2BiVQA: Double Bi-LSTM based Video Quality Assessment of UGC Videos

Ahmed Telili, Sid Ahmed Fezza, Wassim Hamidouche, Hanene F. Z. Brachemi Meftah

2BiVQA: Double Bi-LSTM based Video Quality Assessment of UGC Videos

Ahmed Telili, Sid Ahmed Fezza, Wassim Hamidouche, *Member, IEEE* and Hanene F. Z. Brachemi Meftah

***Abstract*—Recently, with the growing popularity of mobile devices as well as video sharing platforms (e.g., YouTube, Facebook, TikTok, and Twitch), User-Generated Content (UGC) videos have become increasingly common and now account for a large portion of multimedia traffic on the internet. Unlike professionally generated videos produced by filmmakers and videographers, typically, UGC videos contain multiple authentic distortions, generally introduced during capture and processing by naive users. Quality prediction of UGC videos is of paramount importance to optimize and monitor their processing in hosting platforms, such as their coding, transcoding, and streaming. However, blind quality prediction of UGC is quite challenging because the degradations of UGC videos are unknown and very diverse, in addition to the unavailability of pristine reference. Therefore, in this paper, we propose an accurate and efficient Blind Video Quality Assessment (BVQA) model for UGC videos, which we name 2BiVQA for double Bi-LSTM Video Quality Assessment. 2BiVQA metric consists of three main blocks, including a pre-trained Convolutional Neural Network (CNN) to extract discriminative features from image patches, which are then fed into two Recurrent Neural Networks (RNNs) for spatial and temporal pooling. Specifically, we use two Bi-directional Long Short Term Memory (Bi-LSTM) networks, the first is used to capture short-range dependencies between image patches, while the second allows capturing long-range dependencies between frames to account for the temporal memory effect. Experimental results on recent large-scale UGC VQA datasets show that 2BiVQA achieves high performance at lower computational cost than most state-of-the-art VQA models. The source code of our 2BiVQA metric is made publicly available at: https://github.com/atelili/2BiVQA.**

***Index Terms*—Blind video quality assessment, user-generated content, deep learning, Bi-LSTM, spatial pooling, temporal pooling.**

I. Introduction

Currently, video represents the majority of Internet traffic. According to Cisco's recent report [1], it now accounts for around 82% of global Internet traffic. Some of this traffic is generated by streaming video providers like Netflix, Amazon Prime Video, etc. Usually, the content they provide has been created by experts using professional capture devices and in a controlled environment, known as Professionally Generated Content (PGC). PGC videos are pristine high-quality videos that reach a certain level of perfection, making them suitable candidates as references in Video Quality Assessment (VQA) process. On the other hand, User-Generated Content (UGC) accounts for a significant portion of video traffic, which is collected and shared over social media and other video-sharing platforms, such as Facebook, Youtube, TikTok and Twitch. This content is typically captured by nonprofessional users using their own capture devices (e.g., smartphones) and under different shooting conditions. Unlike PGC videos, UGC videos may suffer from multiple authentic distortions that can be introduced during acquisition. Moreover, compression and transmission distortions are still introduced before uploading to the hosting platform. UGC distortions are unpredictable, more diverse, intermixed, and the unavailability of a pristine reference makes the prediction of UGC video quality very challenging. Thus, there is a great need to accurately assess the quality of UGC videos in order to optimize and monitor their processing in hosting platforms, such as their coding, transcoding and streaming.

For VQA, the most reliable technique is to perform a subjective quality evaluation. In subjective tests, a panel of human viewers is asked to rate the quality of stimuli displayed and assessed under a particular protocol and viewing conditions [2]. However, subjective tests are time-consuming, costly, and they cannot be used in real-time applications. As an alternative, objective quality measures have been developed to automatically predict the quality of videos. Depending on the required amount of the reference information, objective VQA metrics can be divided into three categories: Full Reference Video Quality Assessment (FR-VQA), Reduced Reference Video Quality Assessment (RR-VQA) and No Reference Video Quality Assessment (NR-VQA). FR-VQA quality metrics require the presence of the entire pristine video frames to compare against in order to compute the quality score. However, adopting such a strategy for UGC videos is not consistent, since the videos uploaded to the hosting platform have already undergone distortions due to acquisition and compression, making them not suitable as reference videos. Thus, no reference or blind VQA metrics remain the obvious solution that solves the UGC-VQA issue. Although most recent Blind Image Quality Assessment (BIQA)/Blind Video Quality Assessment (BVQA) methods achieve good performance on synthetic distortion datasets [3], their performance on UGC videos remains far from satisfactory [4]–[8], and predicting the quality of UGC videos is still challenging and unsolved problem.

Recently, with the massive growth in social media, attention has moved more towards building an accurate and efficient BVQA model suitable for UGC content, which allows achieving more intelligent analysis and processing in various applications [9]. Hence, in recent years, researchers have deployed considerable efforts into the development of in-the-wild UGC datasets such as KoNViD-1k [5], LIVE-VQC [6], and YouTube-UGC [7], to cite a few examples. These UGC datasets differ significantly from synthetic distortion datasets

A. Telili and W. Hamidouche are with Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France (e-mail: atelili@insa-rennes.fr and whamidouche@insa-rennes.fr).

SA. Fezza and H. F. Z. Brachemi Meftah are with National Higher School of Telecommunications and ICT, Oran, Algeria (e-mail: sfezza@ensttic.dz).

by a varied type and a wide range of distortions, but also by the fact that the distortion is not uniformly distributed over the spatial and temporal domains, resulting in fluctuating video quality.

The existing metrics do not consider or consider insufficiently this last aspect. They do not take into account how non-uniform distortions affect the overall frame quality score and how adjacent frames, from past and future, impact the perceived quality of the current frame. Typically, existing metrics use the mean as a pooling strategy, which is not a good representation of the spatial-temporal quality distributions. In this regard, we use a data-driven deep-learning approach in the proposed metric to enhance the VQA.

It is obviously desirable to have accurate video quality metrics for the UGC videos. Thus, in this work, we propose an efficient model for UGC-VQA, termed 2BiVQA for double Bi-LSTM Video Quality Assessment. The main contributions of this paper can be summarized as follows:

- We propose an accurate and efficient BVQA metric for UGC that performs the quality assessment in line with the Human Visual System (HVS). The components of 2BiVQA include a Convolutional Neural Network (CNN) for spatial feature extraction and two Recurrent Neural Networks (RNNs) for capturing spatial-temporal dependencies. We show that pre-training the features extraction module on an in-the-wild Image Quality Assessment (IQA) dataset significantly improves the performance of 2BiVQA.
- We leverage two RNNs, namely Bi-directional Long Short Term Memory (Bi-LSTM) networks, for both spatial and temporal pooling, which allows our model to take into account the characteristics of UGC videos as well as the HVS behavior.
- We conduct experiments on three widely-used UGC-VQA datasets to demonstrate the effectiveness of 2BiVQA. Experimental results show that the proposed 2BiVQA metric achieves competitive performance with State-Of-The-Art (SOTA) methods and provides the best generalization capability, even at a low computational cost.

The rest of this paper is organized as follows. Section II reviews related work, then Section III presents the proposed 2BiVQA model. The performance of our model is assessed and analyzed in Section IV. Finally, Section V concludes the paper.

II. Related work

Given the unavailability of pristine sources, FR-VQA metrics cannot well predict the perceptual quality of UGC videos. Thus, in this section, we focus on BVQA methods, as these methods are the most suitable for providing UGC video quality estimation. BVQA methods can be grouped into two categories, whether their relevant features are extracted from the input video based on conventional handcrafted techniques or deep learning-based models.

A. Handcrafted feature-based BVQA models

The earliest BIQA/BVQA methods were mostly distortion-specific QA algorithms, which address a specific type of distortion, such as blur [10], [11], blockiness [12], [13], ringing [14], banding [15], [16] and noise [17], [18] or targeted multiple types of distortion [19]–[21]. Later on, the most successful handcrafted features-based BVQA models mainly rely on learning approaches [22], [23], using a set of relevant perceptual features combined with a regression model to predict quality scores [24]–[27].

