Subjective and Objective Quality Assessment of 2D and 3D Foveated Video Compression in Virtual Reality

Article *in* IEEE Transactions on Image Processing · June 2021

DOI: 10.1109/TIP.2021.3087322

CITATIONS
2

READS
77

5 authors, including:

Yize Jin
University of Texas at Austin
18 PUBLICATIONS **59** CITATIONS
SEE PROFILE

Meixu Chen
University of Texas at Austin
7 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Todd Goodall
University of Texas at Austin
23 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Alan Bovik
University of Texas at Austin
919 PUBLICATIONS **102,699** CITATIONS
SEE PROFILE

Subjective and Objective Quality Assessment of 2D and 3D Foveated Video Compression in Virtual Reality

Yize Jin, Meixu Chen, *Student Member, IEEE*, Todd Goodall, Anjul Patney, Alan C. Bovik, *Fellow, IEEE*

***Abstract*—In Virtual Reality (VR), the requirements of much higher resolution and smooth viewing experiences under rapid and often real-time changes in viewing direction, leads to significant challenges in compression and communication. To reduce the stresses of very high bandwidth consumption, the concept of foveated video compression is being accorded renewed interest. By exploiting the space-variant property of retinal visual acuity, foveation has the potential to substantially reduce video resolution in the visual periphery, with hardly noticeable perceptual quality degradations. Accordingly, foveated image / video quality predictors are also becoming increasingly important, as a practical way to monitor and control future foveated compression algorithms. Towards advancing the development of foveated image / video quality assessment (FIQA / FVQA) algorithms, we have constructed 2D and (stereoscopic) 3D VR databases of foveated / compressed videos, and conducted a human study of perceptual quality on each database. Each database includes 10 reference videos and 180 foveated videos, which were processed by 3 levels of foveation on the reference videos. Foveation was applied by increasing compression with increased eccentricity. In the 2D study, each video was of resolution 7680×3840 and was viewed and quality-rated by 36 subjects, while in the 3D study, each video was of resolution 5376×5376 and rated by 34 subjects. Both studies were conducted on top of a foveated video player having low motion-to-photon latency (∼50ms). We evaluated different objective image and video quality assessment algorithms, including both FIQA / FVQA algorithms and non-foveated algorithms, on our so called LIVE-Facebook Technologies Foveation-Compressed Virtual Reality (LIVE-FBT-FCVR) databases. We also present a statistical evaluation of the relative performances of these algorithms. The LIVE-FBT-FCVR databases have been made publicly available and can be accessed at https://live.ece.utexas.edu/research/LIVEFBTFCVR/index.html.**

***Index Terms*—foveation, subjective video quality, Virtual Reality, subjective study, stereoscopic 3D, foveated video compression, objective video quality, visual acuity.**

I. Introduction

VIRTURAL Reality (VR) has experienced a substantial growth in popularity, due to recent advancements in consumer head-mounted displays (HMDs) and associated computing hardware technologies. While cable-tethered headsets for personal computers such as the HTC Vive, Oculus rift, and Microsoft Hololens remain popular, standalone, untethered headsets like the Oculus Quest are even more successful because of the freedom of movement they allow. Owing to greatly increasing numbers of consumer applications, virtual and augmented reality traffic is expected to increase 12-fold by 2022, as compared to 0.33 exabytes per month in 2017 [1]. While gaming has largely driven the VR space, immersive and 360° videos are gaining wider acceptance and in the future are expected to drive significant increases in demand for bandwidth consumption.

To capture omnidirectional scenes, immersive videos are usually generated by 360° cameras containing multiple well-synchronized and calibrated lenses. The video frames obtained from each lens are then stitched into various formats, such as equirectangular projection (ERP), and cubemap (CMP) [2]. While immersive videos provide higher degrees of freedom and richer visual information, their bandwidth consumption is much higher than traditional videos. Moreover, efficiency of communicating immersive videos to HMDs is limited both by bandwidth and the need for high resolution displays. The resolutions of mainstream HMDs range from 1K×1K to 2K×2K per eye, and their fields of view (FOVs) range from 90° to 130°. To match the resolutions of the HMD, the resolutions of immersive videos to be displayed expand by more than 4-fold, from at least 4K×2K (UHD) up to (currently) 8K×4K. Yet, the maximum resolution of the human eyes is about 120 pixels per degree (ppd), while the HMD screen resolution equates to $10 \sim 20$ ppd. Hence, higher screen resolutions are desirable, but this would require even higher bandwidths. At the same time, delivering smooth, real-time experiences even during rapid changes in viewing direction requires low motion-to-photon latency, further constraining optimization of immersive video streaming.

One way to remedy the aforementioned problems is by developing foveated processing protocols, an idea that is again gaining traction. Similar to the way that chroma subsampling takes advantage of the reduced bandwidth of visual chrominance signals relative to luminance, foveation exploits the reduced visual acuity in the visual periphery relative to the foveal region. Foveated video compression first gained attention more than two decades ago, but there was no driving need for the technology at the time [3]–[6]. Foveated image / video quality assessment (FIQA / FVQA) models were also integrated into foveated compression algorithms to control their performance [7]–[10]. Due to the availability of consumer eyetrackers that can be easily incorporated into HMDs, there is an increasing research interest in the potential of foveation,

Manuscript created August 24, 2020;

Y. Jin, M. Chen, and A. C. Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, 78712 USA e-mail: yizejin@utexas.edu; chenmx@utexas.edu; bovik@ece.utexas.edu

While working on this project, T. Goodall and A. Patney were with Facebook Technologies.

and foveated compression algorithms [11]–[14] that build on modern video codec standards like H.264 / AVC [15] and H.265 / HEVC [16].

As foveated compression algorithms evolve, there is an increasing need for foveated image / video quality assessment (FIQA / FVQA) algorithms that can be used to assess and control compression. Towards advancing progress in this direction, and recognizing that there are no existing foveated video quality databases addressing compression that are publicly and freely available, we designed and created two databases of foveated / compressed immersive VR videos, rated by human subjects, which we will refer to as the LIVE-FBT-FCVR databases. One of the databases contains 2D content, while the other contains stereoscopic 3D content. The new databases contain diverse contents and encompass important features: 1) To smoothly sample the space of the FOV, three levels of foveation were applied on the content in both databases; 2) to reduce aliasing and fully make use of the screen resolution inside the HMD, the VR videos in the 2D database are of spatial resolution 7680×3840, while those in the 3D VR database are of 5376×5376; 3) we systematically combined compression distortion with video foveation, both of which affect foveated video quality as viewed by foveated eyes; 4) to ensure smooth, foveated visual experiences, we designed a foveated video player having low motion-to-photon latency ($\sim 50ms$). On each database, we conducted a human subjective study of foveated + compressed video quality, against which we evaluated a variety of leading IQA / VQA and FIQA / FVQA algorithms.

The rest of the paper is organized as follows: Section II studies related work on foveated video quality assessment. Section III discusses design choices made in the construction of the databases. Section IV describes our subjective testing methodology, and the ways we processed the collected data. In Section V, the quality prediction performances of leading IQA / VQA models are compared and analyzed on the new databases. Finally Section VI concludes the paper along with some remarks on possible future research directions.

II. Related Work

A. Subjective Quality Assessment

Traditional VQA databases such as LIVE VQA [17], LIVE MOBILE [18], CSIQ-VQA [19] and CDVL [20] have been used to greatly advance the development of objective VQA algorithms. Other databases dedicated to the study of video quality of experience (QoE), such as the LIVE NFLX [21] and LIVE Mobile Stall Video Databases [22], [23], have also played an important role in the design of improved video streaming services. Recently, a subjective database of audio-visual signals (LIVE-SJTU A/V-QA database [24]) was designed to study multimodal audio-video quality perception. Towards improving VR experiences, important questions need to be addressed: How can immersive IQA / VQA databases be used to facilitate the development of objective VR IQA / VQA algorithms, and, can they be used to achieve significant bandwidth savings in immersive VR systems, especially, those designed for video streaming?

Towards answering the questions, VR researchers have developed several databases that include VR-specific features. A testbed for conducting subjective studies on immersive contents was proposed in [25], and a pilot experiment on JPEG compression distortions was conducted. A 4K (4096×2048) immersive image database called CVIQD was described in [26]. CVIQD contains 165 compression distorted images generated from 5 pristine images, including JPEG, H.264 / AVC, and H.265 / HEVC. In [27], [28], CVIQD was expanded to include 16 reference images and 528 distorted / compressed images. In [29], an omnidirectional IQA (OIQA) database was proposed, containing 16 reference images of resolutions ranging from 11332×5666 to 13320×6660, and 320 distorted images with 4 types of impairments: JPEG compression, JPEG2000 compression, Gaussian blur, and Gaussian noise. In [30], a stereo 3D database was proposed, containing 450 distorted 3D immersive images generated from 15 pristine images, impaired by Gaussian noise, Gaussian blur, downsampling, VP9 compression, HEVC compression, and VR-specific stitching distortions. In [31], an immersive VQA database comprising 48 sequences downloaded from YouTube and VRCun was proposed, containing sequences varying from 3K (2880×1440) to 8K (7680×3840). In [32], another immersive VQA database called IVQAD 2017 was described, containing 10 reference 4K videos resolution captured with an Insta360 camera, from which 225 distorted videos were generated by applying spatial downsampling, temporal downsampling, and compression distortions.

While these databases are valuable tools for understanding immersive video quality, none of them address the great potential of incorporating foveation into bandwidth-hungry immersive VR systems.

B. Objective Quality Assessment

In practice, objective IQA / VQA algorithms serve as a substitute for subjective quality assessment (QA). Generally, objective QA models are classified into three categories: full reference (FR), reduced reference (RR), and no reference (NR). In our context, we also consider whether an algorithm belongs to non-foveated (traditional) or foveated QA categories.

While the PSNR and MSE are notorious for their poor correlation with subjective quality scores [33], perceptually based FR IQA algorithms such as SSIM [34], MS-SSIM [35], VIF [36], and FSIM [37] exhibit much better performance on predicting picture quality. In scenarios when the reference images are absent or not available, natural scene statistics (NSS) based NR IQA models, which capture deviations of distorted scene statistics from those of pristine images are often quite effective [38]–[41]. Another class of NR IQA models, BPRI and BMPRI [42], [43] use a pseudo-reference image (PRI) generated from the distorted image to attempt to facilitate measurement of the severity of distortions.

Some early FR IQA models used for VR are based on PSNR, such as WS-PSNR [44], CPP-PSNR [45], S-PSNR [46]. SSIM-based 360° IQA models were also developed to capture VR perceptual quality, such as S-SSIM [47] and

(a) Coco (b) EER (c) Stadium (d) Starbucks (e) Waterfall

(f) Microsoft (g) Oasis (h) Playground (i) Sanjacinto (j) Sculpture

Fig. 1: Sample frames of the reference videos in the 2D database.

(a) Bar_3d (b) Bridge_3d (c) Domain_3d (d) Dunk_3d (e) EER_3d

(f) Maintower_3d (g) Microsoft_3d (h) Oasis_3d (i) Street_3d (j) Sculpture_3d

Fig. 2: Sample frames of the reference videos in the 3D database.

