Recentley, 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 article, 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 to extract discriminative features from image patches, which are then fed into two Recurrent Neural Networks for spatial and temporal pooling. Specifically, we use two Bi-directional Long Short-term Memory 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.

CCS Concepts: • Computing methodologies → Perception; • Human-centered computing → Visualization design and evaluation methods;

Additional Key Words and Phrases: Blind video quality assessment, user-generated content, deep learning, Bi-LSTM, spatial pooling, temporal pooling

ACM Reference format:
Ahmed Telili, Sid Ahmed Fezza, Wassim Hamidouche, and Hanene F. Z. Brachemi Meftah. 2023. 2BiVQA: Double Bi-LSTM-based Video Quality Assessment of UGC Videos. ACM Trans. Multimedia Comput. Commun. Appl. 20, 4, Article 100 (December 2023), 22 pages.
https://doi.org/10.1145/3632178

Authors’ addresses: A. Telili and W. Hamidouche, Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, 20 Av. des Buttes de Coesmes, 35700 Rennes, France, 35000; e-mails: [atelili, whamidouche]@insa-rennes.fr; S. A. Fezza and H. F. Z. Brachemi Meftah, National Higher School of Telecommunications and ICT, BP 1518 El M’Naouer, Route Es-Senia, 31000 Oran, Algeria, e-mails: sfezza@ensttic.dz, hbrachemi@inntic.dz.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/12-ART100 $15.00
https://doi.org/10.1145/3632178

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 20, No. 4, Article 100. Publication date: December 2023.
1 INTRODUCTION

Currently, video represents the majority of Internet traffic. According to Cisco’s recent report [7], 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, and so on. 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. However, 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 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 [5]. 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 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 [65], their performance on UGC videos remains far from satisfactory [21, 69, 78, 81, 87], 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 [103]. Hence, in recent years, researchers have deployed considerable efforts into the development of in-the-wild UGC datasets such as KoNViD-1k [21], LIVE-VQC [69], and YouTube-UGC [81], to cite a few examples. These UGC datasets differ significantly from synthetic distortion datasets 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 article 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 article is organized as follows. Section 2 reviews related work, then Section 3 presents the proposed 2BiVQA model. The performance of our model is assessed and analyzed in Section 4. Finally, Section 5 concludes the article.

2 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.

2.1 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 [41, 80], blockiness [44, 83], ringing [9], banding [77, 82], and noise [2, 54] or targeted multiple types of distortion [17, 39, 45]. Later on, the most successful handcrafted features-based BVQA models mainly rely on learning approaches [43, 97], using a set of relevant perceptual features combined with a regression model to predict quality scores [12, 32, 52, 56].

The most popular BIQA/BVQA algorithms are based on Natural Scene Statistics (NSS) [59], 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 [48], BRISQUE [48]), Discrete Cosine Transform (DCT) (BLIINDS [60] and BLIINDS-II [61]), and Discrete Wavelet Transform (DWT) (BIQI [51], DIIVINE [52], C-DIIVINE [102]). These metrics have been expanded to the VQA task using the space-time natural video statistics models [35, 62, 70]. For instance, in Reference [38], 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 [101]), the joint statistics of the gradient magnitude and Laplacian of Gaussian responses in the spatial domain (GM-LOG [90]) and the gradient domain of LAB color transforms (HIGRADE [32]). 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 Reference [12], 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 [93], the authors proposed to extend it to video signal V-CORNIA [89]. 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. [42, 45] 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 Reference [45], 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 MPRI.

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 [64, 76, 89].

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 Reference [49]. 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 [40] 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 Reference [62], 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. [35] 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) [31] 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) [78], which is based on features selection among top-performing BIQA/BVQA models such as BRISQUE, HIGRADE, TLVQM, and so on. To select the most relevant features, Random Forest (RF) is used to remove the less significant features. Finally, a Support Vector Machine (SVM) is used to regress the final features vector into a quality score. Kancharla et al. [26] 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 Reference [75], 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 [105].

2.2 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 [18, 68], object detection [4, 13], and image segmentation [37, 58]. Recently, with the release of several larger IQA/VQA datasets [11, 14, 44, 81], 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 [27, 28] or (2) leveraging pre-trained deep CNNs on large datasets like ImageNet [8], 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. [27], 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 [72, 91], which achieved quite good performance. For video, however, very few methods based on deep learning have been dedicated to BVQA.