The most popular BIQA/BVQA algorithms are based on Natural Scene Statistics (NSS) [28], extracted from either spatial domain or transform domain. NSS refer to the hypothesis that the natural scenes form a minor subspace within the space of all conceivable signals. These NSS are altered by the presence of distortions, so they were widely used to blindly measure the quality of images/videos. Successful models relying on NSS are derived from the spatial domain (NIQE [29], BRISQUE [29]), Discrete Cosine Transform (DCT) (BLIINDS [30] and BLIINDS-II [31]) and Discrete Wavelet Transform (DWT) (BIQI [32], DIIVINE [24], C-DIIVINE [33]). These metrics have been expanded to the VQA task using the space-time natural video statistics models [34]–[36]. For instance, in [37], an NSS-based method was proposed for BIQA, which consists in extracting NSS features from the local binary pattern (LBP) map and the mean subtracted and contrast normalized (MSCN) coefficients of the image. Thus, the extracted NSS features along with other features related to perceptual characteristics constitute the quality-aware features for quality estimation. Other extensions of NSS have been proposed including in log-derivative and log-Gabor spaces (DESIQUE [38]), the joint statistics of the gradient magnitude and Laplacian of Gaussian responses in the spatial domain (GM-LOG [39]) and the gradient domain of LAB color transforms (HIGRADE [25]). HIGRADE is based on both gradient orientation information extracted from the gradient structure tensor and gradient magnitude. Once these features are extracted, a mapping is performed from the feature space to the Mean Opinion Score (MOS) scores using the Support Vector Regression (SVR). In [26], the authors proposed a bag of feature-maps approach. In which several feature maps are derived from multiple color spaces and transform domains, then scene statistics from each of these maps are extracted. The obtained results demonstrated the relevance of the extracted features for the quality prediction of images corrupted by complex mixtures of authentic distortions. Motivated by the success of unsupervised feature learning for BIQA in CORNIA [40], the authors proposed to extend it to video signal V-CORNIA [41]. In V-CORNIA, frame-level features are extracted by unsupervised feature learning, and a SVR is subsequently used to learn a mapping from feature space to frame-level quality scores. Finally, an overall video quality score is derived via a hysteresis temporal pooling. Min *et al.* [21], [42] introduced a new concept called Pseudo-Reference Image (PRI) and developed a PRI-based BIQA framework (BPRI). Unlike traditional reference image, which is assumed to have a perfect quality, PRI is generated from the

same distorted image and intentionally subjected to the highest distortion. Notably, this framework employs PRI to predict image quality by formulating distortion-specific metrics that evaluate diverse types of distortion, including blockiness, sharpness and noisiness. These metrics evaluate the structural similarity between a distorted image and its corresponding PRI. Thus, a highly distorted image will have a higher degree of similarity with the corresponding PRI. The concept of PRI has been extended to Multiple PRIs (MPRI) in [21], which are obtained by further degrading the distorted image in several ways and to certain degrees, and then comparing the similarities between the distorted image and its MPRIs.

To leverage these rich IQA metrics for VQA context, a straightforward approach is to compute the quality score of each frame and then pool them into an overall video quality score. The most adopted temporal pooling method is the average, however, this approach does not take into account the temporal change and quality fluctuation. This is why more advanced temporal pooling strategies have been proposed [41], [43], [44].

However, performing VQA cannot be based solely on spatial information, i.e., based only on IQA metrics, since temporal information such as motion plays a crucial role in the perception of quality/distortion in the video and must be taken into account. Therefore, unlike simply extending IQA methods to assess video quality using a pooling strategy, other methods have attempted to include temporal information directly in their models. For instance, a completely blind metric called Video Intrinsic Integrity and Distortion Evaluation Oracle (VIIDEO) was proposed in [45]. VIIDEO is based on a set of perceptually relevant temporal video statistic models of video frame difference signals. Inter-subband correlations over local and global time spans were used to quantify the degree of distortion in the video and thus predict the quality score. Manasa and Channappayya [46] proposed estimating perceptual quality by estimating statistical irregularities in optical flow using features at the patch and frame levels. V-BLIINDS has been proposed in [36], which includes a spatiotemporal NSS model of DCT coefficient statistics, as well as a motion model that quantifies motion coherency in the video. Li *et al.* [34] proposed a BVQA based on the spatiotemporal statistics of videos in the 3D-DCT domain, which allows describing the spatial and temporal regularities of local space-time regions simultaneously. Two-Level Approach for No-Reference Consumer Video Quality Assessment (TLVQM) [47] is another handcrafted features-based BVQA method relying on a two-level approach for features extraction. First, Low Complexity Features (LCF) are calculated at a rate of one frame per second over the entire video sequence, then the LCF are utilized to select a set of representative subset of frames for calculating High Complexity Features (HCF). Finally, both low and high complexity features are aggregated as a single feature vector representing the whole video sequence by using SVR as a regression model. A more recent fusion-based BVQA model is VIDeo quality EVALuator (VIDEVAL) [4], which is based on features selection among top-performing BIQA/BVQA models such as BRISQUE, HIGRADE, TLVQM, etc. To select the most relevant features, Random Forest (RF) is used to remove the less significant features. Finally, a Support Vector Machine (SVM) with a linear kernel is trained to regress the final features vector into a quality score. Kancharla *et al.* [48] proposed a BVQA method, which includes a bandpass filter model of the visual system to evaluate the temporal quality and a weighted NIQE module to estimate the frame-level spatial quality. Finally, the global video quality score is computed by the average of the spatial quality and the temporal quality. All previous methods tried to predict the average quality perceived by end users, known as MOS. Differently, in [49], the authors proposed to model how a single observer perceives the media quality using a neural network instead of predicting the MOS. The training of a neural network relies on the observer's ratings collected from subjective experiments to mimic his quality judgment, which implicitly accounts for his individual characteristics such as user expectations and personality that have an impact on quality of experience [50].

B. Deep learning-based BVQA models

In recent years, deep CNNs have shown outstanding performance in a wide range of computer vision tasks such as image classification [51], [52], object detection [53], [54] and image segmentation [55], [56]. Recently, with the release of several larger IQA/VQA datasets [7], [13], [57], [58], deep CNNs have been extensively explored to solve image/video quality assessment problem. Yet, due to the lack of large-scale IQA/VQA datasets, it is quite challenging to train a deep CNN from scratch to reach a competitive performance. To overcome the limitation of small data size, two solutions have been used in the literature: 1) performing a patch-wise training to increase data samples [59], [60], or 2) leveraging pre-trained deep CNNs on large datasets like ImageNet [61], then performing fine-tuning on target IQA/VQA datasets.

The first adoption of a CNN model to the problem of IQA was made by Kang *et al.* in [59], where a CNN was used to blindly predict the image quality score. They combined feature learning and regression in end-to-end optimization without using handcrafted features. Following this work, considerable deep learning-based BIQA methods have been proposed [62], [63], which achieved quite good performance. For video, on the other hand, very few methods based on deep learning have been dedicated to BVQA.

For instance, Ahn *et al.* [64] proposed a BVQA metric based on a deep CNN model named DeepBVQA, which includes various spatial and temporal cues. In DeepBVQA, a pre-trained CNN model for IQA is used to extract spatial features from each frame, and temporal sharpness variation is exploited to extract temporal features. Then, these spatial and temporal features are combined into a video feature to be regressed to a final quality score. Another deep learning-based VQA model was proposed in [65], which consists of a 3D-CNN to extract spatio-temporal features followed by a Long Short Term Memory (LSTM) to predict the perceived video quality. A multi-task CNN framework, named V-MEON, was proposed in [66] that predicts both the quality score and codec type of a video. V-MEON is based on 3D-CNN network to extract spatio-temporal features from a video, followed by the codec

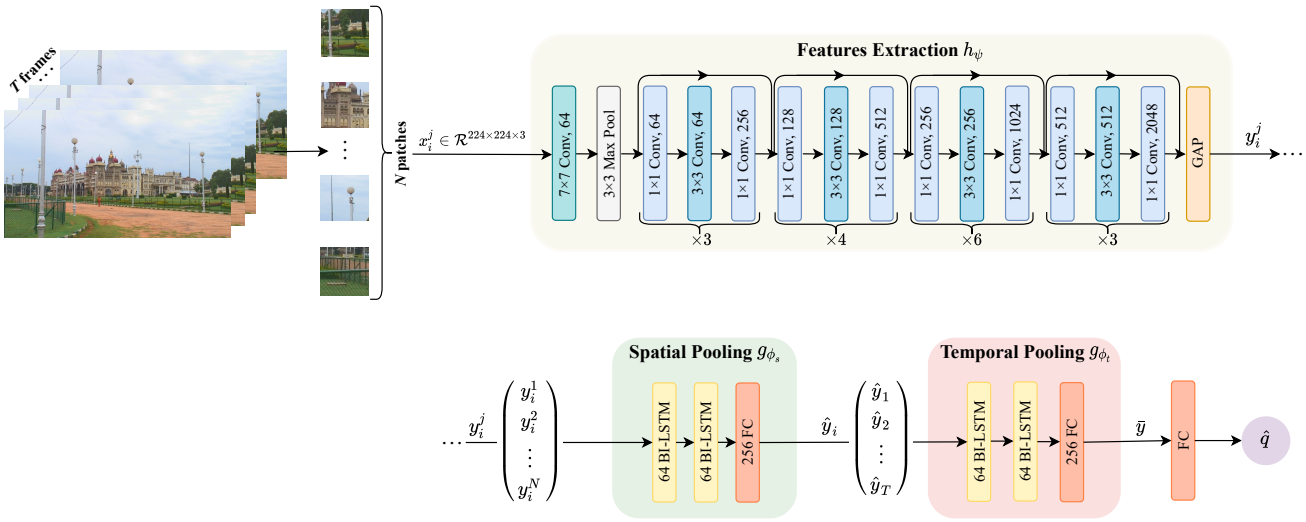

Fig. 1: The overall framework of the proposed 2BiVQA metric. The features extraction module is used to extract spatial features $\boldsymbol{y}_i^j$ from patches $\boldsymbol{x}_i^j$. The spatial and temporal pooling modules are used to aggregate features into a final vector $\bar{\boldsymbol{y}}$ while accounting for HVS behavior. Finally, the regression module uses the final vector $\bar{\boldsymbol{y}}$ to predict the quality score $\hat{q}$.

classifier and the quality predictor that are jointly optimized. VSFA [67] model also uses a CNN, pre-trained on image classification tasks, as a features extraction module, and then it uses Gated Recurrent Unit (GRU) and a subjectively-inspired temporal pooling layer to output the video quality score. Next, an improved version of VSFA, named MDVSFA, was proposed in [68]. MDVSFA uses a mixed dataset training strategy for training a single VQA model with multiple datasets. Yi *et al.* [69] proposed a modified VGG-16 network with non-local layers to learn the global relationship of spatial features, which can be regarded as a kind of attention mechanism. In addition, they combined GRU and a temporal pooling layer to model the temporal-memory effects.