SSIM360 [48]. Recently, deep learning has been applied in VR IQA problems. A VR-specific blind IQA model, MC360IQA [28], used a hyper structure on a ResNet34 [49] network along with an image quality regressor to fuse features from intermediate layers of the ResNet. Another deep learning model, DeepVR-IQA [50], used adverserial learning to improve the performance of their blind IQA predictor, whereby a descriminator was designed to distinguish predicted scores from the ground-truth scores.

IQA models can also be used to predict video quality when applied on a frame-by-frame basis, where the temporal information of videos is not considered. To capture temporal distortions as well as spatial distortions, a variety of models have been proposed. An early VQA model called the Video Quality Metric (VQM) calculates quality features on local spatial-temporal (S-T) regions, including temporal features (mean and standard deviation) extracted from frame differences [51]. An FR algorithm called the MOVIE index [52] represents temporal artifacts by modeling the responses of motion sensitive neurons in extra-cortical area MT [53]. The Video Multimethod Assessment Fusion (VMAF) [54] combines features obtained from VIF [36], DLM [55], and frame differences, using a Support Vector Regressor (SVR).

General-purpose NR VQA algorithms have proven difficult to design, due to the high complexity of temporal distortions and the absence of reference information. RR VQA algorithms predict distorted video quality given a reduced amount of information from the reference video. These include NSS-based models such as RRED [57], STRRED [58], and Speed-QA [59].

Progress have also been made on the development of NR VQA algorithms. V-BLIINDS [60] employs natural video statistics (NVS) and a model of motion coherency to characterize video quality. The authors of [61] model spatial-temporal natural video statistics in a 3D discrete cosine transform (DCT) domain, and use them to predict video quality. The Two Level Video Quality Model (TLVQM) [62] utilizes low- and high-complexity features to predict video quality, achieving high performance on the CVD2014 [63], KoNViD-1K [64], and Live-Qualcomm datasets [65].

While there has been extensive research on non-foveated IQA / VQA models, progress on the development of FIQA / FVQA models has been limited. An early FR model called the Foveated Wavelet Quality Index (FWQI) [66] measures foveated image quality by combining an eccentricity-dependent contrast sensitivity function (CSF) model [6] with a visually detectable noise threshold model [67]. The Foveated PSNR (FPSNR) and foveated weighted signal-to-noise ratio (FWSNR) models [69] use curvilinear coordinate systems to model foveation. In [70], the authors defined a Foveation-based Content Adaptive SSIM (FA-SSIM) index, which extends the popular SSIM to account for foveated viewing. A recently developed NR FVQA model called Space-Variant BRISQUE (SVBRISQUE) achieves state-of-the-art (SOTA) performance using NSS features and a neural noise model to predict the quality of immersive videos [68].

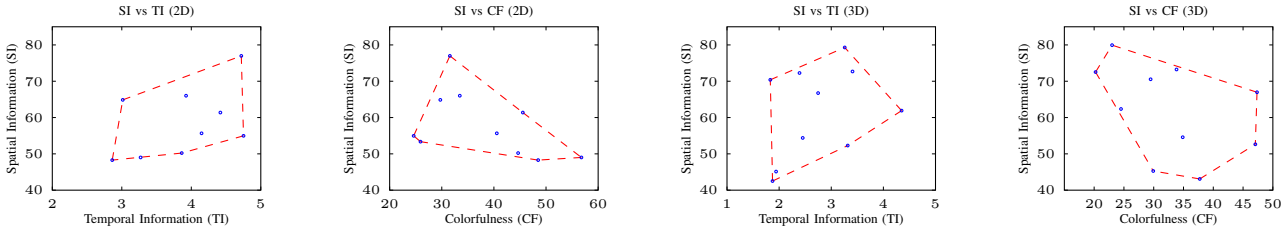


Fig. 3: Spatial Information (SI), Temporal Information (TI), and Colorfulness (CF) measurements on the 2D and 3D databases.

III. LIVE-FBT-FCVR Databases

A. Video Capture

We employed an Insta360 Pro camera to capture the immersive videos. The camera supports a maximum resolution of 7680×3840@30fps on 2D scenes, and 6400×6400@30fps on 3D scenes. To reach this resolution, the videos were captured using the six lens systems and compressed with HEVC, then stitched into a single immersive video, on which a second compression (HEVC) was applied. To minimize compression artifacts, we chose high target bitrates for the compression processes so that we could use the stitched videos as references. Since, in practice, the FOV is affected by the interpupillary distance (IPD) and by the amount of eye relief, 4K resolution is not sufficient to avoid the need for antialiasing. Given the trade-off between aliasing reduction and computational complexity, we used 7680×3840@30fps as the resolution of the 2D video contents, and 5376×5376@30fps for the 3D videos.

For each database, we selected 10 high-quality, diverse reference videos of duration 10s each, captured in Austin, Texas, as shown in Figs. 1 and 2. The videos were stored in YUV 4:2:0 8 bit ERP format. We computed three popular content measurements on all the videos: the Spatial Information (SI), Temporal Information (TI), and Colorfulness (CF) of the reference videos as in Fig 3. SI measures the spatial activity in each luminance frame using Sobel kernels, TI measures the temporal variations of luminance frames by frame differencing, [71], and CF measures the variety and intensity of colors in the videos [72]. The plots illustrate the diversity of scene complexity and colorfulness, but also a limited range in temporal activity, since we did not capture or include videos having large object or camera motions, both to reduce stitching errors and the likelihood of induced motion sickness in the VR environment.

B. Test Sequences

Foveated distortions are characterized by a perceptual quality falloff with increasing eccentricity. In foveated compression / streaming algorithms, this space-variant property is usually implemented by dividing the FOV into two or three concentric, annular regions, on which are applied different levels of foveation, assigning greater quantization factors or lower resolution to the outer regions [11]–[13]. We deployed three regions / levels of foveation to model the falloff in quality, in a manner that could be reasonably implemented by multiple compression quantization parameters (QPs). Seeking to find insights into the proper selection of QPs of both the foveal and peripheral regions in foveated compression algorithms, we used the globally-deployed VP9 codec to create compression distortions. At each level of foveation, we used the VP9 constant quantization mode (Q mode), by specifying the same *-qmin* and *-qmax* parameters in the FFmpeg libvpx-vp9 encoder.

TABLE I: Quantization factors and annular radii.

2D	Quantization $-crf$	51	56	60	63
	radii (radian)	0.08	0.16	0.24	0.32
3D	Quantization $-crf$	51	56	60	63
	radii (radian)	0.1	0.2	0.3	0.4

The design of test VR sequences having foveation / compression distortions involves some unique difficulties. Unlike traditional VQA studies, where the distortion level of a content is determined using a single parameter, the distortion of foveated and compressed videos are determined both by the inner radius of each region and by the level of compression distortion within the region. By using three levels of foveation, the distortions are determined by two inner radii and three QPs. Because of the curse of dimensionality, which heavily impacts the duration of the study, we limited the number of distortion parameters to five.

We created test sequences in three steps. We first sampled the space of compressed videos using 4 QP values ($-crf$ in VP9), yielding 5 levels of compression distortion (including the references), which were determined to have perceptually discriminable levels of distortion when viewed in VR. Second, we divided the FOV into one central, three annular, and one peripheral region, hence 4 radii overall. The selected QPs and radii are shown in Table I. Each foveated / distorted video was created by choosing 3 of 5 compression levels (including the references), and 2 of 4 radii, as shown in Fig. 4. Thus, the highest quality is obtained by selecting $[ref, -crf\,51, -crf\,56]$ as the 3 compression levels in both the 2D and 3D databases, where ref indicates the reference video, and by selecting $[0.24, 0.32]$ as the 2 radii for the 2D database, and $[0.3, 0.4]$ for the 3D database. The lowest quality, however, is obtained when selecting $[-crf\,56, -crf\,60, -crf\,63]$ as the compression distortions (for both databases), and $[0.08, 0.16]$ / $[0.1, 0.2]$ as the radii for the 2D / 3D databases, respectively. The radii were chosen such that the quality range of the test sequences was perceptually broad, i.e. the test sequences having the highest quality would have nearly the same appearance as their corresponding reference videos, while those having

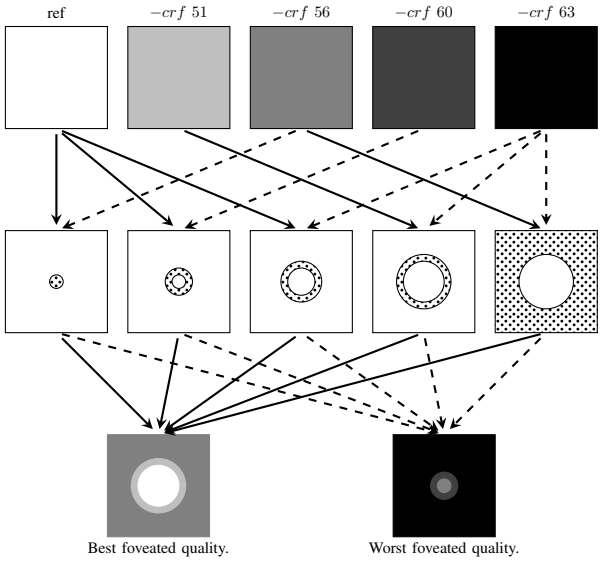


Fig. 4: Ilustration of methods of creating foveated / compressed videos. The inner radii define concentric regions as shown in the second row. The distorted videos are defined by selecting inner radii separating the multiple adjacent foveation regions. The solid arrows indicate the best foveated quality possible, while the dashed arrows indicate the worst foveated quality possible.

the lowest quality would present very poor quality. Finally, given the restriction that distortion increases (quality descends) from the foveal region to the periphery, and that the inner radius is always smaller than the outer radius, there were in total 60 possible combinations of QPs and radii. It was not possible to use all of these, since it would impractically increase the duration of the study, hence we randomly sampled 18 combinations from each content, yielding 180 distorted videos in each database. However, to ensure that a sufficiently wide range of quality would be sampled for each content, we first divided the 60 combinations into 5 broad quality groups, based on a visual comparison by the study authors: Excellent (E), Good (G), Fair (F), Bad (B). We then randomly selected 3, 4, 4, 4, and 3 combinations from the 5 quality groups, respectively.

C. Design Choices and Features of the database

Next we explain a number of design particulars that helped shape the database.

1) VP9 Compression: We selected VP9 codec to apply compression distortions to the test videos. VP9 is one of the most widely used video codecs, and is exemplar of the increasing popularity of royalty-free video coding standards. While the successor AV1 has recently become available, it is not yet deployed in HMDs, and it is reasonable to expect that the coding artifacts produced by these deeply related technologies are perceptually similar.

2) Quantization Parameters: Compression artifacts are often less noticeable in VR environments than when viewed on traditional devices. This may be a result of downsampling in HMDs, which can reduce blocking artifacts [73]. To better represent compression distortions, an aggressive quantization scheme was defined to produce five levels of distortions that are generally perceptually distinguishable in VR. This allows for less labeling ambiguity and more successful model building, as we have discovered in many past studies.

3) Combinations of Quantization Factors and Radii: The most significant difference between the new LIVE-FBT-FCVR databases and traditional databases is that compression distortions were applied in a systematic foveated way, yielding a wide variety of test sequences representative of plausible combinations of distortion severities and foveal-to-peripheral gradations.

