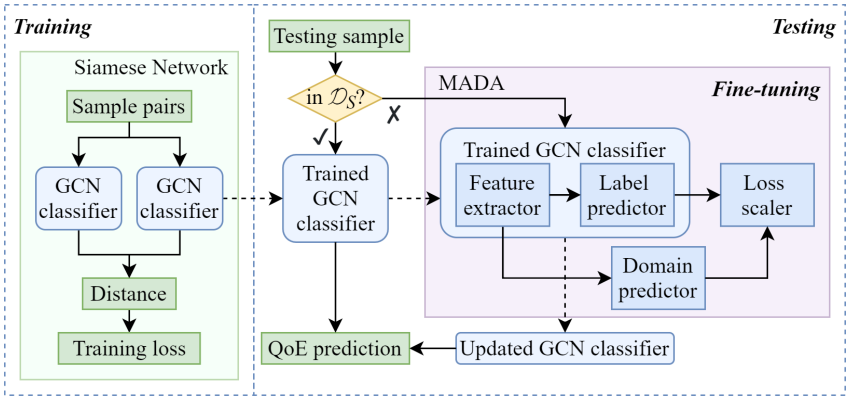


Figure 9. The overall architecture of EyeQoE.

domain predictor works in an adversarial way. With the high-dimensional features as the input, it aims to decide if the given graph presentation belongs to a source domain or a target domain. Ideally, the domain predictor, once properly trained, cannot distinguish between them. It indicates that our model's inference performances over existing videos and unseen videos are almost the same. The loss scaler computes the loss of label prediction and domain prediction and aggregates them into the final loss value $\mathcal{L}_{tot}$

$$\mathcal{L}_{tot} = \sum_{k=1}^{K}\left(\frac{\gamma^k}{n^k}\sum_{i=1}^{n^k}\mathcal{L}_y^{k,i}\right) - \lambda\left(\sum_{k=1}^{K}\left(\frac{\gamma^k}{n^k}\sum_{i=1}^{n^k}\mathcal{L}_d^{k,i}\right) + \frac{1}{n^{K+1}}\sum_{i=1}^{n^{K+1}}\mathcal{L}_d^{K+1,i}\right).$$

Here $\mathcal{L}_y^{k,i}$ and $\mathcal{L}_d^{k,i}$ ($k \in [1, K]$) stand for the label prediction loss and the domain prediction loss over a sample from source domain $\mathcal{D}_S^k$. $\mathcal{L}_d^{K+1,i}$ stands for the domain prediction loss over a sample from the target domain $\mathcal{D}_T$. n^k and n^{K+1} represent the number of samples of $\mathcal{D}_S^k$ and $\mathcal{D}_T$, respectively. K is the total number of source domains. λ is a parameter that controls the balance between label prediction loss and domain prediction loss. γ^k stands for the similarity between $\mathcal{D}_S^k$ and $\mathcal{D}_T$. It is calculated as the *cosine similarity* between content metrics of videos from these two domains. The content metrics include colorfulness, luminance, and motion as mentioned above. Basically, two domains that share a higher similarity in their videos tend to exhibit similar prediction performance through a trained model. Hence, the prediction loss of each source domain contributes to the total loss with a different weight determined by γ^k: A source domain of a larger γ^k has a more prominent impact.

In the MADA network, the GCN classifier is initiated with parameters derived from the offline training phase, whereas parameters of the domain predictor are set as random values. MADA is triggered with the arrival of an unseen video out of the source domains. The trained GCN classifier is then fine-tuned through multiple rounds of iterations, where backpropagation is performed and all weights are updated through the gradient descent algorithm. We will examine in Section 7.3 with details regarding the efficiency of the fine-tuning process.

6.4 Piecing All Together

Figure 9 outlines the overall architecture of EyeQoE. The core component is a GCN classifier designed to infer the QoE score given the subject's graph-structured eye-based cues. To handle the issue of subjects and visual stimuli heterogeneity, we enhance our GCN classifier with a Siamese network which consists of two identical GCN classifiers. Sample pairs are carefully selected and used to train the classifier. Almeida-Pineda algorithm [53], a gradient-based optimization method, is adopted. In the testing stage, given a new sample, EyeQoE first

Table 2. Summary of the video set.

Video	Category	Color.	Lumin.	Motion	Projection	Source
Bar	PB	High	Low	Med.	ERP	Vimeo
Boat	FA	High	Med.	Med.	ERP	Vimeo
Bunnies	FA	High	Med.	Low	EAC	YouTube
City	TE	High	High	Low	ERP	Vimeo
Dance	En	Med.	High	Med.	EAC	YouTube
Girl	PB	Med.	Low	High	ERP	Vimeo
Lions	En	High	High	Med.	EAC	YouTube
Ski	S	Low	High	Low	ERP	Vimeo
Snowmobile	S	Low	High	High	ERP	Vimeo
Waterfall	En	High	Low	Low	ERP	Vimeo

Table 3. Participant demographic information.

Gender	#	Age	#	Eye color	#	Eye wear	#	Experience	#
Female	21	18-23	26	Brown	33	None	19	No	34
Male	28	24-29	13	Blue	6	Glasses	22	Yes	16
N/A	1	30-35	9	Hazel	3	Colorless contact	7		
		>35	2	Other	8	Colored contact	2		

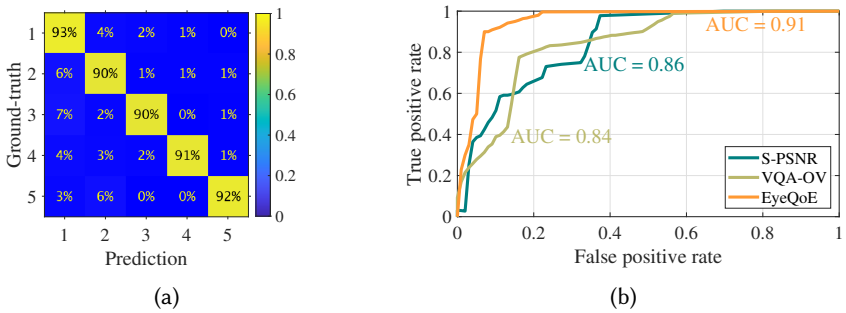
examines if it belongs to any of the source domains. If yes, it indicates that the corresponding video type has been covered during training. Hence, the trained GCN classifier is applied directly for QoE inference. Otherwise, the video is deemed from the target domain. Then our proposed MADA is applied to fine-tune the GCN classifier with the new sample. Finally, the QoE is derived by feeding the sample into the updated classifier.

7 EVALUATION

7.1 Settings

Experiment setup. We implement EyeQoE on a PC running Windows 10 operating system. It is equipped with an Intel Core i7-7820X processor and GeForce RTX 2080 graphic cards. An HTC Vive Pro VR headset is used to provide the VR environment and render videos to subjects. A Pupil Labs eye tracker is embedded inside the VR headset to capture subjects' eye movements. The VR headset is connected to the PC via a USB cable. EyeQoE is implemented using the Keras 2.3.0 library built on top of the TensorFlow 2.0 framework. The Adam optimizer [39] is employed for optimizing the training process.

