IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 16, 2022, accepted April 6, 2022, date of publication April 14, 2022, date of current version April 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3167446

No-Reference Video Quality Assessment Using Distortion Learning and Temporal Attention

KOFFI KOSSI[1], STÉPHANE COULOMBE[2], (Senior Member, IEEE), CHRISTIAN DESROSIERS[2], AND GHYSLAIN GAGNON[1], (Member, IEEE)
[1]Department of Electrical Engineering, École de technologie supérieure, Université du Québec, Montreal, QC H3C 1K3, Canada
[2]Department of Software and IT Engineering, École de technologie supérieure, Université du Québec, Montreal, QC H3C 1K3, Canada

Corresponding author: Koffi Kossi (koffi-segnedji.kossi.1@ens.etsmtl.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), and in part by Dr. Richard J. Marceau Chair on Wireless IP Technology for Developing Countries, particularly through our industrial partner, Media5 Corporation.

ABSTRACT The rapid growth of video consumption and multimedia applications has increased the interest of the academia and industry in building tools that can evaluate perceptual video quality. Since videos might be distorted when they are captured or transmitted, it is imperative to develop reliable methods for no-reference video quality assessment (NR-VQA). To date, most NR-VQA models in prior art have been proposed for assessing a specific category of distortion, such as authentic distortions or traditional distortions. Moreover, those developed for both authentic and traditional distortions video databases have so far led to poor performances. This resulted in the reluctance of service providers to adopt multiple NR-VQA approaches, as they prefer a single algorithm capable of accurately estimating video quality in all situations. Furthermore, many existing NR-VQA methods are computationally complex and therefore impractical for various real-life applications. In this paper, we propose a novel deep learning method for NR-VQA based on multi-task learning where the distortion of individual frames in a video and the overall quality of the video are predicted by a single neural network. This enables to train the network with a greater amount and variety of data, thereby improving its performance in testing. Additionally, our method leverages temporal attention to select the frames of a video sequence which contribute the most to its perceived quality. The proposed algorithm is evaluated on five publicly-available video quality assessment (VQA) databases containing traditional and authentic distortions. Results show that our method outperforms the state-of-the-art on traditional distortion databases such as LIVE VQA and CSIQ video, while also delivering competitive performance on databases containing authentic distortions such as KoNViD-1k, LIVE-Qualcomm and CVD2014.

INDEX TERMS Video quality assessment, no reference, transfer learning, multi-task learning, attention mechanism, authentic distortion, traditional distortion, content-aware, NR-VQA.

I. INTRODUCTION

For manufacturers and telecommunications service providers, the recent increase of video-driven data consumption has led to the challenge of delivering better video services. It has also created a pressing need to monitor and control the video quality to maximize their benefits [1]. As a result, VQA has drawn increasing attention from researchers in the field. VQA, which aims to predict the perceptual quality of a video, remains a fundamental problem in many video processing tasks such as video acquisition, compression and transport [2]–[4]. Like IQA (Image Quality Assessment), there are *subjective* and *objective* VQA approaches. Subjective VQA is the most reliable of the two, however its high cost and complexity to prepare and run tests involving humans makes this approach impractical for automated quality assessement. On the other hand, objective VQA uses computational models to predict the video quality in line with the perception of the human visual system (HVS). Existing objective VQA methods can be classified into full-reference VQA (FR-VQA) [5]–[8], reduced-reference VQA (RR-VQA) [9], [10] and no-reference VQA (NR-VQA) [11]–[15] based on the accessibility of the corresponding reference when estimating a video's quality. Compared to

The associate editor coordinating the review of this manuscript and approving it for publication was Usama Mir.

FR-VQA and RR-VQA, which require all or part of the information from reference videos, NR-VQA is highly beneficial as the reference video is not required [16]. Therefore, NR-VQA models are better suited for many practical applications, such as real-time monitoring of the received video quality at a streaming client.

According to recent studies, the legacy NR-VQA models don't perform well on videos containing authentic or natural distortions such as videos in KoNViD-1k dataset [17]. As in [18], we call *authentic/in-capture distortions*, those occurring during acquisition and *traditional/post-capture distortions*, those generated in a controlled lab such as compression and transmission distortions (e.g. packet loss). Authentic videos are also referred as *in-the-wild*.

Consequently, there is a need to design approaches that can perform well on a broader range of data. To design a robust objective VQA method, it is thus important to consider the different types of distortions that can impact video quality. However, the complexity of temporal visual characteristics and content-dependent video compression artifacts make NR-VQA a very challenging task.

In this paper, we propose a novel deep learning NR-VQA method based on multi-task learning and temporal attention to predict the video quality for both authentic and traditional distortions databases. Our method combines content-aware and distortion features extracted in two different CNN branches of the network (see Fig. 1) and incorporates them into a Gated Recurrent Unit (GRU) network coupled with a temporal attention mechanism.

To demonstrate the superiority of our method, we conduct experiments on five publicly-available databases, namely CVD2014 [19], KoNViD-1k [17], LIVE-Qualcomm [20], LIVE VQA [21], and CISQ video [22]. Compared to state-of-the-art approaches, our proposed method can predict more consistently the video quality for authentic/in-capture databases (CVD2014, KoNViD-1k, LIVE-Qualcomm) as well as for traditional/post-capture databases (LIVE VQA, CISQ video).

Additionally, we perform an ablation study to further validate our approach and verify the advantage of its main components. Finally, we evaluate our model's computational complexity and observe a good trade-off between high accuracy and computational efficiency for our deep learning method.

The main contributions of this work are as follows:

- We present a novel deep learning approach for objective NR-VQA, based on multi-task learning where the distortion of individual frames in a video and the overall quality of the video are predicted by a single neural network. Our approach also leverages temporal attention to select the frames of a video sequence which contribute the most to its perceived quality. Compared to recent models such as CNN-TLVQM [23] and RAPIQUE [24], which are a mixture of hand-crafted and CNN feature extraction models, our method uses a deep learning strategy for extracting all features. Moreover, while recent models such as CoINVQ [25] use several branches for feature extraction, and aggregate temporal features using average pooling, our method uses a GRU that can better model temporal dynamics in the video and its impact on overall quality. Experiments showed that standard pooling strategies like average pooling are not well suited for non temporally uniform distortions such as transmission errors.
- A special pooling designed with a weighted sum of mean and attention pooling. This enhanced pooling mechanism considers both temporally-local quality and global quality. This is achieved via a novel combination of attention-based pooling, focusing on frames having a greater impact on perceived quality, and average pooling. The benefit of our pooling mechanism, compared to three other techniques that reflect the human judgement of quality (minimum, temporal and recency pooling), is confirmed in Table 5 of this manuscript.
- We introduce a distortion network designed as a complementary feature extraction branch to improve the video quality prediction, specially in case of traditional distortion. From the state-of-the-art, we present one of the first deep learning model for a range of databases containing traditional and authentic distortions. Recent models are mostly designed for authentic video distortion and don't take into consideration traditional video distortions such as distortions related to wireless transmission. Authors of [18] have recently designed a hand-crafted model for a wide variety of databases containing traditional and authentic distortions but the performance of their model on authentic databases is not competitive with state-of-the-art approaches.
- The proposed method achieves state-of-the-art performance for both authentic and traditional distortion databases, outperforming existing current approaches for traditional distortion databases (LIVE VQA, CSIQ video) while also providing accuracy on par with top-ranked approaches for authentic distortion databases (CVD2014, KoNViD-1k and LIVE-Qualcomm).

The rest of this paper is organized as follows. In Section II, we present related works. In Section III, we describe our proposed NR-VQA method. Our experiments and results are presented and discussed in Section IV. Finally, in Section V, we conclude and suggest some future works.

II. RELATED WORKS

Recent NR-VQA methods, which are mostly learning-based, can be roughly divided in two groups: those based uniquely on spatial image-level features and those that also account for temporal information between the frames in the video.