IV. Subjective Studies

A. Interface Design and Real-time Foveation

The design of the subject interface required careful handling of the system latency [74], which is the time elapsed between the change in gaze direction and the completion of foveated rendering. In [75], it was suggested that a total system latency of $50 \sim 70$ms could be tolerated, due to the saccadic omission of the HVS. Since we aimed to develop a database that would provide smooth (albeit distorted) foveated viewing experiences, it was crucial to control the system latency to ensure smooth playback.

In the interface, the foveated videos were rendered in real time based on measurements of the subjects' gaze directions. This was made possible since, instead of compressing / foveating the videos in real time, as would be required during application, we pre-compressed the ERP videos using the QPs in Table I. Then to create the foveated experience, we created a foveated video player which was able to read 3 raw / pre-compressed YUV videos and 2 radii from disk, corresponding to three levels of foveation, and transferred them to GPU for foveated rendering by a fragment shader. To achieve this, we relied on a VideoClarity ClearView system equipped with SSD Redundant Arrays of Independent Disks (RAIDs), supporting a sequential reading speed of 10GB/s. Then, the 3 YUV video frames were merged / foveated using the 2 radii by the fragment shader, and finally displayed inside the HMD. The YUV videos were strictly synchronized at frame level to avoid any temporal artifacts during playout, and none were observed. A more detailed description of the foveated video player can be found in [78]. To remove perceptual edge artifacts between the adjacent levels of foveation, linear blending of the content across the sharp foveation boundaries was used:

$$b(x,y) = \begin{cases} \frac{e-e_i+w}{w}, & \text{if } e_i - w < e < e_i, \\ 0, & \text{if } e \leq e_i - w, \\ 1, & \text{if } e \geq e_i, \end{cases} \tag{1}$$

where $i \in 1, 2$ indexes the two boundaries between the 3 levels / regions of foveation, w is the blending width, and $e = \sqrt{(x-x_0)^2 + (y-y_0)^2}$ is the eccentricity of (x, y) with respect to the gaze point (x_0, y_0). For both 2D and 3D databses, we fixed $w = 0.02$ radians for the inner boundary,

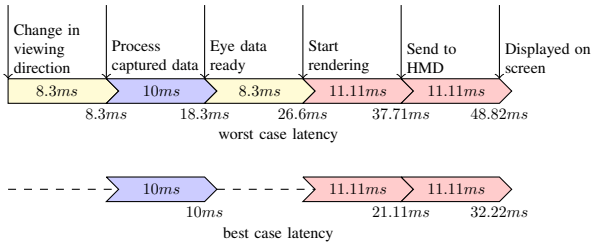


Fig. 5: Best and worst case latency of the system.

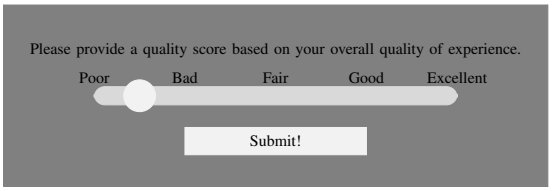


Fig. 6: Rating bar used by the human subjects.

and $w = 0.04$ radians for the outer boundary. The blended pixels were then calculated by

$$I_b(x, y) = b \cdot I_{out}(x, y) + (1 - b) \cdot I_{in}(x, y), \qquad (2)$$

where $I_{out}(x, y)$ and $I_{in}(x, y)$ denote the co-located pixels at (x, y) from contents outside and inside the boundary, respectively.

By estimating both the best and the worst case system latencies, we ensured that our system satisfied the requirements suggested in [75]. An HTC Vive HMD integrated with a Tobii Pro VR eye tracker was employed in the study. The refresh rate of the HMD screen is 90fps, while the sampling frequency of the eye tracker is 120Hz. After a change in gaze direction, the idealized best case would occur when the eye tracker immediately captures the change, while the worst case would occur when the change occurs immediately after the last time sample. The latency in the two extremes would be 0ms and 8.3ms, respectively. The time expended capturing the eye status and data processing by the eye tracker is about 10ms, after which the gaze data is available to the fragment shader. The data could arrive $0 \sim 8.3$ms before the submission of Direct3D [79] calls, and after that, 11.11ms is expended rendering and another 11.11ms sending the rendered image to the HMD panels prior to display [80]. Overall, the latency is about 32ms in the best case and 49ms in the worst case, as illustrated in Fig. 5.

The interface was built using Unity Game Engine, and the foveated video player was compiled into dynamic link libraries (DLL), and then integrated into Unity as native plug-ins. The Tobii VR Unity SDK was employed for calibration and processing of the gaze data [81].

B. Subjective Testing Design

The subjective study utilized a Single stimulus protocol [82], where the subjects recorded scores on a continuous quality scale, ranging from 0 to 1, where 0 denotes the worst quality.

Both of the LIVE-FBT-FCVR databases (2D and 3D) were randomly divided into two sessions, with each session containing 90 of the 180 distorted videos and 10 "hidden" reference videos. To balance the display of distorted videos between the two sessions, the 90 distorted videos were created by randomly selecting 9 of the 18 distorted versions of each content. To avoid the effects of contextual or memory comparisons, videos of the same contents were forced to be located at least three videos apart in the presentations. Care was also taken to avoid any bias owing to a specific order of the sequences, by randomly generating a playlist for each subject. Since the duration of each video is 10s, and the subjects, on average, required less than 10s to assign each score, the total duration of each session averaged less than 35 minutes.

For each subject, the two sessions were separated by at least 24 hours apart to avoid fatigue in the second session. During each session, subjects could terminate the experiment at any time if they felt the need. After the playback of each video, subjects rated the VR video quality using the continuous rating bar shown in Fig 6. The rating bar was marked with Likert labels ranging from "Poor" to "Excellent" to facilitate anchoring the rating process, and subjects could use their controllers to select and submit a score without taking off the headset. The subjects were informed that they could assign their ratings anywhere along the continuous scale. The rating bar was attached to a virtual canvas in HMD local coordinates, so that it remained on the center of the FOV regardless of head movements.

C. Subjects Training

A total of 76 subjects were recruited to participate in the subjective tests, all of them undergraduate students at The University of Texas at Austin, aged between 20 to 30 years, and unfamiliar with video quality assessment and video distortions. Among them, 38 participated in the 2D study, while 38 participated in the 3D study, and no subjects participated in both studies. At the beginning of each study, the Snellen test was conducted to ensure that each subject had normal or corrected-to-normal visual acuity. Subjects were also asked if they were prone to discomfort or nausea when exposed to a VR environment. Prior to the 3D study, the subjects also participated in a RanDot Stereo test of 3D perception. Surprisingly, no subject was rejected as a consequence of screening. The subjects were also asked to adjust the IPD of the HTC Vive HMD to alleviate any discomfort. Subjects having IPDs outside of the range of the HMD ($60.3mm \sim 73.7mm$) were allowed to participate in the study, with the awareness that they could terminate the test if they wanted to.

Before the first session of each study, each subject was orally briefed regarding the purpose of the study and presented with detailed instructions in written form. Then, a training session was conducted to help familiarize the subjects with the system. For the 2D / 3D studies, 12 / 10 training sequences were used, which were not included in the database. The quality range of these videos was similar to the quality range of the test videos, giving the subjects a sense of what they would see in the formal sessions.

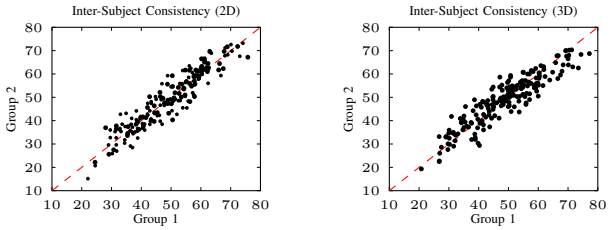


Fig. 7: Scatter plots of MOS from two non-overlapping, equal-size groups of subjects. Left: Inter-subject consistency of the 2D database, SROCC=0.936. Right: Inter-subject consistency of the 3D database, SROCC=0.915.

At the beginning of each session, each subject was guided through an eye tracker calibration process. During this process, the subjects would stare at five red dots that were sequentially displayed at regular spatial intervals. As each dot was displayed and fixated, the gaze direction of the subject was recorded and used to calibrate the eye tracker. During the testing phase, the subjects were instructed to rate the videos based on their own judgments of perceived quality, without expressing any preference of the contents. The subjects were also instructed to view as much as possible of the 360° environment, by moving their eyes and head during the playback of each video.

D. Data Processing

We calculated both subjective Mean Opinion Scores (MOS) and Difference Mean Opinion Scores (DMOS) from the recorded subject ratings. Within each database, denote s_{ijk} as the subjective score given by the i^{th} subject, on the j^{th} foveated video, during the k^{th} session, where j_{ref} is the corresponding reference video. To compute MOS, the Z-scores were first computed per session:

$$\mu_{ik}^{MOS} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} s_{ijk}, \tag{3}$$

$$\sigma_{ik}^{MOS} = \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} (s_{ijk} - \mu_{ik}^{MOS})^2}, \tag{4}$$

$$z_{ijk}^{MOS} = \frac{s_{ijk} - \mu_{ik}^{MOS}}{\sigma_{ik}^{MOS}}, \tag{5}$$

wherein N_{ik} denotes the number of distorted videos viewed by the i^{th} subject in session k. Since the reference videos were rated twice by each subject, the corresponding Z-scores from the two sessions were averaged:

$$z_{ij_{ref}}^{MOS} = \frac{1}{2} \sum_{k=1,2} z_{ij_{ref}k}^{MOS}. \tag{6}$$

To compute DMOS, the differences between the scores of each distorted video and the corresponding hidden reference video was computed,

$$d_{ijk} = s_{ijk} - s_{ij_{ref}k}. \tag{7}$$

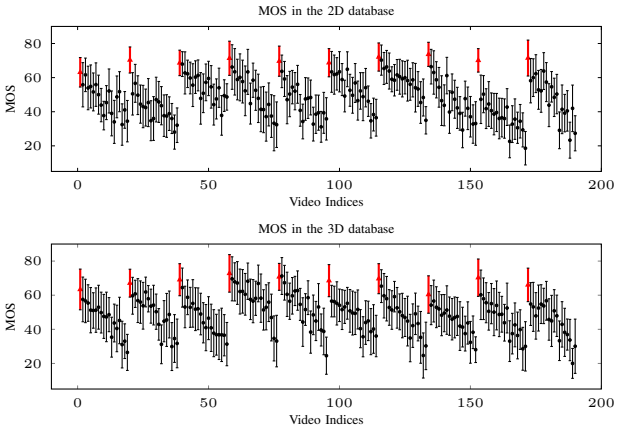


Fig. 8: MOS in the 2D (top) and 3D (bottom) LIVE-FBT-FCVR databases. The MOS of reference videos are highlighted in red.

Then, Z-scores were computed within each session,

$$\mu_{ik}^{DMOS} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} d_{ijk}, \tag{8}$$

$$\sigma_{ik}^{DMOS} = \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} (d_{ijk} - \mu_{ik}^{DMOS})^2}, \tag{9}$$

$$z_{ijk}^{DMOS} = \frac{d_{ijk} - \mu_{ik}^{DMOS}}{\sigma_{ik}^{DMOS}}. \tag{10}$$

The Z-scores from the two sessions were then merged by dropping the index k. Over 99% of the Z-scores were found to lie within the range [-3,3]. Subject rejection was performed following the procedure in [82]. Finally, the Z-scores were mapped to the range [0,100]:

$$z_{ij}^{'} = \frac{100(z_{ij}+3)}{6}, \tag{11}$$

where z_{ij} are Z-scores of MOS or DMOS.

