Estimating the Temporal Impairment of Streaming Video: A Dual-Granularity Guided Approach

Guandong Yu, Xiangyang Zhu, Qingbo Wu, *Member, IEEE,* Kede Ma, *Member, IEEE,* King Ngi Ngan, *Fellow, IEEE,* Hongliang Li, *Senior Member, IEEE,* Gaoxiong Yi

***Abstract*—The past decade has witnessed the progress of the model in predicting quality of experience (QoE). Most of these QoE models are developed for traditional video-on-demand streaming services. Nowadays, with the explosive growth of real-time video communication, users' higher expectation of conversation fluency brings more challenges to the modeling ability for temporal distortion. Inspired by the successful application of deep learning technology in sequence-based tasks, the latest studies resort to Three-Dimensional Convolutional Neural Networks (3D-CNN) or Long-Short Term Memory (LSTM) to extract temporal features. Although these methods can be a supplement to spatial feature extractor, the hard inductive bias of 3D-CNN and LSTM essentially limits their performance ceiling in QoE tasks. In this paper, we propose a dual-granularity guided QoE evaluation approach, namely DG-QoE, which explicitly attends to the impact of playback stalling on perceived quality at two different granular levels. To describe local and global temporal features of video, we introduce a set of gating parameters at pixel level and frame level, respectively. By jointly adjusting them, we give DG-QoE the ability to capture features closely related to perceived quality. Furthermore, we employ CNN as feature extractor for single video frame due to its remarkable ability to capture strong visual features. Eventually, the overall QoE prediction could be derived by combining both aspects. Detailed experimental results on benchmark QoE databases demonstrate the superiority of DG-QoE over the representative state-of-the-art metrics.**

***Index Terms*—quality of experience, streaming video, dual-granularity, deep learning.**

I. Introduction

STREAMING videos have become the dominant contributor to the Internet traffic[9], owing to the proliferation of multimedia applications and increasing maturity of wireless communication technology. According to the Cisco Visual Networking Index (VNI), video traffic will occupy 71% of all consumed bandwidth by 2022. Concurrent with the steady rise in user demands on streaming video service is the scarcity of network resources and instability of available bandwidth, in particular on mobile networks. For instance, a drop of available bandwidth may result in a higher compression ratio and/or a lower encoding resolution, leading to compression and scaling artifacts, respectively[30]. Compression artifacts are usually perceived as blocky regions or local flicker, while scaling artifacts are presented as visible blocking, blurring and/or halo artifacts. All these artifacts are spatial, i.e., the temporal aspect of video playback is uninterrupted. On the other hand, network perturbations may also cause packet loss or packet delay.

G. Yu, Q. Wu King Ngi Ngan and Hongliang Li are with School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: gdyu.tony@gmail; qbwu,knngan,hlli,fmmeng@uestc.edu.cn).

Gaoxiong Yi is with the Media Lab, Tencent, Shenzhen 518000, China. (e-mail:)

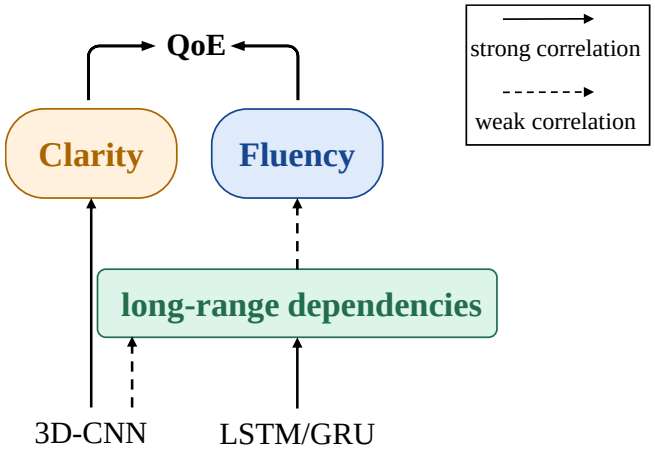


Fig. 1: The main influencing factors of QoE and the existing corresponding modeling methods. QoE can be damaged due to distortions in two dimensions, namely spatial distortion affecting playback clarity and temporal distortion affecting playback fluency. The modeling of spatial distortion using pre-trained neural network has been proved to be successful in IQA tasks. However, the modeling of temporal distortion in the existing deep learning-based methods relies on LSTM or GRU designed to capture long-distance dependence, which is inconsistent with the fact that fluency judgment depends only on adjacent frames.

A usual decoder concealment method is the repetition of the last correctly received frame until the correct new frame is received, which visually results in frame dropping or frame delaying. Such temporal artifacts can even originate at the encoder side due to errors that can appear during the coding procedure.

Both spatial and temporal artifacts impair user Quality of Experience (QoE), i.e., the overall level of user satisfaction while viewing streaming content[27]. Given similar pricing schemes, it will affect the customers' choice between different competing providers. Consequently, a counter provider needs to be able to observe and react quickly on quality problems, at best before the customers perceive them and consider churn. Meanwhile, reducing operational costs with limited bandwidth is also a problem they need to consider. In this case, they have to frequently make decisions on the bandwidth provision and storage resource allocation, based on the trade-

offs between operational costs and service quality perceived by end users. Therefore, understanding if and how video quality affects viewing experience to guide resource allocation is of significance[26], [46], [29], [20].

As end viewers are the final recipients and arbiters of video services, it is most reliable to obtain subjective opinions directly from them, that is, mean opinion score (MOS) in general. However, it depends on a subjective experiment with enough subjects in a well-controlled laboratory environment, which is time-consuming and laborious in practical applications. Hence, objective QoE models are needed to imitate human beings to automatically assess QoE.

QoE is affected by many complex and sometimes inaccessible factors, all of which can be roughly divided into two categories, representing spatial clarity and temporal fluency, respectively. Previous studies have proved from different perspectives that people care more about clarity than fluency[4], [2], [16], [13], which explains the bad performance of early models that simply correlate the statistics of rebuffering events to QoE[45], [37], [24]. To overcome the limitations of the rebuffering centric QoE models, recent efforts take full-reference video quality assessment (FR-VQA) models[30], [39], [41] as the presentation quality measure, achieving highly competitive performance on existing benchmarks[5], [17], [33]. Unfortunately, as the primary contributor to model performance, FR-VQA can not work when processing user-generated content (UGC) videos[21] and in real-time video communication scenarios such as live video broadcasting[7] and video conferencing, because the reference videos are unavailable.

Playback smoothness is another important factor in user QoE. In most existing QoE models, this is measured by statistics related to freezing events, such as the number, duration and frequency of rebuffering. As FR-QVA relies on reference information, it limits the practicability of these models when the access to the application programming interface (API) of a player is declined[24]. More importantly, it is highly difficult for these QoE models to generalize to diverse impairments only by hand-crafted mapping functions. Considering that the video also has sequence attributes, researchers have recently applied the competitive deep learning technology used in sequence-based tasks, such as Long-Short Term Memory (LSTM)[19], [48], [18] and Three-Dimensional Convolutional Neural Networks (3D-CNN)[50], to streaming video QoE tasks. Although these additional modules have been proved to gain in feature extraction, they are not designed specifically for QoE tasks, which leads to doubts about their generalization and interpretability. A reasonable conjecture is that the hard inductive bias set specifically for other tasks may cause lower performance upper bound. On the other hand, the resulting computational overhead is unacceptable for practical application requirements.