For instance, Ahn et al. [1] 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 Reference [96], which consists of a 3D-CNN to extract spatio-temporal features followed by a LSTM to predict the perceived video quality. A multi-task CNN framework, named V-MEON, was proposed in Reference [36] 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 classifier and the quality predictor that are jointly optimized. VSFA [33] 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 Reference [34]. MDVSFA uses a mixed dataset training strategy for training a single VQA model with multiple datasets. Yi et al. [94] 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 [104], 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. [79] 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 Reference [71], 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 Reference [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. [66] 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 Reference [47], 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 [85], 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 [86], 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 [85], it was also adopted in Reference [24]. However, instead of applying the same sampling strategy to all types of videos, in Reference [24], 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 [6, 46], because the audio information can significantly influence human judgment/perception. Thus, in Reference [6], 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.

3 PROPOSED DOUBLE BI-LSTM VIDEO QUALITY ASSESSMENT METHOD

Let us consider a video sequence \( V \) as a set of \( T \) consecutive frames: \( V = \{x_1, x_2, \ldots, x_T\} \). The problem of UGC-VQA is defined as a function \( m \) that predicts a quality score \( \hat{q} \) from a video sequence \( V \) as follows:

\[
\hat{q} = m(x_1, x_2, \ldots, x_T).
\]

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.

3.1 Features Extraction

CNN features have been shown to correlate well with perceptual judgments [99] and represent good candidates for human perception-related applications [3, 10, 25, 92, 99]. 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 [78] showed that well-known deep CNN feature descriptors (e.g., ResNet-50 [18], VGG-16 [100], 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 [8] 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 2 and 3). In the following, we first introduce the backbone model, and then we detail the feature extraction process.
**ResNet-50:** A variant of ResNet model that introduced a concept allowing us 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 [18]. 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.

**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 the other hand. Thus, for each frame $x_i$, a sliding window is used to extract $N$ patches $x_j^i, \forall j \in \{1, \ldots, N\}, \forall i \in \{1, \ldots, T\}$, with a stride slightly smaller than the patch size ($224 \times 224$). Then, these patches are fed into the ResNet-50 model pre-trained on ImageNet for the extraction of spatial features $y_j^i$ from the input patch $x_j^i$ as follows:

$$y_j^i = h_\psi(x_j^i),$$

where $h_\psi$ represents the parametric function of the feature extraction model with training parameters $\psi$.

### 3.2 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 [84].

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 $(x_1^i, x_2^i, \ldots, x_N^i)^T$ 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 [63] 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 [20], then we introduce Bi-LSTM, and finally, the learning strategy of the proposed spatial pooling module is described.

**LSTM** (Long short-term memory) [20]: 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 [19]. Each Long Short Term Memory (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:

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

where \( \sigma \) 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) \[63\]: 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-directional Long Short Term Memory (Bi-LSTM) is far better than regular LSTM in many fields, like forecasting time series \[67\], phoneme classification \[16\], speech recognition \[15\], and so on. 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 \( (y^1_i, y^2_i, \ldots, y^N_i)^T \) of a frame \( i \) is fed into the spatial pooling module, expressed as

\[
\hat{y}_i = g_{\phi_s}(y^1_i, y^2_i, \ldots, y^N_i), \quad \forall i \in \{1, \ldots, T\},
\]

where \( g_{\phi_s} \) is the parametric function of the spatial pooling module with the training parameters \( \phi_s \).

**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 \( D_S \) with learning task \( T_S \) to initialize its parameters before fine-tuning it for a new learning task \( T_T \) in the target domain \( D_T \), where a domain \( D \) consists of a feature space \( X \) and a marginal probability distribution \( P(X) \), where \( X = \{s_1, \ldots, s_n\} \in X \).

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 \[22\], which is a large-scale in-the-wild IQA dataset, regardless of the remaining module as an IQA model, as illustrated in Figure 4. 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.

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

### 3.3 Temporal Pooling

Aggregating quality features of frames into an overall video score is one of the main still unresolved challenges in VQA [31, 34, 55, 64, 76]. 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 [29, 53, 55, 89]. 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, 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 \((\dot{y}_1, \dot{y}_2, \ldots, \dot{y}_T)\) into a global feature vector \(\ddot{y}\) 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

\[
\ddot{y} = g_{\phi_t}(\dot{y}_1, \dot{y}_2, \ldots, \dot{y}_T),
\]

where \(g_{\phi_t}\) is the parametric function of the temporal pooling module with the training parameters \(\phi_t\).

### 3.4 Quality Regression

After extracting quality-aware features and aggregating them into a single vector \(\ddot{y}\), 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:

\[
\hat{q} = \zeta(\ddot{y}),
\]

where \(\zeta\) denotes the FC layer.