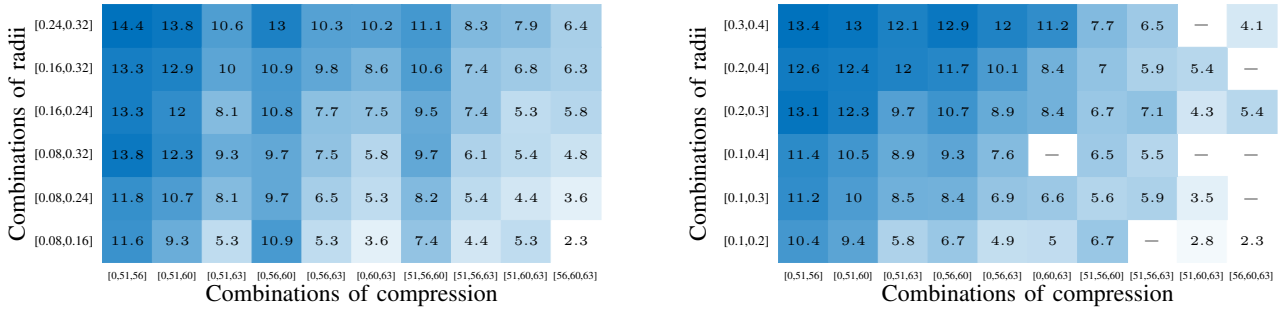
Among the 38 subjects who participated in the 2D study, 2 of them did not finish both sessions, while 6 / 3 of the remaining 36 subjects included in the MOS / DMOS calculations were rejected, respectively. In the 3D study, 4 of the 38 the subjects did not finish both sessions, while 7 / 4 of the remaining 34 subjects included in the MOS / DMOS calculations were rejected. The MOS were found to lie in the ranges [18.61, 73.64] and [20.02, 72.80] in the 2D and 3D databases, respectively. The DMOS were found to lie within the ranges [22.76, 70.28], and [25.04, 68.24], in the 2D and 3D databases, respectively.

E. Validation of Results

1) Inter-Subject Consistency: The inter-subject consistency was explored by randomly dividing the subjects into two disjoint and equal groups, then measuring the Spearman Rank Correlation Coefficient (SROCC) correlation of the MOS values computed from these two groups. We performed the random division 1000 times, and the ranges of correlations

TABLE II: Directions used in the evaluation framework.

Longitude	0	$\pi/6$	$\pi/3$	$\pi/2$	$2\pi/3$	$5\pi/6$	0	$\pi/6$	$\pi/3$	$\pi/2$	$2\pi/3$	$5\pi/6$	0	$\pi/6$	$\pi/3$	$\pi/2$	$2\pi/3$	$5\pi/6$
Latitude	$\pi/4$	$\pi/4$	$\pi/4$	$\pi/4$	$\pi/4$	$\pi/4$	$\pi/2$	$\pi/2$	$\pi/2$	$\pi/2$	$\pi/2$	$\pi/2$	$3\pi/4$	$3\pi/4$	$3\pi/4$	$3\pi/4$	$3\pi/4$	$3\pi/4$



(a) Maps of Mean Ranked Opinion Scores (MROS) of test videos in the 2D database.

(b) Maps of Mean Ranked Opinion Scores (MROS) of test sequences in the 3D database.

Fig. 9: Comparison of MROS observed for different compression / radii combinations. Tables of MROS from the (a) 2D and (b) 3D databases. The trends of the subjective quality scores could be observed by comparing the column / row combinations of compression / radii.

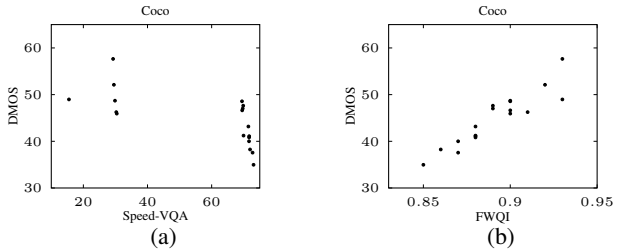


Fig. 10: Example of inconsistent SROCC and PLCC on the 2D database. (a) Scatter plot of Speed-VQA prediction vs. DMOS of content "Coco". (b) The scatter plots of FWQI prediction vs. DMOS on "Coco".

were found to be [0.90, 0.96] / [0.88, 0.94], with median values 0.94 / 0.92 on the 2D / 3D databases. A high degree of subject consistency was observed between the randomly divided groups on both databases, despite the complex viewing conditions introduced by the VR environment, 3D stereo vision, and foveation. The scatter plots of MOS from two groups are shown in Fig. 7.

2) Intra-Subject Consistency: We also measured intra-subject consistency by calculating the SROCC between the Z-scores assigned by each individual subject against MOS [83]. The median correlations on the 2D / 3D database were found to be 0.746 / 0.706, a reasonable degree of intra-subject agreement.

F. Analysis of Opinion Scores

The obtained MOS of the test videos are plotted in Fig 8. The results show that a wide range of foveated / compressed video quality was sampled. The error bars show that the outcomes of the 3D study contain greater uncertainty than those from the 2D study.

To explore the relationships between the scores reported on the combinations of compression distortions and foveation radii, we ranked the Z-scores (DMOS) assigned by each subject on each content, averaged the ranked indices across all subjects, and finally mapped the averaged indices referred to as "Mean Ranked Opinion Scores" or MROS back to the table of all combinations, as shown in Fig. 9, where "–" indicates that the combination was not sampled. By comparing the rows / columns in both maps, on can observe trends in the scores reported for changing combinations of compression / radii.

The maps obtained for the 2D (Fig. 9a) and the 3D (Fig. 9b) databases reveal the expected result that higher scores were assigned to foveated videos having less severe compression artifacts and larger radii (upper left corner of each map), with lowering scores towards the bottom right corners. Compare corresponding rows in the two maps, the relative quality scores may be observed to be in good general agreement. In a few instances, there is disagreement, which may be due to the introduction of 3D and the different display resolutions used for the 2D and 3D studies.

V. Objective Quality Metrics

We evaluate a wide variety of QA algorithms on the newly created LIVE-FBT-FCVR databases. As in [76], four criteria were adopted for evaluation: Pearson's linear correlation coefficient(PLCC), Spearmlan's rank order correlation coefficient (SROCC), Kendall's rank order correlation coefficient (KROCC), and root mean square (RMSE). DMOS were used for evaluating FR / RR algorithms, and MOS were used for training and evaluating NR algorithms. A four-parameter logistic non-linearity was employed before calculating PLCC and RMSE [77]:

$$Q(x) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp(-\frac{x-\beta_3}{|\beta_4|})} \tag{12}$$

A. Evaluation Framework

To recover the foveated experience and enable the comparison of algorithms, simulating the real-time foveation scheme

TABLE III: Performance comparison of FR / RR models on the 2D database for each content and for different quality ranges (underlined). "High" indicates the distorted contents assigned the highest 30% DMOS, "Median" indicates the following 40%, and "Low" indicates the remaining 30%. The best values are boldfaced.

		PSNR	SSIM	MS-SSIM	VIF	S-RRED	Sp-IQA	FSIM	ST-RRED	Sp-VQA	VMAF	FWQI	FASSIM
Coco	PLCC↑	0.6776	0.6638	0.6666	0.8622	0.6388	0.6465	0.6751	0.6607	0.3587	0.7241	**0.9154**	0.8317
	SROCC↑	0.9340	0.9340	0.9319	**0.9381**	0.9195	0.9092	0.9195	0.9319	0.9257	0.9381	0.9133	0.8369
EER	PLCC↑	0.7751	0.8236	0.7858	0.8932	0.7563	0.7504	0.7597	0.7633	0.7649	0.7898	**0.9133**	0.7827
	SROCC↑	0.9381	0.9401	**0.9443**	0.9401	0.9401	0.9340	0.9401	0.9381	0.9340	0.9422	0.9030	0.7585
Stadium	PLCC↑	0.8560	0.8408	0.8466	0.9198	0.8312	0.8305	0.8329	0.8166	0.8312	0.8786	**0.9686**	0.8360
	SROCC↑	0.9505	0.9505	0.9505	0.9587	0.9443	0.9340	0.9505	0.9505	0.9463	0.9587	**0.9628**	0.8349
Starbucks	PLCC↑	0.8003	0.8343	0.8083	0.8671	0.7824	0.7802	0.7894	0.7870	0.7701	0.8109	**0.9005**	0.7346
	SROCC↑	0.9174	0.9174	0.9174	0.9216	0.9133	0.9133	0.9133	0.9133	0.9133	0.9154	**0.9257**	0.7626
Waterfall	PLCC↑	0.8066	0.8203	0.8165	0.9117	0.7958	0.7972	0.8007	0.7996	0.7812	0.8385	**0.9389**	0.8960
	SROCC↑	0.9649	**0.9794**	0.9649	0.9628	0.9484	0.9505	0.9401	0.9567	0.9463	0.9752	0.9608	0.8803
Microsoft	PLCC↑	0.7730	0.7915	0.7793	0.8329	0.7716	0.7725	0.7686	0.7725	0.7666	0.7813	**0.9080**	0.8239
	SROCC↑	0.8824	0.8844	0.8885	0.8927	0.8617	0.8617	0.8555	0.8906	0.8617	0.8968	**0.9071**	0.7750
Oasis	PLCC↑	0.6528	0.7240	0.6675	0.7517	0.6345	0.6284	0.6428	0.6355	0.6188	0.6640	**0.9050**	0.8360
	SROCC↑	0.9257	0.9319	0.9319	0.9546	0.9112	0.7296	0.8989	0.9154	0.9174	0.9257	**0.9670**	0.7626
Playground	PLCC↑	0.8638	0.8727	0.8747	0.9183	0.8678	0.8682	0.8674	0.8690	0.8526	0.8940	**0.9455**	0.8251
	SROCC↑	0.9711	0.9773	0.9711	0.9670	0.9670	0.9690	0.9670	0.9670	**0.9856**	0.9856	0.9690	0.7998
Sanjacinto	PLCC↑	0.5386	0.5642	0.5470	0.7022	0.5073	0.5006	0.5239	0.5018	0.5085	0.6001	**0.8948**	0.8151
	SROCC↑	0.7668	0.7853	0.7668	0.8142	0.5934	0.4964	0.7193	0.7152	0.7028	0.8184	**0.9278**	0.8204
Sculpture	PLCC↑	0.7572	0.8014	0.7667	0.8441	0.7321	0.7306	0.7471	0.7327	0.7090	0.7754	**0.9276**	0.8612
	SROCC↑	0.9133	0.9216	0.9133	0.9133	0.9133	0.9092	0.9133	0.9092	0.9133	**0.9236**	0.9133	0.8596
High	PLCC↑	0.4461	0.3456	0.4504	0.5207	0.5680	0.5686	0.5738	0.5522	0.5858	0.5769	**0.6618**	0.3917
	SROCC↑	0.4517	0.3813	0.4735	0.5480	0.5832	0.5878	0.6193	0.5909	0.6033	0.5763	**0.6729**	0.4069
Median	PLCC↑	0.4246	**0.5225**	0.5121	0.4796	0.4649	0.4204	0.4125	0.2811	0.2737	0.4340	0.3172	0.3017
	SROCC↑	0.3987	**0.4938**	0.4891	0.4779	0.4526	0.4085	0.4266	0.3090	0.2593	0.4476	0.2890	0.2776
Low	PLCC↑	0.2645	0.2266	0.2159	0.4015	0.2847	0.2845	0.2511	0.3424	0.2331	0.4108	0.5084	**0.5897**
	SROCC↑	0.3004	0.2776	0.2144	0.4576	0.2585	0.2815	0.2386	0.3569	0.2277	0.4981	0.5123	**0.6183**
Overall	PLCC↑	0.6941	0.7260	0.7288	**0.8102**	0.7896	0.7760	0.7712	0.6922	0.6584	0.8047	0.7906	0.7573
	SROCC↑	0.6954	0.7191	0.7243	0.8068	0.7885	0.7866	0.7808	0.7010	0.6238	**0.8103**	0.7848	0.7418

TABLE IV: Results of F-test performed on the residuals between model predictions and DMOS values on the 2D database. Each entry in the table is a codeword consisting of 14 symbols, where the first 10 symbols indicate the 10 video contents, the next 3 denote the high, median, and low content quality ranges, and the final symbol denotes the overall performance. A symbol value of "0" indicates the model in the row is statistically superior to the one in the column, a value of "1" indicates statistically inferior, and a value of "-" indicates equivalent.