Another challenge is the limited capacity for subjective QoE measurements as well as the size of existing QoE datasets. Compared with other tasks with tens of thousands of training samples, the largest QoE dataset contains only less than 1500[14], which is even less than 500 in other QoE datasets[17], [15], [16], [3], [2]. To a certain extent, this requires us to make better use of information in such limited databases to develop models with superior performance.

To address these issues, we aim to design a QoE prediction framework with the following design guidelines and objectives. First, the framework should rely solely on decoded video frames without any other information. Second, additional parameterized modules should be kept to the minimal or even be avoided when possible, unless they have direct and clear significance with the QoE task. Moreover, the framework should exploit inductive prior to reduce dependence on a large number of training data while ensuring robustness of the model. Under these guidelines, as illustrated in Figure 2, we propose DG-QoE, a representation learning framework for streaming video QoE prediction in an end-to-end manner. To achieve a comprehensive description of spatio-temporal distortion, a two-stream architecture is designed to model the complex interaction from two aspects. Taking into account the different aspects of two streams and the reduction of computational overhead, we adopt a lower temporal resolution on the spatial stream and the opposite on the temporal stream. With the decoded frame and absolute residual maps as input, two modules for modeling spatial and temporal distortion calculate the scores representing video clarity and fluency, respectively. Finally, the model outputs the weighted sum of the above two scores as the overall video quality.

To demonstrate the performance of proposed method, we conduct experiments on four benchmark datasets covering a broad set of video contents, encoder configurations, network conditions, ABR algorithms, and viewing devices. Our method is compared with five state-of-the-art methods, and its superior performance is proved by the experimental results in all considered scenarios.

In summary, this paper makes the following key contributions:

- We propose DG-QoE, a novel deep learning-based QoE evaluation framework for streaming video. Without requiring buffer or other manifest information, our method relying only on the decoded video can be applied in a variety of scenarios such as real-time audio-video communication.
- We introduce gating parameters at pixel and frame levels to capture local and global temporal features, respectively. To this effect, we propose a module which can adaptively adjust to learn the impact of playback stalling on QoE rather than relying on empirical fixed parameters.
- The newly proposed module brings considerable performance gains, as well as negligible parameter increases, to facilitate deployment in practical applications.

The rest of the paper is organized as follows. We review the relevant literature in Section II. The proposed DG-QoE framework is depicted in Section III. The details of experimental results are discussed in Section IV. In the end, concluding remarks are given in Section V.

II. Related Works

Recent years have witnessed the constant progresses of objective QoE model[4]. A description of the existing QoE models is shown in Table I. Assuming rebuffering dominates the

TABLE I: Comparison of objective QoE models

QoE model	Input		Distortion modeling	
	Spatial	Temporal	Spatial	Temporal
Mok2011[37]	–	rebuffer statistics	–	linear
Liu2012[32]	bitrate	rebuffer statistics	linear	linear
FTW[24]	–	rebuffer statistics	–	eponential
Xue2014[47]	QP	rebuffer statistics	linear	logarithmic
Yin2015[49]	bitrate	rebuffer statistics	linear	linear
Bentaleb2016[5]	decoded frames	rebuffer statistics	linear	linear
Spiteri2016[42]	bitrate	rebuffer statistics	logarithmic	linear
VideoATLAS[1]	decoded frames	rebuffer statistics	VQA	SVR
P.1203[40]	bitrate, resolution	rebuffer statistics	random forset	random forset
SQI[17]	decoded frames	rebuffer statistics	VQA	linear
KSQI[13]	decoded frames	rebuffer statistics	VQA	non-parametric
DeepQoE[50]	decoded frames	rebuffer statistics	3D-CNN,embedding	3D-CNN
Kwong2021[28]	decoded frames	decoded frames	NSS	LSTM
TRR-QoE[8]	decoded frames	decoded frames	CNN	self-attention
DG-QoE(Ours)	decoded frames	decoded frames	CNN	DG

viewing experience, the earliest QoE models simply correlate the statistics of stalling events to QoE[37], [24]. These works propose linear or exponential equation mapping rebuffering-related metrics to QoE, so they can most directly reflect the fluency of video playback. However, the complete negligence of presentation quality reduces the relevance of these QoE models to perceptual quality. To overcome the limitations of the rebuffering-centric QoE models, many studies propose to complement rebuffering-related metrics with average bitrate or quantization parameter (QP) as the input to the quality prediction model[32], [47], [49], [42], [40]. Nevertheless, the performance of these model depends on the assumption that every single bit contributes equally to the video quality, which is is fundamentally flawed according to the rate-distortion theory[10], and may deteriorate in different compression, transmission and reproduction systems[11], [44]. When bitrate and QP are replaced by state-of-the-art VQA models[30], [39], [41] as the presentation quality measure[5], [17], [33], the ability of QoE model to predict perceived quality is further enhanced. Duanmu *et al.* [17] combined FR quality prediction algorithms with initial loading delay and rebuffering, forming the QoE model named Streaming Quality Index (SQI). Bampis *et al.* [1] propose a machine learning-based framework referred to as Video ATLAS, which combines perceptually video quality, rebuffering-related factors and memory-related functions to predict the end user QoE.

Despite the demonstrated success, most of aforementioned QoE models focus on video-on-demand services, where the reference video information, including manifest, are easily available. However, when these access are limited, especially in real-time audio-video communication scenarios, the above work will fail. Recent work has resorted to deep learning technology to expand its success in other tasks and get rid of dependence on reference information. Considering a video is an image sequence, CNN is widely adopted in QoE model. However, due to the lack of temporal feature extraction, the temporal distortion can not be perceived, and the model fail to achieve impressive performance only relying on CNN. To get rid of this dilemma, Zhang *et al.*[50] use 3D convolutional neural network to extract generalized spatio-temporal features. Recurrent network proved successful in sequence-based tasks such as Long Short-Term Memory (LSTM)[48] is also used to learn the sequence of spatial features as spatio-temporal features and to capture the temporal dependencies involved in the time-varying QoE[28], [19]. Although the features extracted by these additional modules can be used as a supplement to spatial features, the lack of targeted inductive bias still leads to limited performance gains. An intuitive and explainable method is to hard code the characteristics reflecting the video playback fluency into the network structure and replace the borrowed modules. To address this void, we binarize the residual maps of video frames on two different fine-grained level through the gating parameters, and directly extract the features related to playback fluency. Our design complements CNN and can easily implement end-to-end training.

