IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 August 2024, accepted 18 September 2024, date of publication 26 September 2024, date of current version 7 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3468336

RESEARCH ARTICLE

No-Reference Video Quality Assessment Using Transformers and Attention Recurrent Networks

KOFFI KOSSI[1,3], STÉPHANE COULOMBE[2,3], (Senior Member, IEEE), AND CHRISTIAN DESROSIERS[2,3], (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
[3]International Laboratory on Learning Systems (ILLS), McGill–ÉTS–Mila–CNRS–CentraleSupelec, Université Paris Saclay, Montreal, QC H3H 2T2, Canada

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

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT In recent years, numerous studies have investigated the development of methods for video quality assessment (VQA). These studies have predominantly focused on specific types of video degradation tailored to the application of interest. However, natural videos or recent videos generated by users (UGC) present complex distortions that are not easy to model. Consequently, most current VQA approaches struggle to achieve high performance when applied to these videos. In this paper, we propose a novel Transformer-based architecture that extracts spatial distortion features and spatio-temporal features from videos in two specialized branches. The spatial distortion branch leverages a transfer learning strategy where a standard ViT is pre-trained using a masked autoencoder (MAE) self-supervised learning task, and then fine-tuned to predict the distortion type of corrupted images from the CSIQ database. The features from this branch capture degradation at the level of individual frames. On the other hand, the second branch employs a 3D Shifted Windows Transformer (Swin-T) to extract spatio-temporal features across multiple frames. Once again, we use transfer learning to extract rich features by pre-training this 3D Swin-T model on a video dataset for human action recognition. Finally, a temporal memory block hinged on an attention recurrent neural networks is proposed to predict the final video quality score from the spatio-temporal sequence of features. We evaluate the performance of our method on two popular UGC databases, namely KoNViD-1k and LIVE-VQC. Results show it outperforms state-of-the-art models on the KoNViD-1k database, achieving a SROCC performance of 0.927 and a PLCC of 0.925, while also delivering highly competitive results on the LIVE-VQC database.

INDEX TERMS CNN, distortion, Swin-Transformers, UGC, video quality assessment, vision Transformers, ViT, NLP.

I. INTRODUCTION

As predicted by Cisco, video consumption has rapidly increased in the last years [1]. The recent rise of flex office and telework has further accelerated this trend, especially in the consumption of user-generated content (UGC) videos. This type of video, which can be used for employee training, skill acquisition, and providing leisure activities to name a few, has become prominent in today's video landscape, with individuals capturing and streaming their own content on popular social media platforms such as YouTube, Facebook, TikTok, and Twitter [2].

For video and telecommunication service providers, the billions of videos flowing over the Internet and their infrastructure need to be monitored and analyzed in order to increase customer satisfaction and maximize profits. However, an important problem with UGC videos is that they do not have any *pristine* reference available. Furthermore, such videos are captured using a variety of cameras and smartphones, each introducing its own set of combined distortions, such as defocus blur, color saturation, and noise. Given their diverse content and the presence of complex

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

distortions, predicting the quality of UCG videos becomes increasingly challenging.

In recent years, No-Reference Video Quality Assessment (NR-VQA) has emerged as a field of intense research, with computer vision researchers predominantly relying on deep learning algorithms for their models. Convolutional Neural Networks (CNNs) have so far stood as the dominant deep learning architecture for this task. However the recent success of Transformers in Natural Language Processing (NLP) has prompted researchers to explore their application in NR-VQA. To enhance the performance of NR-VQA models, researchers have begun integrating the key element of Transformers, namely the self-attention mechanism, into CNNs [3], or even replacing the CNN backbone entirely with Transformers [4].

In this paper, we introduce a novel NR-VQA method based on Transformers to predict the quality of UGC videos. The main contributions of this work are as follows:

- We propose a two-branch network for video quality assessment (VQA) exploiting specialized Transformer architectures and transfer learning to extract features modeling distortion within individual frames and spatial-temporal information across multiple frames. The first branch, which is based on a standard Visual Transformer (ViT), is pre-trained from IQA data containing diverse types of spatial distortions to learn a robust representation of these distortions. In contrast, the second branch leverages a 3D-Swin Transformer, pre-trained on a human action recognition dataset, to capture spatio-temporal features at different scales. An attention Gated Recurrent Unit (GRU)-based temporal-memory block, taking both types of features as input, is also proposed to capture long-range dependencies influencing the final perceptual quality score.
- Our work is among the first to study and compare the performance of distortion networks based on CNNs and ViTs in the context of video quality prediction.