	PSNR	SSIM	MS-SSIM	VIF	S-RRED	Sp-IQA	FSIM	ST-RRED	Sp-VQA	VMAF	FWQI	FASSIM
PSNR	----------	----------0---	----------	----------0--0	----------0	----------0	----------0	----------	----------	----------0	-00-0000000--0	----------
SSIM	----------1---	----------	----------	----------0	---0-0--------	---0-0--------	---0-0--------	---0-0----1---	-----0----1-0-	---0-0-------0	--00000000--0-	------0-----0-
MS-SSIM	----------	----------	----------	----------0	----------	----------	----------	----------	----------1-0-	----------0	-000000000----	----------0-
VIF	----------1--1	----------1	----------1	----------	1----------	1----------	1----------	1----------1	1---------1--1	----------	--0--00-00----	1----------
S-RRED	----------1	---1-1---------	----------	0----------	----------	----------	----------	----------1	----------1	----------	000-0-0000----	--------0-----
Sp-IQA	----------1	---1-1---------	----------	0----------	----------	----------	----------	----------1	----------1	----------	000-0-0000----	--------0-----
FSIM	----------1	---1-1---------	----------	0----------	----------	----------	----------	----------1	----------1--1	----------	000-0-0000----	--------0-----
ST-RRED	----------	---1-1----0---	----------	0----------0	----------0	----------0	----------0	----------	----------	----------0	-00-0-00000--0	--------0-----
Sp-VQA	----------	-----1----0-1-	----------0-1-	0---------0--0	----------0	----------0	----------0--0	----------	----------	----------0--0	00000-00000--0	------0------0
VMAF	----------1	---1-1-------1	----------1	----------	----------	----------	----------	----------1	---------1--1	----------	-00-0-0-00----	----------
FWQI	-11-1111111--1	--11111111--1-	-111111111----	--1--11-11----	111-1-1111----	111-1-1111----	111-1-1111----	-11-1-11111--1	11111-11111--1	-11-1-1-11----	----------	111--1-1------
FA-SSIM	----------	-------1-----1-	----------1-	0----------	--------1-----	--------1-----	--------1-----	--------1-----	-------1------1	----------	000--0-0------	----------

described in Section IV-A, we adopted a viewport-based assessment framework, as discussed in [78] to simulate the foveated images presented inside the HMD. First, we selected 18 3D directions, and created a viewport video for each direction. The directions are listed in geographic coordinates in Table II. The FOV of each viewport was set to $90°$, and the resolution was set to 1024×1024. Finally, a foveated viewport video was created for each viewing direction in the 2D database, and two foveated videos, associated with the left and the right eye, were constructed for the 3D database, using the same combinations of compression and region radii.

B. Evaluation of FR and RR Algorithms

1) Non-foveated IQA algorithms: We first tested 7 non-foveated IQA algorithms on both databases: PSNR, SSIM [34], MS-SSIM [35], VIF[1] [36], S-RRED [58], Speed-IQA [59] and FSIM [37]. To accommodate these IQA algorithms within our evaluation framework, we computed the score of each frame on each viewport video, then averaged (pooled) scores across all frames and viewports into one final score. On the stereo 3D videos, the predictions from left and right viewport videos were simply averaged.

2) Non-foveated VQA algorithms: We also included three non-foveated VQA algorithms: ST-RRED [58], Speed-VQA [59], and VMAF[2] [54]. The score on each viewport video were also computed and averaged into one final score.

3) Foveated IQA algorithms: Finally, we implemented the legacy foveated models FWQI [66] and FA-SSIM [70], and evaluated their performances. For FA-SSIM, we set the hyper-parameters γ and β to 30 and 1, respectively, on both the 2D and 3D databases.

To evaluate the FR / RR algorithms, we computed the PLCC, SROCC, KROCC, and RMSE of the predicted quality scores against DMOS, and reported only PLCC and SROCC

[1]We used the pixel domain implementation of VIF: https://live.ece.utexas.edu/research/Quality/index_algorithms.htm

[2]We used the pretrained VMAF model from https://github.com/Netflix/vmaf

TABLE V: Performance comparison of FR / RR models on the 3D database for each content and for different quality ranges (underlined). "High" indicates the distorted contents assigned the highest 30% DMOS, "Median" indicates the following 40%, and "Low" indicates the remaining 30%. The best values are boldfaced.

		PSNR	SSIM	MS-SSIM	VIF	S-RRED	Sp-IQA	FSIM	ST-RRED	Sp-VQA	VMAF	FWQI	FASSIM
Bar	PLCC↑	0.4977	0.5823	0.5059	0.7647	0.4444	0.2728	0.4731	0.4555	0.1886	0.5590	0.8704	**0.9192**
	SROCC↑	0.6821	0.7049	0.6821	0.7131	0.6512	0.3994	0.6801	0.6821	0.4324	0.7110	0.8287	**0.8895**
Bridge	PLCC↑	0.5201	0.7045	0.5306	0.7909	0.4387	0.5548	0.4714	0.4556	0.5894	0.6186	**0.8656**	0.8639
	SROCC↑	0.6140	0.6305	0.6120	0.8287	0.6120	0.5686	0.6120	0.6120	0.6326	0.6305	0.8411	**0.8915**
Domain	PLCC↑	0.6589	0.6752	0.6641	0.7964	0.6149	0.6773	0.6166	0.6221	0.5894	0.6702	0.9098	**0.9352**
	SROCC↑	0.7998	0.7998	0.7998	0.8225	0.7812	0.7152	0.7998	0.7812	0.7234	0.7998	0.8865	**0.9008**
Dunk	PLCC↑	0.5819	0.6510	0.5908	0.7392	0.5788	0.5168	0.5874	0.5647	0.4434	0.5927	**0.8006**	0.7625
	SROCC↑	0.7193	0.7276	0.7193	0.7420	0.7090	0.5728	0.7193	0.7193	0.6120	0.7090	0.7626	**0.8244**
EER	PLCC↑	0.7203	0.7312	0.7263	0.8048	0.7021	0.7872	0.6921	0.7036	0.7061	0.7368	0.9080	**0.9272**
	SROCC↑	0.8246	0.8555	0.8246	0.8700	0.8184	0.9133	0.8225	0.8246	0.9236	0.8308	**0.9257**	0.8977
Maintower	PLCC↑	0.5920	0.6661	0.5975	0.7925	0.5749	0.4973	0.5890	0.5515	0.3344	0.6927	**0.8642**	0.7810
	SROCC↑	0.7069	0.7131	0.7049	0.8184	0.6471	0.5067	0.6636	0.7069	0.5170	0.7564	**0.8349**	0.7990
Microsoft	PLCC↑	0.7101	0.7671	0.7175	0.8252	0.6980	0.5678	0.7000	0.7093	0.5673	0.7354	0.8901	**0.9165**
	SROCC↑	0.8782	0.8782	0.8782	0.8927	0.8782	0.6739	0.8782	0.8782	0.6780	0.8782	0.8968	**0.9338**
Oasis	PLCC↑	0.5985	0.6864	0.6093	0.7228	0.5723	0.7041	0.5781	0.5672	0.5375	0.6324	0.8486	**0.9153**
	SROCC↑	0.7668	0.7998	0.7668	0.8080	0.7482	0.7131	0.7647	0.7647	0.7007	0.7874	0.8080	**0.8771**
Redhouse	PLCC↑	0.6368	0.6295	0.6349	0.8013	0.6304	0.5019	0.6353	0.6392	0.3754	0.6745	0.8529	**0.9066**
	SROCC↑	0.7255	0.7564	0.7255	0.7895	0.7090	0.6409	0.7255	0.7255	0.6409	0.7564	0.8390	**0.9333**
Sculpture	PLCC↑	0.6323	0.6274	0.6293	0.7241	0.6074	0.2161	0.6160	0.6101	0.5148	0.6343	0.8275	**0.8684**
	SROCC↑	0.7172	0.7296	0.6904	0.7668	0.6904	0.4964	0.6904	0.6904	0.5212	0.7668	0.7895	**0.8449**
High	PLCC↑	0.1166	0.1290	0.2917	**0.4556**	0.3826	0.2913	0.3344	0.3135	0.2866	0.2528	0.2391	0.3695
	SROCC↑	0.0701	0.1695	0.2454	**0.4113**	0.3348	0.2765	0.2927	0.2852	0.3528	0.2950	0.2672	0.3235
Median	PLCC↑	0.1968	0.2612	0.2566	0.4011	0.3276	0.1406	0.3096	0.2920	0.2124	0.3499	0.4037	**0.4306**
	SROCC↑	0.2259	0.2535	0.2939	**0.4185**	0.2937	0.1288	0.3261	0.2780	0.1123	0.3714	0.4180	0.3903
Low	PLCC↑	0.2428	0.1633	0.1866	0.1590	0.2135	0.3449	0.2395	0.3132	0.2067	0.4072	**0.6500**	0.3692
	SROCC↑	0.2751	0.2295	0.2617	0.2604	0.2595	0.2861	0.2916	0.4123	0.2303	0.4621	**0.6072**	0.3805
Overall	PLCC↑	0.4379	0.4184	0.5337	0.6536	0.5604	0.5084	0.5904	0.5706	0.4666	0.6362	**0.8041**	0.7549
	SROCC↑	0.4418	0.4429	0.5531	0.6765	0.5744	0.4959	0.6237	0.5846	0.4816	0.6522	**0.7841**	0.7401

TABLE VI: Results of F-test performed on the residuals between model predictions and DMOS values on the 3D database. Each entry in the table is a codeword consisting of 14 symbols, where the first 10 symbols indicate the 10 video contents, the next 3 denote the high, median, and low content quality ranges, and the final symbol denotes the overall performance. A symbol value of "0" indicates the model in the row is statistically superior to the one in the column, a value of "1" indicates statistically inferior, and a value of "-" indicates equivalent.