Dataset. All source videos are downloaded from two major platforms of 360-degree videos, YouTube and Vimeo. The original version is of 4K resolution and 25 fps frame rate. The videos cover a wide range of genres, such as nature, sports, and city view. To facilitate the experiment, each video is of a 25-second duration without significant scene changes. Each source video is subject to two types of distortions, including resolution degradation and stalling. For the former, we use the JM reference implementation of the H.264 scalable video codec (SVC) to compress the 4K original videos into lower resolutions such as 2K, 1080p and 720p. For the latter, we add freeze frames to simulate stalling in three different versions: 8 stalls each lasting 1 second, 4 stalls each lasting 2 seconds, and 2 stalls each lasting 4 seconds.



(a) (b)

Figure 10. Overall performances of EyeQoE and the comparison with two existing QoE models. (a) Confusion matrix of EyeQoE's predictions. (b) ROC curves of different approaches.

A data collection campaign is conducted over three months. 50 subjects are recruited. They are from a university in the United States, most of them are international students from multiple different countries and nations. Table 3 summarizes the demographic information of the participants. The diversity is observed in the gender, age, eye color, eye wear, and VR experience. They are asked to wear a VR headset to watch 360-degree videos of different qualities and give a score from 1 to 5 that best describes their experience after watching each video. Original, uncompressed reference videos are randomly placed amongst the set of videos shown, although the subjects are unaware of their presence. The score that subjects give these references is representative of the bias that the subject carries. By subtracting the reference video scores from those for the distorted videos, the biases are compensated for yielding differential scores for each distorted video. We divide the data collection into two separate sessions, each lasting no more than one hour, to avoid the discomforts caused by watching the immersive videos too long. The interval between two sessions is at least 24 hours. We further implement a UI via Unity, the most widely used VR development platform, to facilitate the data collection.

We did a literature review over the existing open-sourced datasets. As none of them meets our need, we decided to collect our own dataset. We have now publicized it on https://github.com/MobiSec-CSE-UTA/EyeQoE_Dataset.git.

7.2 Overall Performance

Figure 10a exhibits the confusion matrix of EyeQoE's prediction results. Rows represent the ground truth from 1 to 5, whereas columns represent the prediction results. Values on the diagonal are the success rate, i.e., the percentage of predicted results that EyeQoE gets right. The result is promising as the success rate is above 90% for all QoE values. Besides, we observe that EyeQoE achieves slightly better performance when predicting low and high QoE scores (1 and 5). It may be attributed to the fact that users generally perform well in distinguishing between the best- and worst-quality videos, while the boundaries for the medium ones tend to be vague in labeling.

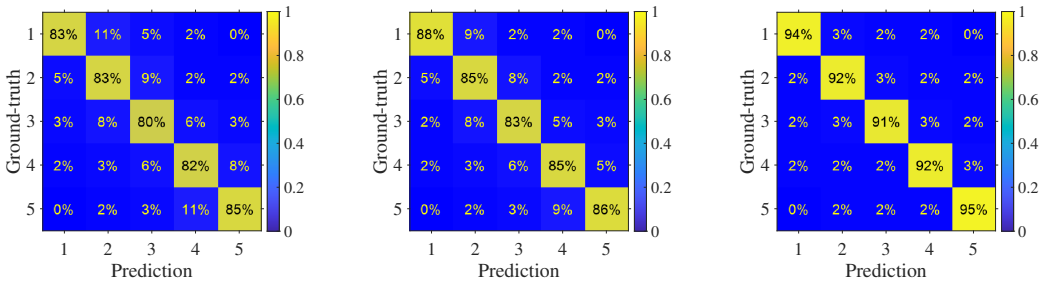
Comparison with state-of-the-art. We compare the performance of EyeQoE with two state-of-the-art solutions for 360-video QoE assessment: S-PSNR [80] and VQA-OV [43]. S-PSNR is a video-centric model; it is built upon the classic PSNR model but further takes into account the pixel distortion issue in projection. VQA-OV belongs to the human factor incorporated model; its main idea is to assign weights on the pixel-wise distortion in calculating the PSNR, where the weights reflect the subject's visual attention on the video.

The ROC curve for each model is depicted in Figure 10b. It is a classic metric to see how a model balances between true positives and false positives. Ideally, the model is expected to have a steep ROC curve to deliver an

accurate inference. Clearly, EyeQoE outperforms the other two with the largest AUC (area under the curve) of 0.91. In comparison, those for S-PSNR and VQA-OV are merely 0.86 and 0.84, respectively. S-PSNR and VQA-OV fail to counter critical factors, such as cybersickness, immersiveness, and fatigue, in QoE assessment. In contrast, rather than exhaustively enumerating and considering all possible impact factors for QoE assessment, EyeQoE leverages ocular behaviors as an indicator to reveal the subject's perceived QoE.

Advantage of GCN-based model in QoE assessment. We further compare the accuracy performance between the GCN + Siamese network and prior works, S-PSNR and VQA-OV. Particularly, the GCN + Siamese network is an ablation version of EyeQoE by removing the domain adaption component. Since none of the above models include domain adaption, the performance should demonstrate the superiority of our GCN-based design. Figure 11 shows the confusion matrices produced by each approach. Apparently, GCN + Siamese yields the best performance among the three. Its diagonal line has larger values, meaning more accurate assessments are produced. For all QoE values, GCN + Siamese maintains a success rate above 91%, whereas the S-PSNR and VQA-OV acquire much lower success rates, ranging from 80% to 88%.

The reasons that the proposed model outperforms S-PSNR and VQA-OV can be summarized as follows. First, our method leverages ocular behaviors, which are neglected by state-of-the-art designs; these cues offer valuable information of a user's QoE as validated in Section 4. Second, by applying GCN on graphs formed by fixations and saccades, we are able to exploit the temporal dependency and content dependency of the eye-based cues by inspecting temporal adjacent and similar activities. The "node-link" structure of the irregular non-Euclidean graphs implies that only graph learning techniques are suitable to explore these dependencies. Third, the Siamese network used during training automatically extracts the most relevant features and eliminates subjects and visual stimuli heterogeneity.

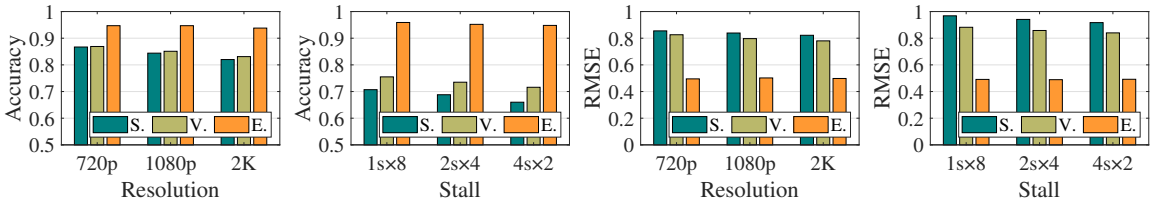


(a) S-PSNR (b) VQA-OV (c) GCN + Siamese

Figure 11. Advantage of GCN-based model - confusion matrix.