III. Method

In this section, we discuss details of the proposed DG-QoE model whose overall framework is presented in Figure2. Our network is composed of three components: frames sampling and pre-processing, distortion modeling and quality fusion. Given a distorted video to be evaluated, we conduct two sampling and corresponding pre-processing operations at the same time to obtain the video frame and residual maps, which respectively contain the spatial and temporal information of the video. They are then fed to modules Spatial Perceiving Module (SPM) and Temporal Perceiving Module (TPM), respectively. SPM extracts deep visual features and outputs a metric on spatial dimension. Complementarily, TPM is responsible for sensing video temporal consistence and generating the corresponding metric. The overall QoE prediction is finally obtained by the weighted summation of both aspects, where the weight is an optimized parameter. In the following, we would like to elaborate each component.

A. Frame Sampling and Pre-processing

The first phase is sampling, in which redundant frames will be discarded. As the content changes little from frame to frame, videos have high informational redundancy compared with their image counterparts. Many previous works take all frames for training, resulting in computational inefficiency problem on account of the redundant information existing among consecutive frames. In this work, we introduce a sparse

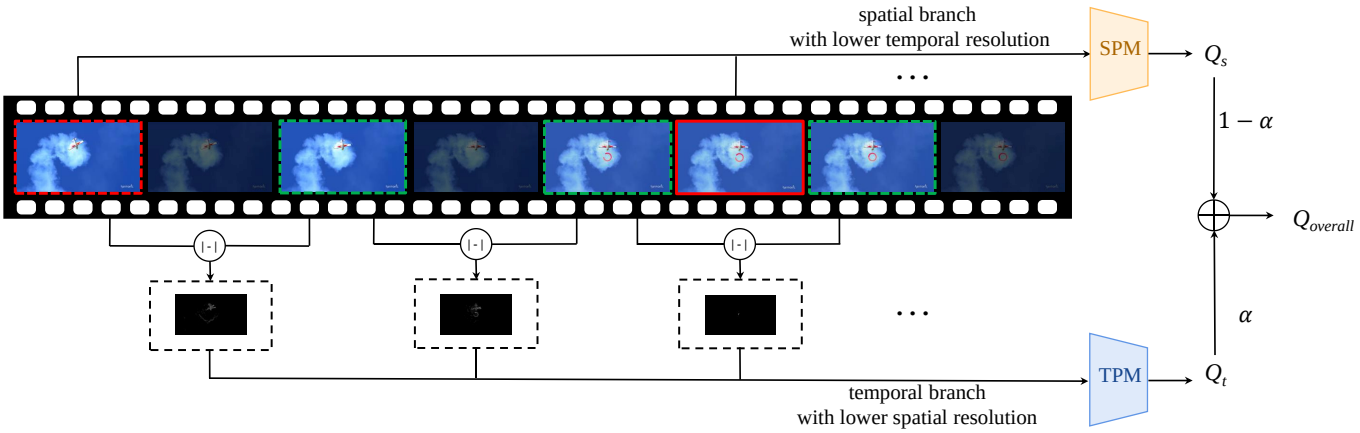


Fig. 2: The overall architecture of DG-QoE. It contains three phases: frame sampling and pre-processing, distortion modeling and quality fusion. Given a decoded video, we first sample two frame sequences at different spatial and temporal resolutions. To depict distortion from both spatial and temporal aspects, one group is fed directly to the SPM while the other is fed to the TPM after the residual procedure. The final QoE could be calculated by the weighted sum of the two. The video is adopted from[2]

sampling strategy in both spatial and temporal dimensions instead of dense frames with full resolution, providing focused and complementary frame information for subsequent modules. Specifically, higher spatial resolution and lower temporal resolution are reasonable and necessary in order to fully capture spatial information of frames while reducing computational overhead. Complementarily, sampling frames with lower spatial resolution and higher temporal resolution is intended to focus on the consistency of playback. Figure 2 shows an example of this sampling strategy.

In order to preserve the information related to playback fluency more explicitly and exclude the irrelevant content information, we apply the residual operation in temporal stream. First of all, we convert the sampled frame into a gray scale to reduce the subsequent computational overhead. Then absolute value of the pixel level differences between decoded frames are calculated by:

$$R^i_{j,k} = \left| I^i_{j,k} - I^{i-1}_{j,k} \right|, \tag{1}$$

where $I^i_{j,k}$ and $R^i_{j,k}$ are the value of the pixel on the position (j,k) of the i-th sampled frame and residual map, respectively. So far, the frame sampling and pre-processing of the video are completed, and the obtained two sets of frame sequences are sent to SPM and TPM, respectively.

B. Distortion Modeling

Given decoded frame sequences and residual maps, the goal of distortion modeling sub-network is to extract discriminative features through SPM and TPM respectively and generate two complementary quality metrics, namely Q_s and Q_t.

1) SPM: The SPM is designed to estimate the presentation quality of frame sequences, which is a vital factor affecting the final QoE. A mass of works have proved the success of CNN trained with large amounts of data in visual tasks such as classification, which has also been well extended in the field of quality evaluation. Encouraged by the positive results, we consider the expansion of pre-trained model to capture the distortion information of frame sequences. The structure of SPM is illustrated in Figure 3.

Mathematically, given a total of T_s frames sampled in spatial streaming, we feed the video frame $\mathbf{f}_t(t = 1, 2, ..., T_s)$ into a pre-train CNN model without fully-connected (FC) layer originally trained for classification tasks, to output the feature vector $\mathbf{v}_t(t = 1, 2, ..., T_s)$, that is:

$$\mathbf{v}_t = \mathrm{CNN}(\mathbf{f}_t). \tag{2}$$

Afterwards, we perform dimension reduction using a single FC layer in the optimization process jointly as followed:

$$q_t = \mathbf{W}_{s1}\mathbf{v}_t + \mathbf{b}_{s1}, \tag{3}$$

where $\mathbf{W}_{s1}$ and $\mathbf{b}_{s1}$ are the weight and bias parameters in the single FC layer.

In practice, the scale of the overall QoE score varies with the predefined requirements. To reduce this impact and improve the versatility of the model, we further normalize q_t to (0,1) by sigmoid function, i.e.,

$$\widehat{q}_t = \frac{1}{1 + e^{-q_t}}. \tag{4}$$

Taking the average of the above results, the overall spatial characteristics can be expressed as:

$$Q_s = \frac{1}{T_s} \sum_{t=1}^{T_s} \widehat{q}_t. \tag{5}$$

2) TPM: Frame freezing is a very common kind of temporal impairment that can be observed in videos streamed over error prone networks and can seriously damage the user QoE. When access to manifest and buffer information is limited, traditional frame difference dependent methods can detect freezing frames by pre-setting multiple thresholds based on experience. It is unrealistic to select a common threshold, while setting these thresholds according to different scenarios is

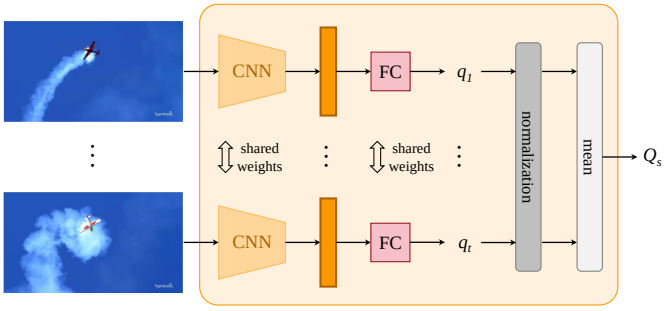


Fig. 3: The structure of the SPM. Each sampled frame is represented by a pre-trained CNN followed by a fully-connected layer and is normalized to (0,1). Considering the overall video frame quality, global average pooling is applied to output spatial metric Q_s.