	PSNR	SSIM	MS-SSIM	VIF	S-RRED	Sp-IQA	FSIM	ST-RRED	Sp-VQA	VMAF	FWQI	FASSIM
PSNR	----------	----------	----------	----------0	----------	----------11--	----------	----------	-0----------	----------0	000-00000----0	000-0--00----0
SSIM	----------	----------	----------	----------0	----------1---	----------11--	----------1---	----------1---	-0----------	----------11-0	000000000---00	000-0--00----0
MS-SSIM	----------	----------	----------	----------	----------	----------1---	----------	----------	-0----------	----------	000-00000---00	000-0--00----0
VIF	----------1	----------1	----------	----------	----------	----------1--1	----------	----------	-----1------1	----------	--0-0-------00	0-0-0--------0
S-RRED	----------	----------0---	----------	----------	----------	----------	----------	----------	-0----------	----------	000-00-00----0	000-0--00-0--0
Sp-IQA	----------00--	----------00--	----------0---	----------0--0	----------	----------	----------0	----------	----------0---	----------00	0-0-00-0000-00	0-0-0--0000--0
FSIM	----------	----------0---	----------	----------	----------	----------1	----------	----------	-0----------	----------	000-00-0-----0	000-0--0--0--0
ST-RRED	----------	----------0---	----------	----------	----------	----------	----------	----------	-0----------	----------	000-00-00----0	000-0--00-0--0
Sp-VQA	-1----------	-1----------	-1----------	-----0------0	-1----------	----------1---	-1----------	-1----------	----------	----------	0-0-00-00---00	0-0-0--00----0
VMAF	----------1	----------00-1	----------	----------	----------	----------11	----------	----------	----------	----------	000-0--------0	000-0-----0--0
FWQI	111-11111----1	111111111---11	111-11111---11	--1-1-------11	111-11-11----1	1-1-11-1111-11	111-11-1-----1	111-11-11----1	1-1-11-11---11	111-1--------1	----------	----------1-
FA-SSIM	111-1--11----1	111-1--11----1	111-1--11----1	1-1-1--------1	111-1--11-1--1	1-1-1--1111--1	111-1--1--1--1	111-1--11-1--1	1-1-1--11----1	111-1-----1--1	----------0-	----------

in Table III and Table V for the 2D and the 3D database, respectively, since KROCC and RMSE were observed to follow similar trends. The scatter plots of each model against DMOS are shown in Fig. 11 and Fig. 12.

In Table III and Table V, both the overall performance, per-content performance, and performance in high, median, and low quality ranges are compared. We employed the logistic non-linearity in Eq. 12 to map the predicted scores of each model to the range of DMOS before computing the overall performance (PLCC), and computed per-content PLCC and PLCC for different quality ranges without further mapping. The "High" quality range were distorted contents labeled by the highest 30% of DMOS, "Median" denotes the following 40%, and "Low" denotes the lowest 30%.

As shown in Table III, when tested on the 2D database, the overall performance of a non-foveated model, VIF, was better than that of other models, including the foveated models, FWQI and FA-SSIM. Overall, VIF, FWQI, and S-RRED were the three best performing models. However, when analyzed on a per-content basis, VIF generally performed worse than FWQI. It may also be observed that the SROCC and PLCC of the non-foveated models were generally not consistent (except VIF), while opposite is observed of the foveated models (FWQI and FA-SSIM). This is because the non-foveated models generally had difficulties distinguishing the perceptual relevance of heavily foveated contents, particularly in peripheral regions. Hence they failed to distinguish between perceptually different foveated videos, yielding stucked columns of scatter points, as in Fig. 10. Similar effects may be seen in the all-model (FR and RR) plots in Fig. 11 among the non-foveated models. The foveated models perform well over the low quality ranges, since the importance of quality in the foveal / near-foveal regions are given greater weight. It is also interesting to observe that Speed-VQA and STRRED delivered lower performance than their spatial-only counterparts.

As shown in Table V and Fig. 12, the 3D database, FWQI, FA-SSIM, and VIF were the three best performing models overall. Similar misaligned SROCC and PLCC plots were

TABLE VII: Comparison of NR VQA models on the 2D and 3D databases. The highest values are boldfaced.

	Methods	SROCC↑	KROCC↑	PLCC↑	RMSE↓
2D	BRISQUE	0.797±0.22	0.639±0.18	0.708±0.18	9.60±3.29
	SVBRISQUE	**0.900±0.11**	**0.736±0.12**	**0.884±0.10**	6.91±2.53
	NIQE	0.605±0.32	0.457±0.24	0.675±0.31	**6.47±2.27**
	V-BLIINDS	0.440±0.25	0.327±0.20	0.431±0.25	11.11±2.07
	TLVQM	0.509±0.36	0.381±0.26	0.470±0.36	10.38±3.09
3D	BRISQUE	0.751±0.19	0.587±0.15	0.699±0.17	9.11±2.97
	SVBRISQUE	**0.875±0.12**	**0.695±0.12**	**0.877±0.12**	**5.99±1.79**
	NIQE	0.732±0.19	0.570±0.15	0.781±0.17	6.59±2.00
	V-BLIINDS	0.391±0.24	0.283±0.18	0.300±0.22	9.33±1.50
	TLVQM	0.696±0.21	0.517±0.17	0.699±0.21	7.82±2.43

observed on most of the non-foveated models.

Comparing the performances of models on the two databases, the non-foveated models all experienced a significant performance decrease, while the foveated models were robust on both databases. While the reasons for this are manifold, one of the most may be that: since the predictions of non-foveated models are much more heavily impacted by the heavily distorted periphery, the correlations between peripheral quality and ground truth perceptual quality (DMOS) largely determines the performance of non-foveated models. As may be observed from the Tables in Fig. 9, on the 2D database, MOS / DMOS were much more affected by the most peripheral qualities, but much less so on the 3D database. This suggests the possibility that the perceived depths of non-fixated (likely background) regions were less attended to, i.e. a sort of attentional depth masking.

C. Statistical Evaluation

As in [17], we evaluated the possible statistical superiority of each FR / RR model over every other one based on F-tests between objective models. By assuming that the distribution of the residual errors between the predictions of an objective model and the DMOS follows a Gaussian distribution, the ratios between the variances of residual errors between two objective models follow an F distribution. An F-test was then conducted, the null hypothesis being that the variances of the two models were equal. The possible statistical superiority of one model over another was determined at the 95% significance level. The results of the statistical significance tests on the 2D and 3D databases can be found in Table IV and Table VI, respectively.

The results on the 2D database show that the FR FWQI model was mostly statistically superior to the other models. On the 3D database, the results of the F-test also indicate that FWQI is statistically superior than the other compared models overall.

D. NR Algorithms

We compared 5 NR algorithms on both the 2D and 3D databases: BRISQUE [39], NIQE [40], SVBRISQUE [68], V-BLIINDS [60], and TLVQM [62]. BRISQUE, SVBRISQUE, V-BLIINDS, and TLVQM, were learned using a Support Vector Regressor (SVR) with radial basis function [84].

Among the NR algorithms, SVBRISQUE is a recent model specific to NR FVQA, whereby space-variant NSS were deployed to capture perceptual distortions occurring at different eccentricities. In the model, traditional GGD and AGGD models [39] were extended to space-variant GGD and AGGD models. An assumption of local smoothness was used to estimate local NSS parameters, thereby supplying space-variant eccentricity-dependent quality-aware features. In addition, a neural noise model was deployed to capture uncertainties in visual processing, and to reduce instabilities introduced by image saturation. Finally, an SVR was learned to predict subjective scores (MOS).

TABLE VIII: Median and standard deviation of performances of FR VQA models on the 2D and 3D databases over 45 random iterations of 80-20 train-test splits. The highest values are boldfaced.

	Methods	SROCC↑	KROCC↑	PLCC↑	RMSE↓
2D	PSNR	0.827±0.19	0.653±0.17	0.769±0.19	6.06±1.49
	SSIM	0.827±0.21	0.674±0.18	0.773±0.18	6.35±1.04
	MS-SSIM	0.842±0.21	0.684±0.16	0.801±0.18	6.09±1.10
	VIF	0.877±0.17	0.726±0.16	0.845±0.16	4.53±1.45
	S-RRED	0.884±0.08	0.726±0.10	0.819±0.09	4.86±0.96
	Sp-IQA	0.881±0.07	0.716±0.10	0.808±0.08	5.10±0.86
	FSIM	0.887±0.08	0.733±0.09	0.811±0.10	4.83±1.03
	ST-RRED	0.829±0.09	0.684±0.11	0.771±0.08	5.69±0.92
	Sp-VQA	0.806±0.12	0.621±0.13	0.735±0.09	6.41±1.02
	VMAF	**0.896±0.10**	**0.737±0.12**	0.829±0.12	**4.56±1.24**
	FWQI	0.884±0.09	0.726±0.11	**0.877±0.08**	5.00±0.54
	FASSIM	0.710±0.12	0.558±0.12	0.778±0.11	5.39±1.28
3D	PSNR	0.460±0.41	0.326±0.31	0.467±0.42	7.23±1.90
	SSIM	0.461±0.33	0.347±0.25	0.484±0.34	7.78±1.75
	MS-SSIM	0.597±0.23	0.432±0.19	0.579±0.23	7.04±1.16
	VIF	0.735±0.19	0.558±0.17	0.726±0.18	6.07±1.04
	S-RRED	0.541±0.19	0.421±0.16	0.509±0.17	6.91±1.11
	Sp-IQA	0.520±0.26	0.379±0.20	0.513±0.23	8.09±1.48
	FSIM	0.617±0.18	0.474±0.15	0.527±0.16	6.92±0.90
	ST-RRED	0.611±0.19	0.463±0.15	0.554±0.18	6.75±1.03
	Sp-VQA	0.516±0.24	0.430±0.20	0.553±0.23	6.86±1.38
	VMAF	0.668±0.17	0.495±0.15	0.636±0.17	6.24±0.90
	FWQI	0.815±0.15	**0.653±0.14**	**0.867±0.11**	4.53±0.59
	FASSIM	**0.830±0.17**	0.646±0.17	0.858±0.18	**4.51±1.25**

In each case, the model features were first computed on each viewport video. For BRISQUE, the features from each viewport video were obtained by averaging per-frame features, then averaged across the 18 viewports (36 viewport videos for the 3D database). We chose the hyperparameters of SVBRISQUE exactly as in [68]. Each database was randomly divided into a training set, containing 80% of the sequences, and a test set, containing the remaining 20%, with no overlapping contents between the two subsets. This random division was conducted 1000 times, and the median performance figures reported in Table VII. For NIQE, we computed the predicted scores on each viewport frame, then averaged the scores across all 300 frames and 18 viewports (36 for the 3D database).

As may be observed in Table VII, SVBRISQUE achieved the best quality prediction performance by wide margins. It is interesting that the non-foveated NR models were more robust across databases, in contrast to the FR / RR models. This robustness could have been provided by the SVR.