To demonstrate the superiority of our method, we conduct experiments on two popular and publicly available video databases namely KoNViD-1k [5] and LIVE VQC [6] which contain respectively natural videos and videos with complex in-capture distortions. Our method outperforms state-of-the-art (SOTA) models on the KoNViD-1k database while delivering competitive results on the LIVE-VQC database.

The rest of our paper is organized as follows. In Section II, we present the related works. We then describe our proposed NR-VQA method in section III. Afterwards, we present and discuss our experimental results in section IV. Finally, in Section V, we conclude and suggest some future works.

II. RELATED WORKS

In this section, we organize NR-VQA approaches from the literature into two groups based on whether deep learning techniques are used to extract features. We provide an overview of relevant studies within each group.

A. NON-DEEP LEARNING METHODS

Non-deep learning methods do not automatically extract features from images or sequences of frames in videos. The most dominant ones are based on the theory of natural scene statistics (NSS) and were originally developed for image quality assessment (IQA). Later, these methods were modified for use in VQA by combining the quality scores of individual frames to generate an overall quality rating for the entire video. Given the multitude of extracted features, machine learning algorithms are commonly employed to predict frame quality based on these handcrafted features. Examples of these techniques include NIQE [7], BRISQUE [8], CORNIA [9], V-BLIINDS [10], VIIDEO [11], SACONVA [12] and others [13], [14]. For instance, in V-CORNIA [15], the authors utilized the Support Vector Regression (SVR) algorithm to predict frame-level quality scores. They subsequently aggregated these scores using temporal pooling to derive the overall quality score for the entire video. Recent popular handcrafted methods include ChipQA [16], TLVQM [17] and VIDEVAL [18]. The TLVQM method extracts low complexity and high complexity features such as spatial activity, exposure or sharpness by handcrafted means and predict the video quality for each video by applying SVR and Random Forest (RF) algorithms on these features. In the same way, VIDEVAL ensembles different handcrafted features to model the diverse authentic distortions for predicting the UGC video quality. However, the distortions typically found in natural videos are quite complex and cannot be well captured with handcrafted methods.

B. DEEP LEARNING METHODS

In the past decade, deep learning has revolutionized various scientific fields and most of recent NR-VQAs methods are based on this powerful approach. Considering the latest developments in computer vision, we can classify these deep learning-based methods into two categories depending on the main algorithm used when automatically extracting the features in the video: CNN-based and Transformers-based methods.

1) CNN-BASED METHODS

With the outstanding performance of CNNs in computer vision, researchers have adopted them as the main backbone for their models. Compared to previous approaches, CNNs offer an automated way to extract image features and generally outperform classical or handcrafted methods. For example, TLVQM, which was primarily designed as a handcrafted method, obtained a significant performance gain when their authors replaced manually extracted features by automatic extraction performed with CNNs [19]. In the realm of VQA models that leverage CNN, we can mention VSFA [20], RAPIQUE [21], HEKE [22], PVQ [23], MLSP [24], RIRNet [25], and others [26], [27], [28], [29], [30].