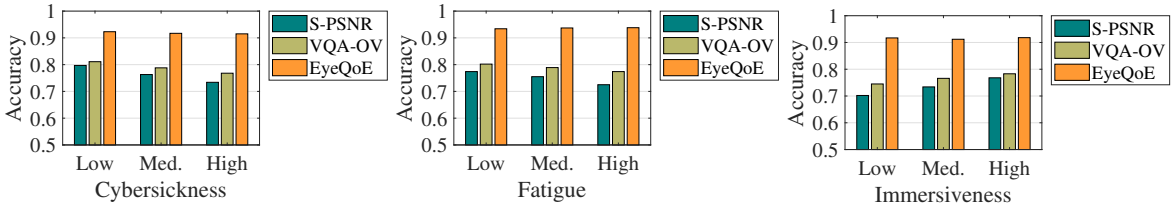
Performance over videos of different distortions. We further investigate the efficacy of EyeQoE over 360-degree videos of various distortions in Figure 12. Two kinds of distortions are examined, resolution and stalls. Figure 12a shows the assessment accuracy by varying the resolution from 720p to 2K. The accuracy of EyeQoE is all above 0.928. Besides, the performance variance under different settings is almost unnoticeable. This is the same case in Figure 12b-12d. Hence, EyeQoE delivers consistent performance for videos of various distortions. EyeQoE outperforms the other two schemes in all cases, especially the stalling distortion. Recall that S-PSNR and VQA-OV measure video QoE through pixel distortions and are thus incapable of reflecting video quality degradation caused by stalling events.

Impact of subjective factors. Now we evaluate EyeQoE's performance subject to cybersickness, fatigue, and immersiveness. Results are illustrated in Figure 13. EyeQoE exhibits high accuracy across various conditions. It



(a) Accuracy vs. resolutions. (b) Accuracy vs. stalls. (c) RMSE vs. resolutions. (d) RMSE vs. stalls.

Figure 12. Impact of distortion types on prediction performances. S: S-PSNR; V: VQA-OV; E: EyeQoE.



(a) Accuracy vs. cybersickness. (b) Accuracy vs. fatigue. (c) Accuracy vs. immersiveness.

Figure 13. Impact of cybersickness, fatigue, and immersiveness on prediction performances.

implies that eye-based cues serve as effective indicators of viewer's perceived QoE. Besides, EyeQoE outperforms the other two models, S-PSNR and VQA-OV, by a clear margin. As discussed, neither S-PSNR nor VQA-OV considers the above subjective factors in QoE modeling. It also explains why their performances become even worse under a high level of cybersickness, fatigue, and immersiveness.

Handling longer videos. EyeQoE is designed in the following way to accommodate longer videos. First, if a video contains multiple scenes, it is divided into several segments, each having one scene. In this way, we obtain S segments of the target video. Then, the subject's eye-based cues during each segment are structured as one graph and fed into the trained model. The QoE for that segment is thus derived. To aggregate the QoE's from S segments, previous works apply either uniform averaging (e.g., [70]) or weighted averaging (e.g., [22, 76]). EyeQoE follows the latter one, where the overall QoE of the video is a weighted average of the QoE for each segment as follows:

$$Q_{total} = \frac{\sum_{i=1}^{S} w(i) Q_i}{\sum_{i=1}^{S} w(i)} \tag{5}$$

where Q_i is the QoE output for the i-th segment and $w(\cdot)$ stands for the weight determined by the segment duration and the subject's memory factor. The rationale behind the second design is that a subject's perceived experience over segments rendered later contributes more to the overall QoE [8, 22]. In this way, the temporal dependencies are preserved within each segment.

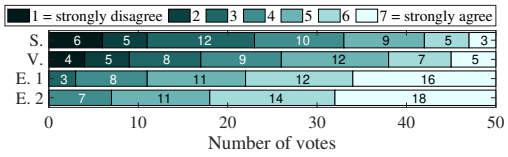
To further evaluate EyeQoE's performance on longer videos with frequent scene changes, we use 5 long 360-degree videos from YouTube. Table 4 lists the duration and scene rate (number of scenes per minute) of these videos as well as the corresponding performance of EyeQoE. We observe that the video duration does not affect much on EyeQoE's performance. However, as the scene rate increases, the overall performance experiences slight degradation with lower accuracy and higher RMSE. Since a long video is divided into multiple segments each with one scene, a higher scene rate thus leads to segments with shorter duration. Hence, the number of features extracted would be reduced, which in turn affects the performance of EyeQoE.

Table 4. Performance on long videos.

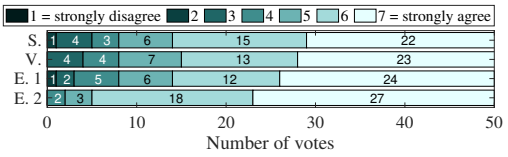
Video	Duration (min)	Scene rate	Seg. count	Accuracy	RMSE
City view	4:00	10.00	8	0.88	0.77
Coaster	5:32	0.72	12	0.93	0.39
Crime scene	22:24	0.76	48	0.90	0.32
Haydee	2:01	2.98	6	0.90	0.79
Viking village	2:09	0.47	4	0.95	0.45

Table 5. EyeQoE's performance on different video categories.

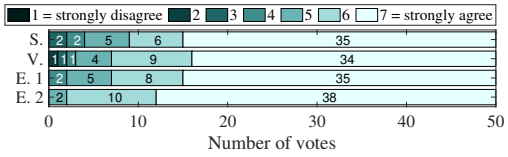
Domain	HHL	HML	LHL	...	LLH	MLM	HLH
Accuracy	0.94	0.93	0.93	...	0.92	0.91	0.91
RMSE	0.49	0.50	0.50	...	0.54	0.57	0.57



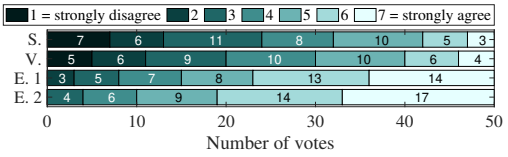
(a) S1. The result is accurate and meets my perceived QoE score.



(b) S2. I feel physically comfortable without dizziness or sore eyes caused by this model.



(c) S3. This model does not interrupt or distract my experience of the video watching.



(d) S4. I would like to have this model to rate QoE scores for me for practical use.

Figure 14. Survey results (S. = S-PSNR, V. = VQA-OV, E. 1 = EyeQoE before the experiment. E. 2 = EyeQoE after the experiment).

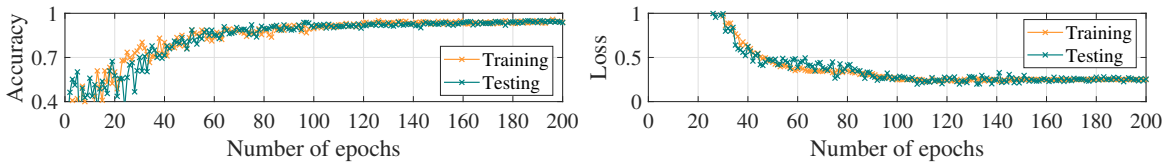
7.3 Micro Benchmarks

Impact of the training ratio. The impact of the training ratio on the performances of EyeQoE is analyzed. As presented in Table 6, the performance is enhanced steadily as the size of the training dataset increases. It indicates that EyeQoE has robust data scalability. Meanwhile, the performance improvement becomes marginal as the ratio surpasses 60%.