### 4 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.

### 4.1 Datasets

To train, fine-tune and test the proposed model, three UGC-VQA datasets were considered, including KoNVid-1K [21], LIVE-VQC [69], and YouTube-UGC [81]. The features of these three datasets
Table 1. Summary of the Considered UGC-VQA Datasets

| Database       | # Videos | Resolution | Time | Label       | Range       |
|----------------|----------|------------|------|-------------|-------------|
| KoNViD-1k [21] | 1,200    | 540p       | 8 s  | MOS+σ       | [1,5]       |
| LIVE-VQC [69]  | 585      | 240p–1,080p| 10 s | MOS         | [0,100]     |
| YouTube-UGC [81]| 1,380    | 360p–4k    | 20 s | MOS+σ       | [1,5]       |

are summarized in Table 1. For YouTube-UGC, we excluded 57 grayscale videos for a fair comparison as in Reference [78]. 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) [57, 78]:

\[
q' = 5 - 4 \times ((5 - q)/4 \times 1.1241 - 0.0993),
\]

\[
q' = 5 - 4 \times ((100 - q)/100 \times 0.7132 + 0.0253),
\]

where Equations (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 [22] is used to train the spatial pooling module separately. This dataset contains 10,073 in-the-wild images with a resolution of 1,024 × 768 pixels.

### 4.2 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:

\[
L_g(x) = \beta + \frac{\alpha - \beta}{1 + \exp(-x + y/|\delta|)},
\]

### 4.3 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 [30] with an initial learning rate of \(1e - 4\), batch size of 16 and the Mean Squared Error (MSE) as loss function \(\ell_2\):

\[
\ell_2(q^I, \hat{q}^I) = \frac{1}{L} \sum_{l=1}^{L} (q^I_l - \hat{q}^I_l)^2,
\]

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 20, No. 4, Article 100. Publication date: December 2023.
Table 2. Performance of the Ablation Study for the Spatial Pooling Module on the KonIq-10k Dataset

| Backbone        | # 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     | 1,664                | 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        | 2,048                | 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  | 2,560                | 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  |

Bold entries indicate the top three performing methods, while the best method is underlined.

where $q^l$, $\hat{q}^l$, 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 $1 \times 10^{-4}$ and MES as a loss function $\ell_2(q, \hat{q})$ computed between $q$ 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.\(^1\)

4.4 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 [100], Densenet169 [23], ResNet-50 [18], and EfficientNetB7 [74]. 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 2, where some interesting observations can be made.

\(^1\)https://github.com/atelili/2BiVQA
Table 3. Performance of the Ablation Study on the KoNViD-1k Dataset

| Backbone   | Temporal pooling  | SROCC ↑ | PLCC ↑ | KRCC ↑ | RMSE ↓ | #parameters       |
|------------|-------------------|---------|--------|--------|--------|-------------------|
| VGG16      | Mean              | 0.776/0.711 | 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              | 0.750/0.800 | 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         |

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.

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 feature 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 3. 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.

4.5 Performance Evaluation and Comparison

To assess the performance of the proposed 2BiVQA metric, we compared it with ten BIQA models (BRISQUE [48], NIQE [50], ILNIQE [98], BMPRI [45], StairIQA [72], HIGRADE [32], CORNIA [93], HOSA [88], FRIQUEE [12], PaQ-2-PiQ [95], and KonCept512 [22]), and five BVQA models (V-BLIINDS [62], FAST-VQA [85], TLVQM [31], VIDEVAL [78], and RAPIQUE [79]). In addition, two...
deep CNN models (VGG-19 [68] and ResNet-50 [18]) using transfer learning were benchmarked. Among these methods, NIQE and ILNIQE are completely blind, because they do not 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 4 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 5 provides a 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.

4.6 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 6, from which 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.
Table 4. Performance Comparison of Evaluated BVQA Models on the four UGC-VQA Datasets