E. Comparing FR and NR Algorithms

To enable comparisons between FR and NR algorithms, we applied the same NR evaluation procedure to the FR algorithms. First, we randomly selected 2 of the 10 contents, then computed the performance of each FR algorithm on all

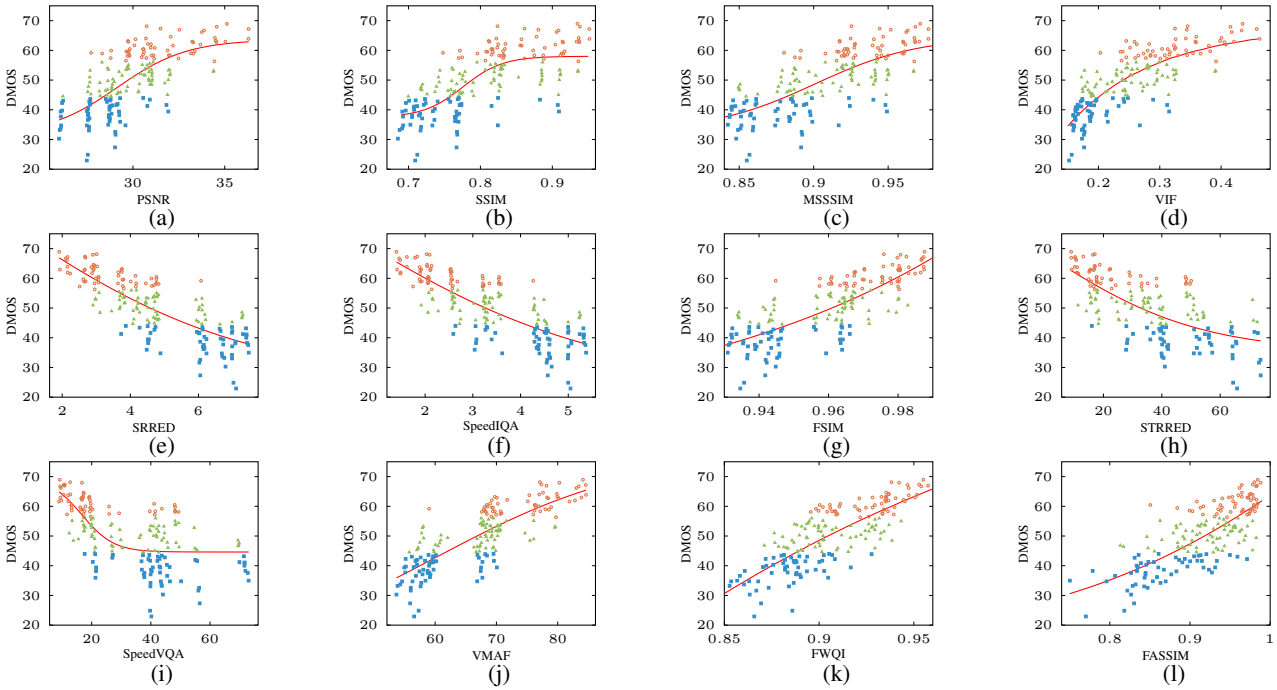


Fig. 11: Scatter plots of all of the compared objective FR and RR VQA scores vs. DMOS on all videos in the new LIVE-FBT-FCVR 2D database. Red, green, and blue points indicates high, median, and low quality ranges, respectively. The red curve indicates the best fitting logistic function

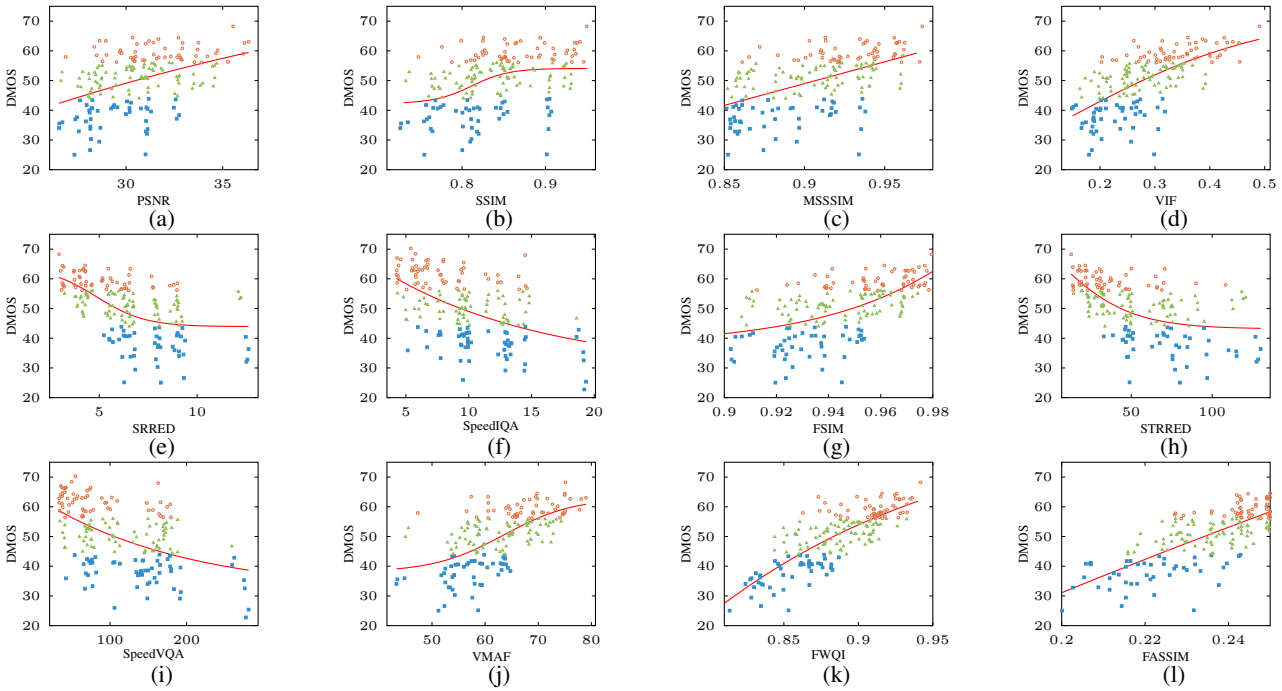


Fig. 12: Scatter plots of all of the compared objective FR and RR VQA scores vs. DMOS on all videos in the new LIVE-FBT-FCVR 3D database. Red, green, and blue points indicates high, median, and low quality ranges, respectively. The red curve indicates the best fitting logistic function

distorted versions of the selected contents. We repeated this process over $\binom{10}{2} = 45$ unique splits. We also used 1000 random train-test splits to match the NR procedure, and the results were essentially the same. Finally, we report the median and the standard deviation of performance in TABLE VIII.

It may be observed that, on both databases, the performance of FR algorithms when adopting the NR evaluation procedure was higher than when evaluated on all the distorted videos. On the 2D database, the non-foveated FR algorithms generally obtained higher performances than the NR algorithms. It may be observed that the non-foveated FR model, VMAF, the foveated FR model, FWQI, and the foveated NR model, SVBRISQUE, were the top three models. On the 3D database, it may be observed that the non-foveated NR algorithms, however, were generally better than the FR algorithms. The reason may be, as explained in Section V-D, that the SVR was able to learn attentional depth masking on the 3D database. It may also be observed that SVBRISQUE was still the best performing model in terms of SROCC, KROCC, and PLCC.

VI. Conclusion

We created a 2D and a stereo 3D VR database of foveated / compressed videos, each containing 10 diverse contents and 180 distorted immersive videos derived from the 10 reference videos. A 2D / 3D subjective study including 38 / 38 subjects was then conducted on the videos. The resulting LIVE-FBT-FCVR databases are unique in terms of the high resolution, foveation distortion, and VR environment. We also presented an evaluation of the performances of a wide variety of objective algorithms on both databases.

A distinguishing feature of our database is that the foveation distortion was considered as a combination of different levels of compression and foveation radii. The results of the subjective evaluations show that, in the 2D study, subjective quality was more affected by peripheral quality, while in the 3D study, the subjective quality was largely affected by foveal quality.

The results of the objective VQA algorithm comparisons provide insights into future algorithm development. In particular, the shortcomings of traditional (non-foveated) VQA algorithms were laid bare.

We believe that the new LIVE-FBT-FCVR databases will benefit the development of future FVQA algorithms, and help facilitate the development of protocols to reduce bandwidth consumption by immersive video streaming services. We also believe that the databases will help understanding of the relationships between the space-variant vision system and the perceptual quality of foveated videos.

Acknowledgment

The authors thank Facebook Technologies for fruitful discussions and supports.

References

[1] Cisco Corp. (Dec. 2018), *Cisco Visual Networking Index (VNI) Complete Forecast Update, 2017-2022.* [Online]. Available: https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/1211_BUSINESS_SERVICES_CKN_PDF.pdf

[2] JVET, *Algorithm descriptions of projection format conversion and video quality metrics in 360Lib.* [Online]. Available: https://mpeg.chiariglione.org/standards/exploration/future-video-coding/n16699-algorithm-descriptions-projection-format-conversion

[3] P. L. Silsbee, A. C. Bovik and D. Chen, "Visual pattern image sequence coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, no. 4, pp. 291-301, Aug. 1993.

[4] T. H. Reeves and J. A. Robinson. "Adaptive foveation of MPEG video," *Fourth ACM International Conference on Multimedia*, New York, 1997.

[5] P. Kortum and W. S. Geisler, "Implementation of a foveated image coding system for image bandwidth reduction," *SPIE Conference on Human Vision and Electronic Imaging*, 1996.

[6] W. S. Geisler and J. S. Perry, "Real-time foveated multiresolution system for low-bandwidth video communication," *SPIE Conference on Human Vision and Electronic Imaging*, 1998.

[7] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1397-1410, 2001.

[8] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 243-254, 2003.

[9] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Transactions on Image Procesing*, vol. 10, no. 7, pp. 977-992, 2001.

[10] Chia-Chiang Ho, Ja-Ling Wu, and Wen-Huang Cheng, "A practical foveation-based rate-shaping mechanism for MPEG videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 11, pp. 1365-1372, 2005.

[11] J. Ryoo, K. Yun, D. Samaras, S. R. Das, and G. Zelinsky, "Design and evaluation of a foveated video streaming service for commodity client devices," *ACM International Conference on Multimedia Systems*, New York, 2016.

[12] M. F. Romero-Rondón, L. Sassatelli, F. Precioso, and R. Aparicio-Pardo, "Foveated Streaming of Virtual Reality Videos," *ACM Multimedia Systems Conference*, New York, 2018.

[13] H. Kim, J. Yang, M. Choi, J. Lee, S. Yoon, Y. Kim, and W. Park. "Eye tracking based foveated rendering for 360 VR tiled video," *ACM Multimedia Systems Conference*, New York, 2018.

[14] G. K. Illahi, T. V. Gemert, M. Siekkinen, E. Masala, A. Oulasvirta, and A. Ylä-Jääski. "Cloud Gaming with Foveated Video Encoding," *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 1, 7 (March 2020).

[15] H. Kalva, "The H.264 Video Coding Standard," *IEEE MultiMedia*, vol. 13, no. 4, pp. 86-90, Oct.-Dec. 2006.

[16] G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[17] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427-1441, June 2010.

[18] A. K. Moorthy, L. K. Choi, A. C. Bovik and G. de Veciana, "Video quality assessment on mobile devices: subjective, behavioral and objective studies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652-671, Oct. 2012.