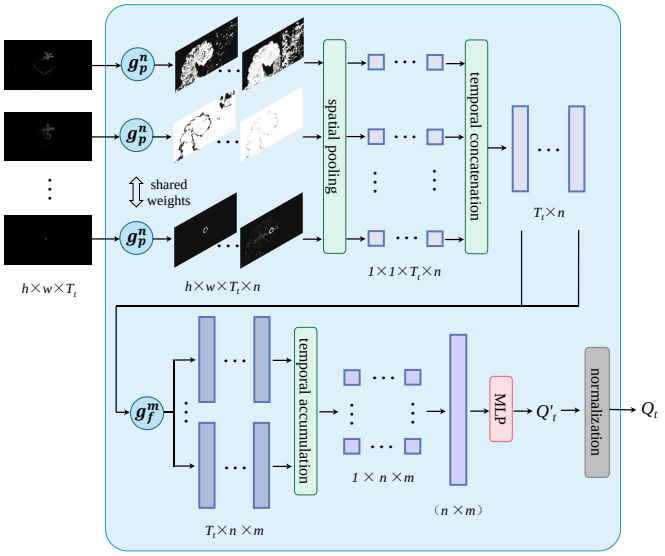


Fig. 4: The structure of the TPM. Two sets of gating parameters $\boldsymbol{g_p^n}$ and $\boldsymbol{g_f^m}$ binarize residual maps from pixel level and frame level, respectively. The gating parameters are optimized jointly with the network. The resulting feature maps are pooled and spliced to form a vector and then fed to a MLP. Finally, the temporal metric Q_t is obtained by normalization.

cumbersome and complex. Here we introduce an interpretable module called TPM, encoding traditional heuristic methods to network structure for end-to-end learning. The structure of TPM is illustrated in Figure 4.

a) Pixel Gating Parameters $\boldsymbol{g_p^n}$: The biggest challenge in modeling the temporal impairment is to distinguish between smooth and stalled videos with slow movement. Based on absolute residual map, we propose a simple and effective modification.

For human visual system, calculating the absolute difference between the two pixel values is much more difficult than judging whether the two values are consistent in perception. Inspired by this, assuming that the human eyes can not distinguish two pixels whose pixel value difference is less than $g_p (0 < g_p \leq 255)$, we perform an approximate binary operation to the residual map:

$$\widehat{R}^i_{j,k} = \frac{255}{1 + e^{-\tau_p (R^i_{j,k} - g_p)}}, \tag{6}$$

where τ_p is a scale factor. In this way, the pixel value with intensity less than g_p in R approaches 0, and the rest approaches 255. With properly set threshold parameter, the residual maps of slow-moving stalled frames and normal frames can be clearly distinguished by spatial pooling:

$$E^i = \frac{\sum_{j=1}^{M} \sum_{k=1}^{N} \widehat{R}^i_{j,k}}{M \times N}, \tag{7}$$

where E^i, M and N are the expectation, height and width of the i-th residual map, respectively. The greater the difference between two frames, the greater the E of their residual map.

Compared with the traditional method, our improvement focus more on whether the pixel value changes, rather than how much it changes. This essentially helps to distinguish the slow motion video frames that are played smoothly from the stalled ones, because the area where the pixels value change in the former is much larger. Figure 5 visualizes this efficiency through specific threshold parameters.

Appropriate parameter setting shows its impressive effect. Furthermore, to enhance the perception range of the model and get rid of its dependence on empirical settings, we integrate it into the network through a set of learnable gating parameters and procure end-to-end training. In the example in Figure 5, the three gating parameters $\boldsymbol{g_p^3} = \{g_{p1}, g_{p2}, g_{p3}\}$ are added to parameter set of DG-QoE and initialized as 1, 2 and 3. This is a set of gating parameters at pixel level, so called $\boldsymbol{g_p^n}$, where n represents the number of parameters in the set.

b) Frame Gating Parameters $\boldsymbol{g_f^m}$: Then we draw the E^i of residual maps in a video into a chord graph, which is shown in Figure 6. It can be clearly seen that even in hard samples such as slow motion, there are some thresholds that can distinguish between smooth and stalled frames. Thanks to the previously designed $\boldsymbol{g_p^n}$, the selection of threshold becomes more robust and the filtering principle becomes obvious.

Similar to the binary principle mentioned in $\boldsymbol{g_p^n}$, we don't care what the specific value of E is. All we need is to judge whether the change of frame is small enough to be difficult to detect by the human eye through E, that is, to simulate people's definition rules for freezing events. Therefore, we perform a similar binary operation to E by:

$$\widehat{E}^i = \frac{1}{1 + e^{-\tau_f (g_f - E^i)}}, \tag{8}$$

where τ_f is a scale factor. In this way, when the difference between two frames is small enough, that is, E is less than g_f, $\widehat{E}$ tends to 1 and otherwise tends to 0. Multiple parameters are set to improve the model's ability to perceive and discriminate the temporal damage from multiple dimensions. In the example in Figure 6, the three gating parameters $\boldsymbol{g_f^3} = \{g_{f1}, g_{f2}, g_{f3}\}$ are added to parameter set of DG-QoE and initialized as 1, 1.5 and 2. This is a set of gating parameters at frame level, so called $\boldsymbol{g_f^m}$, where m represents the number of parameters in the set.

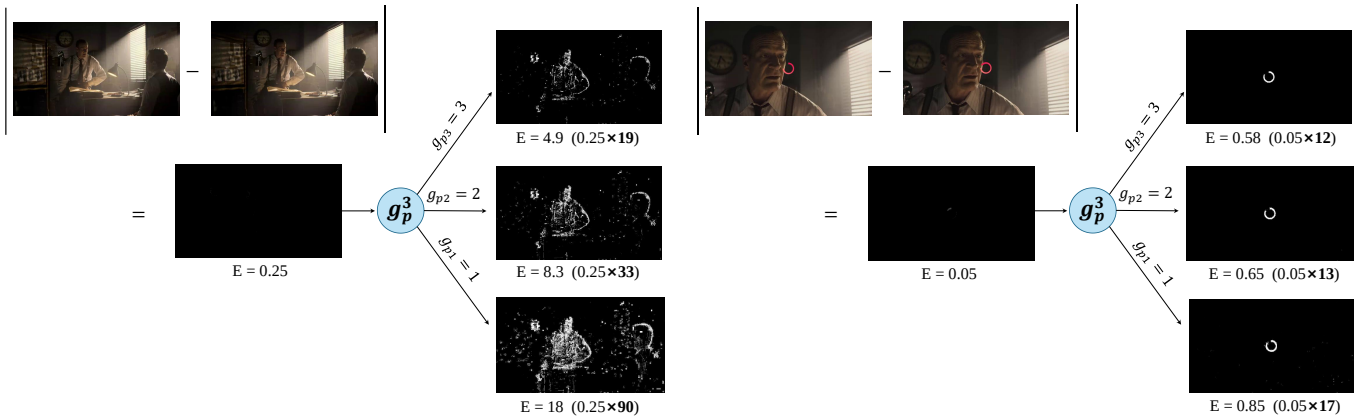


Fig. 5: Illustration of the difference of residual maps between smooth and stalled videos with slow movement under different thresholds. The original residual maps of the two are very similar, and the difference becomes obvious after the binarization procedure. E represents the expectation of residual map intensity value. The video is adopted from[2]