Image-based NR-VQA methods share roots with image quality assessment methods and thus involve the analysis of natural scene statistics (NSS) [26], [27]. The supporting theory of NSS is that certain statistical properties of natural images are highly related with how the HVS processes these images, thus image quality can be obtained by

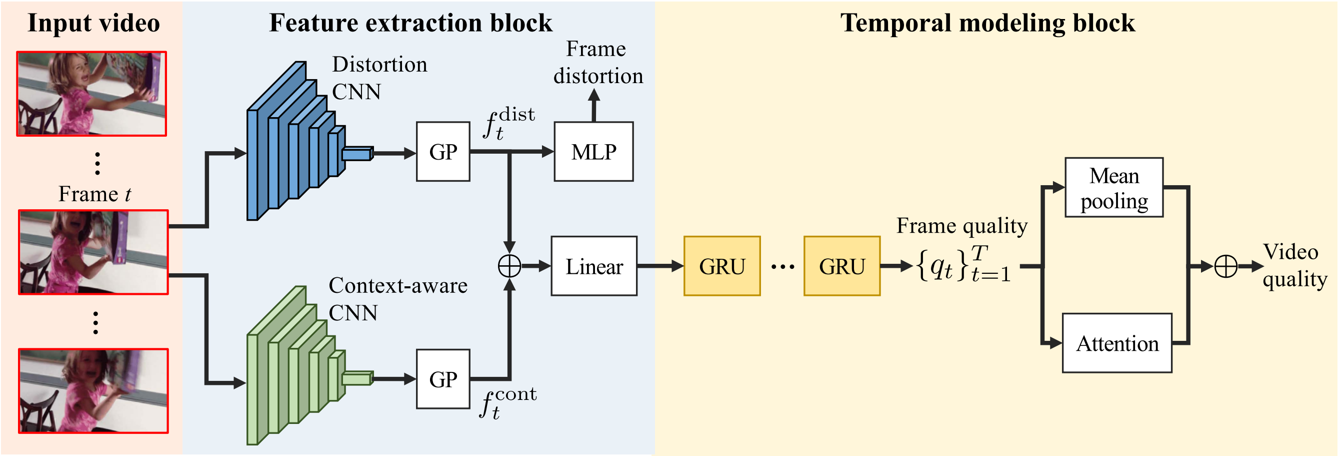

FIGURE 1. **The architecture of the proposed method which consists of two main blocks: the feature extraction and the temporal modeling blocks. The feature extraction block extracts simultaneously, from each video frame, both the distortion and content-aware features. The second block is a GRU network coupled with an attention mechanism and mean pooling for selecting relevant frames and improving the robustness of the model. GP: Global Average Pooling.**

measuring the deviations from these statistics [28]. These image-level approaches have been extended to videos by evaluating such statistics at the frame level and aggregating them to get a quality score for the entire video. Examples of such approaches include Naturalness Image Quality Evaluator (NIQE) [11], COdebook Representation for No-Reference Image Assessment (CORNIA) [29], Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [15], Feature maps based Referenceless Image QUality Evaluation Engine (FRIQUEE) [30], and High dynamic range Image GRADient based Evaluator (HIGRADE) [31].

On the other hand, few approaches in the literature take directly into consideration the temporal aspect of videos. The best known learning-based model in this domain is Video BLIINDS (V-BLIINDS) [32], which extends the image-based metrics designed for NIQE [11] by adding temporal motion information and time-frequency characteristics of the video. The temporal features are extracted using block-based motion estimation and Discrete Cosine Transform (DCT) coefficients computed from frame differences. This approach has been referred as the baseline method against which most NR-VQA methods are compared. Another well-known machine-learning NR-VQA model is the Video COdebook Representation for No-reference Image Assessment (V-CORNIA) [33]. This frame-feature learning approach uses Support Vector Regression (SVR) to first predict quality at the frame level, and then applies temporal pooling on frame-level qualities to obtain the overall video level score.

Recently, Deep Neural Networks (DNN) have been applied to NR-VQAs. For example, SACONVA [34] uses a three-dimensional (3D) shearlet transform to extract frame-level features which allows capturing spatio-temporal quality features. A Convolutional Neural Network (CNN) and a logistic regression model are then employed respectively to expand these features and obtain the quality scores. The COnvolutional neural network and Multi-regression based Evaluation (COME) [35] approach separates the problem of extracting spatio-temporal quality features in two parts. First, a CNN is used on the CSIQ database to extract spatial quality features, based on max pooling and the standard deviation of activations in the final layer. Temporal quality features are then obtained as standard deviations of motion vectors in the video. Lastly, the predictions of two SVR models are combined with a Bayes classifier to predict the final quality score. The Video Multi-task End-to-end Optimized neural Network (V-MEON) [36] approach, which is the video version of MEON [37], predicts video quality with a multi-task framework that jointly estimates the perceptual quality of a video and predicts its codec type using spatio-temporal features extracted from a 3D CNN.

Two recently-developed methods against which most methods are compared, VSFA [13] and TLVQM [38], are used as state-of-the-art baselines in our study. VSFA [13] extracts content-aware features from a CNN pre-trained on ImageNet [39] and uses a GRU to model the long-term dependencies between different frames in the video. Additionally, a subjectively inspired temporal pooling model is proposed to consider the hysteresis effect observed in human judgments of time varying video quality [40]. Our proposed method differs from VSFA in two important ways. First, while VSFA predicts the video quality directly, we employ a multi-task learning strategy where the distortion type of individual images is predicted in a separate branch of the network. As in self-supervised representation learning approaches [41], this branch is trained with easy-to-obtain labels, i.e. images corrupted with different types of distortion, to learn a representation in a pre-training phase. In a second phase, this image-wise representation is combined with the features computed in another branch of the network for predicting the quality of the whole video. Compared to VSFA, which uses a pre-defined strategy to combine the quality scores of separate frames, our method uses learned attention to find frames in the sequence having the greatest impact on perceived quality.

The Two Level Video Quality Model (TLVQM) [38] model adopts a hierarchical feature extraction approach for predicting the video quality. Specifically, two types of features are extracted: low complexity features characterizing global information of the full sequence and high complexity features such as spatial activity, exposure or sharpness which are extracted from a small representative subset of frames. Compared to TLVQM, where spatial and temporal features are extracted together using hand-crafted techniques, our method learns the spatial and temporal characteristics of videos with two neural network branches.

The above-mentioned studies have focused on a specific type of distortion and trained models on a same category of distortion databases. For example, recent state-of-the-art methods like VSFA and TLVQM are designed specifically for in-capture distortion databases and tested only on this type of data. This limits the adoption of these algorithms in application settings where distortion can arise during the acquisition as well as transmission processes. Thus, there is a need to develop novel NR-VQA algorithms that consider all types of distortions such as in-capture/authentic as well as post-capture/traditional distortions. Apart from the two state-of-the-art baselines used in this study, we compared the performance of our method to some recent published NR-VQA approaches such as RAPIQUE [24], PVQ [42], CoINVQ [25] and CNN-TLVQM [23], where the latter is an improvement of the TLVQM model [38].

III. PROPOSED METHOD

In this section, we present the details of our deep learning NR-VQA method, which combines content-aware and distortion features. These combined features are sent to a fully-connected (FC) linear layer reducing their dimensionality and integrated into a GRU network which models the temporal inter-dependencies. Finally, a temporal attention mechanism and a traditional average pooling are coupled to this GRU network to select the frames of a video sequence that contribute mostly to the perceived quality and improve the robustness of our model. The overall architecture of the proposed method is shown in Fig. 1. We detail each part in the following sections.

A. FEATURE EXTRACTION

Recent studies on image distortion and HVS have motivated our approach for feature extraction. Comparing NR-IQA to NR-VQA models, deep learning is widely used for the former task while only a few NR-VQA deep learning models have been proposed to date [43]. Moreover, improvements are often observed when combining the distortion and content-aware features for estimating NR-IQA [37], [44] and NR-VQA [73]. Thus, we take advantage of these studies and extract these two types of features in our study.

1) DISTORTION FEATURES

It is well known that image distortion strongly affects the final quality score. The more severe the image distortion is, the lower the quality score will be. Recently, it was revealed that deep neural network (DNN) features are distortion-sensitive [45], [46], and NR-IQA/VQA methods began to incorporate networks for predicting distortion in their model [37], [47], [73]. Additionally, it was shown that DNN layers of increasing depth learn features of growing complexity. Hence, features computed in the first layers typically resemble the output of Gabor filters or color blobs, while features in deeper layers correspond to semantic entities such as circular objects with a specific texture or even faces [48].