Impact of training epochs. To determine whether the model has been trained properly, we monitor the training process in Figure 15. Figure 15a shows the accuracy with respect to the number of epochs. Note that one epoch is when an entire training dataset is passed both forward and backward through the model once. The accuracy quickly increases to 0.90 and becomes converged after around 80 epochs. Figure 15b plots the loss value, another indicator of whether the model is properly trained. It is considered as the "price" paid for assessment

Table 6. EyeQoE's performance regarding the training ratio.

Training ratio (%)	10	20	30	40	50	60	70	80
Accuracy	0.71	0.83	0.85	0.86	0.89	0.93	0.93	0.93
RMSE	1.07	0.90	0.77	0.61	0.60	0.51	0.52	0.50

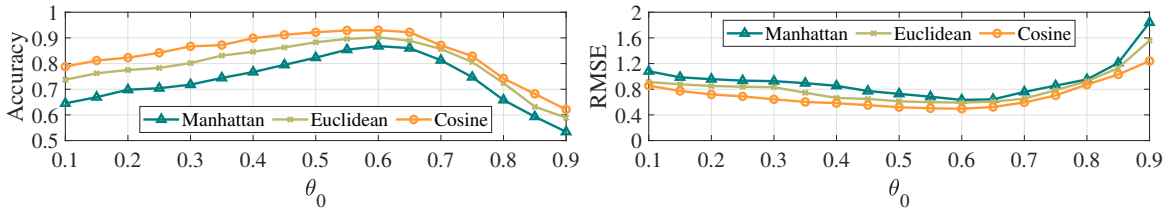


(a) Accuracy vs. number of epochs (b) Loss vs. number of epochs

Figure 15. Training and testing performance over different number of epochs.

inaccuracy. As shown, loss tends to be stable after 100 epochs. Combining the results above, it is sufficient to set 100 epochs for training in our case.

Impact of graph construction metrics. Now we evaluate the performance of EyeQoE given different graph construction metrics. To construct a comprehensive graph, similarity is computed between any two fixations to decide if an edge is added. We employ three different similarity metrics: Manhattan similarity, Euclidean similarity, and cosine similarity. They are classic metrics widely adopted for graph modeling [11]. We also examine the impact of threshold θ_0. Recall that an edge is added if $\theta > \theta_0$. A shown in Figure 16, cosine similarity leads to the best overall performance among the three similarity metrics. We also find that EyeQoE achieves its best performance with accuracy = 0.93 and RMSE = 0.50 at $\theta_0 = 0.6$. Basically, a too-large value of θ_0 would fail to exploit content-dependency between fixations, while a too-small value would introduce unnecessary noise to learning.



(a) Accuracy vs. θ_0 (b) RMSE vs. θ_0

Figure 16. Impact of similarity metrics and θ_0's on prediction performances.

Performance of domain adaption. Next we evaluate the performance of EyeQoE on domain adaptation. The fine-tuning process is executed via the proposed MADA with the arrival of an unseen video. The impact of domain space κ is examined. Recall that κ represents the total number of domains, i.e., video types under consideration. In the experiment, three values are adopted $\kappa \in \{8, 27, 64\}$. They are derived by dividing the space of video content metrics, i.e., colorfulness/luminance/motion, into 2, 3, and 4 levels, respectively ($\{8, 27, 64\} = \{2^3, 3^3, 4^3\}$). In the setting, n_T is equal to 0, 5, 10, and 15. Particularly, $n_T = 0$ means the trained model (over existing samples) is directly applied to an unseen video, while $n_T = 5$ means 5 samples in the target domain are used to fine-tune the

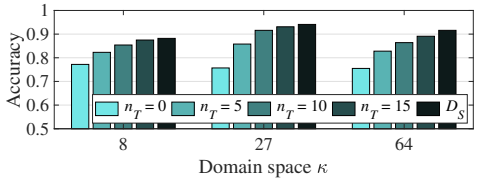
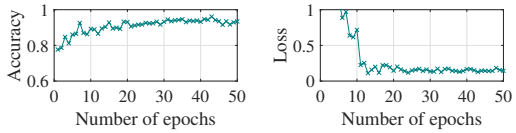
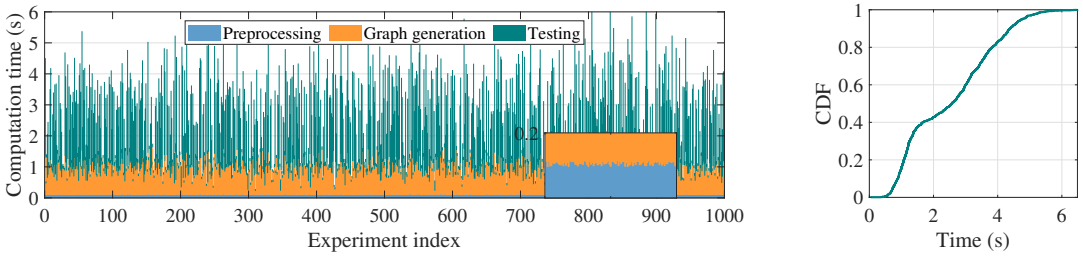


Figure 17. Performance of domain adaption with different configurations.



(a) Accuracy vs. epochs (b) Loss vs. epochs

Figure 18. Convergence of MADA with respect to training epochs.



(a) Stacked computation time.

(b) CDF of total computation time.

Figure 19. Computation latency of QoE prediction.

model. For each n_T value, the same number of samples are randomly picked from each source domain to form the inputs alongside the target domain samples. For comparison, we also test the prediction accuracy on source domains, denoted as D_S in Figure 17. This means that the new video belongs to a source domain, and the trained GCN classifier is directly applied without using MADA.

As demonstrated in Figure 17, the best overall performance is achieved when $\kappa = 27$ among the three values. In general, a too coarse categorization, i.e., small κ, would fail to capture the uniqueness of each domain. On the other hand, too fine-grained categorization, i.e., a large κ, would reduce the number of samples in each domain and thus result in over-fitting. Both cases affect the test accuracy. We also investigate the impact of n_T. A larger n_T is found to produce higher accuracy, since more samples allow the model to fine-tune its parameters in more rounds to better adapt to the target domain. Meanwhile, it also implies more videos from the same target domain to collect. Fortunately, the accuracy already reaches 0.92 with $n_T = 10$. We thus claim that EyeQoE can deliver satisfactory prediction performance for unforeseen videos within 10 samples of the same type. Figure 18 shows the fine-tuning process with respect to the number of epochs. Both the accuracy and the loss value become stable after about 20 epochs. It indicates that the domain adaption can quickly converge.