| Dataset         | KonViD-1k | LIVE-VQC |
|-----------------|-----------|----------|
|                 | SROCC     | PLCC     | KRRCC | RMSE   | SROCC     | PLCC     | KRRCC | RMSE   |
| BRSIQE [48]     | 0.656     | 0.657    | 0.476 | 0.481  | 0.592     | 0.638    | 0.416 | 13.100 |
| NIQE [50]       | 0.541     | 0.553    | 0.379 | 0.533  | 0.595     | 0.628    | 0.425 | 13.110 |
| ILNIQE [98]     | 0.526     | 0.540    | 0.369 | 0.540  | 0.503     | 0.543    | 0.355 | 14.148 |
| StairIQA [72]   | 0.767     | 0.778    | 0.570 | 0.881  | 0.740     | 0.787    | 0.547 | 10.597 |
| BMPIR [45]      | 0.519     | 0.522    | 0.354 | 0.661  | 0.502     | 0.586    | 0.350 | 13.558 |
| HIGRADE [32]    | 0.720     | 0.726    | 0.531 | 0.439  | 0.610     | 0.633    | 0.439 | 13.027 |
| FRIQUEE [12]    | 0.747     | 0.748    | 0.550 | 0.425  | 0.657     | 0.700    | 0.477 | 12.198 |
| CORNIA [93]     | 0.716     | 0.713    | 0.523 | 0.448  | 0.671     | 0.718    | 0.484 | 11.832 |
| HOSA [88]       | 0.765     | 0.766    | 0.569 | 0.414  | 0.687     | 0.741    | 0.503 | 11.353 |
| VGG-19 [68]     | 0.774     | 0.784    | 0.584 | 0.395  | 0.656     | 0.716    | 0.472 | 11.783 |
| ResNet-50 [18]  | 0.801     | 0.810    | 0.610 | 0.374  | 0.663     | 0.720    | 0.478 | 11.591 |
| KonCept512 [22] | 0.734     | 0.748    | 0.542 | 0.426  | 0.664     | 0.727    | 0.479 | 11.626 |
| PaQ-2-PiQ [95]  | 0.613     | 0.601    | 0.433 | 0.514  | 0.643     | 0.668    | 0.456 | 12.619 |
| V-BLINDS [62]   | 0.710     | 0.703    | 0.518 | 0.459  | 0.693     | 0.717    | 0.507 | 11.765 |
| TLVQM [31]      | 0.772     | 0.768    | 0.577 | 0.410  | 0.798     | 0.802    | 0.608 | 10.145 |
| VIDEVAL [78]    | 0.783     | 0.780    | 0.584 | 0.402  | 0.752     | 0.751    | 0.563 | 11.100 |
| RAQique [79]    | 0.807     | 0.815    | 0.618 | 0.364  | 0.741     | 0.765    | 0.557 | 10.665 |
| FAST-VQA [85]   | 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 |
|-----------------|--------------|--------------|
|                 | SROCC        | PLCC         | KRRCC | RMSE  | SROCC        | PLCC         | KRRCC | RMSE  |
| BRSIQE [48]     | 0.382        | 0.395        | 0.263 | 0.591 | 0.569        | 0.586        | 0.403 | 0.561 |
| NIQE [50]       | 0.237        | 0.277        | 0.160 | 0.617 | 0.462        | 0.477        | 0.322 | 0.611 |
| ILNIQE [98]     | 0.291        | 0.330        | 0.198 | 0.605 | 0.459        | 0.474        | 0.321 | 0.611 |
| StairIQA [72]   | 0.753        | 0.744        | 0.558 | 0.597 | 0.777        | 0.780        | 0.586 | 0.480 |
| BMPIR [45]      | 0.295        | 0.372        | 0.199 | 0.636 | 0.505        | 0.525        | 0.351 | 0.678 |
| HIGRADE [32]    | 0.737        | 0.721        | 0.547 | 0.447 | 0.739        | 0.736        | 0.547 | 0.467 |
| FRIQUEE [12]    | 0.765        | 0.757        | 0.568 | 0.416 | 0.756        | 0.755        | 0.565 | 0.454 |
| CORNIA [93]     | 0.597        | 0.605        | 0.421 | 0.513 | 0.676        | 0.697        | 0.484 | 0.494 |
| HOSA [88]       | 0.602        | 0.604        | 0.425 | 0.513 | 0.695        | 0.708        | 0.503 | 0.489 |
| VGG-19 [68]     | 0.702        | 0.699        | 0.509 | 0.456 | 0.732        | 0.748        | 0.539 | 0.461 |
| ResNet-50 [18]  | 0.718        | 0.709        | 0.522 | 0.453 | 0.755        | 0.774        | 0.561 | 0.438 |
| KonCept512 [22] | 0.587        | 0.594        | 0.410 | 0.513 | 0.660        | 0.676        | 0.475 | 0.509 |
| PaQ-2-PiQ [95]  | 0.265        | 0.293        | 0.177 | 0.615 | 0.472        | 0.482        | 0.324 | 0.608 |
| V-BLINDS [62]   | 0.559        | 0.555        | 0.389 | 0.535 | 0.654        | 0.659         | 0.473 | 0.520 |
| TLVQM [31]      | 0.669        | 0.659        | 0.481 | 0.484 | 0.727        | 0.734         | 0.534 | 0.470 |
| VIDEVAL [78]    | 0.778        | 0.773        | 0.583 | 0.404 | 0.796        | 0.793         | 0.603 | 0.426 |
| RAQique [79]    | 0.761        | 0.762        | 0.561 | 0.406 | 0.808        | 0.818         | 0.614 | 0.407 |
| FAST-VQA [85]   | **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 |

The underlined and boldfaced entries indicate the best and top three performers on each dataset for each performance measure, respectively.
Table 5. 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 |

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.