More recently [70], Zhang *et al.* introduced a Multi-Dimensional VQA (MD-VQA) metric aimed at assessing the visual quality of compressed UGC live videos. This method evaluates the quality in terms of semantic, distortion, and motion aspects. Tu *et al.* [71] proposed a hybrid method, named Rapid and Accurate Video Quality Prediction Evaluator (RAPIQUE), which uses both handcrafted and deep CNN-based high-level features. RAPIQUE is based on two modules, a NSS features extractor module, which extracts both spatial and temporal features, and a deep CNN features extractor (ResNet-50) which extracts deep high-level features. Finally, a regressor model is used to map the extracted features to a quality score. Another deep learning-based VQA model was proposed in [72], including a 2D-CNN to extract quality-aware spatial feature representation from raw pixels of the video frames, as well as a 3D-CNN dedicated to the extraction of motion features, followed by a Multi-layer Perceptron (MLP) regression module to map these features into chunk-level quality scores, and finally, temporal average pooling is used to derive the video-level quality score. In [73], the authors proposed to hierarchically add the feature maps from intermediate layers into the final feature maps and calculate their global mean and standard deviation as the feature representation. Thus, covering the full range of visual features from low-level to high-level. Subsequently, Fully Connected (FC) and temporal pooling are used for the quality regression. In the same way, Shen *et al.* [74] proposed a BVQA method with spatio-temporal feature fusion and hierarchical information integration. Their metric consists of three stages: a multiscale feature extraction network that extracts spatio-temporal features, a hierarchical spatio-temporal fusion network that integrates intermediate feature information, and finally, a quality regression network that predicts the video quality. In [75], a completely BVQA metric has been proposed. This metric consists of a self-supervised multiview contrastive learning approach, which captures the joint distributions of frame differences with frames and optical flow. Wu *et al.* proposed a BVQA metric called FAST-VQA [76], which relies on a new sampling technique, Grid Mini-patch Sampling (GMS). GMS divides a video into spatially non-overlapping grids, randomly selects a mini-patch from each grid, and then assembles and temporally aligns these mini-patches to construct fragments. After sampling, the resultant fragments are fed into the Fragment Attention Network (FANet) to obtain the final video quality score. The same authors have also introduced the DOVER method [77], which provides video quality prediction from aesthetic and technical perspectives. Specifically, DOVER metric consists of two branches, each dedicated to focusing on one perspective. Finally, the overall quality score is obtained by a subjectively-inspired fusion of the predictions from the two perspectives. Given the success of the patch-sampling mechanism proposed in FAST-VQA [76], it was also adopted in [78]. However, instead of applying the same sampling strategy to all types of videos, in [78], the authors proposed to first classify the video into three content types, according to the professionalism of the produced content. Then, based on this classification, different spatial and temporal sampling strategies are applied, thus making it possible to build a unified VQA model.

All these described works only consider visual information, but some recent works have also included audio information

in the QA process via a multimodal approach [79], [80], because the audio information can significantly influence human judgment/perception. Thus, in [80], the authors first proposed a novel UGC Audio-Visual Quality Assessment (AVQA) database, which includes UGC audio and video sequences. Then, a deep learning-based approach was proposed, which includes four modules: a visual feature extraction module, an audio feature extraction module, a temporal pooling module, and finally an audio-visual fusion module that combines the features of the two modalities and provides the final score.

III. Proposed Double Bi-LSTM Video Quality Assessment Method

Let us consider a video sequence $\mathcal{V}$ as a set of T consecutive frames: $\mathcal{V} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$. The problem of UGC-VQA is defined as a function m that predicts a quality score $\hat{q}$ from a video sequence $\mathcal{V}$ as follows:

$$\hat{q} = m(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T). \tag{1}$$

To address this problem, we propose a BVQA metric called 2BiVQA for double Bi-LSTM Video Quality Assessment. As illustrated in Figure 1, the framework of the proposed 2BiVQA is composed of four main modules: features extraction, spatial pooling, temporal pooling, and finally, a quality regression module. These four modules are integrated to form an end-to-end BVQA metric. Each of the four modules will be described in detail in the following sections.

A. Features extraction

CNN features have been shown to correlate well with perceptual judgments [81] and represent good candidates for human perception-related applications [81]–[85]. The performance of CNN strongly depends on the number of training samples. However, existing UGC-VQA datasets are much smaller compared to the typical computer vision datasets with millions of samples. Thus, it is very difficult to train a deep CNN from scratch while achieving competitive quality prediction performance, since the model can be prone to over-fitting problem. Nevertheless, the authors of [4] showed that well-known deep CNN feature descriptors (e.g., ResNet-50 [52], VGG-16 [86], etc.) pre-trained on other vision tasks like image classification are transferable to UGC IQA/VQA problems, and they can achieve outstanding performance. In our contribution, we opted for the ResNet-50 pre-trained on ImageNet [61] as the backbone model to extract spatial features. Even though several CNN models can be used, we obtained the best performance with ResNet-50 (see Tables II and III). In the following, we first introduce the backbone model, and then we detail the feature extraction process.

ResNet-50: is a variant of ResNet model that introduced a concept allowing to train ultra-deep neural networks that can include hundreds and even thousands of layers. In fact, theoretically, a deeper neural network is able to learn more complex features, which should lead to better performance. However, in practice, the ultimate effect of adding more and more layers is increasing the training error. This problem is known as the degradation problem [52]. Residual blocks introduced in ResNet aim to solve this issue. It includes shortcut connections to perform identity mapping, which allows a deeper model to have no higher training error than its shallower counterpart.

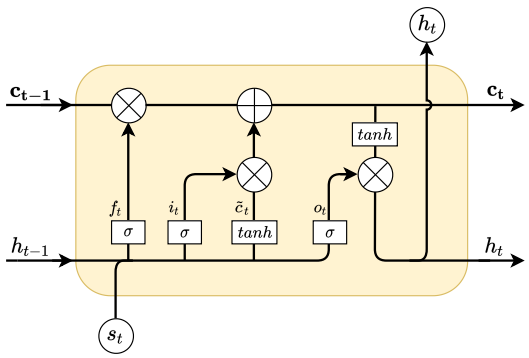

Fig. 2: The internal structure of the LSTM cell.

Features extraction process: one issue with using pre-trained models is their standard input shape. Two solutions can be envisaged to overcome this constraint, either resizing the input frame or dividing it into several patches. The first technique can affect the perceptual quality or attenuate the intensity of pre-existing artifacts. Therefore, we opted for the second solution, which solves the problem of a standard size input on the one hand, and avoids over-fitting with the limited dataset on another hand. Thus, for each frame $\boldsymbol{x}_i$ a sliding window is used to extract N patches $\boldsymbol{x}_i^j$, $\forall j \in \{1, \ldots, N\}, \forall i \in \{1, \ldots, T\}$, with a stride slightly smaller than the patch size (224×224). Then, these patches are fed into the ResNet-50 model pre-trained on ImageNet for the extraction of spatial features $\boldsymbol{y}_i^j$ from the input patch $\boldsymbol{x}_i^j$ as follows:

$$\boldsymbol{y}_i^j = h_\psi(\boldsymbol{x}_i^j), \tag{2}$$

where h_ψ represents the parametric function of the feature extraction model with training parameters ψ.

B. Spatial pooling

Once the features have been extracted from each patch, we need to aggregate them into one vector per frame. To do this, it is essential to take into account that UGC distortions are not uniformly distributed. In addition, the local distortion visibility is influenced by its surrounding regions, which can either emphasize or mask the perception of distortion. Moreover, the perceptual quality of HVS varies over the spatial domain [87].