To implement our distortion prediction network, we leverage the high performance of CNNs trained on ImageNet [39], and choose the ResNet-50 architecture [49] as backbone network. We also adopt a transfer learning strategy and fix the parameters of the first layer of the pre-trained ResNet-50, training only the following layers to learn the type of distortion. This fine tuning strategy, which uses visual features from early layers accelerates the training and leads also to a better generalization.

We employ a self-supervised approach to learn a useful representation from low-cost labels in a pre-training phase. We train our distortion network with the most apparent distortion database (CSIQ) in contrast to the work of [73], where the authors used in-capture/authentic distortion data to train their model. We selected this image database because of its excellent results with Most Apparent Distortion (MAD) or what is most apparent to the human observer [50]. The distortions used in the CISQ image database are called *the most apparent distortions* and are JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink Gaussian noise, additive white Gaussian noise, and Gaussian blurring. The CSIQ image database contains 30 reference images distorted with six types of distortions, each at four to five different levels of distortion. Moreover, we tested our model with the KADID-10k [51] image database (which contains 10,125 distorted images grouped in 25 distortion types) and obtained almost the same performance as with the CSIQ image database. Actually, KADID-10k contains distortions selected from the TID2013 [52] image database and some new authentic distortions. However, as supported by studies from MLSP [53], pretrained CNNs on ImageNet are already robust enough to predict videos impacted by authentic distortions.

Our distortion network is trained to predict the type of distortion that was applied to an input image. As mentioned above, distorted images can be generated in large quantity and at almost no cost compared to having humans rate the quality of videos. Since the distortion in images affects their perceived quality, we use the representation learned in the self-supervised pre-training phase to boost the learning of the downstream NR-VQA task. Toward this goal, we truncate the distortion prediction network at the last convolutional layer and use the output of this last layer as additional features for

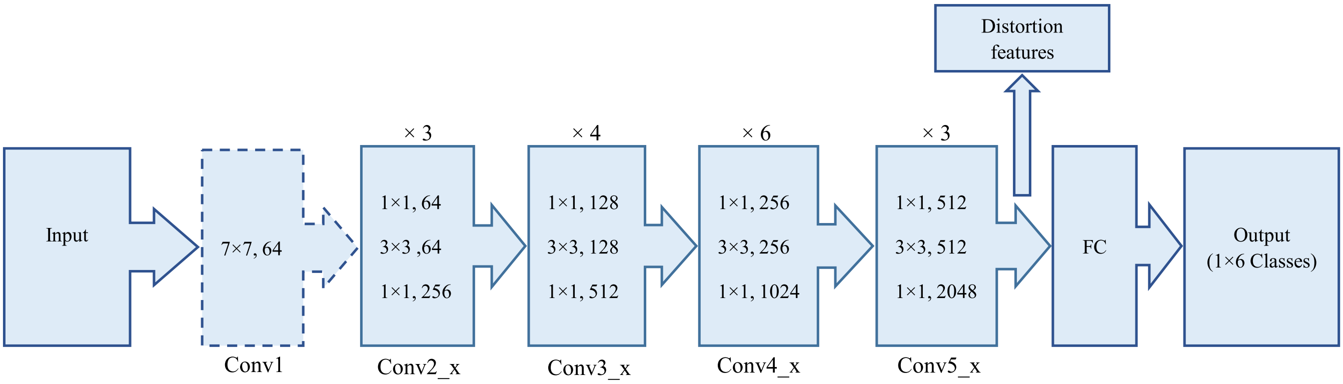

FIGURE 2. The architecture of the distortion feature extraction network with ResNet-50 as base model using for example a standard input image size. The conv1 is frozen and the blocks of conv2_x, conv3_x, conv4_x and conv5_x are trained for distortion prediction.

the NR-VQA network. Similar to [73], during training, we do not resize the input images to avoid introducing additional artifacts. Thus, our network is trained on images having the same resolution as those used for collecting the subjective quality in the CSIQ image database [50]. Fig. 2 shows the architecture of the distortion prediction network. Additional details are provided in Section IV-A2.

2) IMAGE CONTENT-AWARE FEATURES

Numerous studies have shown that the human judgment of visual video quality is content related. For example, it was found that two compressed images with the same compression ratio may have different subjective quality if they contain different scenes [54], [55]. Hence, it is important to take into consideration the content of images when designing the NR-VQA model. Furthermore, our ablation study also confirms the improvement with the content-aware network.

As in the distortion prediction network, we use the ResNet-50 network pre-trained on ImageNet as backbone for extracting our content-aware features.

For a given video containing T frames, we feed a frame I_t, $(t = 1, \ldots, T)$ simultaneously to the distortion prediction and content-aware CNN networks. The output features vectors f_t^{dist} and f_t^{cont} from each of these CNNs are obtained by truncating the networks to the last convolutional block (see Fig. 2) and applying spatial global average (mean) pooling. This generates a representation of size $m \times n \times 2048$, where 2048 is the number of feature maps and $m = n = 1$ after pooling:

$$f_t^{\text{dist}} = \text{GP}(\text{CNN}_{\text{dist}}(I_t)) \quad (1)$$

$$f_t^{\text{cont}} = \text{GP}(\text{CNN}_{\text{cont}}(I_t)) \quad (2)$$

Here, GP is the global average pooling and $t \in \{1, \ldots, T\}$. Finally, we concatenate the distortion features f_t^{dist} and the content-aware features f_t^{cont}. The result features f_t^{dc} is thus obtained as

$$f_t^{\text{dc}} = f_t^{\text{dist}} \oplus f_t^{\text{cont}} \quad (3)$$

where $\oplus$ is the concatenation operator and $t \in \{1, \ldots, T\}$.

B. MODELING TEMPORAL EFFECTS

Unlike in IQA, another important challenge in designing the VQA model is effectively modeling the temporal information of videos. In this study, we achieve this by implementing two separate techniques. First, we use GRU layers to capture long-term dependencies between frames in the video. Second, we employ a temporal attention mechanism to select the most relevant frames for predicting the overall video quality.

1) TEMPORAL MODELING

Recurrent Neural Networks (RNNs) have shown a great potential for tackling various sequences modeling tasks in machine learning [56]. In our study, we select the GRU model, which is a simplified version of the Long Short Term Memory (LSTM) model. Unlike LSTMs, a GRU merges the input and forget gates of LSTM and simplifies them with an update gate. Thus, GRUs have fewer parameters than LSTMs, which makes their training easier and lowers the computational requirements [57]. Like LSTMs, GRUs can alleviate the vanishing and exploding gradient problems of the traditional RNN model.

Since the features extracted and combined from the two CNNs networks (i.e., f_t^{dc}) are of high dimension, they cannot be used directly as input to the GRU network. To alleviate this problem, we perform a dimension reduction step using a linear (fully-connected) operation:

$$x_t = W_{xf} f_t^{\text{dc}} \quad (4)$$

where W_{xf} are the learned parameters of the linear model. After this dimensional reduction to a size of 128, the features x_t, $(t = 1, \ldots, T)$ are sent to the GRU. We consider the hidden states of the GRU as the integrated features, where the initial state is given by h_0 and the previous state by h_{t-1}. The current hidden state h_t is computed as

$$z_t = \sigma(W_{zx} x_t + U_{zh} h_{t-1}) \quad (5)$$

$$r_t = \sigma(W_{rx} x_t + U_{rh} h_{t-1}) \quad (6)$$

$$c_t = \tanh\left(W_{cx} x_t + U_{ch}(r_t \otimes h_{t-1})\right) \quad (7)$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes c_t \quad (8)$$

where σ is the sigmoid function and $\otimes$ is an element-wise multiplication operator. z_t, r_t and c_t are respectively the update gate, reset gate and candidate activation. W_{zx}, W_{rx}, W_{cx}, U_{zh}, U_{rh} and U_{ch} are the related weight matrices.