4.7 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 @ 3.70 GHz ×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.
Table 6. 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 7. Average Runtime Comparison Evaluated on 1,080p Videos from YouTube-UGC

| Method     | Deep Learning | Framework | Time (s) |
|------------|---------------|-----------|----------|
|            |               | 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 |

We consider videos from YouTube-UGC at HD resolution (1,920 × 1,080), then we recorded the average runtime in seconds, as shown in Table 7. For better illustration, Figure 6 shows the scatter plots of SROCC versus runtime. It may be observed that FAST-VQA and BIQA models are faster than other methods, while VGG19, RAPIQUE, and 2BiVQA are relatively comparable.

5 CONCLUSION

In this article, 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.
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] Sewoong Ahn and Sanghoon Lee. 2018. Deep blind video quality assessment based on temporal human perception. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP’18). IEEE, 619–623.

[2] Aishy Amer and Eric Dubois. 2005. Fast and reliable structure-oriented video noise estimation. IEEE Trans. Circ. Syst. Video Technol. 15, 1 (2005), 113–118.

[3] S. A. Amirshahi, M. Pedersen, and S. X. Yu. 2016. Image quality assessment by comparing CNN features between images. J. Imag. Sci. Technol. 60, 6 (2016), 60410–1.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal speed and accuracy of object detection. Retrieved from https://arxiv.org/abs/2004.10934

[5] ITU Recommendation BT. 2012. Methodology for the Subjective Assessment of the Quality of Television Pictures. Int. Telecommun. Union 6 (2012).

[6] Y. Cao, X. Min, W. Sun, and G. Zhai. 2023. Subjective and objective audio-visual quality assessment for user generated content. In IEEE Transactions on Image Processing, Vol. 32, 3847–3861.

[7] U Cisco. 2020. Cisco annual internet report (2018–2023) white paper. Cisco, San Jose, CA.

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

[9] Xiaojun Feng and Jan P. Allebach. 2006. Measurement of ringing artifacts in JPEG images. In Digital Publishing, Vol. 6076. International Society for Optics and Photonics, 60760A.

[10] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu. 2017. Deepsim: Deep similarity for image quality assessment. Neurocomputing 257 (2017), 104–114.

[11] Deepti Ghadiyaram and Alan C. Bovik. 2015. Massive online crowdsourced study of subjective and objective picture quality. IEEE Trans. Image Process. 25, 1 (2015), 372–387.

[12] Deepti Ghadiyaram and Alan C. Bovik. 2017. Perceptual quality prediction on authentically distorted images using a bag of features approach. J. Vision 17, 1 (2017), 32–32.

[13] Ross Girshick. 2015. Fast r-CNN. In Proceedings of the IEEE International Conference on Computer Vision. 1440–1448.
2BiVQA: Double Bi-LSTM-based Video Quality Assessment of UGC Videos

[14] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. 2021. KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. *IEEE Access* 9 (2021), 72139–72160.

[15] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 273–278.

[16] Alex Graves and Jürgen Schmidhuber. 2005. Frame-wise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* 18, 5-6 (2005), 602–610.

[17] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenhui Zhang. 2014. Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Trans. Broadcast.* 60, 3 (2014), 555–567.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[19] Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncert. Fuzz. Knowl.-Based Syst.* 6, 02 (1998), 107–116.

[20] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.

[21] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KonVID-1k). In *Proceedings of the 9th International Conference on Quality of Multimedia Experience (QoMEX’17)*. IEEE, 1–6.

[22] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.* 29 (2020), 4041–4056.

[23] Gao Huang, Zhang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.

[24] Xinhui Huang, Chunyi Li, Abdelhak Bentaleb, Roger Zimmermann, and Guangtao Zhai. 2023. XGC-VQA: A unified video quality assessment model for user, professionally, and occupationally-generated content. Retrieved from https://arXiv:2303.13859

[25] J. Johnson, A. Alahi, and L. Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*. Springer, 694–711.

[26] Parimala Kancharla and Sumohana S. Channappayya. 2022. Completely blind quality assessment of user generated video content. *IEEE Trans. Image Process.* 31 (2022), 2623–274. https://doi.org/10.1109/TIP.2021.3130541

[27] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1733–1740.

[28] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C. Bovik. 2017. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Process. Mag.* 34, 6 (2017), 130–141.