Thus, to mimic the HVS behavior as well as to account for UGC distortion features, unlike processing each patch independently, we consider the entire sequence of patches $(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, ..., \boldsymbol{x}_i^N)^T$ of a frame i at once. To achieve this, we use a sequence model that can efficiently capture the dependencies between the patches of a frame. Specifically, we design our spatial pooling using Bi-LSTM network [88] as the sequence model, which provides the ability to deal with dependencies across patches.

In the following, we first explain the internal mechanisms of LSTM [89], then we introduce Bi-LSTM, and finally, the learning strategy of the proposed spatial pooling module is described.

LSTM (Long short term memory) [89]: is one of the most popular RNNs, designed to deal with long time-dependencies. It allows solving the diminishing and exploding gradient problems in long structures [90]. Each LSTM cell consists of an input gate i_t, a forget gate f_t, an output gate o_t, a candidate cell state $\tilde{c}_t$, a cell state c_t, and a hidden state h_t, as shown in Figure 2. i_t is used to determine the information to store in the current cell state c_t, while f_t determines the thrown away information. o_t decides the information to be passed to the current hidden state h_t, which is computed as follows:

$$\begin{aligned} i_t &= \sigma(W^{(i)} \cdot (h_{t-1} \oplus x_t) + b^{(i)}), \\ f_t &= \sigma(W^{(f)} \cdot (h_{t-1} \oplus x_t) + b^{(f)}), \\ o_t &= \sigma(W^{(o)} \cdot (h_{t-1} \oplus x_t) + b^{(o)}), \\ \tilde{c}_t &= \tanh(W^{(c)} \cdot (h_{t-1} \oplus x_t) + b^{(c)}), \\ c_t &= f_t \times c_{t-1} + i_t \times \tilde{c}_t, \\ h_t &= o_t \times \tanh(c_t), \end{aligned} \tag{3}$$

where σ is the sigmoid function and $\oplus$ is the concatenation operator. $+$ and $\times$ are the element-wise addition and product operations, respectively. $W^{(x)}$ and $b^{(x)}$ represent the weight matrix and the bias vector of gate x, respectively.

Bi-LSTM (Bi-directional long short term memory) [88]: is a stack of two independent LSTMs. This structure allows the network to take both backward and forward information in consideration. It has been proved that Bi-LSTM is far better than regular LSTM in many fields, like forecasting time series [91], phoneme classification [92], speech recognition [93], etc. However, to the best of our knowledge, Bi-LSTM has not yet been considered in IQA/VQA problems.

The architecture of the spatial pooling module is shown in Figure 3. The module is composed of two Bi-LSTM layers with K cells each, followed by a FC layer with 256 nodes. We have found that using this architecture with $K = 64$ leads to the best results in our experiments. The feature vector $(\boldsymbol{y}_i^1, \boldsymbol{y}_i^2, \ldots, \boldsymbol{y}_i^N)^T$ of a frame i is fed into the spatial pooling module, expressed as:

$$\hat{\boldsymbol{y}}_i = g_{\phi_s}(\boldsymbol{y}_i^1, \boldsymbol{y}_i^2, \ldots, \boldsymbol{y}_i^N), \quad \forall i \in \{1, \ldots, T\}, \tag{4}$$

where g_{ϕ_s} is the parametric function of the spatial pooling module with the training parameters ϕ_s.

Pre-training technique: transfer learning is a powerful machine learning technique. Here, we perform pre-training followed by fine-tuning, which is a widely-used transfer learning paradigm. Pre-training refers to training a model in a specific source domain $\mathcal{D}_S$ with learning task $\mathcal{T}_S$ to initialize its parameters before fine-tuning it for a new learning task $\mathcal{T}_T$ in the target domain $\mathcal{D}_T$, where a domain $\mathcal{D}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where $X = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \in \mathcal{X}$.

In our approach, the spatial pooling module is trained in this way in two separate stages. It is first pre-trained using the KonIQ-10k dataset [94], which is a large-scale in-the-wild IQA dataset, regardless of the remaining module as an IQA model. We assume that previously described complex behaviors and characteristics of HVS and UGC distortions, respectively, are embedded in the subjective quality dataset. Thus, we aim to transfer the knowledge acquired by the model after training on the KonIQ-10k dataset, encoded in the weights of the model, to exploit it for the target UGC-VQA task. Moreover, this pre-training step has the advantage of presenting to the model more diverse content by leveraging the larger authentic IQA dataset.

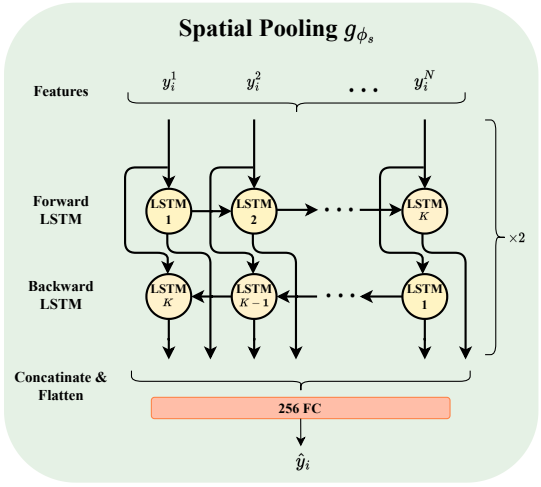

Fig. 3: The architecture of the spatial pooling module.

Finally, the spatial pooling module is fine-tuned using the subjective video quality scores with the rest of the modules in the second stage.

C. Temporal pooling

Aggregating quality features of frames into an overall video score is one of the main still unresolved challenges in VQA [43], [44], [47], [68], [95]. In fact, the human quality judgment at a late frame can definitely be affected by the previous frames, which is widely known as the temporal-memory effect [41], [95]–[97]. According to this temporal behavior of the HVS, low-quality frames leave more impressions on the viewer. For instance, if most of the frames are of high quality, and only a few frames are of low quality, humans generally determine that the video is of low quality. Most of the previously developed VQA metrics focus much more on the accuracy of quality scores at the frame level, disregarding the impact of adjacent frames, from past and future, on the subjective quality of the current frame.

Therefore, in order to take into account the temporal variation of distortions as well as the temporal-memory effects of human perception, we propose a novel temporal pooling method using Bi-LSTM network to aggregate frame-level features $(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_T)$ into a global feature vector $\bar{\boldsymbol{y}}$ for the entire video sequence. As described previously, Bi-LSTM networks have the ability to take both backward and forward information into consideration, which makes it possible to capture long-range dependencies between frames like the temporal-memory effect. Similar to spatial pooling, this module is composed of two Bi-LSTM layers with 64 cells each, followed by a FC layer with 256 nodes. The temporal pooling module is defined as:

$$\bar{\boldsymbol{y}} = g_{\phi_t}(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \ldots, \hat{\boldsymbol{y}}_T), \tag{5}$$

TABLE I: Summary of the considered UGC-VQA datasets.

Database	#Video	Resolution	Time	Label	Range
KoNViD-1k [5]	1200	540p	8s	MOS+σ	[1,5]
LIVE-VQC [6]	585	240p-1080p	10s	MOS	[0,100]
YouTube-UGC [7]	1380	360p-4k	20s	MOS+σ	[1,5]

where g_{ϕ_t} is the parametric function of the temporal pooling module with the training parameters ϕ_t.

D. Quality regression

After extracting quality-aware features and aggregating them into a single vector $\bar{y}$, we need to map these features to the final video quality score q. Here, we used one node FC as a regression model with a linear activation function. Therefore, we obtain the final video quality score as follows:

$$\hat{q} = \zeta(\bar{\boldsymbol{y}}), \tag{6}$$

where ζ denotes the FC layer.

IV. Experimental results

In this section, we first define the experimental setups, including the description of datasets, the evaluation methods and the implementation details. Then, we present the results of ablation studies, the comparison with SOTA, the generalization capability, and finally, the complexity evaluation.

A. Datasets

To train, fine-tune and test the proposed model, three UGC-VQA datasets were considered, including KoNViD-1K [5], LIVE-VQC [6] and YouTube-UGC [7]. The features of these three datasets are summarized in Table I. For YouTube-UGC, we excluded 57 grayscale videos for a fair comparison as in [4]. We also used these three datasets to create a fourth dataset, which is the union of them after MOS calibration using the Iterative Nested Least Squares Algorithm (INSLA) [4], [98]:

$$q' = 5 - 4 \times ((5 - q)/4 \times 1.1241 - 0.0993)\,, \tag{7}$$
$$q' = 5 - 4 \times ((100 - q)/100 \times 0.7132 + 0.0253)\,, \tag{8}$$

where Eqs. (7) and (8) are used for calibrating KoNViD-1K and LIVE-VQC, respectively, while YouTube-UGC is selected as the anchor dataset. q' and q denote the adjusted and the original scores, respectively. The formed dataset is referred to in the following as All-Combined.

In addition, the KonIQ-10k IQA dataset [94] is used to train the spatial pooling module separately. This dataset contains 10,073 in-the-wild images with a resolution of 1024×768 pixels.