The GRU captures the temporal dependencies among the features extracted from each frame of the sequence. Actually, each GRU receives the features (x_t) as an input and outputs its hidden state. The GRU output could be seen as a selective memory of past hidden states and this underlines the long-term dependencies of the final output. Moreover, the spatial and temporal correlations are jointly learned through optimization. Thus, the quality of each frame is predicted by the spatio-temporal features. In our study, for simplicity in terms of computations and memory, we select the standard GRU with a single layer, i.e no stacked GRU. The hidden size, which is the amount of information stored, is set to 32.

Finally, the attention mechanism described in the next subsection is used through the video sequence to select important frames for the quality prediction.

2) ATTENTION MECHANISM AND POOLING

Soft attention [58] has been used with great success in various vision tasks such as image captioning and emotion classification [59]–[61]. Following this principle, we add a temporal attention mechanism to the GRU network, which estimates the importance of each frame for predicting the overall perceptual video quality. This is achieved by computing attention weights α_t that define how much each frame should be considered in the final output.

The predicted quality, q_t, for each frame t is then calculated as

$$q_t = \sigma(W_{qh} h_t + b_q) \in [0, 1]. \tag{9}$$

W_{qh} and b_q are respectively the weights and bias parameters and they are jointly learned with all the other components of the system.

Denoting as T the total number of frames, for each frame at instant t, the attention weights are computed as follows:

$$\alpha_t = \frac{\exp(q_t)}{\sum_{j=1}^{T} \exp(q_j)}. \tag{10}$$

A common problem with attention mechanisms is that they often focus on a limited set of temporal or spatial characteristics, which can make the model less robust in terms of generalization performance when the contents of the training and testing videos are quite different.

To improve our model's robustness, we combine temporal attention with a standard mean pooling strategy which considers all frames in the video. For the T frames in the video, the overall video quality Q is finally calculated as

$$Q = \beta \sum_{t=1}^{T} \alpha_t q_t + (1-\beta) \sum_{t=1}^{T} \frac{q_t}{T} \tag{11}$$

where $\beta \in [0, 1]$ determines the relative importance of temporal attention and mean pooling. In our experiments, we empirically set $\beta = 0.5$, giving equal importance to these two techniques.

IV. EXPERIMENTAL RESULTS

This section first describes the experimental setup, including the VQA databases and evaluation criteria used for evaluation as well as the implementation details of our method. To validate the superiority of the proposed method, we then conduct four experiments: comparison on individual databases, cross databases evaluation, ablation study and computational efficiency analysis.

A. EXPERIMENTAL SETUP

1) DATABASES

Five publicly available databases, namely CVD2014 [19], KoNViD-1k [17], LIVE-Qualcomm [20], LIVE VQA [21] and CSIQ video [22], are employed to validate the performance of the proposed method. We present the characteristics of each one in this subsection.

CVD2014 (Camera Video Database) [19] consists of 234 videos of resolutions 640 × 480 and 1280 × 720 recorded using 78 different cameras. Each video captures one of five different scenes and presents distortions related to the video acquisition process such as sharpness, luminance, saturation, and graininess. The length of the trimmed videos is 10–25 s with 11–31 fps. The realignment mean opinion scores (MOS) lay in the range of −6.50 to 93.38.

KoNViD-1k (Konstanz Natural Video Database) [17] contains 1200 natural videos with authentic distortions sampled according to six specific attributes from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) database [62]. The resulting database contains video sequences that are representative of a wide variety of contents and authentic distortions. The videos are 8 s at a resolution of 960 × 540 with 24/25/30 fps. The MOS ranges from 1.22 to 4.64.

LIVE-Qualcomm (LIVE-Qualcomm Mobile In-Capture Video Quality Database) [20] contains 208 videos of resolution 1920 × 1080 captured by 8 different smartphones. These videos have a length of 15 s with 30 fps and are affected by six common in-capture distortion categories which are color, exposure, focus, sharpness, stabilization and artifacts related distortions. The realignment MOS ranges from 16.56 to 73.64.

LIVE VQA (LIVE Video Quality Assessment Database) [21] contains 150 degraded videos distorted from 15 reference videos using four different distortion types which are H.264 compression, MPEG-2 compression and two transmission-based distortions defined as H.264 compressed bit streams transmitted over error-prone IP and wireless networks respectively. Additionally, these distortions are applied with different levels of severity. These videos are of resolution 768 × 432 with lengths between 8-10 s and 25/50 fps. The average differential MOS ranges from 30.94 to 81.16.

CSIQ video database [22] contains 216 videos with a resolution of 832 × 480. Each video has a duration of 10 s, and span a range of 24, 25, 30, 50, and 60 fps. In total, six types of distortions are used to build this database which consists of four compression-based distortions (H.264 compression, H.265 compression, Motion-JPEG compression, Wavelet-based compression using the Snow codec) [63] and two transmission-based distortions (H.264 videos subjected to simulated wireless transmission loss and additive white noise [50]). The average differential MOS ranges from 14.48 to 82.80.

The performance of our proposed model is evaluated on the datasets described above. Similar to the state-of-the-art methods, the performance of our model is evaluated in terms of SROCC (Spearman Rank Order Correlation Coefficient), PLCC (Pearson's Linear Correlation Coefficient), and RMSE (Root Mean Square Error) between the predicted scores and the mean opinion scores. As recommended by the Video Quality Expert Group (VQEG) [64] for adjusting scaling and non-linearity effects between subjective scores (or MOS) and objective scores, PLCC is calculated after performing a non-linear logistic fitting between subjective scores (s) and objective scores (o). The non-linear transform $f(o)$ used in this study is given by:

$$f(o) = \frac{\tau_1 - \tau_2}{1 + e^{-\frac{o-\tau_3}{\tau_4}}} + \tau_2. \tag{12}$$

The parameters τ_1 to τ_4 are fitting parameters initialized with $\tau_1 = s_{\max}$, $\tau_2 = s_{\min}$, $\tau_3 = \mu_o$, $\tau_4 = \sigma_o/4$, $s_{\min}$, $s_{\max}$ are the minimum and maximum subjective scores, and μ_o, σ_o are the mean and standard deviation of the objective scores.

2) IMPLEMENTATION DETAILS

Our model is implemented using the PyTorch framework [65] and comprises two training blocks: a feature extraction block and a temporal modeling block (see Fig. 1). The feature extraction block contains two CNN branches, the content-aware CNN and distortion prediction CNN. We use a ResNet-50 pre-trained on ImageNet as backbone for both CNNs. The content-aware features are extracted using the content-aware CNN. The distortion prediction network is trained with CSIQ image database to estimate the type of distortion, keeping the parameters of the first layer (conv1) as fixed. We use cross-entropy as loss function and the Adam optimizer [66], training the model for 200 epoch with a learning rate of 0.0001.

The temporal modeling block receives as inputs the concatenated features of the two CNNs (content-aware and distortion) and is trained to estimate the quality score of the entire video. Inside this temporal modeling block, a dimension reduction step is first performed using a linear projection. These reduced-size features are then fed to the GRU for estimating the quality score of each video frame. Finally, a temporal attention mechanism and a mean pooling are employed to aggregate the scores predicted for each frame. To train the temporal modeling block, we use an L1 loss between the aggregated score predicted for the video and the ground-truth score, and employ the Adam optimizer with an initial learning rate of 0.0001 and a batch size of 4.

The GRU in our model has a single layer and a hidden state size of 32. The ground-truth MOS are scaled in the range [0, 1] using the min–max scaling.

B. PERFORMANCE ON INDIVIDUAL DATABASE

To facilitate the analysis, we separate the results on the five databases into two groups based on the categories of distortions, i.e. authentic and traditional distortions databases. We used 80% of the data for training and the remaining 20% for testing.

For a fair comparison with the state-of-the-art methods selected in this study, following [38], we run 100 different random splits with 80% of the sequence for training and 20% for testing in each simulation. The same random splits were used to evaluate all tested methods and we followed the same procedures as used in our baselines for training our model.