[29] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. 2018. 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’18)*. 219–234.

[30] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR’15)*, Yoshua Bengio and Yann LeCun (Eds.).

[31] Jari Korhonen. 2019. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.* 28, 12 (2019), 5923–5938.

[32] Debabrata Kundu, Deepti Ghadiyaram, Alan C. Bovik, and Brian L. Evans. 2017. No-reference quality assessment of tone-mapped HDR pictures. *IEEE Trans. Image Process.* 26, 6 (2017), 2957–2971.

[33] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2351–2359.

[34] Dingquan Li, Tingting Jiang, and Ming Jiang. 2021. Unified quality assessment of in-the-wild videos with mixed datasets training. *Int. J. Comput. Vision* 129, 4 (2021), 1238–1257.

[35] Xuelong Li, Qun Guo, and Xiaoqiang Lu. 2016. Spatiotemporal statistics for video quality assessment. *IEEE Trans. Image Process.* 25, 7 (2016), 3329–3342.

[36] Weili Liu, Andrew Rabinovich, and Alexander C. Berg. 2015. ParseNet: Looking wider to see better. Retrieved from https://arxiv.org/abs/1506.04579

[37] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. 2015. ParseNet: Looking wider to see better. Retrieved from https://arxiv.org/abs/1506.04579

[38] Yutao Liu, Ke Gu, Xiu Li, and Yongbing Zhang. 2020. Blind image quality assessment by natural scene statistics and perceptual characteristics. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 3 (2020), 1–91.

[39] Yanan Lu, Fengying Xie, Tongliang Liu, Zhiqiu Jiang, and Dacheng Tao. 2015. No reference quality assessment for multiply-distorted images based on an improved bag-of-words model. *IEEE Signal Process. Lett.* 22, 10 (2015), 1811–1815.
[40] K. Manasa and Sumohana S. Channappayya. 2016. An optical flow-based no-reference video quality assessment algorithm. In Proceedings of the IEEE International Conference on Image Processing (ICIP’16). IEEE, 2400–2404.
[41] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. 2002. A no-reference perceptual blur metric. In Proceedings of the International Conference on Image Processing. Vol. 3. IEEE, III.
[42] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen. 2017. Blind quality assessment based on pseudo-reference image. IEEE Trans. Multimedia 20, 8 (2017), 2049–2062.
[43] Xiongkuo Min, Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, Patrick Le Callet, and Chang Wen Chen. 2021. Screen content quality assessment: Overview, benchmark, and beyond. ACM Comput. Surveys 54, 9 (2021), 1–36.
[44] Xiongkuo Min, Kede Ma, Ke Gu, Guangtao Zhai, Zhou Wang, and Weisi Lin. 2017. Unified blind quality assessment of compressed natural, graphic, and screen content images. IEEE Trans. Image Process. 26, 11 (2017), 5462–5474.
[45] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang. 2018. Blind image quality estimation via distortion aggravation. IEEE Trans. Broadcast. 64, 2 (2018), 508–517.
[46] Xiongkuo Min, Guangtao Zhai, Jiannao Zhou, Mylene C. Q. Farias, and Alan Conrad Bovik. 2020. Study of subjective and objective quality assessment of audio-visual signals. IEEE Trans. Image Process. 29 (2020), 6054–6068.
[47] Shankhanil Mitra and Rajiv Soundararajan. 2022. Multiview contrastive learning for completely blind video quality assessment of user generated content. In Proceedings of the 30th ACM International Conference on Multimedia. 1914–1924.
[48] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21, 12 (2012), 4695–4708.
[49] Anish Mittal, Michele A. Saad, and Alan C. Bovik. 2015. A completely blind video integrity oracle. IEEE Trans. Image Process. 25, 1 (2015), 289–300.
[50] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2012. Making a “completely blind” image quality analyzer. IEEE Signal Process. Lett. 20, 3 (2012), 209–212.
[51] Anush Krishna Moorthy and Alan Conrad Bovik. 2010. A two-step framework for constructing blind image quality indices. IEEE Signal Process. Lett. 17, 5 (2010), 513–516.
[52] Anush Krishna Moorthy and Alan Conrad Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Trans. Image Process. 20, 12 (2011), 3350–3364.
[53] Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. 2009. Considering temporal variations of spatial visual distortions in video quality assessment. IEEE J. Select. Top. Signal Process. 3, 2 (2009), 253–265.
[54] Andrey Norkin and Neil Birkbeck. 2018. Film grain synthesis for AV1 video codec. In Proceedings of the Data Compression Conference. IEEE, 3–12.
[55] Jincheol Park, Kalpana Seshadrinathan, Sanghoon Lee, and Alan Conrad Bovik. 2012. Video quality pooling adaptive to perceptual distortion severity. IEEE Trans. Image Process. 22, 2 (2012), 610–620.
[56] Soo-Chang Pei and Li-Heng Chen. 2015. Image quality assessment using human visual DOG model fused with random forest. IEEE Trans. Image Process. 24, 11 (2015), 3282–3292.
[57] Margaret H. Pinson and Stephen Wolf. 2003. An objective method for combining multiple subjective data sets. In Proceedings of the Conference on Visual Communications and Image Processing, Vol. 5150. International Society for Optics and Photonics, 583–592.
[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, 234–241.
[59] Daniel L. Ruderman and William Bialek. 1994. Statistics of natural images: Scaling in the woods. Phys. Rev. Lett. 73, 6 (1994), 814.
[60] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. 2010. A DCT statistics-based blind image quality index. IEEE Signal Process. Lett. 17, 6 (2010), 583–586.
[61] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. IEEE Trans. Image Process. 21, 8 (2012), 3339–3352.
[62] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. 2014. Blind prediction of natural video quality. IEEE Trans. Image Process. 23, 3 (2014), 1352–1365.
[63] Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45, 11 (1997), 2673–2681.
[64] Kalpana Seshadrinathan and Alan C. Bovik. 2011. Temporal hysteresis model of time varying subjective video quality. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’11). IEEE, 1153–1156.
[65] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. 2010. Study of subjective and objective quality assessment of video. IEEE Trans. Image Process. 19, 6 (2010), 1427–1441.
2BiVQA: Double Bi-LSTM-based Video Quality Assessment of UGC Videos