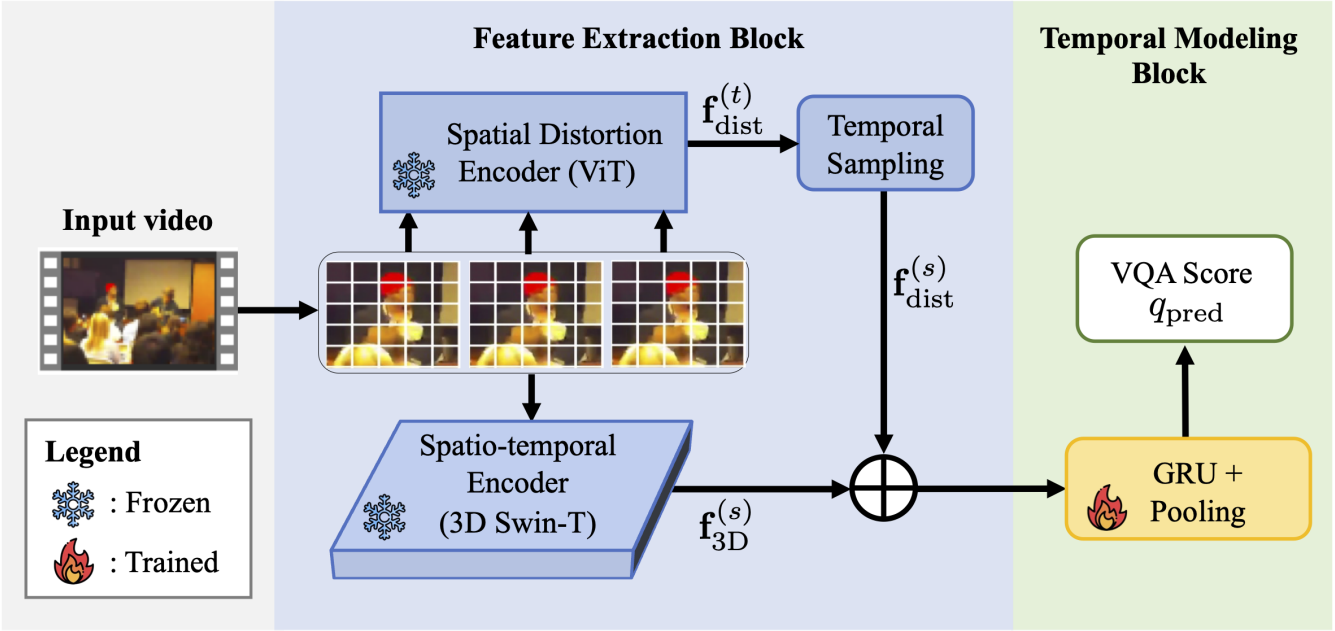

FIGURE 1. The architecture of the proposed method. Our proposed model is composed of two blocks namely the features extraction block and the temporal modeling block. The features extraction block extracts both the spatial distortion characteristics (per frame with the ViT encoder) and the spatio-temporal characteristics (per group of $M = 3$ frames with the 3D Swin Transformers) in the videos. These features are sent to a temporal modeling block consisting essentially of an attention GRU network coupled with mean pooling (embedded in the temporal memory block), for predicting the quality of the entire video. Note that before combining these two characteristics, a temporal sampling operation is applied to the frames from the spatial distortion features.

Due to the multitude of features extracted by CNNs, machine learning algorithms such as SVR can be used for predicting the overall quality. Deep learning algorithms such as recurrent neural networks can also be employed with CNN features to predict frame per frame quality. For instance, the authors of VSFA [20] extract content-aware features from a pre-trained CNN and introduce them into a GRU network which models the long-term dependencies between different frames. In the same way, the work in [30] presents a deep learning model where distortion and content features are first extracted by CNNs and then fed to a recurrent neural network coupled to predict the video quality. Moreover, to capture more diverse representations, the author in [31] extracted and combined features from seven different pre-trained CNNs.

2) TRANSFORMERS-BASED METHODS

While CNN-based methods are still dominant in computer vision, Transformers have become the standard architecture for sequence-to-sequence modeling thanks to their self-attention mechanism that can capture long-range dependencies in the data. Initially proposed for Natural Language Processing (NLP) tasks, researchers started to integrate the self-attention mechanisms found in Transformers into CNN layers [32]. In their seminal work introducing the Vision Transformer (ViT) [33], Dosovitskiy et al. successfully adapted Transformers to computer vision tasks. Since then, a backbone shift has been underway for vision models, with ViTs replacing CNNs in a broad range of computer vision studies [34].

The work in [4] showed that a pure Transformer architecture, similar to those employed in NLP, could achieve SOTA results for image classification. In the field of IQA, Yang et al. proposed a multi-dimension attention network called MANIQA which combines a standard ViT for feature extraction with Swin Transformer blocks modeling spatial attention hierarchically [35]. Their approach achieved impressive performance on popular image databases such as KADID-10k [36] and TID2013 [37]. Recently, methods utilizing ViTs, namely MaxVQA [38] and FAST-VQA [39], have significantly surpassed SOTA NR-VQA methods employing CNNs.

III. PROPOSED METHOD

In this work, we exploit two categories of Transformers to solve the VQA problem. Our model utilizes both a ViT encoder and Swin Transformer models to extract spatial distortion and spatio-temporal features, respectively. These features are combined and fed into a temporal-memory block, which consists of an attention GRU network and mean pooling layers. This block helps capture interdependencies among frames, allowing for the prediction of the entire video's quality. The architecture of our proposed method is illustrated in Figure 1, with further details provided in the following subsections.