1) AUTHENTIC VQA DATABASES

In authentic VQA databases, the distortions are introduced by the camera and the processing software during capture. These types of distortions include blurriness, insufficient color representation, over/under-exposure, focus, sharpness and stabilization related distortions. Three authentic distortion databases, namely CVD2014 [19], KoNViD-1k [17] and LIVE-Qualcomm [20], are selected in this part. We compared the performance of our proposed method with popular state-of-the-art methods: NIQE [11], BRISQUE [15], V-BLIINDS [32], HIGRADE [31], FRIQUEE [30], TLVQM [38] and VSFA [13]. Additionally, CNN-TLVQM [23], which is an improvement of TLVQM [38], recently reported top performance on the KoNViD-1k database. We included this method in all our comparisons tables. For other recently published methods such as RAPIQUE [24], PVQ [42] and CoINVQ [25], we just reported their performance on KoNViD-1k in this section as the authors did not report their performance on the other databases.

Also, to have a fair comparison, we re-simulated the VSFA method by using the 80:20 split adopted in our study. The TLVQM and the other state-of-the-art methods are already simulated using this split. We also re-simulated CNN-TLVQM on the three authentic databases using the authors' pre-trained CNN model and MATLAB R2020a [67]. The source code of MLSP [53] is not publicly available, thus we cannot reproduce their model with the 80:20 split adopted in our work. As in recent studies, we evaluated the performance of our proposed method in terms of SROCC, PLCC and RMSE. In Table 1, we report the mean performance (and standard deviation) of compared methods for each database, in terms of SROCC, PLCC and RMSE. We also computed the overall performance of tested methods following the strategy of the original VSFA paper [13], where the performances for each database are combined using a weighted average, and the weight of a database is proportional to its number

TABLE 1. Performance results on in-capture distortion databases. In each column, the best, and second-best values are respectively marked in boldface, and underlined. Note that * are performances taken from paper [38] and † from the methods' original papers. Other results were reproduced using the authors' code.

Method	Overall Performance			CVD2014		
	SROCC ↑	PLCC ↑	RMSE ↓	SROCC ↑	PLCC ↑	RMSE ↓
NIQE [11]*	0.39 (±0.07)	0.40 (±0.06)	4.24 (±1.88)	0.58 (±0.10)	0.61 (±0.09)	17.1 (±1.5)
BRISQUE [15]*	0.57 (±0.06)	0.58 (±0.06)	4.10 (±0.45)	0.63 (±0.10)	0.64 (±0.10)	16.9 (±2.2)
V-BLIINDS [32]*	0.62 (±0.06)	0.61 (±0.60)	3.72 (±0.46)	0.70 (±0.09)	0.71 (±0.09)	15.2 (±2.2)
HIGRADE [31]*	0.73 (±0.04)	0.72 (±0.04)	3.44 (±0.37)	0.74 (±0.06)	0.76 (±0.06)	14.2 (±1.5)
FRIQUEE [30]*	0.75 (±0.04)	0.76 (±0.04)	2.99 (±0.29)	0.82 (±0.05)	0.83 (±0.04)	12.0 (±1.2)
TLVQM [38]*	0.79 (±0.03)	0.79 (±0.03)	2.81 (±0.33)	0.83 (±0.04)	0.85 (±0.04)	11.3 (±1.3)
VSFA [13]	0.81 (±0.03)	0.82 (±0.03)	2.70 (±0.03)	0.88 (±0.03)	0.88 (±0.03)	10.3 (±1.2)
CNN-TLVQM [23]	**0.82** (±0.03)	**0.83** (±0.02)	**2.55** (±0.03)	0.86 (±0.04)	0.88 (±0.03)	10.3 (±1.1)
Proposed	0.81 (±0.03)	0.82 (±0.03)	2.56 (±0.03)	**0.89** (±0.03)	**0.90** (±0.03)	**9.4** (±1.4)

Method	KoNViD-1k			LIVE-Qualcomm		
	SROCC ↑	PLCC ↑	RMSE ↓	SROCC ↑	PLCC ↑	RMSE ↓
NIQE [11]*	0.34 (±0.05)	0.34 (±0.05)	0.61 (±0.03)	0.46 (±0.13)	0.48 (±0.12)	10.7 (±1.3)
BRISQUE [15]*	0.56 (±0.05)	0.57 (±0.04)	0.52 (±0.02)	0.55 (±0.10)	0.54 (±0.10)	10.3 (±0.9)
V-BLIINDS [32]*	0.61 (±0.04)	0.58 (±0.05)	0.53 (±0.03)	0.60 (±0.10)	0.67 (±0.09)	9.2 (±1.0)
HIGRADE [31]*	0.73 (±0.03)	0.72 (±0.03)	0.44 (±0.02)	0.68 (±0.08)	0.71 (±0.08)	8.6 (±1.1)
FRIQUEE [30]*	0.74 (±0.03)	0.74 (±0.03)	0.43 (±0.02)	0.74 (±0.07)	0.78 (±0.06)	7.6 (±0.8)
TLVQM [38]*	0.78 (±0.02)	0.77 (±0.02)	0.41 (±0.02)	0.78 (±0.07)	0.81 (±0.06)	7.1 (±1.0)
VSFA [13]	0.80 (±0.02)	0.81 (±0.02)	0.39 (±0.02)	0.77 (±0.07)	0.79 (±0.06)	7.5 (±0.9)
CNN-TLVQM [23]	**0.82** (±0.02)	**0.82** (±0.02)	**0.36** (±0.02)	**0.81** (±0.05)	**0.83** (±0.03)	**6.73** (±0.8)
RAPIQUE [24]†	0.80	**0.82**	**0.36**	–	–	–
PVQ [42]†	0.79	0.79	–	–	–	–
CoINVQ [25]†	0.76	0.77	0.41	–	–	–
Proposed	0.80 (±0.02)	0.80 (±0.02)	0.39 (±0.02)	0.78 (±0.07)	0.81 (±0.06)	7.4 (±0.9)

of videos. As can be seen, our proposed method achieves the second-best overall performance in terms of prediction correlation and accuracy (SROCC, PLCC and RMSE), not far behind CNN-TLVQM, which yields the best overall performance, and with comparable performance (SROCC and PLCC) as VSFA. Although our method and CNN-TLVQM present a similar RMSE, CNN-TLVQM gives slightly higher SROCC and PLCC than our approach. Moreover, on the KoNViD-1k database, RAPIQUE has a performance comparable to our method in terms of SROCC, while presenting better PLCC and RMSE.

In summary, from Table 1, our method achieves SROCC, PLCC, and RMSE results placing it among the top three compared methods. Thus our approach is competitive with the state-of-the-art methods on authentic distortion databases.

2) TRADITIONAL VQA DATABASES

The distortions found in the traditional VQA databases are introduced during a compression or transmission process, which is why they are also called post-capture distortion databases. In this study, we have selected CSIQ video [22] and LIVE VQA [21] as post-capture databases. We used the same procedure as with the authentic databases to perform the evaluation, i.e. running 100 different random splits with a 80:20 ratio between training and test examples. For comparison, we selected the state-of-the-art methods found in the literature that performed well on traditional VQA databases such as V-BLINDS [32], SACONVA [34], V-MEON [36], VIIDEO [12] and NIQE [11].

Since VSFA, TLVQM, CNN-TLVQM, and RAPIQUE were not tested on traditional VQA databases, to have a fair comparison, we simulated these methods on the selected traditional distortion databases. Following the main studies on traditional VQA databases, we calculated the median of the 100 performance values obtained for VSFA, TLVQM, CNN-TLVQM, RAPIQUE and our proposed method.

As the source code for SACONVA is not publicly available, we only reported the results from their article. Table 2 gives the performance of tested methods in terms of median SROCC and PLCC.

We observe that our method outperforms all other approaches by a large margin, on both the CSIQ video and LIVE VQA databases. For CSIQ video, our method gives SROCC and PLCC improvements of 0.05 and 0.04, respectively, compared to the second best approach SACONVA. Likewise, for LIVE VQA, it achieves a boost of 0.03 in SROCC and PLCC with to the second-ranked approach which is again SACONVA. As expected, approaches for this type of databases, in particular SACONVA,

TABLE 2. Performance results on traditional distortion databases: CSIQ VIDEO and LIVE VQA. Note that * are performances taken from paper [18], [34] and † from the methods' original papers. Other results were reproduced using the authors' code.