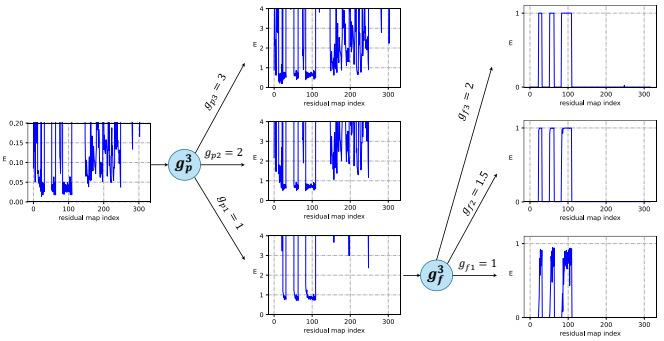


Fig. 6: Expectation value of all residual maps in a video with slow movement. The pixel-level binarization helps to distinguish the smooth playback frames from the freezing ones.

In order to describe the overall temporal characteristics, we accumulate all E^i of frames in the video by:

$$E_v = \sum_{i=1}^{T_t} \widehat{E}^i, \tag{9}$$

where T_t is the number of residual maps in temporal streaming. The longer the stalling duration, the greater the E_v of the video.

c) Dual-Granularity Modeling: As shown in Figure 4, the proposed TPM module relies on the above gating parameters to model the temporal distortion of the video. Given a total of T_t residual maps, each of them will be approximately binarized into n maps by $\boldsymbol{g_p^n}$. Expectations are then calculated to characterize inter-frame differences at different perceptual fine-grained levels. We concatenate the results along temporal dimension to obtain the representation of T_t residual maps of a video over n perceptual scales. Afterwards, $\boldsymbol{g_f^m}$ approximately binarize the obtained features and expand their dimensions by m times. Accumulative pooling preserves feature responses while reducing the temporal dimension. The feature points are flattened into vector $\mathbf{V}$, where each point corresponds to the temporal information represented by the parameter pair (g_{pn}, g_{fm}). Finally, $\mathbf{V}$ is fed into a multi-layer perceptrons (MLP) to output a measurement representing the temporal characteristics:

$$Q'_t = \mathbf{W}_{t2}\delta(\mathbf{W}_{t1}\mathbf{V} + \mathbf{b}_{t1}) + \mathbf{b}_{t2}, \tag{10}$$

where $\mathbf{W}_{t1}$, $\mathbf{W}_{t2}$, $\mathbf{b}_{t1}$, $\mathbf{b}_{t2}$ are the parameters of the FC layers aiming to reduce the feature dimensions, and $\delta(\cdot)$ denotes the ReLU[23] function. Similar to Q_s, Q_t is normalized by:

$$Q_t = \frac{1}{1 + e^{-Q'_t}}. \tag{11}$$

C. Quality Fusion

a) overall QoE prediction: To fully describe the overall perceived quality of video, the weighted sum strategy is applied to two complementary Q_s and Q_t:

$$Q_{overall} = (1 - \alpha)Q_s + \alpha Q_t, \tag{12}$$

where $\alpha \in (0, 1)$ is a parameter jointly optimized with the network to adaptively balance the contribution of Q_s and Q_t to QoE.

b) loss function: Given a mini-batch size of Z, we denote a mini batch of training samples by$\{(\mathbf{X}^{(z)}, y^{(z)})\}_{z=1}^{Z}$, where $\mathbf{X}^{(z)}$ and $y^{(z)}$ represent the z-th raw input video clip and the corresponding ground truth, respectively. During training process, we adopt smooth L_1 loss[22] as:

$$\mathcal{L}_{sL1} = \begin{cases} 0.5(q^{(z)} - y^{(z)})^2 & if\ \left|q^{(z)} - y^{(z)}\right| < 1 \\ \left|q^{(z)} - y^{(z)}\right| - 0.5 & otherwise \end{cases}, \tag{13}$$

where $q^{(z)}$ denotes predicted scores generated by the model. Considering that human beings are more consistent producing rankings of perceptual quality rather than absolute scores[36], we add a supplementary loss function as the Pearson linear correlation coefficient (PLCC) loss [31] between $\{q^{(z)}\}$ and

TABLE II: PLCC between the objective QoE model prediction and MOS on the benchmark datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average
Mok2011[37]	0.311	0.501	0.172	0.032	0.254
Liu2012[32]	0.521	0.731	0.612	0.293	0.539
FTW[24]	0.291	0.586	0.323	0.147	0.337
Yin2015[49]	0.341	0.686	0.742	0.341	0.528
Bentaleb2016[5]	0.743	0.893	0.625	0.682	0.734
Spiteri2016[42]	0.612	0.731	0.798	0.685	0.706
VideoATLAS[1]	0.132	0.643	0.385	0.672	0.458
P.1203[40]	0.324	0.818	0.769	0.636	0.637
SQI[17]	0.759	0.910	0.673	0.717	0.765
KSQI[13]	0.753	0.905	0.794	0.720	0.793
DeepQoE[50]	0.869	0.912	0.813	0.771	0.841
DG-QoE(Ours)	**0.922**	**0.954**	**0.893**	**0.829**	**0.895**

TABLE III: SRCC between the objective QoE model prediction and MOS on the benchmark datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average
Mok2011[37]	0.335	0.516	0.152	0.052	0.264
Liu2012[32]	0.438	0.733	0.598	0.473	0.561
FTW[24]	0.325	0.549	0.182	0.084	0.285
Yin2015[49]	0.441	0.689	0.742	0.541	0.603
Bentaleb2016[5]	0.651	0.891	0.719	0.686	0.737
Spiteri2016[42]	0.493	0.721	0.794	0.665	0.668
VideoATLAS[1]	0.075	0.671	0.472	0.675	0.473
P.1203[40]	0.419	0.828	0.793	0.672	0.678
SQI[17]	0.649	0.912	0.695	0.712	0.582
KSQI[13]	0.658	0.896	0.774	0.703	0.758
DeepQoE[50]	0.791	0.915	0.809	0.752	0.817
DG-QoE(Ours)	**0.903**	**0.955**	**0.879**	**0.804**	**0.875**