[19] P. V. Vu and D. M. Chandler, "Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *Journal of Electronic Imaging*, vol. 23, no. 1, p. 013016, 2014.

[20] M. H. Pinson, "The consumer digital video library," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 172–174, 2013.

[21] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron and A. C. Bovik, "Study of Temporal Effects on Subjective Video Quality of Experience," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5217-5231, Nov. 2017.

[22] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, 2014, pp. 989-993.

[23] D. Ghadiyaram, J. Pan and A. C. Bovik, "A subjective and objective study of stalling events in mobile streaming bideos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 183-197.

[24] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias and A. C. Bovik, "Study of Subjective and Objective Quality Assessment of Audio-Visual Signals," in IEEE Transactions on Image Processing, vol. 29, pp. 6054-6068, 2020.

[25] E. Upenik, M. Řeřábek and T. Ebrahimi, "Testbed for subjective evaluation of omnidirectional visual content," *Picture Coding Symposium (PCS)*, Nuremberg, 2016.

[26] W. Sun, K. Gu, G. Zhai, S. Ma, W. Lin and P. Le Calle, "CVIQD: Subjective quality evaluation of compressed virtual reality images," *IEEE International Conference on Image Processing (ICIP)*, Beijing, 2017, pp. 3450-3454.

[27] W. Sun et al., "MC360IQA: The Multi-Channel CNN for Blind 360-Degree Image Quality Assessment," 2019 IEEE International Symposium on Circuits and Systems (ISCAS), 2019, pp. 1-5.

[28] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan and S. Ma, "MC360IQA: A Multi-channel CNN for Blind 360-Degree Image Quality Assessment," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 1, pp. 64-77, Jan. 2020.

[29] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang and X. Yang, "Perceptual Quality Assessment of Omnidirectional Images," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1-5.

[30] M. Chen, Y. Jin, T. Goodall, X. Yu and A. C. Bovik, "Study of 3D virtual reality picture quality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 89-102, Jan. 2020.

[31] M. Xu, C. Li, Y. Liu, X. Deng and J. Lu, "A subjective visual quality assessment method of panoramic videos," *IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, 2017.

[32] H. Duan, G. Zhai, X. Yang, D. Li and W. Zhu, "IVQAD 2017: An immersive video quality assessment database," *International Conference on Systems, Signals and Image Processing (IWSSIP)*, Poznan, 2017.

[33] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at Signal Fidelity Measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98-117, Jan. 2009.

[34] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

[35] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, CA, USA, pp. 1398-1402, Vol.2, 2003.

[36] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430-444, Feb. 2006.

[37] L. Zhang, L. Zhang, X. Mou and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378-2386, Aug. 2011.

[38] D. L. Ruderman, "The statistics of natural images," *Netw. Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

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

[40] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209-212, March 2013.

[41] L. Zhang, L. Zhang and A.C. Bovik, "A feature-enriched completely blind local image quality analyzer," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579-2591, August, 2015.

[42] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang and C. W. Chen, "Blind Quality Assessment Based on Pseudo-Reference Image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049-2062, Aug. 2018.

[43] X. Min, G. Zhai, K. Gu, Y. Liu and X. Yang, "Blind Image Quality Estimation via Distortion Aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508-517, June 2018.

[44] S. Yule, A. Lu, and Y. Lu, "WS-PSNR for 360 video objective quality evaluation," MPEG Joint Video Exploration Team, 116, 2016.

[45] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," *SPIE Optics and Photonics for Information Processing X*, vol. 9970, 2016, Art. no. 99700C.

[46] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," *IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[47] S. Chen, Y. Zhang, Y. Li, Z. Chen and Z. Wang, "Spherical Structural Similarity Index for Objective Omnidirectional Video Quality Assessment," *IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1-6.

[48] Facebook, "Quality assessment of 360 video view sessions," Accessed: Apr. 12, 2019. [Online]. Available: https://code.fb.com/videoengineering/quality-assessment-of-360-video-view-sessions/

[49] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[50] H. G. Kim, H. Lim and Y. M. Ro, "Deep Virtual Reality Image Quality Assessment With Human Perception Guider for Omnidirectional Image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917-928, April 2020.

[51] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312-322, Sept. 2004.

[52] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335-350, Feb. 2010.

[53] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Res.*, vol. 38, no. 5, pp. 743–761, Mar 1998.

[54] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, *Toward a practical perceptual video quality metric*. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html.

[55] S. Li, F. Zhang, L. Ma, and K. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[56] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," *International Conference on Neural Information Processing Systems*, pp. 855-861, 1999.

[57] R. Soundararajan and A. C. Bovik, "RRED indices: Reduced reference entropic differencing framework for image quality assessment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, pp. 1149-1152.

[58] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684-694, April 2013.

[59] C. G. Bampis, P. Gupta, R. Soundararajan and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333-1337, Sept. 2017.

[60] M. A. Saad, A. C. Bovik and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352-1365, March 2014.

[61] X. Li, Q. Guo and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3329-3342, July 2016.

[62] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," in IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5923-5938, Dec. 2019.

[63] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen and J. Häkkinen, "CVD2014—A database for evaluating no-reference video quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3073-3086, July 2016.

[64] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," *Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2017, pp. 1–6.

[65] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061–2077, Sep. 2018.

[66] Z. Wang, A. C. Bovik, L. Lu, and J. L. Kouloheris, "Foveated wavelet image quality index," *SPIE Applications of Digital Image Processing XXIV*, vol. 4472, pp. 42-52, 2001.

[67] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1164-1175, Aug. 1997.

[68] Y. Jin, T. Goodall, A. Patney, and A. C. Bovik, "A Foveated Video Quality Assessment Model Using Space-Variant Natural Scene Statistics," *IEEE International Conference on Image Processing*, submitted, 2021.

[69] Sanghoon Lee, M. S. Pattichis and A. C. Bovik, "Foveated video quality assessment," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 129-132, March 2002.

[70] S. Rimac-Drlje, G. Martinović and B. Zovko-Cihlar, "Foveation-based content Adaptive Structural Similarity index," *International Conference on Systems, Signals and Image Processing*, Sarajevo, 2011, pp. 1-4.

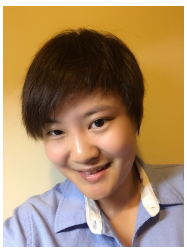
[71] ITU-T Recommendation P.910, *Subjective video quality assessment methods for multimedia applications.* [Online]. Available: https://www.itu.int/rec/T-REC-P.910-200804-I.

[72] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *SPIE Conference on Human Vision and Electronic Imaging*, pp. 87–95, 2003.

[73] Y. Zhang et al., "Subjective panoramic video quality assessment database for coding applications," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 461-473, June 2018.

[74] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder, "Foveated 3D graphics," *ACM Trans. Graph.*, vol. 31, no. 6, 2012.
[75] R. Albert, A. Patney, D. Luebke, and J. Kim, "Latency requirements for foveated rendering in virtual reality," *ACM Trans. Appl. Percept.*, vol. 14, no. 4, 2017.
[76] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Transactions on Image Processing*, submitted.
[77] VQEG, "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment," 2000.
[78] Y. Jin, M. Chen, T. Goodall, Z. Wan, and A. C. Bovik, "Study of 2D foveated video quality in virtual reality," *SPIE Applications of Digital Image Processing*, pp. 18-26, 2020.
[79] Microsoft, *Direct3D* [Online]. Available: https://docs.microsoft.com/en-us/windows/win32/direct3d
[80] A. Vlachos, Valve. (Mar. 2015), *Advanced VR Rendering* [Online]. Available: http://media.steampowered.com/apps/valve/2015/Alex_Vlachos_Advanced_VR_Rendering_GDC2015.pdf
[81] Tobii Pro SDK, [Online]. Available: https://www.tobiipro.com/product-listing/vr-integration/
[82] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," *Int. Telecommun. Union* (2012).
[83] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 541–558, 2013.
[84] C. C. Chang and C. J. Lin. *LIBSVM: A Library for Support Vector Machines [Online]*. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Yize Jin received the B.S. degree in Physics, and the M.S. degree in Microelectronics from Fudan University, Shanghai, China, in 2014 and 2017 respectively. He is currently pursuing Ph.D. degree with the Laboratory for Image and Video Engineering, The University of Texas at Austin. His research interests include image and video processing, visual perception, machine learning, and virtual reality.

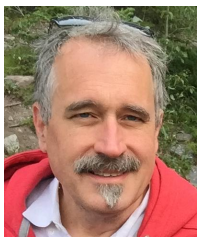


Meixu Chen received the B.Eng degree in Information Engineering from Xi'an Jiaotong University, China, and the M.S. degree in Electrical and Computer Engineering from The University of Texas at Austin, in 2016 and 2019, respectively. She is currently pursuing the Ph.D. degree from the Department of Electrical and Computer Engineering at The University of Texas at Austin. Since 2017, she has been a Research Assistant at the Laboratory for Image and Video Engineering, The University of Texas at Austin. Her research interests focus on image and video processing, machine learning, and perception.

Todd Goodall earned his Ph.D. in Electrical and Computer Engineering from the University of Texas at Austin in May 2018. His research interests include the statistical modeling of images and videos, design of image and video quality assessment algorithms, and visual perception. In addition, he is the recipient of the NDIA UWD Academic Fellowship for 2012-2013 and the Engineering Foundation Endowed Graduate Presidential Scholarship for 2015-2016. He also won the best paper award at the Picture Coding Symposium in 2018.

Anjul Patney is a Principal Research Scientist in NVIDIA's Human Performance and Experience research group, based in Redmond, Washington. Previously and while working on this project, he was a Research Scientist at Facebook Reality Labs (2019-2021) and a Senior Research Scientist in Real-Time Rendering at NVIDIA (2013-2019). He received a Ph.D. from UC Davis in 2013, and B.Tech. from IIT Delhi in 2007.

Anjul's research areas include visual perception, computer graphics, machine learning, and virtual/augmented reality. His recent work led to advances in deep-learning for real-time graphics (co-developed DLSS 1.0), perceptual metrics for spatiotemporal image quality, foveated rendering for VR graphics, and redirected walking in VR environments.



Alan C. Bovik (F '95) is the Cockrell Family Regents Endowed Chair Professor at The University of Texas at Austin. His research interests include image processing, digital photography, digital television, digital streaming video, social media, and visual perception. For his work in these areas he has been the recipient of the 2019 Progress Medal from The Royal Photographic Society, the 2019 IEEE Fourier Award, the 2017 Edwin H. Land Medal from The Optical Society, a 2015 Primetime Emmy Award for Outstanding Achievement in Engineering Development from the Television Academy, a 2020 Technology and Engineering Emmy Award from the National Academy for Television Arts and Sciences, and the Norbert Wiener Society Award and the Karl Friedrich Gauss Education Award from the IEEE Signal Processing Society. He has also received about 10 'best journal paper' awards, including the 2016 IEEE Signal Processing Society Sustained Impact Award. His books include The Essential Guides to Image and Video Processing. He co-founded and was longest-serving Editor-in-Chief of the IEEE Transactions on Image Processing, and also created/Chaired the IEEE International Conference on Image Processing which was first held in Austin, Texas, 1994.