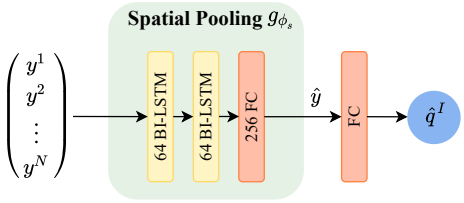

Fig. 4: The architecture of the spatial pooling module for pre-training.

B. Evaluation methods

Since there is no defined size of the training (or test) set for KoNViD-1k, LIVE-VQA, and YouTube-UGC, and as convened, we randomly split each dataset into two non-overlapping subsets, 80% for training and 20% for testing. We performed k fold iterations, and the median performance on the test sets is reported.

We considered four standard measures to assess the performance of the proposed model, including Spearman Rank Order Correlation Coefficient (SROCC) and Kendall Rank Correlation Coefficient (KRCC), which are prediction monotonicity measures, and Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error (RMSE), which are prediction accuracy measures. Before calculating PLCC and RMSE, we performed a nonlinear four-parametric logistic regression to match the predicted score to the subject score as follows:

$$L_g(x) = \beta + \frac{\alpha - \beta}{1 + \exp(-x + \gamma/|\delta|)}. \tag{9}$$

C. Training process

The training is conducted in two steps: first, the spatial pooling module is pre-trained, then the spatial and temporal pooling modules are trained end-to-end.

For the pre-training step, each image from KonIQ-10k dataset is divided into N patches. Then, the features extracted from the patches $(y^1, y^2, \ldots, y^N)^T$ are fed into the spatial pooling module for training. A FC layer with an identity activation function is added as a regressor head to predict the image quality score $\hat{q}^I$, as illustrated in Figure 4.

The pre-training process is performed with 200 epochs using Adam optimizer [99] with an initial learning rate of $1e-4$, batch size of 16 and the Mean Squared Error (MSE) as loss function ℓ_2:

$$\ell_2(q^I, \hat{q}^I) = \frac{1}{L}\sum_{l=1}^{L}\left(q_l^I - \hat{q}_l^I\right)^2, \tag{10}$$

where $q^I, \hat{q}^I$ and L represent the ground truth, the predicted image quality score and the batch size, respectively.

The second stage of the training process is to map the frame-level feature vectors into the global quality score. For this purpose, the spatial pooling module is fine-tuned at the same time as the temporal pooling module on the target UGC-VQA datasets. This is done using the same hyper-parameters as the pre-training step: 200 epochs with an initial learning rate of $1e-4$ and MSE as a loss function $\ell_2(q, \hat{q})$ computed between

TABLE II: Performance of the ablation study for the spatial pooling module on the KonIQ-10k dataset. Bold entries indicate the top three performing methods, while the best method is underlined.

Model	#Features per patch	Spatial pooling	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
VGG16	512	Concatenate	0.821	0.829	0.628	0.378
		Mean	0.818	0.829	0.622	0.381
		LSTM	0.875	0.894	0.689	0.249
		Bi-LSTM	0.876	0.900	0.693	0.242
Densenet169	1664	Concatenate	0.833	0.850	0.639	0.293
		Mean	0.825	0.839	0.623	0.373
		LSTM	0.871	0.898	0.691	0.245
		Bi-LSTM	**0.878**	**0.902**	**0.698**	**0.240**
ResNet50	2048	Concatenate	0.856	0.871	0.669	0.275
		Mean	0.856	0.858	0.664	0.275
		LSTM	**0.906**	**0.924**	**0.737**	**0.214**
		Bi-LSTM	**0.910**	**0.925**	**0.742**	**0.213**
EfficientNetB7	2560	Concatenate	0.763	0.741	0.570	0.414
		Mean	0.791	0.816	0.593	0.391
		LSTM	0.851	0.873	0.664	0.272
		Bi-LSTM	0.853	0.878	0.668	0.266

q and $\hat{q}$, which represent the ground truth and the predicted video quality scores, respectively.

Note that during the pre-training and fine-tuning steps, the weights of the backbone model are frozen. To support the principle of reproducible research and fair comparison, an implementation of the 2BiVQA metric is made publicly available for the research community[1].

D. Ablation studies

To justify the choice and highlight the contribution of each component in the proposed model, we conducted ablation studies on the following aspects. The first study aims to select the best backbone model to extract reliable perceptual features. We considered four well-known deep CNN models: VGG16 [86], Densenet169 [104], ResNet-50 [52] and EfficientNetB7 [105]. Also, we tested four spatial pooling methods: simple concatenation, arithmetic mean, LSTM, and Bi-LSTM. The result of this first study is reported in Table II, where some interesting observations can be made. First, ResNet-50 is able to extract the most significant perceptual features and achieves better performance than the other CNN models. Second, RNN models (LSTM and Bi-LSTM) are the pooling methods that obtained the highest correlation scores allowing a 6.09% improvement in terms of SROCC compared to the classical pooling methods, including concatenation and arithmetic mean. Finally, this study shows that the best combination for the IQA is ResNet-50 as a features extraction model and Bi-LSTM as a spatial pooling method.

In the second study, we investigated the effect of pre-training the spatial pooling module, and we also tested several temporal pooling methods, including arithmetic mean, harmonic mean, geometric mean, LSTM, and Bi-LSTM. We depict the results of this second study in Table III. It is important to note that this study is conducted on KoNViD-1K dataset with a randomly 80%-20% split over only one iteration to avoid a huge training time. The results show that pre-training the spatial pooling module on KonIQ-10k dataset significantly improves the prediction performance, for instance in terms of SROCC by 3.04%. Moreover, the results indicate that Bi-LSTM is the best-performing temporal pooling method, showing its effectiveness in capturing long-range dependencies between frames.

E. Performance evaluation and comparison

To assess the performance of the proposed 2BiVQA metric, we compared it with ten BIQA models (BRISQUE [29], NIQE [100], ILNIQE [101], BMPRI [21], StairIQA [63], HIGRADE [25], CORNIA [40], HOSA [102], FRIQUEE [26], PaQ-2-PiQ [103] and KonCept512 [94]), and five BVQA models (V-BLIINDS [36], FAST-VQA [76], TLVQM [47], VIDEVAL [4] and RAPIQUE [71]). In addition, two deep CNN models (VGG-19 [51] and ResNet-50 [52]) using transfer learning were benchmarked. Among these methods, NIQE and ILNIQE are completely blind because they don't require any training. The rest of the methods were trained and tested under the same conditions as our proposed model. For VGG-19 and ResNet-50 models, the frame-level scores are obtained using two FC layers with 256 and 1 nodes, respectively. For all considered BIQA models, we extend them for VQA by averaging the separate frame quality scores to obtain the overall video quality score.

Table IV shows the performance of these methods on the four considered datasets. We can notice that most of the BIQA metrics, except those CNN-based, provide low performance, which indicates that the temporal-related features are substantial for VQA, and using a simple average pooling is not sufficient to achieve high performance. We can also observe that CNN-based BIQA approaches, i.e., VGG-19 and ResNet-50, perform well on larger datasets (KoNViD-1k, YouTube-UGC and All-Combined), showing the superiority of the data-driven deep-learning approaches over handcrafted feature-based ones when trained with sufficient dataset size.

On KoNViD-1K, BVQA methods generally provide acceptable results, while our 2BiVQA model achieves the second-highest performance, outperforming the majority of recent SOTA models, with FAST-VQA ranking as the best performer. On LIVE-VQC, which contains many mobile videos showing huge camera motions, 2BiVQA consistently ranks within the top three performers based on evaluation metrics. TLVQM method also yields competitive scores on this dataset, thanks to its many heavily designed motion-relevant features. On YouTube-UGC, RAPIQUE, 2BiVQA, FAST-VQA and VIDEVAL metrics achieve the best correlation scores, outperforming by fare the other BVQA models. Finally, for the largest dataset (All-Combined), 2BiVQA delivers the third-highest performance, slightly outperformed by RAPIQUE and FAST-VQA.

Although our 2BiVQA method does not outperform FAST-VQA in terms of correlation metrics, it has significant advantages in terms of training efficiency. Table V provides a

[1] https://github.com/atelili/2BiVQA

TABLE III: Performance of the ablation study on the KoNViD-1k dataset. Each entry is presented as spatial pooling without/with pre-training on the KonIQ-10k dataset. Bold entries indicate the top three performing methods, while the best method is underlined.