Computation latency. We now examine the computation latency of QoE prediction over one video. All the operations include the preprocessing of eye-based cues, graph generation, and testing (including MADA for domain adaption). Figure 19a gives the stacked computation latency of each operation. Among the three, testing incurs the largest overhead, about 1.51 s on average. It is due to the fine-tuning for domain adaption. The average latency for preprocessing and graph generation is 0.11 s and 0.88 s, respectively. Figure 19b further illustrates the CDF of the total computation latency of one QoE prediction. The average value is 2.5 s, with 90% of measurements lower than 4.2 s. It indicates that a subject's QoE score can be derived shortly, in a couple of seconds, after a 360-degree video is finished displaying. This duration is comparable to that from the prevalent QoE collection

solution, in which users are asked to provide QoE scores manually; yet, EyeQoE is executed automatically without human involvement.

Impact of subject-dependent features on QoE assessment. Different subjects may be impacted in various ways. To investigate the significance and distinction of impact factors, we correlate several objective and subjective factors with the QoE scores from the collected data. Specifically, objective factors such as video resolution and stalling events are directly derived from the preprocessed videos, whereas subjective factors, including cybersickness, fatigue, and immersiveness, are collected during the experiments by confirming with the subjects about their corresponding subjective feelings. Figure 20 demonstrates the result, from which we make the following observations. First, among all the listed impact factors, stalling events and cybersickness are the most critical factors, as different cybersickness levels result in the most distinct QoE distributions, and that low QoE scores are induced whenever stalling events occur. Second, QoE scores highly concentrate with different levels of stalling events. For example, 88% of videos with 8 stalls are rated with QoE as 1; the variance of QoE scores is $\sigma^2 = 0.10$. Similarly, 64% and 76% of videos with 4 and 2 stalls are rated with QoE as 2 and 3 ($\sigma^2 = 0.34$ and 0.25), respectively. This means that with the same levels of stalling events, more subjects perceive similar QoE, which indicates that this factor brings a common significance across various subjects. In contrast, immersiveness results in a relatively even distribution of QoE scores, suggesting that this factor is distinct across different users.

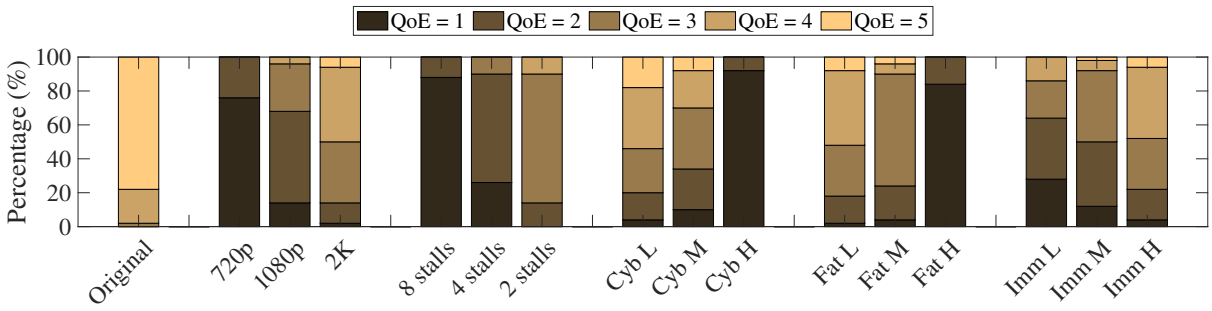


Figure 20. Correlation between impact factors and QoE.

8 DISCUSSION AND FUTURE WORK

In this section, we discuss several limitations of this work and present our future research directions.

Extra cost introduced by domain adaption. Domain adaption is activated only for unseen videos; that is, the process will be bypassed when the type of videos that are covered in the training process. Hence, no extra training cost is incurred. For unseen videos, domain adaption does cause certain training cost. To quantify it, we have evaluated the time consumption for domain adaption in the experiment. Figure 19a presents the stacked computation time of EyeQoE's all major processes, including preprocessing of eye-based cues, graph generation, and testing. Specifically, testing is conducted over both seen and unseen videos. The latter includes the domain adaption operation. We observe that the testing time ranges between 0.1 s and 5.1 s, among which larger values tend to associate with unseen videos due to the domain adaption.

In current multimedia services, user's QoE is mainly obtained by asking people to rate their perceived quality via surveys or self-reports. However, such procedures are inconvenient and may even be annoying for the users. EyeQoE intends to automate the entire process by constructing a QoE assessment model. User's perceived QoE would be generated and collected automatically. In this sense, timing is not the main consideration of our design. Still, according to the above result, QoE assessment for unseen videos (including domain adaption) can be done

within 5.1 s, which is satisfactory for real-world implementation. Of course, it would be even more desirable if the latency can be further shortened. We plan to investigate this possibility in our future work.

Enhancing the prediction accuracy of EyeQoE. This work demonstrates the feasibility of using eye-based cues for QoE assessment. While the overall accuracy performance is satisfactory, there is still room for improvement. To this end, we plan to pursue two potential directions. The first one is to combine EyeQoE with traditional objective quality of service (QoS) metrics such as bandwidth, latency, video quality, etc. Specifically, we will integrate the QoS metrics as new dimensions alongside the eye-based cues as the inputs of our QoE model. The graph modeling will be revised accordingly with the introduction of additional inputs. The selection of QoS metrics will be carefully determined. They should be practical to collect at VR terminals and play positively in enhancing EyeQoE's accuracy. In the other direction, we intend to combine EyeQoE with other existing QoE models for 360-degree videos. The hypothesis is that QoE models capturing a greater diversity of potentially informative features might improve the overall model robustness when included. We plan to apply *ensemble methods* [56, 59] over multiple representative QoE models and EyeQoE to derive the aggregated prediction results. Comparison will be made with each single model over the prediction accuracy.

Reducing QoE prediction latency for unseen videos. Under the current design, online QoE predictions over unseen videos are executed at the level of seconds. The latency is mainly caused by the graph generation and the domain adaption process, i.e., fine-tuning the trained QoE model. While this value is practically acceptable for pure QoE collection, it would be too large to support real-time QoE-aware service management, which can benefit applications such as adaptive 360-degree video streaming [77]. Essentially, service providers can timely adjust streaming strategies, such as resolutions, rendering speed, and scheduling priority, in accordance with the viewer's QoE estimated in real-time. As our future work, we plan to investigate the feasibility of forecasting viewer's perceived QoE a short period ahead of time, which then better tolerates the prediction latency. There is an important observation that viewer's subjective feelings typically do not change suddenly. For instance, one's cybersickness and fatigue are gradually accumulated as prolonged exposure in a VR environment. Such temporal dependencies can be exploited for QoE forecasting.

Other approaches for adaption to unseen videos. A critical challenge of this work is to adapt the QoE model, trained by existing video clips, to unseen videos. Aside from domain adaption as adopted here, another interesting future direction is to leverage few-shot learning [15, 73]. We frame the challenge as a few-shot learning problem, that is: how to train the GCN classifier such that it can quickly adapt to an unseen video after a few learning iterations with a small number of annotated samples from the same category (Section 6.3) that the unseen video belongs to. Few-shot learning is promising in classifying new data when only a few training samples with supervised information are available and has been successfully applied in language processing [78], text classification [79], and image classification [13].