Method	CSIQ VIDEO		LIVE VQA	
	SROCC ↑	PLCC ↑	SROCC ↑	PLCC ↑
NIQE [11]*	–	–	0.23	0.27
VIIDEO [12]*	-	-	0.62	0.65
V-BLIINDS [32]*	0.86	0.85	0.83	0.84
SACONVA [34]†	0.86	0.87	0.86	0.87
TLVQM [38]	0.74	0.74	0.60	0.62
VSFA [13]	0.76	0.74	0.73	0.75
V-MEON [36]†	0.82	0.82	–	–
CNN-TLVQM [23]	0.81	0.79	0.77	0.79
RAPIQUE [24]	0.76	0.77	0.66	0.72
Proposed	**0.91**	**0.91**	**0.89**	**0.90**

TABLE 3. SROCC results for cross-databases training and testing.

Testing Dataset	Method	Training Dataset		
		KoNViD-1k	LIVE-Qualcomm	CVD2014
CVD2014	TLVQM	0.47	0.55	–
	CNN-TLVQM	0.68	0.26	–
	VSFA	0.69	0.55	–
	Proposed	**0.74**	**0.61**	–
LIVE-Qualcomm	TLVQM	0.47	–	0.43
	CNN-TLVQM	0.61	–	**0.44**
	VSFA	**0.64**	–	0.40
	Proposed	0.61	–	0.36
KoNViD-1k	TLVQM	–	0.53	0.40
	CNN-TLVQM	–	0.11	0.58
	VSFA	–	**0.66**	0.59
	Proposed	–	**0.66**	**0.64**

V-BLINDS and V-MEON, perform better than VSFA, TLVQM, CNN-TLVQM, and RAPIQUE. We also observe a good performance improvement for CNN-TLVQM compared to TLVQM on the traditional distortion databases, while the performance of RAPIQUE is not satisfactory.

Unlike these approaches, our method performs very well on traditional and authentic VQA databases largely due to its self-supervised representation learning step based on distortion prediction. Specifically, for traditional distortions, our proposed method largely outperforms other state-of-the-art approaches.

C. PERFORMANCE ACROSS DATABASES

An important challenge for NR-VQA models based on deep learning is generalizing to data with different characteristics than the training database. Thus, we evaluated the generalization performance of our method in a cross-database scenario using training and testing databases with different contents and types of distortions. For each training database, we took the trained models and used them to estimate the quality scores of the videos from the other databases. We evaluated the cross-database results in terms of SROCC and reported the best performance value obtained for each test database. The performance of our method in this scenario is compared with that of VSFA, TLVQM and CNN-TLVQM.

Table 3 reports the performance of the four tested methods in terms of SROCC, when trained on an authentic distortion database and tested on the remaining two ones. The proposed method obtains the best performance in four of the six training-testing scenarios of this table. Moreover, our method obtains good generalization performance (SROCC) when trained on the KoNViD-1k or the LIVE-Qualcomm database. From Tables 1 and 3, although CNN-TLVQM performs slightly better than our proposed model on the KoNViD-1k and LIVE-Qualcomm databases, our model generalizes better. We believe that this is because it learns all features while CNN-TLVQM also uses hand-crafted features which may not be optimal for all settings. Also described in the literature, KoNViD-1k comprises natural videos with a wide diversity of contents while LIVE-Qualcomm contains videos with rich scenes. Hence, these results illustrate the robustness of our method to training with data having very different characteristics.

Although not presented in Table 3, we also evaluated the generalization ability of tested methods on the CSIQ video and LIVE VQA databases. However we observed a low performance for those scenarios. For example, when our method is trained on the CSIQ video database and tested on LIVE-VQA, it obtains a SROCC of 0.30.

Similarly, we observed a poor generalization performance when the models are trained on authentic VQA database and tested on traditional/post-capture VQA database (and vice versa). Thus, as concluded by some previous studies [38], learning-based VQA models perform poorly when the distortion type in the testing database is almost absent in the training database.

Finally, we compared the generalization performance of our proposed method with some deep learning NR-IQA models. Actually, authors of [68] have tested the deep learning NR-IQA models such as WaDIQaM [69] and SPAQ [70] on VQA databases (KoNViD-1k, LIVE-Qualcomm, and CVD2014) and concluded that their overall performance was not satisfactory due to temporal information being discarded. Furthermore, this article shows that NR-VQA models such as VSFA and CNN-TLVQM present better generalization performance than those deep learning NR-IQA models.

D. ABLATION STUDY

In this section, we analyze the impact on performance of the different components of our model. Firstly, we evaluate the importance of the content-aware network on the proposed model. Secondly, we evaluate the benefit of the proposed strategy for aggregating the quality scores of individual images into a global video score, based on temporal attention and mean pooling. Toward this goal, we compared this strategy against three pooling mechanisms found in the literature. Finally, we compare the performance of our GRU-based method with two competitive temporal memory networks, RNN and LSTM. To avoid bias in the results,

TABLE 4. SROCC and PLCC results for our proposed model vs the same model but without the context-aware network.

Method	KoNViD-1k		LIVE-VQA		CSIQ VIDEO	
	SROCC↑	PLCC↑	SROCC↑	PLCC↑	SROCC↑	PLCC↑
Without context-aware network	0.79	0.80	0.84	0.85	0.89	0.89
Proposed	**0.80**	**0.80**	**0.86**	**0.86**	**0.91**	**0.91**

TABLE 5. SROCC and RMSE results for different pooling methods.

Pooling	KoNViD-1k		CSIQ Video	
	SROCC	RMSE	SROCC	RMSE
Min [71]	0.77 (±0.02)	0.41 (±0.02)	0.81 (±0.06)	9.98 (±1.74)
Recency [72]	0.79 (±0.02)	0.39 (±0.02)	0.88 (±0.04)	7.93 (±0.89)
TP (VSFA) [13]	**0.80** (±0.02)	**0.39** (±0.03)	0.89 (±0.04)	7.76 (±1.33)
Proposed	**0.80** (±0.02)	0.39 (±0.03)	**0.91** (±0.02)	**6.93** (±0.70)

TABLE 6. SROCC and RMSE results for different temporal networks.

Network	KoNViD-1k		CSIQ Video	
	SROCC	RMSE	SROCC	RMSE
RNN	0.80 (±0.02)	0.39 (±0.02)	0.91 (±0.03)	7.15 (±0.98)
LSTM	0.80 (±0.02)	0.39 (±0.02)	0.90 (±0.03)	7.35 (±0.89)
Proposed	**0.80** (±0.02)	0.39 (±0.03)	**0.91** (±0.02)	**6.93** (±0.70)

we selected 20 random splits and reused them for all test scenarios. As our study covers both traditional and authentic distortion databases, our ablation study is conducted using both authentic and traditional databases.

1) STUDY OF CONTENT-AWARE NETWORK

In Table 4, we performed an ablation study by using only the distortion network (i.e. the proposed method without context-aware network) for predicting the video quality.

The results show an improvement when adding the context-aware network. In this study, this improvement is more notable for LIVE-VQA and CSIQ VIDEO databases (traditional distortion databases).

2) STUDY OF POOLING METHODS

We compared the performance of our pooling strategy, which combines temporal attention with mean pooling, against two other pooling techniques that reflect the human judgements of quality [71] and the temporal pooling (TP) designed by authors of VSFA [13]:

- Min pooling, which selects the minimal score across the different frames of a video. This strategy is based on the idea that users rate the overall video quality based on the worst degradation.
- Recency pooling, which is based on the temporal hysteresis effect where users remember poor quality frames in the past and lower the perceived quality scores for the following frames, even when the frame quality has returned to acceptable levels. We fine-tuned the recency parameters for KoNViD-1k database by setting the frame rate to 30 fps, and the memory intensity effect parameter α_r to 0.01. More details can be found in [72].