Model	Temporal pooling	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	#parameters
VGG16	Mean	0.776 / 0.771	0.773 / 0.767	0.588 / 0.575	0.424 / 0.451	1,213,953
	Harmonic	0.776 / 0.773	0.774 / 0.769	0.588 / 0.576	0.424 / 0.453	1,213,953
	Geometric	0.777 / 0.772	0.774 / 0.768	0.588 / 0.576	0.424 / 0.452	1,213,953
	LSTM	0.790 / 0.809	0.799 / 0.829	0.594 / 0.613	0.461 / 0.383	1,820,929
	Bi-LSTM	0.797 / 0.819	0.806 / 0.833	0.606 / 0.622	0.458 / 0.380	2,529,281
DenseNet169	Mean	0.799 / 0.815	0.806 / 0.811	0.601 / 0.624	0.414 / 0.404	1,803,777
	Harmonic	0.800 / 0.816	0.807 / 0.812	0.603 / 0.624	0.415 / 0.403	1,803,777
	Geometric	0.799 / 0.815	0.807 / 0.811	0.602 / 0.624	0.415 / 0.404	1,803,777
	LSTM	0.805 / 0.825	0.821 / **0.839**	0.618 / 0.623	0.392 / **0.373**	2,410,753
	Bi-LSTM	0.810 / 0.825	0.824 / **0.842**	0.621 / 0.626	0.385 / **0.370**	3,050,241
ResNet50	Mean	0.795 / 0.829	0.802 / 0.816	0.601 / 0.631	0.411 / 0.398	2,000,385
	Harmonic	0.796 / 0.828	0.802 / 0.815	0.603 / 0.631	0.409 / 0.400	2,000,385
	Geometric	0.795 / **0.829**	0.802 / 0.815	0.602 / 0.631	0.410 / 0.399	2,000,385
	LSTM	0.825 / 0.827	0.821 / 0.819	0.625 / **0.633**	0.394 / 0.384	2,607,361
	Bi-LSTM	**0.830** / **0.846**	0.820 / **0.840**	**0.634** / **0.652**	0.382 / **0.362**	3,312,385
EfficientNetB7	Mean	0.746 / 0.782	0.766 / 0.780	0.557 / 0.594	0.458 / 0.432	2,262,529
	Harmonic	0.749 / 0.785	0.770 / 0.783	0.559 / 0.597	0.457 / 0.432	2,262,529
	Geometric	0.747 / 0.784	0.768 / 0.782	0.558 / 0.596	0.457 / 0.432	2,262,529
	LSTM	0.752 / 0.800	0.773 / 0.809	0.561 / 0.602	0.451 / 0.398	2,869,505
	Bi-LSTM	0.759 / 0.801	0.776 / 0.814	0.567 / 0.605	0.448 / 0.398	3,508,993

comparison between the characteristics of 2BiVQA and FAST-VQA during training. Notably, 2BiVQA shows approximately 15 times faster training time than FAST-VQA. Furthermore, our approach is efficient in terms of energy consumption, using approximately 30 times less energy than FAST-VQA. In addition, 2BiVQA has fewer model parameters during training than FAST-VQA, thereby simplifying the training process and improving the efficiency of resource allocation.

Figure 5 shows the MOS versus the prediction scores and nonlinear logistic fitted curves for the three best performing models (VIDEVAL, RAPIQUE and 2BiVQA) on the four evaluated datasets. These figures illustrate visually that the performance of 2BiVQA remains stable over the different datasets. Its scatter points are more densely clustered around the fitted curves, which are also more linear, especially for KoNViD-1k, YouTube-UGC, and All-Combined datasets.

F. Cross dataset generalization

A good VQA metric is supposed to generalize to unseen samples. Accordingly, we perform a cross-dataset evaluation by training the three best performing BVQA models on one dataset and testing them on the other datasets. The results are shown in Table VI. From this table, we can observe that the proposed model generalizes well to unseen datasets, and its performance does not depend on the dataset, which represents an essential feature for UGC-BVQA. Notably, 2BiVQA demonstrates superior generalization capability compared to FAST-VQA, RAPIQUE, and VIDEVAL, as indicated by the obtained results. This good generalization of the proposed method, which we believe is primarily due to the separation of the training into two stages, first the pre-training on KonIQ-10k dataset and then the fine-tuning on the target UGC-VQA dataset. The training on this diverse content allows our model to learn a rich feature representation suitable for UGC video quality score prediction. It can also be noted that the cross-domain BVQA methods generalization using YouTube-UGC is the best on average.

G. Complexity and runtime comparison

Computational efficiency is crucial for VQA algorithms, especially in practical deployments. In this regard, we performed runtime comparisons of our model as well as several methods on the same desktop computer equipped with an Intel® Xeon W-2145 CPU @ 3.70GHz $\times$16, 64G RAM, and GeForce RTX 2080 Ti graphics card under Ubuntu 20.04 Long Term Support (LTS) operating system. We used the initially released implementation in MATLAB R2018b and python 3.8.8 for GM-LOG, VIDEVAL, and RAPIQUE metrics. For BRISQUE and NIQE, we used scikit-video python library implementation. FAST-VQA was implemented in PyTorch and the remaining models, namely VGG19 and 2BiVQA, were implemented in TensorFlow. All BIQA models extract features at one frame per second, and then an average pooling was used to get the overall video quality score. We consider videos from YouTube-UGC at HD resolution (1920×1080), then we recorded the average runtime in seconds, as shown in Table VII. For better illustration, Figure 6 shows the scatter plots of SROCC versus runtime. It may be observed that FAST-

TABLE IV: Performance comparison of evaluated BVQA models on the four UGC-VQA datasets. The underlined and boldfaced entries indicate the best and top three performers on each dataset for each performance measure, respectively.

Dataset	KonViD-1k				LIVE-VQC			
Model	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	SROCC ↑	PLCC ↑	KRCC ↑	RMSE↓
BRISQUE [29]	0.656	0.657	0.476	0.481	0.592	0.638	0.416	13.100
NIQE [100]	0.541	0.553	0.379	0.533	0.595	0.628	0.425	13.110
ILNIQE [101]	0.526	0.540	0.369	0.540	0.503	0.543	0.355	14.148
StairIQA [63]	0.767	0.778	0.570	0.881	0.740	0.787	0.547	10.597
BMPRI [21]	0.519	0.522	0.354	0.661	0.502	0.586	0.350	13.558
HIGRADE [25]	0.720	0.726	0.531	0.439	0.610	0.633	0.439	13.027
FRIQUEE [26]	0.747	0.748	0.550	0.425	0.657	0.700	0.477	12.198
CORNIA [40]	0.716	0.713	0.523	0.448	0.671	0.718	0.484	11.832
HOSA [102]	0.765	0.766	0.569	0.414	0.687	0.741	0.503	11.353
VGG-19 [51]	0.774	0.784	0.584	0.395	0.656	0.716	0.472	11.783
ResNet-50 [52]	0.801	0.810	0.610	0.374	0.663	0.720	0.478	11.591
KonCept512 [94]	0.734	0.748	0.542	0.426	0.664	0.727	0.479	11.626
PaQ-2-PiQ [103]	0.613	0.601	0.433	0.514	0.643	0.668	0.456	12.619
V-BLIINDS [36]	0.710	0.703	0.518	0.459	0.693	0.717	0.507	11.765
TLVQM [47]	0.772	0.768	0.577	0.410	**0.798**	**0.802**	**0.608**	**10.145**
VIDEVAL [4]	0.783	0.780	0.584	0.402	0.752	0.751	0.563	11.100
RAPIQUE [71]	**0.807**	**0.815**	**0.618**	**0.364**	0.741	0.765	0.557	10.665
FAST-VQA [76]	**0.846**	**0.854**	**0.638**	**0.337**	**0.792**	**0.844**	**0.633**	**9.904**
2BiVQA	**0.815**	**0.835**	**0.629**	**0.352**	**0.761**	**0.832**	**0.621**	**9.979**

Dataset	YouTube-UGC				All-Combined			
Model	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓	SROCC ↑	PLCC ↑	KRCC ↑	RMSE ↓
BRISQUE [29]	0.382	0.395	0.263	0.591	0.569	0.586	0.403	0.561
NIQE [100]	0.237	0.277	0.160	0.617	0.462	0.477	0.322	0.611
ILNIQE [101]	0.291	0.330	0.198	0.605	0.459	0.474	0.321	0.611
StairIQA [63]	0.753	0.744	0.558	0.597	0.777	0.780	0.586	0.480
BMPRI [21]	0.295	0.372	0.199	0.636	0.505	0.525	0.351	0.678
HIGRADE [25]	0.737	0.721	0.547	0.447	0.739	0.736	0.547	0.467
FRIQUEE [26]	0.765	0.757	0.568	0.416	0.756	0.755	0.565	0.454
CORNIA [40]	0.597	0.605	0.421	0.513	0.676	0.697	0.484	0.494
HOSA [102]	0.602	0.604	0.425	0.513	0.695	0.708	0.503	0.489
VGG-19 [51]	0.702	0.699	0.509	0.456	0.732	0.748	0.539	0.461
ResNet-50 [52]	0.718	0.709	0.522	0.453	0.755	0.774	0.561	0.438
KonCept512 [94]	0.587	0.594	0.410	0.513	0.660	0.676	0.475	0.509
PaQ-2-PiQ [103]	0.265	0.293	0.177	0.615	0.472	0.482	0.324	0.608
V-BLIINDS [36]	0.559	0.555	0.389	0.535	0.654	0.6599	0.473	0.520
TLVQM [47]	0.669	0.659	0.481	0.484	0.727	0.734	0.534	0.470
VIDEVAL [4]	**0.778**	**0.773**	**0.583**	**0.404**	0.796	0.793	0.603	0.426
RAPIQUE [71]	0.761	0.762	0.561	0.406	**0.808**	**0.818**	**0.614**	**0.407**
FAST-VQA [76]	**0.811**	**0.817**	**0.619**	**0.386**	**0.804**	**0.800**	**0.606**	**0.453**
2BiVQA	**0.771**	**0.790**	**0.581**	**0.404**	**0.800**	**0.794**	**0.608**	**0.421**