9 CONCLUSION

In this paper, we present EyeQoE, a novel QoE prediction model for 360-degree videos using subjects' eye-based cues. To extract useful features from the cues, we propose a novel method that models them into graphs and then build a GCN-based classifier to learn over graphs. Our design also involves the Siamese network that deals with learning uncertainty caused by subjects and visual stimuli heterogeneity. A domain adaptation scheme named MADA is further proposed to ensure the efficacy of EyeQoE on unseen videos. A 3-month data collection campaign is carried out to build our own visual-based QoE assessment dataset. Our comprehensive evaluation shows that EyeQoE advances the literature by a suite of new capabilities. First, its best accuracy performance is 92.9% which beats other state-of-the-art models. Second, EyeQoE is capable of capturing various impact factors, such as video stalls and viewer's subjective feelings (e.g., cybersickness, immersiveness, and fatigue), in QoE

prediction, while they are largely overlooked in prior models. Moreover, all the online operations of EyeQoE can be efficiently performed with 90-percentile computation latency within 4.2 seconds.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their insightful comments and suggestions. We are also grateful to NSF (CNS-1943509) for partially funding this research.

REFERENCES

[1] 2005. Copyright. In *Advanced Graphics Programming Using OpenGL*, TOM McREYNOLDS and DAVID BLYTHE (Eds.). Morgan Kaufmann, San Francisco, iv. https://doi.org/10.1016/B978-1-55860-659-3.50030-5
[2] 2021. High-resolution VR Headset for Professionals - Varjo VR-3. https://varjo.com/products/vr-3/
[3] 2021. HTC VIVE Pro Eye. https://www.vive.com/eu/product/vive-pro-eye/overview/
[4] 2021. Neo 2 Neo 2 Eye: All-in-One VR Headset with 6DoF tracking. https://www.pico-interactive.com/us/neo2.html
[5] 2021. Tobii VR: Eye Tracking Technology in Virtual Reality. https://vr.tobii.com/
[6] 2021. VR Eye Tracking For Business: FOVE 0 EYE TRACKING VR DEVKIT. https://fove-inc.com/
[7] Zahid Akhtar, Kamran Siddique, Ajita Rattani, Syaheerah Lebai Lutfi, and Tiago H. Falk. 2019. Why is Multimedia Quality of Experience Assessment a Challenging Problem? *IEEE Access* 7 (2019), 117897–117915. https://doi.org/10.1109/ACCESS.2019.2936470
[8] Christos G. Bampis, Zhi Li, and Alan C. Bovik. 2017. Continuous Prediction of Streaming Video QoE Using Dynamic Networks. *IEEE Signal Processing Letters* 24, 7 (2017), 1083–1087. https://doi.org/10.1109/LSP.2017.2705423
[9] Mathias Benedek, Robert Stoiser, Sonja Annerer-Walcher, and Christof Körner. 2017. Eye Behavior Associated with Internally versus Externally Directed Cognition. *Frontiers in Psychology* 8 (06 2017). https://doi.org/10.3389/fpsyg.2017.01092
[10] Gordon Bill, Elisabeth Whyte, Jason Griffin, and Kathryn Scherf. 2020. Measuring sensitivity to eye gaze cues in naturalistic scenes: Presenting the eye gaze FoCuS database. *International Journal of Methods in Psychiatric Research* 29 (07 2020). https://doi.org/10.1002/mpr.1833
[11] Hongyun Cai, Vincent Zheng, and Kevin Chang. 2017. A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications. *IEEE Transactions on Knowledge and Data Engineering* 30 (09 2017). https://doi.org/10.1109/TKDE.2018.2807452
[12] Raymundo Cassani, Marc-Antoine Moinnereau, and Tiago H. Falk. 2018. A Neurophysiological Sensor-Equipped Head-Mounted Display for Instrumental QoE Assessment of Immersive Multimedia. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2018.8463422
[13] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. 2021. Self-Supervised Learning for Few-Shot Image Classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1745–1749. https://doi.org/10.1109/ICASSP39728.2021.9413783
[14] Sijia Chen, Yingxue Zhang, Yiming Li, Zhenzhong Chen, and Zhou Wang. 2018. Spherical Structural Similarity Index for Objective Omnidirectional Video Quality Assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2018.8486584
[15] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. 2019. A Closer Look at Few-shot Classification.
[16] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. 2020. Unsupervised and Semi-Supervised Domain Adaptation for Action Recognition from Drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
[17] S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. 539–546 vol. 1. https://doi.org/10.1109/CVPR.2005.202
[18] Viviane Clay, Peter König, and Sabine Koenig. 2019. Eye Tracking in Virtual Reality. *Journal of Eye Movement Research* 12 (04 2019). https://doi.org/10.16910/jemr.12.1.3
[19] Gabriela Csurka. 2017. *Domain Adaptation for Visual Applications: A Comprehensive Survey*. https://doi.org/10.1007/978-3-319-58347-1_1
[20] Roberto Irajá Tavares da Costa Filho, Marcelo Caggiani Luizelli, Maria Torres Vega, Jeroen van der Hooft, Stefano Petrangeli, Tim Wauters, Filip De Turck, and Luciano Paschoal Gaspary. 2018. Predicting the Performance of Virtual Reality Video Streaming in Mobile Networks. In *Proceedings of the 9th ACM Multimedia Systems Conference* (Amsterdam, Netherlands) *(MMSys '18)*. Association for Computing Machinery, New York, NY, USA, 270–283. https://doi.org/10.1145/3204949.3204966
[21] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2017. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. arXiv:1606.09375 [cs.LG]
[22] Tho Nguyen Duc, Chanh Minh Tran, Tan Phan-Xuan, and Eiji Kamioka. 2019. Modeling of Cumulative QoE in On-Demand Video Services: Role of Memory Effect and Degree of Interest. *Future Internet* 11 (2019), 171.

[23] Andrew T. Duchowski. 2017. *Eye Tracking Methodology: Theory and Practice*. Springer International Publishing. https://doi.org/10.1007/978-3-319-57883-5

[24] Darragh Egan, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer, and Niall Murray. 2016. An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2016.7498964

[25] Ulrich Engelke, Marcus Barkowsky, Patrick Le Callet, and Hans-Jürgen Zepernick. 2010. Modelling saliency awareness for objective video quality assessment. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. 212–217. https://doi.org/10.1109/QOMEX.2010.5516159

[26] Zesong Fei, Fei Wang, Jing Wang, and Xiang Xie. 2019. QoE Evaluation Methods for 360-Degree VR Video Transmission. *IEEE Journal of Selected Topics in Signal Processing* PP (11 2019), 1–1. https://doi.org/10.1109/JSTSP.2019.2956631

[27] International Organization for Standardization. 2005. *Image Safety - Reducing the Incidence of Undesirable Biomedical Effects Caused by Visual Image Sequences*. ISO. https://books.google.com/books?id=LfAncgAACAAJ

[28] Chiara Galdi and Michele Nappi. 2019. *Eye Movement Analysis in Biometrics*. 171–183. https://doi.org/10.1007/978-981-13-1144-4_8

[29] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research* 17, 1 (Jan. 2016), 2096–2030.