TABLE IV: KRCC between the objective QoE model prediction and MOS on the benchmark datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average
Mok2011[37]	0.281	0.441	0.132	0.032	0.221
Liu2012[32]	0.328	0.521	0.442	0.316	0.332
FTW[24]	0.261	0.236	0.123	0.092	0.178
Yin2015[49]	0.331	0.686	0.539	0.381	0.484
Bentaleb2016[5]	0.473	0.707	0.535	0.501	0.554
Spiteri2016[42]	0.379	0.513	0.598	0.475	0.491
VideoATLAS[1]	0.057	0.483	0.335	0.472	0.337
P.1203[40]	0.304	0.618	0.609	0.469	0.500
SQI[17]	0.482	0.734	0.491	0.512	0.555
KSQI[13]	0.493	0.721	0.586	0.563	0.591
DeepQoE[50]	0.613	0.748	0.627	0.595	0.646
DG-QoE(Ours)	**0.807**	**0.820**	**0.704**	**0.621**	**0.723**

$\{y^{(z)}\}$ in the mini-batch. Mathematically, the PLCC loss is computed by:

$$\mathcal{L}_p = \frac{-\sum_{z=1}^{Z}(q^{(z)}-\overline{q}^{(z)})(y^{(z)}-\overline{y}^{(z)})}{\sqrt{\sum_{z=1}^{Z}(q^{(z)}-\overline{q}^{(z)})^2+\epsilon}\sqrt{\sum_{z=1}^{Z}(y^{(z)}-\overline{y}^{(z)})^2+\epsilon}}, \quad (14)$$

where $\overline{q}^{(z)}$ and $\overline{y}^{(z)}$ denote the mean of $q^{(z)}$ and $y^{(z)}$ across the mini-batch. The ϵ is a constant parameter for numerical stability.

Eventually, the final loss function contains $\mathcal{L}_{sL_1}$ and $\mathcal{L}_p$. Since both loss terms are roughly at the same scale, the training is supervised by the following combined loss:

$$\mathcal{L} = \mathcal{L}_{sL1} + \mathcal{L}_p \quad (15)$$

IV. Experiments

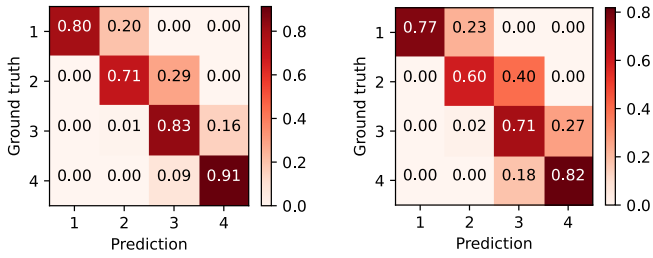
In this section, we first describe the experimental setups, including the QoE dataset, evaluation criteria and implementation detail. We then compare our method with state-of-the-art QoE models. We also conduct an ablation experiment to identify the contribution of key factors in Dg-QoE.

A. Benchmark Databases

We conduct experiments on six publicly available benchmarks to verify that our method is generalizable to general streaming videos, including LIVE-Netflix Video QoE Database, LIVE-NFLX-II Video QoE Database, Waterloo QoE Database, Waterloo-II QoE Database, Waterloo-III QoE Database and Waterloo-IV QoE Database. Their details are summarized as follows:

1) LIVE-Netflix Video QoE Database: LIVE-Netflix Video QoE Database consists of subjective ratings considering 14 source video contents and 112 distorted video sequences obtained by compressing the videos using the H.264 encoder and 8 different playout patterns. A subjective experiment was carried out among 56 subjects.

2) LIVE-NFLX-II Video QoE Database: LIVE-NFLX-II Video QoE Database consists of 15 source videos and a total of 420 distorted sequences (using 7 mobile network traces



(a) DG-QoE. (b) DeepQoE.

Fig. 7: The confusion matrix of the prediction calculated by the proposed DG-QoE and the second-best model (DeepQoE).

and considering 4 client adaptation algorithms). A subjective experiment was carried out among 65 subjects.

3) Waterloo-III QoE Database: Waterloo-III QoE Database consists of 20 source videos and a total of 450 simulated streaming videos with 6 adaptation algorithms under 13 network conditions. All streaming videos are assessed by 34 subjects.

4) Waterloo-IV QoE Database: Waterloo-IV QoE Database consists of 1350 streaming videos (generated from 5 source videos $\times$ 2 encoders $\times$ 9 network traces $\times$ 5 ABR algorithms $\times$ 3 viewing devices). A total of 97 naive subjects participate in the subjective test.

B. Evaluation Criteria

Three criteria are employed for for quantifying the prediction accuracy and evaluating the prediction monotonicity of the models according to the recommendation by the Video Quality Experts Group (VQEG)[43]. We adopt Pearson linear correlation coefficient (PLCC) to evaluate the prediction accuracy, Spearman ranking-order correlation coefficient (SRCC) and Kendell rank correlation coefficient (KRCC) to assess prediction monotonicity. A better objective QoE model should have higher PLCC, SRCC, and KRCC.

C. Implementation Details

From the perspective of lightweight, we choose MobileNetV3-small 0.75[25] pre-trained on ImageNet[12] as the backbone network. We train our model using AdaBound[34] optimizer for 30 epochs with a batch size of 8. The learning rate is set to 3×10^{-4} with a 5-epoch linear warm-up. The weight decay coefficients is set to 1×10^{-4}. Unless otherwise noted, all experiments are conducted on an Ubuntu 18.04 server. The hardware configuration includes Intel® Core™ i9-9900K CPU @3.60GHz $\times$ 16 platform and 2 NVIDIA TITAN Xp GPUs. We use PyTorch[38] to fully take advantage of its dynamic graph design. We set the hyper-parameter τ_p, τ_f and ϵ as 100, 10 and 10^{-10}, respectively.

D. Performance Evaluation

TableII, III and IV show the PLCC, SRCC and KRCC on the benchmark databases, respectively, where the best performers are highlighted with bold face. We evaluate the performance of 11 objective QoE models for streaming videos. The competing algorithms are chosen to cover a diversity of design philosophies, including 8 classic parametric QoE models: Mok2011[37], Liu2012[32], FTW[24], Yin2015[49], Bentaleb2016[5], Spiteri2016[42], SQI[17] and KSQI[13], 3 state-of-the-art learning-based QoE models: VideoATLAS[1], P.1203[40], DeepQoE[50], and the proposed DG-QoE. We adopt the implementation of some of the models publicly available at https://github.com/zduanmu/ksqi and re-implement the rest. For fair comparison, each dataset was randomly partitioned into training and testing data (80/20 split) with non-overlapping content for all the models. We repeat the experiment 10 times based on random splits and report the median values for all compared metrics and the proposed algorithm. Considering that the predicted values and the subjective scores may not share the same scale, we linearly map the MOS ranges of all databases to range [0, 1] to make errors and gradients comparable for different databases.