A. FEATURES EXTRACTION

1) SPATIAL DISTORTION FEATURES

In studies concerning IQA and VQA, it is widely acknowledged that distortions have a considerable impact on visual quality. Researchers, whether employing handcrafted or deep-learning models, aim to incorporate spatial distortion features into their models. This is achieved either by extracting such features from specific databases or by fine-tuning models across various databases. Consequently, numerous studies in NR-VQA emphasize the importance of spatial distortion features when evaluating image and video quality. Furthermore, researchers have demonstrated that features extracted by deep learning networks, particularly CNNs, are highly sensitive to distortions. For instance, in [30], [40], and [41], CNNs are used for spatial distortion feature extraction.

In a recent study, Paul and Chen [42], showed how ViTs offer improved robustness against semantic shifts, common corruptions, and perturbations. Building upon these findings, our work uses ViTs for learning spatial distortion characteristics. We adopt a shallow network with few Transformer layers ($L = 2$) and increase the encoder's robustness using a self-supervised pre-training based on a Masked Auto-Encoder (MAE) with a masked patch ratio of 75%. The motivation for using this pre-training step is that, by reconstructing masked patches from visible ones, the learned features will encode information on the spatial relationships between different regions in the image. The human vision system being very sensitive to changes in these spatial relationships (e.g., a ''normal looking'' region that is out of place will be instantly spotted by a human), the learned features may capture useful information for predicting the perceived visual quality.

After pre-training our spatial distortion encoder with the MAE, we fine-tune it to predict the type of distortion in the input image or frame. As in our previous work [30], we select the CSIQ image database [43] for this task. This database contains 866 images obtained from 30 reference images distorted with six types of distortions. The distortions contained in the CISQ image database are: additive pink Gaussian noise, additive white Gaussian noise, global contrast decrements, JPEG compression, JPEG-2000 compression, and Gaussian blurring. Each video is decomposed into frames or images and each image is decomposed into patches (see more details in Section IV-B). In this work, the convolution projection approach, namely Conv2D, is used for the patchification where the kernel size and stride of the convolution layer are equal to the patch size.

Figure 2 details the components of our distortion network. Let $\mathbf{x}_i$ be the i-th patch of the input image, the final outputs of our ViT encoder are obtained by:

$$\mathbf{z}_0 = [\mathbf{x}_{cls};\ \mathbf{x}_1\mathbf{E};\ \mathbf{x}_2\mathbf{E};\ \ldots;\ \mathbf{x}_n\mathbf{E}] + \mathbf{E}_{pos} \quad (1)$$

$$\hat{\mathbf{z}}_l = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \qquad l = 1 \ldots L \quad (2)$$

$$\mathbf{z}_l = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}_l)) + \hat{\mathbf{z}}_l, \qquad l = 1 \ldots L \quad (3)$$

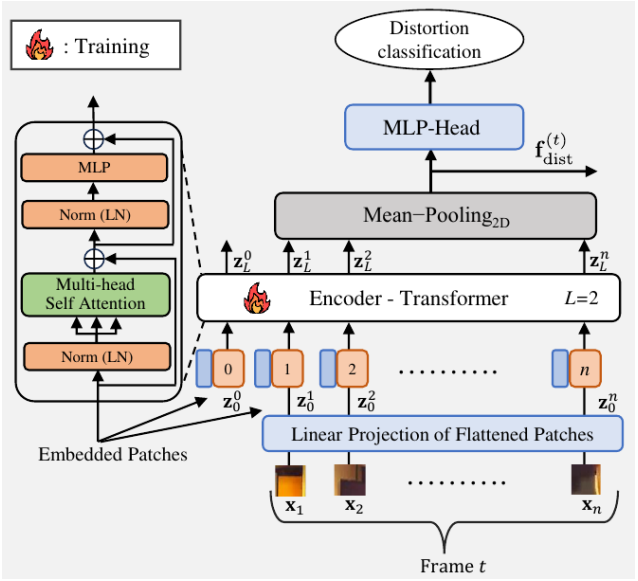

FIGURE 2. The architecture of our spatial distortion encoder based on ViT. The encoder is composed of alternating blocks of Multi-head Self Attention (MSA) and multi-layer perceptron (MLP). Norm (LN): Layer Normalization is applied to the input of each block.