[30] Alan L. V. Guedes, Roberto G. de A. Azevedo, Pascal Frossard, Sérgio Colcher, and Simone Diniz Junqueira Barbosa. 2019. Subjective Evaluation of 360-degree Sensory Experiences. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. 1–6. https://doi.org/10.1109/MMSP.2019.8901743

[31] E. H. Hess. 1975. The role of pupil size in communication. (1975), 110–119.

[32] Tran Huyen, Cuong PHAM, Nam Pham Ngoc, Anh Pham, and Truong Cong Thang. 2018. A Study on Quality Metrics for 360 Video Communications. *IEICE Transactions on Information and Systems* E101.D (01 2018), 28–36. https://doi.org/10.1587/transinf.2017MUP0011

[33] Mohsina Ishrat and Pawanesh Abrol. 2017. Eye movement analysis in the context of external stimuli effect. In *2017 International Conference on Informatics, Health Technology (ICIHT)*. 1–6. https://doi.org/10.1109/ICIHT.2017.7899148

[34] ITU-T. 2007. *Definition of Quality of Experience (QoE)*. https://www.itu.int/md/T05-FG.IPTV-IL-0050/en

[35] Shunichi Kasahara, Shohei Nagai, and Jun Rekimoto. 2015. First Person Omnidirectional Video: System Design and Implications for Immersive Experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video* (Brussels, Belgium) *(TVX '15)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/2745197.2745202

[36] Conor Keighrey, Ronan Flynn, Siobhan Murray, and Niall Murray. 2017. A QoE Evaluation of Immersive Augmented and Virtual Reality Speech Language Assessment Applications. https://doi.org/10.1109/QoMEX.2017.7965656

[37] Donghyun Kim, Sunghwan Choi, Sangil Park, and Kwanghoon Sohn. 2011. Stereoscopic visual fatigue measurement based on fusional response curve and eye-blinks. In *2011 17th International Conference on Digital Signal Processing (DSP)*. 1–6. https://doi.org/10.1109/ICDSP.2011.6004999

[38] Hak Gu Kim, Heoun-Taek Lim, Sangmin Lee, and Yong Man Ro. 2019. VRSA Net: VR Sickness Assessment Considering Exceptional Motion for 360° VR Video. *IEEE Transactions on Image Processing* 28, 4 (2019), 1646–1660. https://doi.org/10.1109/TIP.2018.2880509

[39] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[40] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv e-prints*, Article arXiv:1609.02907 (Sept. 2016), arXiv:1609.02907 pages. arXiv:1609.02907 [cs.LG]

[41] Eric C. Larson, Cuong Vu, and Damon M. Chandler. 2008. Can visual fixation patterns improve image fidelity assessment?. In *2008 15th IEEE International Conference on Image Processing*. 2572–2575. https://doi.org/10.1109/ICIP.2008.4712319

[42] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. 2007. Predicting visual fixations on video based on low-level visual features. *Vision Research* 47, 19 (2007), 2483–2498. https://doi.org/10.1016/j.visres.2007.06.015

[43] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. 2018. Bridge the Gap Between VQA and Human Behavior on Omnidirectional Video: A Large-Scale Dataset and a Deep Learning Model. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 932–940. https://doi.org/10.1145/3240508.3240581

[44] Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, and Xiaoming Tao. 2019. Viewport Proposal CNN for 360deg Video Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[45] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2017. Gated Graph Sequence Neural Networks. arXiv:1511.05493 [cs.LG]

[46] Hantao Liu and Ingrid Heynderickx. 2011. Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data. *IEEE Transactions on Circuits and Systems for Video Technology* 21, 7 (2011), 971–982. https://doi.org/10.1109/TCSVT.2011.2133770

[47] Gebremariam Mesfin, Nadia Hussain, Alexandra Covaci, and Gheorghita Ghinea. 2019. Using Eye Tracking and Heart-Rate Activity to Examine Crossmodal Correspondences QoE in Mulsemedia. *AACM Transactions on Multimedia Computing, Communications, and Applications* 15, 2, Article 34 (June 2019), 22 pages. https://doi.org/10.1145/3303080

[48] Y Morad, H Lemberg, N Yofe, and Y Dagan. 2000. Pupillography as an objective indicator of fatigue. *Current eye research* 21, 1 (July 2000), 535—542. https://doi.org/10.1076/0271-3683(200007)2111-zft535

[49] Jonathan Munro and Dima Damen. 2020. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *CoRR* abs/2001.09691 (2020). arXiv:2001.09691 https://arxiv.org/abs/2001.09691

[50] Muhammad-Sajid Mushtaq and Abdelhamid Mellouk. 2017. 5 - QoE and Power-saving Model for 5G Network. In *Quality of Experience Paradigm in Multimedia Services*, Muhammad-Sajid Mushtaq and Abdelhamid Mellouk (Eds.). Elsevier, 127–160. https://doi.org/10.1016/B978-1-78548-109-3.50005-9

[51] Jonny O'Dwyer, Niall Murray, and Ronan Flynn. 2020. Eye-based Continuous Affect Prediction. arXiv:1907.09896 [cs.HC]

[52] Sean O'Kane. 2021. *Tesla starts using in-car camera for Autopilot driver monitoring - The Verge.* https://www.theverge.com/2021/5/27/22457430/tesla-in-car-camera-driver-monitoring-system

[53] Randall C. O'Reilly. 1996. Biologically Plausible Error-Driven Learning Using Local Activation Differences: The Generalized Recirculation Algorithm. *Neural Computation* 8, 5 (1996), 895–938. https://doi.org/10.1162/neco.1996.8.5.895

[54] Henryk Palus. 2006. Colorfulness of the image: definition, computation, and properties. In *Lightmetry and Light and Optics in Biomedicine 2004*, Katarzyna Kolacz and Jacek Sochacki (Eds.), Vol. 6158. International Society for Optics and Photonics, SPIE, 42 – 47. https://doi.org/10.1117/12.675760

[55] Viswa Mani Kiran Peddinti and Prakriti Chintalapoodi. 2011. Domain Adaptation in Sentiment Analysis of Twitter *(AAAIWS'11-05)*. AAAI Press, 44–49.