In Table 5, we report the average (and standard deviation) SROCC and RMSE obtained for the compared pooling methods, when applied on CSIQ video and KoNViD-1k databases. We see that our pooling strategy achieves the highest SROCC and RMSE on both databases. Notable improvements are especially observed for the CSIQ video database, where our method achieves a 0.02 higher SROCC and a 0.83 lower RMSE than the second best method (temporal pooling of VSFA). The better performances of temporal and recency pooling compared to min pooling confirms the benefit of considering human behaviour in the pooling strategy (i.e., temporal hysteresis effect). However, as shown by its better generalization performance, our model's learned attention is more robust to cross-database differences in terms of content and distortions than the pooling strategy of VSFA, which relies on hand-tuned hyper-parameters.

3) STUDY OF TEMPORAL NETWORK

Next, we evaluate our temporal network by replacing the proposed GRU by a basic RNN or a LSTM. The SROCC and RMSE of our method using different temporal network, on the KoNViD-1k and CSIQ video databases, is reported in Table 6. We find that RNN and LSTM yield a similar performance to our method on the KoNViD-1k database, which can be due to the the short and fixed length (8 seconds) of videos in this database. However, the proposed GRU-based network achieves a higher SROCC and considerably lower RMSE than other approaches on the CSIQ video database (5.7% lower RMSE than the second best approach, LSTM).

E. COMPUTATIONAL EFFICIENCY

Another important consideration when designing NR-VQA methods is the computational efficiency. Some existing NR-VQA approaches offer suitable performance, however they cannot be used in real-life applications due to their high computational complexity. To complete our study, we evaluate the computation performance of our proposed method. We selected twenty representative video sequences from the CVD2014 database, ten with low resolution (640 × 480) and ten with high resolution (1280 × 720). The length of the sequences varies approximately from ten to twenty seconds. We compare the computation times of our proposed method with those of the state-of-the-art baselines selected in this study (VSFA, TLVQM), and with CNN-TLVQM and RAPIQUE. The simulations for all three methods were performed on a desktop computer with NVIDIA Quadro RTX8000 with 4608 CUDA cores.

VSFA and our method were implemented in Python and exploit the PyTorch framework, while the CNN-TLVQM implementation uses only MATLAB and the two other (TLVQM and RAPIQUE) implementations use both

TABLE 7. Average computation times on example videos.

Method	Low Resolution 640×480	High Resolution 1280×720
TLVQM [38]	42.8 sec	112.2 sec
CNN-TLVQM [23]	39.9 sec	95.5 sec
VSFA [13]	8.8 sec	16.4 sec
RAPIQUE [24] (1 fr/sec)	-	**4** sec
Proposed (ResNet-101)	12.1 sec	30.6 sec
Proposed (ResNet-50)	9.6 sec	22.3 sec

MATLAB (for feature extraction) and Python (for regression).

In Table 7, we report the average computation times for TLVQM, VSFA CNN-TLVQM, RAPIQUE and our methods. As shown, our method (based ResNet-50) is about 3 times faster than CNN-TLVQM, which itself is faster than TLVQM. Additionally, our method has average computation times relatively close to VSFA for low-resolution videos. It is however slower than VSFA for high-resolution videos, which require almost 26.5% less runtime compared to our method. This is due to the fact that our method has to compute features in two different CNN branches (distortion prediction network and context-aware network), whereas VSFA only has a single feature extraction branch. Nevertheless, this runtime difference could be reduced by performing in parallel the computations in both branches, at the cost of additional hardware, or by having the two network branches share some of their layers. Compared to VSFA, RAPIQUE shows significant better computation times (see Table 7). However, as can be seen in Table 2, VSFA and RAPIQUE do not perform well on traditional distortion databases. Moreover, our method generalizes better than VSFA method (see Table 3).

V. CONCLUSION

In this paper, we have proposed an objective NR-VQA method for videos affected by both authentic and traditional distortions. The main contributions of our method are three-fold: first, a deep learning based on multi-task learning approach where the distortion of individual frames in a video and the overall quality of the video are predicted by a single neural network; second, a special pooling designed with temporal attention mechanism and average pooling for respectively selecting the frames of a video which contribute the most to its perceived quality and to account for all uncertainties; third, a distortion network designed and used as a complementary network to improve the quality prediction for some VQA databases.

Experiments on five different databases containing videos with authentic and traditional distortions demonstrate the effectiveness of our proposed method in highly-different settings. While state-of-art NR-VQA approaches such as TLVQM, CNN-TLVQM, RAPIQUE and VSFA only perform well on videos with authentic distortion, and give unsatisfactory results on videos with traditional distortion, our method provides competitive performance in both these settings. As a deep learning model, it also has a reasonably good computational complexity.

Despite these promising results, the performance of our model on authentic databases could still be improved. While we did not take into consideration video motion features in this study, which could further boost performance on databases such as LIVE-Qualcomm. In future work, we plan to investigate DNN models which can efficiently extract motion features from videos.

REFERENCES

[1] J. Klink and T. Uhl, "Video quality assessment: Some remarks on selected objective metrics," in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2020, pp. 1–6.

[2] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," 2016, *arXiv:1611.01704*.

[3] M. Ben-Ezra, A. Zomet, and S. K. Nayar, "Video super-resolution using controlled subpixel detector shifts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 977–987, Jun. 2005.

[4] C. Liu and W. T. Freeman, "A high-quality video denoising algorithm based on reliable motion estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Crete, Greece, vol. 6313, Sep. 2010, pp. 706–719.

[5] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.

[6] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2256–2270, Aug. 2019.

[7] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C.-J. Kuo, "A fusion-based video quality assessment (FVQA) index," in *Proc. Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA), Asia–Pacific*, Dec. 2014, pp. 1–5.

[8] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2480–2492, Jun. 2016.

[9] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2012.

[10] K. Zeng and Z. Wang, "Temporal motion smoothness measurement for reduced-reference video quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 1010–1013.

[11] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2012.

[12] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity Oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Nov. 2015.

[13] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2351–2359.

[14] M. A. Saad and A. C. Bovik, "Blind quality assessment of videos using a model of natural scene statistics and motion coherency," in *Proc. Conf. Rec. 46th Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Nov. 2012, pp. 332–336.

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

[16] D. Ghadiyaram, C. Chen, S. Inguva, and A. Kokaram, "A no-reference video quality predictor for compression and scaling artifacts," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3445–3449.

[17] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, and D. Saupe, "The Konstanz natural video database (KoNViD-1k)," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.

[18] S. V. R. Dendi and S. S. Channappayya, "No-reference video quality assessment using natural spatiotemporal scene statistics," *IEEE Trans. Image Process.*, vol. 29, pp. 5612–5624, 2020.

[19] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Hakkinen, "CVD2014—A database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, Jul. 2016.

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

[21] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, ''Study of subjective and objective quality assessment of video,'' *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[22] Laboratory of Computational Perception & Image Quality, Oklahoma State University. (2013). *CSIQ Video Database*. [Online]. Available: http://vision.okstate.edu/csiq/

[23] J. Korhonen, Y. Su, and J. You, ''Blind natural video quality prediction via statistical temporal features and deep spatial features,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3311–3319.

[24] Z. Tu, C.-J. Chen, Y. Wang, N. Birkbeck, B. Adsumilli, and A. Bovik, ''Efficient user-generated video quality prediction,'' in *Proc. Pict. Coding Symp. (PCS)*, Jun. 2021, pp. 1–5.

[25] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, ''Rich features for perceptual quality assessment of UGC videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13430–13439.

[26] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, ''On advances in statistical modeling of natural images,'' *J. Math. Imag. Vis.*, vol. 18, no. 1, pp. 17–33, 2003.

[27] E. P. Simoncelli and B. A. Olshausen, ''Natural image statistics and neural representation,'' *Annu. Rev. Neurosci.*, vol. 24, p. 1193—1216, Mar. 2001.

[28] A. K. Moorthy and A. C. Bovik, ''Blind image quality assessment: From natural scene statistics to perceptual quality,'' *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[29] P. Ye, J. Kumar, L. Kang, and D. Doermann, ''Unsupervised feature learning framework for no-reference image quality assessment,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.

[30] D. Ghadiyaram and A. C. Bovik, ''Perceptual quality prediction on authentically distorted images using a bag of features approach,'' *J. Vis.*, vol. 17, no. 1, p. 32, 2016.