Here, $\mathbf{x}_i\mathbf{E}$ the i-th embedded patch, $\mathbf{x}_{cls}$ (or $\mathbf{z}_0^0$) the class token and $\mathbf{E}_{pos}$ the matrix of position embeddings. $\hat{\mathbf{z}}_l$ and $\mathbf{z}_l$ are respectively the output of the Multi-head Self Attention (MSA) module and the final output of layer l of the ViT. LN is the layer normalization, which is performed before and after each MSA operation. MLP is a multi-layer perceptron, which is a fully connected neural network.

As illustrated in Figure 2, instead of employing the class token, our model applies mean pooling on patch features of the last ViT layer to obtain the spatial distortion features $\mathbf{f}_{\text{dist}}^{(t)}$ for each frame t. The main purpose of these features is to enhance the learning process of the downstream NR-VQA task, following a transfer learning strategy. Hence, during inference we freeze the ViT encoder and, discarding the Multi-layer Perceptron (MLP) head for distortion prediction, use the pooled features $\mathbf{f}_{\text{dist}}^{(t)}$ as additional input to the NR-VQA network.

2) SPATIO-TEMPORAL FEATURES

To take into account the temporal dimension of videos, we select the 3D version of the Swin Transformers [44] which is built with ViT layers that globally connect patches across the spatial and temporal dimensions. Unlike standard ViTs which compute self-attention weights between all pairs of patches, Swin Transformers restrict self-attention computations to local windows forming a hierarchy across multiple layers. Due to this, Swin Transformers can better handle larger image or, as in our case, 3D information [45].

The major components of the Swin Transformers are their Shifted Window MSA (SW-MSA) and Window MSA (W-MSA), which replace the MSA module in the standard ViT. Each SW-MSA module is positioned after the W-MSA.

Similar to the spatial distortion encoder's definition, let $\mathbf{x}_i^{(t)}$ be the i-th 3D patch of frame t. The final output, $\mathbf{z}_l$, of our Swin Transformers is obtained by:

$$\mathbf{z}_0 = [\mathbf{x}_{cls}^{(t)}; \mathbf{x}_1^{(t)}\mathbf{E}; \mathbf{x}_2^{(t)}\mathbf{E}; \ldots; \mathbf{x}_n^{(t)}\mathbf{E}] + \mathbf{E}_{pos} \quad (4)$$

$$\hat{\mathbf{z}}_l = \text{W-MSA}_{3D}\big(\text{LN}(\mathbf{z}_{l-1})\big) + \mathbf{z}_{l-1}, \quad (5)$$

$$\mathbf{z}'_l = \text{MLP}\big(\text{LN}(\hat{\mathbf{z}}_l)\big) + \hat{\mathbf{z}}_l, \quad (6)$$

$$\hat{\mathbf{z}}'_l = \text{SW-MSA}_{3D}\big(\text{LN}(\mathbf{z}'_l)\big) + \mathbf{z}'_l, \quad (7)$$

$$\mathbf{z}_l = \text{MLP}\big(\text{LN}(\hat{\mathbf{z}}'_l)\big) + \hat{\mathbf{z}}'_l, \quad (8)$$

In these equations, $\mathbf{x}_i^{(t)}\mathbf{E}$ and $\mathbf{x}_{cls}^{(t)}$ are, respectively, the i-th embedded 3D patch and the class token of frame t. $\mathbf{E}_{pos}$ is the corresponding matrix of position embeddings. Moreover, $\hat{\mathbf{z}}_l$ and $\hat{\mathbf{z}}'_l$ respectively denote the outputs of the W-MSA$_{3D}$ and SW-MSA$_{3D}$ (3D version of W-MSA and SW-MSA modules).

We consider that each video contains T frames. While the spatial distortion features are extracted for each frame $t \in \{1, \ldots, T\}$, the spatio-temporal features are instead computed for groups of M frames. Therefore, M represents the frame subsampling factor. Denoting as $\mathbf{z}^{(s)}$ the output of the transformer for the s-th group of M frames, with $s \in \{1, \ldots, \lfloor T/M \rfloor\}$, we obtain the spatio-temporal features via a 3D mean pooling operation:

$$\mathbf{f}_{3D}^{(s)} = \text{Mean-Pooling}_{3D}(\mathbf{z}^{(s)}) \quad (9)$$

3) TEMPORAL SAMPLING AND FEATURE CONCATENATION

As the spatio-temporal features $\mathbf{f}_{3D}^{(s)}$ are computed for groups of frames s, they cannot be directly combined with the per-frame spatial distortion features $\mathbf{f}_{\text{dist}}^{(t)}$. To address this issue, we perform a temporal sampling on the output of the distortion network:

$$\mathbf{f}_{\text{dist}}^{(s)} = \text{Temporal-Sampling}(\mathbf{f}_{\text{dist}}^{(t)}) \quad (10)$$

with $\mathbf{f}_{\text{dist}}^{(t)}$ obtained from Figure 2 and the Temporal-Sampling operation for selecting one frame in each group of M frames. In our study, we select the first frame in each group of M frames for spatial distortion feature extraction. In other words, $t \in \{1, (1+M), (1+2M), \ldots, (1+(\lfloor T/M \rfloor - 1)M)\}$. As we only need to compute the features for these sampled frames, the overall cost of running the ViT encoder is reduced by a factor of M. In our experiments, we set $M = 3$. Note that, in general, we can select the n-th frame in each group, with $n \in \{1, 2, 3, \ldots, M\}$.

In the end, the final features $\mathbf{f}_{\text{final}}^{(s)}$ are obtained by concatenating the sampled spatial distortion features and spatio-temporal features for each group $s \in \{1, \ldots, \lfloor T/M \rfloor\}$,

$$\mathbf{f}_{\text{final}}^{(s)} = \mathbf{f}_{\text{dist}}^{(s)} \oplus \mathbf{f}_{3D}^{(s)} \quad (11)$$

where $\oplus$ is the concatenation operator. The dimension of the concatenated features $\mathbf{f}_{\text{final}}^{(s)}$ is 2,048.

B. TEMPORAL MODELING

After the features have been extracted and combined, they are sent to the temporal-memory block which performs three important operations, illustrated in Figure 3.

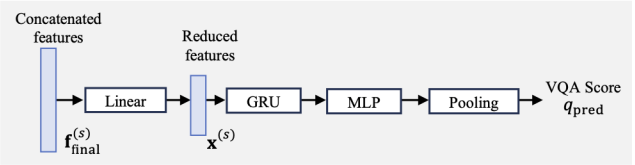

FIGURE 3. The operations performed by the temporal modeling block. GRU: Gated Recurrent Unit.

Firstly, a dimension reduction step is performed on the concatenated features using a fully-connected (linear) layer:

$$\mathbf{x}^{(s)} = \mathbf{W}_{\text{xf}}\mathbf{f}_{\text{final}}^{(s)} \quad (12)$$

where $\mathbf{W}_{\text{xf}}$ are the learned parameters of the linear model. This reduction step is necessary to make the input more manageable for the temporal modeling block.

The features $\mathbf{x}^{(s)}$ of dimensionality $P = 512$ are sent to the attention GRU network whose task is predicting the final perceived quality. By considering the hidden states of the GRU as the integrated features, where the initial state is given by $\mathbf{h}^{(0)}$ and the previous state by $\mathbf{h}^{(s-1)}$, the quality score of each group of frames is predicted as:

$$q^{(s)} = \sigma(\mathbf{W}_{\text{qh}}\mathbf{h}^{(s)} + b_q) \in [0, 1]. \quad (13)$$

$$\text{with } \mathbf{h}^{(s)} = \text{Attention}\big(\text{GRU}(\mathbf{h}^{(s-1)})\big) \quad (14)$$

where $\mathbf{W}_{\text{qh}}$ and b_q are respectively the weights and bias parameters, which are jointly learned with the other parameters of our system. To limit computations and memory, we use the standard GRU with a single layer and, similar to Transformers models [34], compute the attention based on a query (Q), key (K) and value (V) as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V \quad (15)$$

Here, d_k is the dimension of key (K).

Finally, the overall video quality score q_{pred} is obtained by mean pooling:

$$q_{\text{pred}} = \frac{1}{\lfloor T/M \rfloor}\sum_{s=1}^{\lfloor T/M \rfloor} q^{(s)} \quad (16)$$

IV. EXPERIMENTAL STUDY

This section describes our experimental setup and implementation details along with the selected VQA databases and the evaluation criteria used. To confirm the advantages of our proposed method, we then conduct three experiments: comparison on individual databases, cross databases evaluation, ablation study and computational complexity.