[66] Wenhai Shen, Mingliang Zhou, Xingran Liao, Weijia Jia, Tao Xiang, Bin Fang, and Zhaowei Shang. 2022. An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation. IEEE Transactions on Broadcasting 68, 3 (2022), 651–660.

[67] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. The performance of LSTM and BiLSTM in forecasting time series. In Proceedings of the IEEE International Conference on Big Data (Big Data'19). IEEE, 3285–3292.

[68] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR'15), Yoshua Bengio and Yann LeCun (Eds.).

[69] Zeina Sinno and Alan Conrad Bovik. 2018. Large-scale study of perceptual video quality. IEEE Trans. Image Process. 28, 2 (2018), 612–627.

[70] Zeina Sinno and Alan C. Bovik. 2019. Spatio-temporal measures of naturalness. In Proceedings of the IEEE International Conference on Image Processing (ICIP'19). IEEE, 1750–1754.

[71] Wei Sun, Xiongkun Min, Wei Lu, and Guangtao Zhai. 2022. A deep learning based no-reference quality assessment model for ugc videos. In Proceedings of the 30th ACM International Conference on Multimedia. 856–865.

[72] Wei Sun, Xiongkun Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. 2023. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. IEEE J. Select. Top. Signal Process. (2023).

[73] Wei Sun, Tao Wang, Xiongkun Min, Fuwang Yi, and Guangtao Zhai. 2021. Deep learning based full-reference and no-reference quality assessment models for compressed UGC videos. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW'21). IEEE, 1–6.

[74] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning. PMLR, 6105–6114.

[75] Lohic Fotio Tiotsop, Tomas Mizdos, Marcus Barkowsky, Peter Poca, Antonio Servetti, and Enrico Masala. 2022. Mimicking individual media quality perception with neural network based artificial observers. ACM Trans. Multimedia Comput. Commun. Appl. 18, 1 (2022), 1–25.

[76] Zhengzhong Tu, Chia-Ju Chen, Li-Heng Chen, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2020. A comparative evaluation of temporal pooling methods for blind video quality assessment. In Proceedings of the IEEE International Conference on Image Processing (ICIP'20). IEEE, 141–145.

[77] Zhengzhong Tu, Jessie Lin, Yilin Wang, Balu Adsumilli, and Alan C. Bovik. 2020. Bband index: A no-reference banding artifact predictor. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20). IEEE, 2712–2716.

[78] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. UGC-VQA: Benchmarking blind video quality assessment for user generated content. IEEE Trans. Image Process. 30 (2021), 4449–4464.

[79] Zhengzhong Tu, Xiangyu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. 2021. RAPIQUE: Rapid and accurate video quality prediction of user generated content. IEEE Open J. Signal Process. 2 (2021), 425–440. https://doi.org/10.1109/OJSP.2021.3090333

[80] Xin Wang, Baofeng Tian, Chao Liang, and Dongcheng Shi. 2008. Blind image quality assessment for measuring image blur. In Proceedings of the Congress on Image and Signal Processing, Vol. 1. IEEE, 467–470.