From the comparison results, we have several observations. First, not surprisingly, models that only consider temporal distortion (Mok2011[37] and FTW[24]) tend to achieve the lowest performance and poor generalizability. In particular, Mok2011[37]'s lowest performance is less than 0.1 and its highest performance is merely about 0.5. FTW[24] has a similar performance. Therefore, the performance of QoE model is improved to varying degrees when spatial distortion is included as the modeling objective, which becomes the design criterion of the advanced model. Second, thanks to the application of deep learning techniques, recent works (DeepQoE[50] and TRR-QoE[8]) have achieved competitive performance up to 0.9, without requiring reference information or manifest files as traditional methods do. By contrast, the classic QoE models with a fixed parametric form fail to faithfully capture the subjective QoE response on streaming videos with complex distortion patterns. Third, the proposed DG-QoE overwhelmingly surpasses all the competing models with significant margins for PLCC, SRCC and KRCC evaluations cross all benchmark databases. It conforms that the direct modeling only considering the neighbor frame relationship could further promote the QoE prediction compared with the module focusing on long-range dependence.

To further get a sense of the effectiveness of the proposed TRR-QoE, the confusion matrixes for the prediction with the second-best metric are exhibited in Figure 7, where the higher value of an entry (darker color) denotes the stronger the correlation between the row value and the column value. The numbers on both axes indicate that it covers a range of QoE scores, for example, 1 covers the interval from 1 to 2. With fewer outliers observed, the proposed DG-QoE performs favorably to the competing method. We also observe that our model achieve better results across the whole score range, especially when predicting extreme scores (1 and 4). This could be attributed to that the proposed model directly captures the strong discriminant properties for low or high QoE.

TABLE V: Ablation study result on benchmark databases

Loss function	spatial module	temporal nodule	LIVE-NFLX-I			LIVE-NFLX-II			WaterlooSQoE-III			WaterlooSQoE-IV		
			PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
$\mathcal{L}_{sL1}+\mathcal{L}_p$	√	√	0.922	0.903	0.807	**0.954**	**0.955**	**0.820**	**0.893**	**0.879**	**0.704**	**0.829**	**0.804**	**0.621**
$\mathcal{L}_{sL1}+\mathcal{L}_p$	√		0.911	0.889	0.772	0.933	0.935	0.782	0.838	0.805	0.609	0.698	0.671	0.524
$\mathcal{L}_{sL1}+\mathcal{L}_p$		√	0.212	0.497	0.456	0.662	0.677	0.508	0.284	0.344	0.261	0.183	0.154	0.116
$\mathcal{L}_{sL1}$	√	√	0.849	0.745	0.620	0.909	0.916	0.747	0.839	0.801	0.619	0.815	0.796	0.609
$\mathcal{L}_p$	√	√	**0.926**	**0.905**	**0.832**	0.939	0.943	0.798	0.881	0.865	0.688	0.822	0.819	0.635

E. Ablation Study

To demonstrate the importance of each module in our framework, we conduct an ablation study. The effectiveness of the loss function is also included. All results were taken from the median of 10-run-experiments and recorded in TableV.

1) Effectiveness of TPM and SPM: We first show the performance of SPM and TPM of DG-QoE when they work alone. In this ablation experiment, $Q_{overall}$ is represented by Q_s and Q_t, respectively. It can be seen from the first three rows in TableV the performance of the two modules used together is better than that of any one of them used alone. This proves that the two modules of DG-QoE complement each other. We also found that the performance of SPM alone is better than that of TPM alone, which is consistent with the previous experimental results and existing works[4], [2], [16], [13], that is, the perceptual quality of presentation dominates the overall QoE of streaming videos. Another interesting finding is that when faced with more complex and realistic datasets (WaterlooSQoE-III and WaterlooSQoE-IV), the performance drop of the model using TPM alone is much lower than that of SPM alone (0.353 and 0.093 PLCC drop from LIVE-NFLX-II to WaterlooSQoE-III, respectively). A reasonable explanation is that the CNN-based method is competent for the task of modeling spatial distortion, and the upper limit of model performance in more complex scenarios is determined by its ability to model temporal damage. This reveals the effectiveness of the proposed TPM from another perspective, which is also one of the reasons why DG-QoE has greater performance lead on more challenging databases than simple ones (0.29 and 0.68 lead in PLCC on LIVE-NFLX-II and WaterlooSQoE-III, respectively).

2) Effectiveness of PLCC loss function: To the best of our knowledge, most of the works related to quality evaluation use L_1[6] or L_2[35] loss to train the model, except Liu[31]. Here we perform another ablation experiment to explore the effects of different losses. Smooth L_1 loss is considered to combine the advantages of L_1 loss and L_2 loss, so we take smooth L_1 loss as their substitute in the experiment. The performance on benchmark databases are compared in TableV (first and last two rows), from which we can see that the model trained with two losses performs the best. We interpret this as due to complementarity of the two losses. Norm loss forces model to focus on the absolute distance between predicted value and ground truth, while PLCC loss guides the model to pay more attention to ranking. Experimental results also show that model trained with PLCC loss outperforms that with norm loss, which may be because human beings are more consistent producing rankings of perceptual quality rather than absolute scores[36]. Adapting evaluation criteria directly to train models may also help improve performance to some extent.

V. Conclusion

In this paper, we presented a novel objective QoE model to blindly estimate the QoE of streaming videos, namely DG-QoE, by designing an interpretable module and integrating it into an end-to-end optimized network. We introduce two sets of gating parameters optimized with the network to hard-code inductive biases into the architectural structure of network in the form of intensity of residual map. Without cumbersome threshold settings and any reference information such as manifest files, the proposed method outperforms the existing objective QoE models by a sizable margin over various QoE databases according to the experimental results.

Instead of resorting to modules originally designed for other tasks such as 3D-CNN or LSTM, we designed a simple but effective module for sensing video playback fluency from the perspective of task requirements. Inductive bias hard coding is an important reason for performance improvement. However, this may also lead to a potentially lower performance ceiling in special cases such as documents in the video. This is also a new direction which will be explored in future as video conferencing are becoming more and more common.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] Christos G Bampis and Alan C Bovik. Learning to predict streaming video qoe: Distortions, rebuffering and memory. *arXiv preprint arXiv:1703.00633*, 2017.

[2] Christos G Bampis, Zhi Li, Ioannis Katsavounidis, Te-Yuan Huang, Chaitanya Ekanadham, and Alan C Bovik. Towards perceptually optimized end-to-end adaptive video streaming. *arXiv preprint arXiv:1808.03898*, 2018.

[3] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231, 2017.

[4] Nabajeet Barman and Maria G. Martini. Qoe modeling for http adaptive video streaming–a survey and open challenges. *IEEE Access*, 7:30831–30859, 2019.

[5] Abdelhak Bentaleb, Ali C Begen, and Roger Zimmermann. Sdndash: Improving qoe of http adaptive streaming using software defined networking. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1296–1305, 2016.

[6] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.

[7] Pengfei Chen, Leida Li, Yipo Huang, Fengfeng Tan, and Wenjun Chen. Qoe evaluation for live broadcasting video. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 454–458, 2019.

[8] Pengfei Chen, Leida Li, Jinjian Wu, Yabin Zhang, and Weisi Lin. Temporal reasoning guided qoe evaluation for mobile live video broadcasting. *IEEE Transactions on Image Processing*, 30:3279–3292, 2021.

[9] Cisco. Cisco annual internet report (2018–2023) white paper. [Online]. Available:http://timmurphy.org/2009/07/22/line-spacing-in-latex-documents/.