[31] D. Kundu, D. Ghadiyaram, A. C. Bovik, and B. L. Evans, ''No-reference quality assessment of tone-mapped HDR pictures,'' *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2957–2971, Jun. 2017.

[32] M. Saad, A. Bovik, and C. Charrier, ''Blind prediction of natural video quality,'' *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[33] J. Xu, P. Ye, Y. Liu, and D. Doermann, ''No-reference video quality assessment via feature learning,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 491–495.

[34] Y. Li, L. Po, C. Cheung, X. Xu, L. Feng, F. Yuan, and K. Cheung, ''No-reference video quality assessment with 3D shearlet transform and convolutional neural networks,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1044–1057, Jun. 2016.

[35] C. Wang, L. Su, and W. Zhang, ''COME for no-reference video quality assessment,'' in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 232–237.

[36] W. Liu, Z. Duanmu, and Z. Wang, ''End-to-end blind quality assessment of compressed videos using deep neural networks,'' in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 546–554.

[37] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, ''End-to-end blind image quality assessment using deep neural networks,'' *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[38] J. Korhonen, ''Two-level approach for no-reference consumer video quality assessment,'' *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, Dec. 2019.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ''ImageNet: A large-scale hierarchical image database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] K. Seshadrinathan and A. C. Bovik, ''Temporal hysteresis model of time varying subjective video quality,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 1153–1156.

[41] L. Jing and Y. Tian, ''Self-supervised visual feature learning with deep neural networks: A survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[42] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, ''Patch-VQ: 'Patching up' the video quality problem,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14014–14024.

[43] D. Varga, ''No-reference video quality assessment based on the temporal pooling of deep features,'' *Neural Process. Lett.*, vol. 50, no. 3, pp. 2595–2608, Dec. 2019.

[44] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, ''Blind image quality assessment using a deep bilinear convolutional neural network,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[45] S. Dodge and L. Karam, ''Understanding how image quality affects deep neural networks,'' in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.

[46] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, ''The unreasonable effectiveness of deep features as a perceptual metric,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[47] L. He, Y. Zhong, W. Lu, and X. Gao, ''A visual residual perception optimized network for blind image quality assessment,'' *IEEE Access*, vol. 7, pp. 176087–176098, 2019.

[48] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, ''How transferable are features in deep neural networks?'' 2014, *arXiv:1411.1792*.

[49] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] D. M. Chandler, ''Most apparent distortion: Full-reference image quality assessment and the role of strategy,'' *J. Electron. Imag.*, vol. 19, no. 1, Jan. 2010, Art. no. 011006.

[51] H. Lin, V. Hosu, and D. Saupe, ''KADID-10k: A large-scale artificially distorted IQA database,'' in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[52] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. K. Kuo, ''Image database TID2013: Peculiarities, results and perspectives,'' *Signal Process., Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[53] F. Gotz-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. 72139–72160, 2021.

[54] S. Triantaphillidou, E. Allen, and R. Jacobson, ''Image quality comparison between JPEG and JPEG2000. II. scene dependency, scene analysis, and classification,'' *J. Imag. Sci. Technol.*, vol. 51, no. 3, pp. 259–270, May 2007.

[55] E. Siahaan, A. Hanjalic, and J. A. Redi, ''Semantic-aware blind image quality assessment,'' *Signal Process., Image Commun.*, vol. 60, pp. 237–252, Feb. 2018.

[56] A. Graves, A.-R. Mohamed, and G. Hinton, ''Speech recognition with deep recurrent neural networks,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.

[57] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, ''On the properties of neural machine translation: Encoder-decoder approaches,'' 2014, *arXiv:1409.1259*.

[58] D. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2014, *arXiv:1409.0473*.

[59] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' 2015, *arXiv:1502.03044*.

[60] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, ''Residual attention network for image classification,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.

[61] M.-C. Sun, S.-H. Hsu, M.-C. Yang, and J.-H. Chien, ''Context-aware cascade attention-based RNN for video emotion recognition,'' in *Proc. 1st Asian Conf. Affect. Comput. Intell. Interact. (ACII Asia)*, May 2018, pp. 1–6.

[62] B. Thomee, D. Shamma, G. Friedland, B. Elizalde, K. S. Ni, D. N. Poland, D. Borth, and L. Li, ''YFCC100M: The new data in multimedia research,'' *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[63] V. Bajcinovci, M. Vranjes, D. Babic, and B. Kovacevic, ''Subjective and objective quality assessment of MPEG-2, H.264 and H.265 videos,'' in *Proc. Int. Symp. (ELMAR)*, Sep. 2017, pp. 73–77.

[64] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*, VQEG, San Jose, CA, USA, 2000.

[65] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, ''Automatic differentiation in PyTorch,'' in *Proc. NIPS Autodiff Workshop*, 2017, pp. 1–4.

[66] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[67] J. Korhonen, Y. Su, J. You, S. Hicks, and C. Midoglu, ''Reproducibility companion paper: Blind natural video quality prediction via statistical temporal features and deep spatial features,'' in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3622–3626.

[68] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, ''Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1903–1916, Apr. 2022, doi: 10.1109/TCSVT.2021.3088505.

[69] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, ''Deep neural networks for no-reference and full-reference image quality assessment,'' *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[70] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, ''Perceptual quality assessment of smartphone photography,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3674–3683.

[71] S. Rimac-Drlje, M. Vranjes, and D. Zagar, ''Influence of temporal pooling method on the objective video quality evaluation,'' in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, May 2009, pp. 1–5.

[72] 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 *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 141–145.

[73] M. Agarla, L. Celona, and R. Schettini, ''An efficient method for no-reference video quality assessment,'' *J. Imag.*, vol. 7, no. 3, p. 55, Mar. 2021.

KOFFI KOSSI received the degrees both in applied physics and telecommunications engineering from Cadi Ayyad University, Morocco, where he attended classes as an international student, and the M.Eng. degree in software and IT engineering from the École de technologie supérieure (ÉTS), Université du Québec, Montréal, in 2013, where he currently pursuing the Ph.D. degree.

From 2007 to 2017, he worked as a Consultant specialized in audio-video systems and performance, and also, as a part-time University Lecturer for two engineering schools. His research interests include audio–video QoS/QoE, machine learning, deep learning, human visual perception for video communications, and reinforcement learning.

STÉPHANE COULOMBE (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from École Polytechnique de Montréal, Canada, in 1991, and the Ph.D. degree in telecommunications (image processing) from INRS-Telecommunications, Montreal, in 1996.

From 1997 to 1999, he was with the Nortel Wireless Network Group, Montreal. From 1999 to 2004, he worked as a Research Engineer at the Nokia Research Center, Dallas, TX, USA, and as a Program Manager at the Audiovisual Systems Laboratory. He joined the École de technologie supérieure (ÉTS is a constituent of the Université du Québec network), in 2004, where he is currently a Professor with the Department of Software and IT Engineering. His research interests include research and development on video processing and systems, compression, and transcoding.

From 2009 to 2018, he has held the Vantrix Industrial Research Chair in video optimization.

CHRISTIAN DESROSIERS received the Ph.D. degree in computer engineering from Polytechnique Montreal, in 2008.

He was a Postdoctoral Researcher with the University of Minnesota, where he focused on the topic of machine learning. In 2009, he joined as a Professor with the Department of Software and IT Engineering, ÉTS, University of Québec. He is the Co-Director of the Laboratoire d'imagerie, de vision et d'intelligence artificielle, and a member of the REPARTI Research Network. His main research interests include machine learning, image processing, computer vision, and medical imaging.

GHYSLAIN GAGNON (Member, IEEE) received the Ph.D. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2008.

He is currently a Full Professor and the Dean of Research with the École de technologie supérieure, Université du Québec, Montreal, Canada. His research interests include CMOS IC design, digital signal processing, and machine learning with various applications. He is highly inclined toward collaborative research with industry and was a recipient of the 2020 Partnership and Innovation Award from ADRIQ. From 2013 to 2020, he was the Director of the LACIME Research Laboratory—a group of 15 professors and 150 highly dedicated students and researchers in microelectronics, digital signal processing, and wireless communications.