A. DATABASES AND PERFORMANCE MEASURES

In this study, we selected two publicly available and widely used video databases for UGC studies: KoNViD-1k and LIVE-VQC. These databases respectively contain natural videos and real-world mobile photography.

1) KoNViD-1k (KONSTANZ NATURAL VIDEO DATABASE)

Reference [5] is a collection of 1,200 videos of resolution 960×540 sampled according to six specific attributes from the Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset. The resulting database contains video sequences that are representative of a wide variety of contents and authentic distortions. The duration of each video is 8 seconds at 24/25/30 fps. Their mean opinion scores (MOSs) have been collected through a crowdsourcing experiment and range from 1.22 to 4.64.

2) LIVE-VQC (LIVE VIDEO QUALITY CHALLENGE)

Reference [6] is a database containing 585 videos of unique content, captured by 101 different devices (the majority of these were smartphones), with a wide range of complex authentic distortions. Predominant resolutions are 404×720, 1024×720, and 1920×1080. The average duration of these videos is 10 seconds. Similar to KoNViD-1k, the subjective scores were collected via crowdsourcing and a total of 4,776 unique participants produced more than 205,000 opinion scores. The scores span between 0 and 100.

The performance of our proposed model is evaluated on these two datasets. Similar to the SOTA methods, the performances are evaluated in terms of Spearman Rank Order Correlation Coefficient (SROCC), and Pearson's Linear Correlation Coefficient (PLCC). Note that our PLCC is calculated after performing a non-linear logistic fitting between MOS or subjective scores (s) and objective scores (o) [46]:

$$f(o) = \frac{\alpha_1 - \alpha_2}{1 + e^{-\frac{o-\alpha_3}{\alpha_4}}} + \alpha_2. \tag{17}$$

The parameters α_1 to α_4 are adjustment parameters initialized with $\alpha_1 = s_{\max}$, $\alpha_2 = s_{\min}$, $\alpha_3 = \mu_o$, $\alpha_4 = \sigma_o/4$, $s_{\min}$, $s_{\max}$ are the minimum and maximum subjective scores, and μ_o, σ_o are the mean and standard deviation of the objective scores.

B. EXPERIMENTAL SETUP

Our model is implemented using the PyTorch [47] framework and comprises two blocks: a feature extraction block and a temporal modeling block.

The feature extraction block extracts the spatial distortion and spatio-temporal features from two different vision Transformers, namely Encoder ViT (distortion network) and 3D Swin-T (see Figure 1). The distortion encoder is designed and trained from end to end to predict the type of distortion within the input image, while the 3D Swin-T is a pre-trained network which extracts the spatio-temporal features in groups of frames.

Our distortion network or ViT encoder is trained on the CSIQ image database [43]. This image database contains 6 types of distortion and a total of 866 distorted images. Each image is decomposed into patches of size 16×16 pixels; the projected features have a shape of $[H/16, W/16, D]$, where H, W and D denote respectively the image height, width and embedding dimension. The patchification is performed by Conv2D with the kernel size and stride of convolution layer equal to 16 (patch size). The flattened image patches or feature map is then fed into the encoder-Transformer after adding the position embedding. The parameters of our Transformer are D and H, the number of heads (set to 3).

Our distortion network is composed of only $L = 2$ ViT layers (shallow network) and a mean pooling (Mean-Pooling$_{2D}$) is used to average the features of all patches for the classification. To train the distortion encoder, a classification head (MLP-Head) which consists of two fully connected (FC) layers with a dropout layer in between, is used to predict the type of distortion in each frame or image. To train the network, we use the cross-entropy as loss function measuring the distance between the predicted image quality distribution and the ground-truth distribution; the Adam optimizer [48] is used to minimize this loss function and the model is trained on 100 epochs with a learning rate of 0.001.

For the spatio-temporal extraction, we select the 3D Swin-T, which is a pre-trained 3D Swin Transformers on the Kinetics video database [49] for temporal features extraction.

The temporal-memory block receives as input the concatenated features ($\mathbf{f}^{(s)}_{\text{final}}$) and is trained to estimate the quality score of the entire video. As mentioned earlier, before this concatenation, a temporal sampling module is applied on the outputs of the distortion encoder in order to select the first frame in each group of $M = 3$ frames for the concatenation.