[10] L. Davisson. Rate distortion theory: A mathematical basis for data compression. *IEEE Transactions on Communications*, 20(6):1202–1202, 1972.

[11] J. De Cock, Z. Li, M. Manohara, and A. Aaron. Complexity-based consistent-quality encoding in the cloud. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1484–1488, 2016.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[13] Zhengfang Duanmu, Wentao Liu, Diqi Chen, Zhuoran Li, Zhou Wang, Yizhou Wang, and Wen Gao. A knowledge-driven quality-of-experience model for adaptive streaming videos. *arXiv preprint arXiv:1911.07944*, 2019.

[14] Zhengfang Duanmu, Wentao Liu, Zhuoran Li, Diqi Chen, Zhou Wang, Yizhou Wang, and Wen Gao. Assessing the quality-of-experience of adaptive bitrate video streaming, 2020.

[15] Zhengfang Duanmu, Kede Ma, and Zhou Wang. Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach. *IEEE Transactions on Image Processing*, 27(12):6135–6146, 2018.

[16] Zhengfang Duanmu, Abdul Rehman, and Zhou Wang. A quality-of-experience database for adaptive video streaming. *IEEE Transactions on Broadcasting*, 64(2):474–487, 2018.

[17] Zhengfang Duanmu, Kai Zeng, Kede Ma, Abdul Rehman, and Zhou Wang. A quality-of-experience index for streaming video. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):154–166, 2016.

[18] Hossein Ebrahimidinaki, Shervin Shirmohammadi, Emil Janulewicz, and David Cote. Forecasting video qoe with deep learning from multivariate time-series. *IEEE Open Journal of Signal Processing*, pages 1–1, 2021.

[19] Nagabhushan Eswara, S. Ashique, Anand Panchbhai, Soumen Chakraborty, Hemanth P. Sethuram, Kiran Kuchi, Abhinav Kumar, and Sumohana S. Channappayya. Streaming video qoe modeling and prediction: A long short-term memory approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):661–673, 2020.

[20] Guanyu Gao, Huaizheng Zhang, Han Hu, Yonggang Wen, Jianfei Cai, Chong Luo, and Wenjun Zeng. Optimizing quality of experience for adaptive bitrate streaming via viewer interest inference. *IEEE Transactions on Multimedia*, 20(12):3399–3413, 2018.

[21] Deepti Ghadiyaram, Janice Pan, Alan C. Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2018.

[22] Ross B. Girshick. Fast R-CNN. *CoRR*, abs/1504.08083, 2015.

[23] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[24] Tobias Hoßfeld, Raimund Schatz, Ernst Biersack, and Louis Plissonneau. Internet video delivery in youtube: From traffic measurements to quality of experience. In *Data Traffic Monitoring and Analysis*, pages 264–301. Springer, 2013.

[25] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[26] Asiya Khan, Lingfen Sun, and Emmanuel Ifeachor. Qoe prediction model and its application in video quality adaptation over umts networks. *IEEE Transactions on Multimedia*, 14(2):431–442, 2012.

[27] Katrien de Moor Ann Dooms Sebastian Egger et al. Kjell Brunnström, Sergio Ariel Beker. *Qualinet White Paper on Definitions of Quality of Experience*, 2013.

[28] Ngai-Wing Kwong, Sik-Ho Tsang, Yui-Lam Chan, Daniel Pak-Kong Lun, and Tsz-Kwan Lee. No-reference video quality assessment metric using spatiotemporal features through LSTM. In Masayuki Nakajima, Jae-Gon Kim, Wen-Nung Lie, and Qian Kemao, editors, *International Workshop on Advanced Imaging Technology (IWAIT) 2021*, volume 11766, pages 453 – 458. International Society for Optics and Photonics, SPIE, 2021.

[29] Chenglin Li, Laura Toni, Junni Zou, Hongkai Xiong, and Pascal Frossard. Qoe-driven mobile edge caching placement for adaptive video streaming. *IEEE Transactions on Multimedia*, 20(4):965–984, 2018.

[30] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. Toward a practical perceptual video quality metric. *The Netflix Tech Blog*, 6(2), 2016.

[31] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM Multimedia*, pages 546–554, 2018.

[32] Xi Liu, Florin Dobrian, Henry Milner, Junchen Jiang, Vyas Sekar, Ion Stoica, and Hui Zhang. A case for a coordinated internet video control plane. In *Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, pages 359–370, 2012.

[33] Yao Liu, Sujit Dey, Fatih Ulupinar, Michael Luby, and Yinian Mao. Deriving and validating user experience model for dash video streaming. *IEEE Transactions on Broadcasting*, 61(4):651–665, 2015.

[34] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate, 2019.

[35] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2018.

[36] Rafał K. Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012.

[37] Ricky KP Mok, Xiapu Luo, Edmond WW Chan, and Rocky KC Chang. Qdash: a qoe-aware dash system. In *Proceedings of the 3rd Multimedia Systems Conference*, pages 11–22, 2012.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[39] Margaret H Pinson and Stephen Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, 50(3):312–322, 2004.

[40] ITUTP Recommendation. 1203.3,"parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport-quality integration module,". *International Telecommunication Union*, 2017.

[41] Abdul Rehman, Kai Zeng, and Zhou Wang. Display device-adapted video quality-of-experience assessment. In *Human Vision and Electronic Imaging XX*, volume 9394, page 939406. International Society for Optics and Photonics, 2015.

[42] Kevin Spiteri, Rahul Urgaonkar, and Ramesh K Sitaraman. Bola: Near-optimal bitrate adaptation for online videos. *IEEE/ACM Transactions on Networking*, 28(4):1698–1711, 2020.

[43] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. [Online]. Available:http://www.vqeg.org/.

[44] Zhou Wang and Abdul Rehman. Begin with the end in mind: A unified end-to-end quality-of-experience monitoring, optimization and management framework. In *SMPTE 2017 Annual Technical Conference and Exhibition*, pages 1–11, 2017.

[45] Keishiro Watanabe, Jun Okamoto, and Takaaki Kurita. Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes. In *Image Quality and System Performance IV*, volume 6494, page 64940P. International Society for Optics and Photonics, 2007.

[46] Yonggang Wen, Xiaoqing Zhu, Joel J. P. C. Rodrigues, and Chang Wen Chen. Cloud mobile media: Reflections and outlook. *IEEE Transactions on Multimedia*, 16(4):885–902, 2014.
[47] Jingteng Xue, Dong-Qing Zhang, Heather Yu, and Chang Wen Chen. Assessing quality of experience for adaptive http video streaming. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.
[48] Xiangyu Yang and Han Hu. Learning-based qoe prediction and optimization for video streaming. In *2021 6th International Conference on Image, Vision and Computing (ICIVC)*, pages 342–346, 2021.
[49] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over http. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pages 325–338, 2015.
[50] Huaizheng Zhang, Linsen Dong, Guanyu Gao, Han Hu, Yonggang Wen, and Kyle Guan. Deepqoe: A multimodal learning framework for video quality of experience (qoe) prediction. *IEEE Transactions on Multimedia*, 22(12):3210–3223, 2020.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.