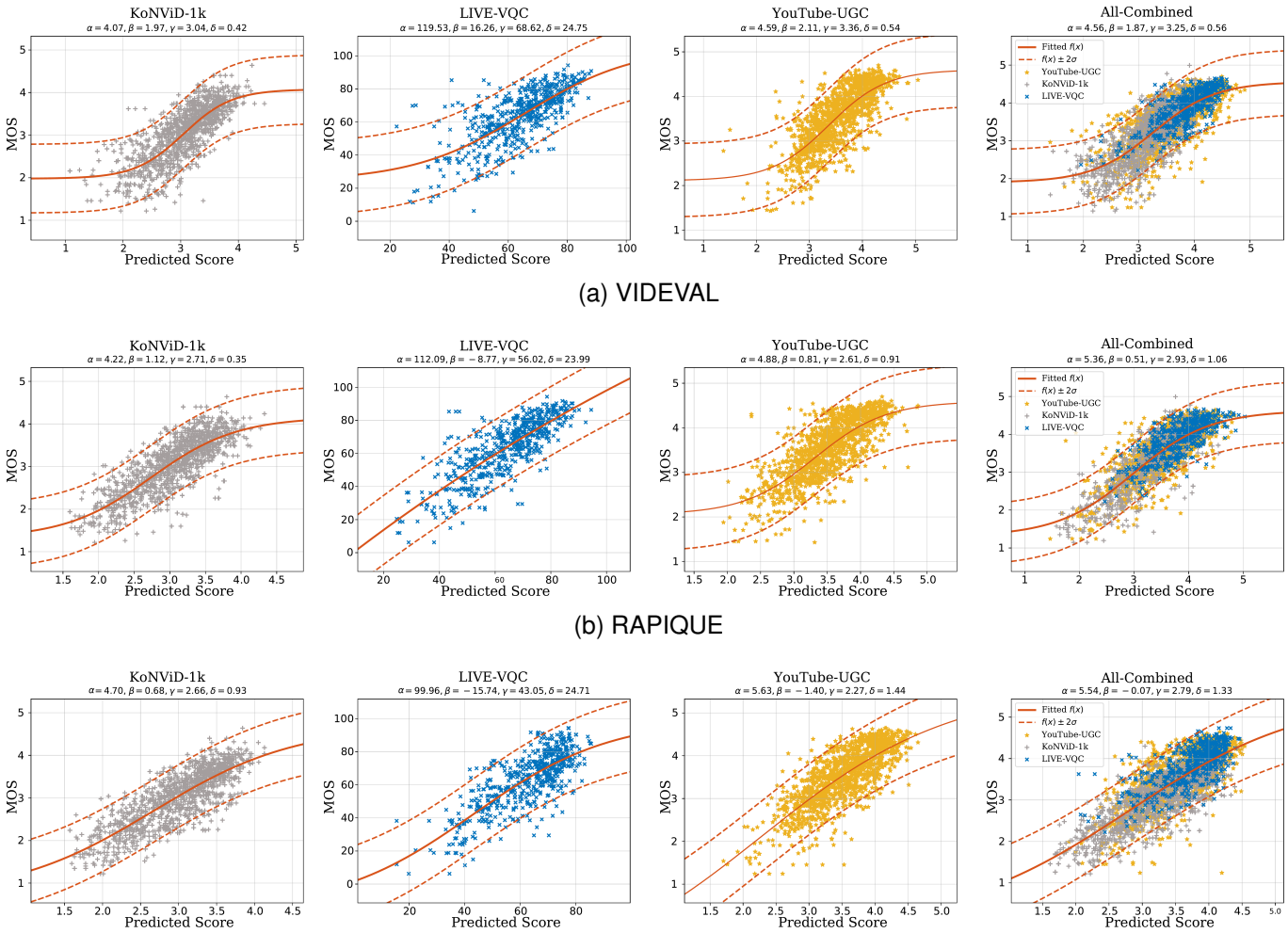

Fig. 5: Scatter plots and nonlinear logistic fitted curves of VIDEVAL, RAPIQUE, and 2BiVQA models versus MOS using k-fold cross-validation on KoNViD-1k, LIVE-VQC, YouTube-UGC, and All-Combined datasets. The logistic model coefficients are given for the three objective metrics tested on the four datasets.

TABLE V: Comparison of 2BiVQA and FAST-VQA characteristics during training.

Model	# parameters	Training time	Energy
FAST-VQA	28,127,901	7920 s	384.92 Wh
2BiVQA	3,246,849	580 s	11.21 Wh

VQA and BIQA models are faster than other methods, while VGG19, RAPIQUE, and 2BiVQA are relatively comparable.

V. Conclusion

In this paper, we proposed an effective BVQA metric for UGC videos, named 2BiVQA for double Bi-LSTM Video Quality Assessment. Our contribution relies on a deep CNN-based model to extract frame-level features and two Bi-LSTM networks for spatial and temporal pooling. Specifically, the first Bi-LSTM network is used to efficiently capture the short-term dependencies between neighboring patches, while the second Bi-LSTM network is exploited to capture long-range dependencies between frames over the entire video. In this way, the proposed 2BiVQA can take into account the features of UGC videos and mimic the behavior of the HVS.

TABLE VI: Cross dataset generalization in terms of SROCC.

Model	Train \Test	KoNViD-1K	LIVE-VQC	YouTube-UGC
VIDEVAL	KoNViD-1K	-	0.604	0.392
	LIVE-VQC	0.644	-	0.277
	YouTube-UGC	0.594	0.388	-
RAPIQUE	KoNViD-1K	-	0.546	0.318
	LIVE-VQC	0.656	-	0.352
	YouTube-UGC	0.582	0.623	-
FAST-VQA	KoNViD-1K	-	0.734	0.373
	LIVE-VQC	0.750	-	0.365
	YouTube-UGC	**0.687**	0.658	-
2BiVQA	KoNViD-1K	-	**0.770**	**0.428**
	LIVE-VQC	**0.753**	-	**0.416**
	YouTube-UGC	0.647	**0.674**	-

TABLE VII: Average runtime comparison evaluated on 1080p videos from YouTube-UGC.

Method	Deep Learning	Framework	Time (Sec.)	
			CPU	GPU
BRISQUE	✗	MATLAB	1.45	✗
NIQE	✗	MATLAB	1.68	✗
GM-LOG	✗	MATLAB	1.77	✗
VIDEVAL	✗	MATLAB	217.2	✗
RAPIQUE	✓	MATLAB	12.6	✗
VGG19	✓	TensorFlow	9.26	7.81
FAST-VQA	✓	PyTorch	13.4	4.9
2BiVQA	✓	TensorFlow	16.2	13.6

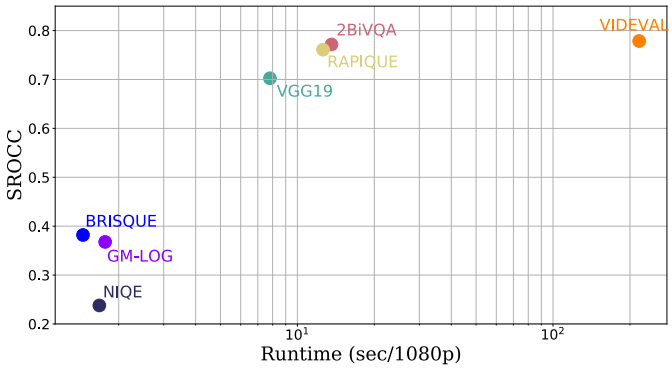

Fig. 6: Scatter plots of SRCC (on YouTube-UGC) of selected BVQA methods versus runtime (on 1080p).

In addition, the training was carried out in two stages to avoid over-fitting with the limited dataset. This training strategy improved feature representation, which significantly increased accuracy performance.

We conducted comprehensive tests on four UGC-VQA datasets. Results showed that 2BiVQA outperforms SOTA methods on two of the considered datasets (KonViD-1k and LIVE-VQC) and achieves competitive performance on YouTube-UGC and All-Combined. We further showed that the performance of the proposed solution is independent of the training dataset and generalizes better on unseen datasets than other BVQA methods, which is a key feature of the UGC VQA problem. Finally, since computational efficiency is crucial for BVQA algorithms, 2BiVQA has achieved a good trade-off between inference runtime, prediction performance and model complexity.