[56] R. Polikar. 2006. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 3 (2006), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

[57] Simone Porcu, Alessandro Floris, Jan-Niklas Voigt-Antons, Luigi Atzori, and Sebastian Möller. 2020. Estimation of the Quality of Experience During Video Streaming From Facial Expression and Gaze Direction. *IEEE Transactions on Network and Service Management* 17, 4 (2020), 2702–2716. https://doi.org/10.1109/TNSM.2020.3018303

[58] Ievgen Redko, Amaury Habrard, Emilie Morvant, Marc Sebban, and Younès Bennani. 2019. 2 - Domain Adaptation Problem. In *Advances in Domain Adaptation Theory*, Ievgen Redko, Amaury Habrard, Emilie Morvant, Marc Sebban, and Younès Bennani (Eds.). Elsevier, 21–36. https://doi.org/10.1016/B978-1-78548-236-6.50002-7

[59] Lior Rokach. 2010. Ensemble-based classifiers. *Artificial Intelligence Review* 33 (02 2010), 1–39. https://doi.org/10.1007/s10462-009-9124-7

[60] Marius Rubo and Matthias Gamer. 2018. Social content and emotional valence modulate gaze fixations in dynamic scenes. *Scientific Reports* 8 (02 2018). https://doi.org/10.1038/s41598-018-22127-w

[61] Débora Salgado, Felipe Martins, Thiago Braga Rodrigues, Conor Keighrey, Ronan Flynn, Eduardo Naves, and Niall Murray. 2018. A QoE assessment method based on EDA, heart rate and EEG of a virtual reality assistive technology system. 517–520. https://doi.org/10.1145/3204949.3208118

[62] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* 20, 1 (2009), 61–80. https://doi.org/10.1109/TNN.2008.2005605

[63] Kalpana Seshadrinathan and Alan Conrad Bovik. 2010. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. *IEEE Transactions on Image Processing* 19, 2 (2010), 335–350. https://doi.org/10.1109/TIP.2009.2034992

[64] Ashutosh Singla, Stephan Fremerey, Werner Robitza, Pierre Lebreton, and Alexander Raake. 2017. Comparison of Subjective Quality Evaluation for HEVC Encoded Omnidirectional Videos at Different Bit-rates for UHD and FHD Resolution. 511–519. https://doi.org/10.1145/3126686.3126768

[65] Ivo Sluganovic, Marc Roeschlin, Kasper B. Rasmussen, and Ivan Martinovic. 2016. Using Reflexive Eye Movements for Fast Challenge-Response Authentication *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 1056–1067. https://doi.org/10.1145/2976749.2978311

[66] Wei Sun, Weike Luo, Xiongkuo Min, Guangtao Zhai, Xiaokang Yang, Ke Gu, and Siwei Ma. 2019. MC360IQA: The Multi-Channel CNN for Blind 360-Degree Image Quality Assessment. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. 1–5. https://doi.org/10.1109/ISCAS.2019.8702664

[67] Yule Sun, Ang Lu, and Lu Yu. 2017. Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video. *IEEE Signal Processing Letters* 24, 9 (2017), 1408–1412. https://doi.org/10.1109/LSP.2017.2720693

[68] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.

[69] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. 2009. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis, Vol. 5478. 337–349. https://doi.org/10.1007/978-3-642-00958-7_31

[70] Huyen T.T. Tran, Nam Pham Ngoc, Tobias Hoßfeld, and Truong Cong Thang. 2018. A Cumulative Quality Model for HTTP Adaptive Streaming. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. https://doi.org/10.1109/QoMEX.2018.8463414

[71] Huyen T. T. Tran, Nam Pham Ngoc, Cuong T. Pham, Yong Ju Jung, and Truong Cong Thang. 2017. A subjective study on QoE of 360 video for VR communication. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*. 1–6. https://doi.org/10.1109/MMSP.2017.8122249

[72] Kazuo Tsubota and Katsu Nakamori. 1993. Dry Eyes and Video Display Terminals. *New England Journal of Medicine* 328, 8 (1993), 584–584. https://doi.org/10.1056/NEJM199302253280817 arXiv:https://doi.org/10.1056/NEJM199302253280817 PMID: 8426634.

[73] Yaqing Wang and Quanming Yao. 2019. Few-shot Learning: A Survey. *CoRR* abs/1904.05046 (2019). arXiv:1904.05046 http://arxiv.org/abs/1904.05046

[74] Xiaoyu Xiu, Yuwen He, Yan Ye, and Bharath Vishwanath. 2017. An evaluation framework for 360-degree video compression. *2017 IEEE Visual Communications and Image Processing (VCIP)* (2017), 1–4.

[75] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. 2019. Assessing Visual Quality of Omnidirectional Videos. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 12 (2019), 3516–3530. https://doi.org/10.1109/TCSVT.2018.2886277

[76] Jingteng Xue, Dong-Qing Zhang, Heather Yu, and Chang Wen Chen. 2014. Assessing quality of experience for adaptive HTTP video streaming. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 1–6. https://doi.org/10.1109/ICMEW.2014.6890604

[77] Abid Yaqoob, Ting Bi, and Gabriel-Miro Muntean. 2020. A Survey on Adaptive 360° Video Streaming: Solutions, Challenges and Opportunities. *IEEE Communications Surveys Tutorials* 22, 4 (2020), 2801–2838. https://doi.org/10.1109/COMST.2020.3006999

[78] Wenpeng Yin. 2020. Meta-learning for Few-shot Natural Language Processing: A Survey. *CoRR* abs/2007.09604 (2020). arXiv:2007.09604 https://arxiv.org/abs/2007.09604

[79] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse Few-Shot Text Classification with Multiple Metrics. *CoRR* abs/1805.07513 (2018). arXiv:1805.07513 http://arxiv.org/abs/1805.07513

[80] Matt Yu, Haricharan Lakshman, and Bernd Girod. 2015. A Framework to Evaluate Omnidirectional Video Coding Schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality*. 31–36. https://doi.org/10.1109/ISMAR.2015.12

[81] Vladyslav Zakharchenko, K. Choi, and J. Park. 2016. Quality metric for spherical panoramic video. In *Optical Engineering + Applications*.

[82] Ramtin Zargari Marandi, Pascal Madeleine, Øyvind Omland, Nicolas Vuillerme, and Afshin Samani. 2018. Eye movement characteristics reflected fatigue development in both young and elderly individuals. *Scientific Reports* 8 (09 2018). https://doi.org/10.1038/s41598-018-31577-1

[83] Ramtin Zargari Marandi, Pascal Madeleine, Øyvind Omland, Nicolas Vuillerme, and Afshin Samani. 2018. Eye movement characteristics reflected fatigue development in both young and elderly individuals. *Scientific Reports* 8 (09 2018). https://doi.org/10.1038/s41598-018-31577-1

[84] S. Zhang, Hanghang Tong, Jiejun Xu, and R. Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6 (2019), 1–23. https://doi.org/10.1186/s40649-019-0069-y

[85] Yufeng Zhou, Mei Yu, Hualin Ma, Hua Shao, and Gangyi Jiang. 2018. Weighted-to-Spherically-Uniform SSIM Objective Quality Evaluation for Panoramic Video. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*. 54–57. https://doi.org/10.1109/ICSP.2018.8652269

[86] Wenjie Zou, Fuzheng Yang, Wei Zhang, Yi Li, and Haoping Yu. 2018. A Framework for Assessing Spatial Presence of Omnidirectional Video on Virtual Reality Device. *IEEE Access* 6 (2018), 44676–44684. https://doi.org/10.1109/ACCESS.2018.2864872