[81] Yilin Wang, Sasi Inguva, and Balu Adsumilli. 2019. YouTube UGC dataset for video compression research. In Proceedings of the IEEE 21st International Workshop on Multimedia Signal Processing (MMSP'19). IEEE, 1–5.

[82] Yilin Wang, Sang-Uk Kum, Chao Chen, and Anil Kokaram. 2016. A perceptual visibility metric for banding artifacts. In Proceedings of the IEEE International Conference on Image Processing (ICIP'16). IEEE, 2067–2071.

[83] Zhou Wang, Alan C. Bovik, and Brian L. Evan. 2000. Blind measurement of blocking artifacts in images. In Proceedings International Conference on Image Processing, Vol. 3. IEEE, 981–984.

[84] Andrew B. Watson and Cynthia H. Null. 1997. Digital images and human vision. In Proceedings of the Electronic Imaging Science and Technology Conference.

[85] Haoning Wu, Chao Feng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenhui Sun, Qiong Yan, and Weisi Lin. 2022. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In Proceedings of the European Conference on Computer Vision. Springer, 538–554.

[86] Haoning Wu, Erli Zhang, Liang Liao, Chao Feng Chen, Jingwen Hou, Annan Wang, Wenhui Sun, Qiong Yan, and Weisi Lin. 2023. 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'23).

[87] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. 2021. Perceptual quality assessment of internet videos. In Proceedings of the 29th ACM International Conference on Multimedia. 1248–1257.

[88] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. 2016. Blind image quality assessment based on high order statistics aggregation. IEEE Trans. Image Process. 25, 9 (2016), 4444–4457.

[89] Jingtao Xu, Peng Ye, Yong Liu, and David Doermann. 2014. No-reference video quality assessment via feature learning. In Proceedings of the IEEE International Conference on Image Processing (ICIP'14). IEEE, 491–495.

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 20, No. 4, Article 100. Publication date: December 2023.
[90] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C. Bovik, and Xiangchu Feng. 2014. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Trans. Image Process.* 23, 11 (2014), 4850–4862.

[91] Xiaohan Yang, Fan Li, and Hantao Liu. 2019. A survey of DNN methods for blind image quality assessment. *IEEE Access* 7 (2019), 123788–123806.

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

[93] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. 2012. Unsupervised feature learning framework for no-reference image quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1098–1105.

[94] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. 2021. Attention based network for no-reference UGC video quality assessment. In *Proceedings of the IEEE International Conference on Image Processing (ICIP’21)*. IEEE, 1414–1418.

[95] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. 2020. 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*. 3575–3585.

[96] Junyong You and Jari Korhonen. 2019. Deep neural networks for no-reference video quality assessment. In *Proceedings of the IEEE International Conference on Image Processing (ICIP’19)*. IEEE, 2349–2353.

[97] Guangtao Zhai and Xiongkuo Min. 2020. Perceptual image quality assessment: A survey. *Sci. China Info. Sci.* 63 (2020), 1–52.

[98] Lin Zhang, Lei Zhang, and Alan C. Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Process.* 24, 8 (2015), 2579–2591.

[99] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

[100] Xiangyu Zhang, Jianhua Zou, Kaiming He, and Jian Sun. 2015. Accelerating very deep convolutional networks for classification and detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 10 (2015), 1943–1955.

[101] Yi Zhang and Damon M. Chandler. 2013. No-reference image quality assessment based on log-derivative statistics of natural scenes. *J. Electron. Imag.* 22, 4 (2013), 043025.

[102] Yi Zhang, Anush K. Moorthy, Damon M. Chandler, and Alan C. Bovik. 2014. C-DIIVINE: No-reference image quality assessment based on local magnitude and phase statistics of natural scenes. *Signal Process.: Image Commun.* 29, 7 (2014), 725–747.

[103] Ying Zhang, Luming Zhang, and Roger Zimmermann. 2015. Aesthetics-guided summarization from multiple user generated videos. *ACM Trans. Multimedia Comput. Commun. Appl.* 11, 2 (2015), 1–23.

[104] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. 2023. MD-VQA: Multi-dimensional quality assessment for UGC live videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1746–1755.

[105] Yi Zhu, Sharath Chandra Guntuku, Weisi Lin, Cheorghita Ghinea, and Judith A. Redi. 2018. Measuring individual video QoE: A survey, and proposal for future directions using social media. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2S (2018), 1–24.

Received 13 April 2023; revised 18 September 2023; accepted 22 October 2023