One future work worth addressing is to extend 2BiVQA to UGC AVQA. This can be achieved by incorporating the audio information into the spatial and temporal pooling blocs.

References

[1] U. Cisco, "Cisco annual internet report (2018–2023) white paper," *Cisco: San Jose, CA, USA*, 2020.

[2] I. R. BT, "500-13," *Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union*, vol. 6, 2012.

[3] 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, 2010.

[4] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Ugc-vqa: Benchmarking blind video quality assessment for user generated content," *IEEE Transactions on Image Processing*, vol. 30, pp. 4449–4464, 2021.

[5] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (konvid-1k)," in *2017 Ninth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[6] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612–627, 2018.

[7] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[8] J. Xu, J. Li, X. Zhou, W. Zhou, B. Wang, and Z. Chen, "Perceptual quality assessment of internet videos," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1248–1257.

[9] Y. Zhang, L. Zhang, and R. Zimmermann, "Aesthetics-guided summarization from multiple user generated videos," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 2, pp. 1–23, 2015.

[10] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proceedings. International conference on image processing*, vol. 3. IEEE, 2002, pp. III–III.

[11] X. Wang, B. Tian, C. Liang, and D. Shi, "Blind image quality assessment for measuring image blur," in *2008 Congress on Image and Signal Processing*, vol. 1. IEEE, 2008, pp. 467–470.

[12] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 3. Ieee, 2000, pp. 981–984.

[13] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5462–5474, 2017.

[14] X. Feng and J. P. Allebach, "Measurement of ringing artifacts in jpeg images," in *Digital Publishing*, vol. 6076. International Society for Optics and Photonics, 2006, p. 60760A.

[15] Y. Wang, S.-U. Kum, C. Chen, and A. Kokaram, "A perceptual visibility metric for banding artifacts," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2067–2071.

[16] Z. Tu, J. Lin, Y. Wang, B. Adsumilli, and A. C. Bovik, "Bband index: a no-reference banding artifact predictor," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2712–2716.

[17] A. Amer and E. Dubois, "Fast and reliable structure-oriented video noise estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 113–118, 2005.

[18] A. Norkin and N. Birkbeck, "Film grain synthesis for av1 video codec," in *2018 Data Compression Conference*. IEEE, 2018, pp. 3–12.

[19] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid no-reference quality metric for singly and multiply distorted images," *IEEE Transactions on Broadcasting*, vol. 60, no. 3, pp. 555–567, 2014.

[20] Y. Lu, F. Xie, T. Liu, Z. Jiang, and D. Tao, "No reference quality assessment for multiply-distorted images based on an improved bag-of-words model," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1811–1815, 2015.

[21] 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, 2018.

[22] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.

[23] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, and C. W. Chen, "Screen content quality assessment: overview, benchmark, and beyond," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–36, 2021.

[24] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[25] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, "No-reference quality assessment of tone-mapped hdr pictures," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2957–2971, 2017.

[26] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of vision*, vol. 17, no. 1, pp. 32–32, 2017.

[27] S.-C. Pei and L.-H. Chen, "Image quality assessment using human visual dog model fused with random forest," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3282–3292, 2015.

[28] D. L. Ruderman and W. Bialek, "Statistics of natural images: Scaling in the woods," *Physical review letters*, vol. 73, no. 6, p. 814, 1994.

[29] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[30] M. A. Saad, A. C. Bovik, and C. Charrier, "A dct statistics-based blind image quality index," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 583–586, 2010.

[31] ——, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *IEEE transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012.

[32] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal processing letters*, vol. 17, no. 5, pp. 513–516, 2010.

[33] Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-diivine: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes," *Signal processing: image communication*, vol. 29, no. 7, pp. 725–747, 2014.

[34] 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, 2016.

[35] Z. Sinno and A. C. Bovik, "Spatio-temporal measures of naturalness," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1750–1754.

[36] 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, 2014.

[37] Y. Liu, K. Gu, X. Li, and Y. Zhang, "Blind image quality assessment by natural scene statistics and perceptual characteristics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–91, 2020.

[38] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *Journal of Electronic Imaging*, vol. 22, no. 4, p. 043025, 2013.

[39] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.

[40] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1098–1105.

[41] J. Xu, P. Ye, Y. Liu, and D. Doermann, "No-reference video quality assessment via feature learning," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 491–495.

[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, 2017.

[43] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 1153–1156.

[44] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "A comparative evaluation of temporal pooling methods for blind video quality assessment," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 141–145.

[45] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2015.

[46] K. Manasa and S. S. Channappayya, "An optical flow-based no-reference video quality assessment algorithm," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2400–2404.

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

[48] P. Kancharla and S. S. Channappayya, "Completely blind quality assessment of user generated video content," *IEEE Transactions on Image Processing*, vol. 31, pp. 263–274, 2022.

[49] L. F. Tiotsop, T. Mizdos, M. Barkowsky, P. Pocta, A. Servetti, and E. Masala, "Mimicking individual media quality perception with neural network based artificial observers," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 1, pp. 1–25, 2022.

[50] Y. Zhu, S. C. Guntuku, W. Lin, G. Ghinea, and J. A. Redi, "Measuring individual video qoe: A survey, and proposal for future directions using social media," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, pp. 1–24, 2018.

[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[53] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934

[54] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[56] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *CoRR*, vol. abs/1506.04579, 2015.

[57] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[58] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," *IEEE Access*, vol. 9, pp. 72 139–72 160, 2021.

[59] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[60] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal processing magazine*, vol. 34, no. 6, pp. 130–141, 2017.

[61] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[62] X. Yang, F. Li, and H. Liu, "A survey of dnn methods for blind image quality assessment," *IEEE Access*, vol. 7, pp. 123 788–123 806, 2019.

[63] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[64] S. Ahn and S. Lee, "Deep blind video quality assessment based on temporal human perception," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 619–623.

[65] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2349–2353.

[66] W. Liu, Z. Duanmu, and Z. Wang, "End-to-end blind quality assessment of compressed videos using deep neural networks." in *ACM Multimedia*, 2018, pp. 546–554.

[67] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[68] ——, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, 2021.

[69] F. Yi, M. Chen, W. Sun, X. Min, Y. Tian, and G. Zhai, "Attention based network for no-reference ugc video quality assessment," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1414–1418.

[70] Z. Zhang, W. Wu, W. Sun, D. Tu, W. Lu, X. Min, Y. Chen, and G. Zhai, "Md-vqa: Multi-dimensional quality assessment for ugc live videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1746–1755.

[71] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "Rapique: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 425–440, 2021.

[72] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proceedings*

of the 30th ACM International Conference on Multimedia, 2022, pp. 856–865.

[73] W. Sun, T. Wang, X. Min, F. Yi, and G. Zhai, "Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[74] W. Shen, M. Zhou, X. Liao, W. Jia, T. Xiang, B. Fang, and Z. Shang, "An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation," *IEEE Transactions on Broadcasting*, 2022.

[75] S. Mitra and R. Soundararajan, "Multiview contrastive learning for completely blind video quality assessment of user generated content," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1914–1924.

[76] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling," in *European Conference on Computer Vision*. Springer, 2022, pp. 538–554.

[77] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[78] X. Huang, C. Li, A. Bentaleb, R. Zimmermann, and G. Zhai, "Xgc-vqa: A unified video quality assessment model for user, professionally, and occupationally-generated content," *arXiv preprint arXiv:2303.13859*, 2023.

[79] X. Min, G. Zhai, J. Zhou, M. C. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing*, vol. 29, pp. 6054–6068, 2020.

[80] Y. Cao, X. Min, W. Sun, and G. Zhai, "Subjective and objective audio-visual quality assessment for user generated content," *IEEE Transactions on Image Processing*, 2023.

[81] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[82] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "Deepsim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, 2017.

[83] X. Yang, F. Li, and H. Liu, "Deep feature importance awareness based no-reference image quality prediction," *Neurocomputing*, vol. 401, pp. 209–223, 2020.

[84] S. A. Amirshahi, M. Pedersen, and S. X. Yu, "Image quality assessment by comparing cnn features between images," *Journal of Imaging Science and Technology*, vol. 60, no. 6, pp. 60 410–1, 2016.

[85] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[86] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 1943–1955, 2015.

[87] A. B. Watson and C. H. Null, "Digital images and human vision," in *Electronic Imaging Science and Technology Conference*, 1997.

[88] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[89] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[90] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[91] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "The performance of lstm and bilstm in forecasting time series," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2019, pp. 3285–3292.

[92] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[93] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[94] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[95] J. Park, K. Seshadrinathan, S. Lee, and A. C. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Transactions on Image Processing*, vol. 22, no. 2, pp. 610–620, 2012.

[96] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 219–234.

[97] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, 2009.

[98] M. H. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," in *Visual Communications and Image Processing 2003*, vol. 5150. International Society for Optics and Photonics, 2003, pp. 583–592.

[99] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[100] 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, 2012.

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

[102] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[103] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3575–3585.

[104] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[105] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.