Inside the temporal-memory block, a dimension reduction step is first performed using a fully connected (FC) or linear projection (the dimension of $\mathbf{f}^{(s)}_{\text{final}}$ is 2,048 since the dimension of $\mathbf{f}^{(s)}_{\text{dist}}$ and $\mathbf{f}^{(s)}_{\text{3D}}$ is 1,024). The reduced-size features are then fed to the attention GRU network for estimating the quality score of each group of frames. In this study, we reduce the features to $P = 512$ and introduce them into the attention GRU with a single layer and a hidden size of 64 (which corresponds to the amount of information stored). Finally, a mean pooling is applied to aggregate the scores predicted for each group of frames. To train the temporal modeling block, we use the $L1$ loss between the scores predicted for the video and the ground-truth scores, and employ the Adam optimizer with an initial learning rate of 0.001 and a batch size of 16. Note that during the training, our ground-truth MOSs are scaled in the range [0, 1] using the min–max scaling.

C. PERFORMANCE ON INDIVIDUAL DATABASE

For a fair comparison with recent SOTA work, such as FAST-VQA [39] and MaxVQA [38] that also use the vision Transformers, we performed 10 simulations using 10 random splits and reported the average of the results. For each split, 80% of the data is used for training and the other 20% is used for testing (split 80:20 for each simulation). Similar to previous works, we evaluated the performance of our model in terms of SROCC and PLCC. Additionally, we compared our work with other popular and best performers works found in the SOTA, namely TLVQM [17], VIDEVAL [18], CNN-TLVQM [19], PVQ [23], RAPIQUE [21] and VSFA [20].

The performance of all the methods are reported in Table 1. We did not report the performance of methods such as RIRNet [25], ChipQA [16] and HEKE [22] because they have been shown [50], [51], [52] to be inferior to PVQ and CNN-TLVQM on authentic VQA databases.

Since the code of FAST-VQA is publicly available, we have re-simulated this method in our Pytorch environment and reported the performance in Table 1. However, although the code of MaxVQA is publicly available, we could not reproduce their individual performance because it requires their database, namely Maxwell, which is not yet publicly available. Therefore, we just reported the performance from their article [38].

TABLE 1. Performance results on KoNViD-1k and LIVE-VQC databases. In each column, the best, and second-best values are respectively marked in boldface, and underlined. Note that * are performances taken from papers [38], [39].

Methods	KoNViD-1k		LIVE-VQC	
	SROCC ↑	PLCC ↑	SROCC ↑	PLCC ↑
TLVQM [17]*	0.773	0.768	0.799	0.803
VIDEVAL [18]*	0.783	0.780	0.752	0.751
RAPIQUE [21]*	0.803	0.817	0.755	0.786
VSFA [20]*	0.773	0.775	0.773	0.795
PVQ [23]*	0.791	0.786	0.827	0.837
CNN-TLVQM [19]*	0.816	0.818	0.825	0.834
FAST-VQA [39]	0.891	0.892	0.827	0.862
MaxVQA [38]*	0.894	0.895	**0.854**	**0.873**
Proposed	**0.927**	**0.925**	0.830	0.840

As can be seen in Table 1, our proposed method achieves the best performance in terms of SROCC and PLCC on the KoNViD-1k database [5], while delivering competitive performance with recent SOTA methods on the LIVE-VQC database [6].

On the KoNViD-1k database, our method shows an improvement of 0.033 in terms of SROCC and of 0.03 in PLCC compared to the SOTA method MaxVQA. The FAST-VQA method ranks third with performances similar to MaxVQA in terms of SROCC and PLCC. The CNN-TLVQM method is ranked fourth and exhibits a large performance drop compared to the top third methods.

On the LIVE-VQC database, our method presents competitive performances in terms of SROCC when compared to PVQ, CNN-TLVQM and FAST-VQA. The best performance on this database is achieved by MaxVQA with differences in terms of SROCC and PLCC of 0.024 and 0.033, respectively, compared to our method which is ranked second in terms of SROCC and third in terms of PLCC (FAST-VQA is the second-best method in terms of PLCC on the LIVE-VQC database).

The higher performance of our method on KoNViD-1k or natural video database can be explained by the effectiveness of Transformers in learning both spatial distortions and spatio-temporal characteristics.

We also notice that the two methods that compete closely with our method, i.e. FAST-VQA and MaxVQA, are also based on the Transformers. However, the use of two complementary branches of Transformers in our method sets it apart. Moreover, while these methods apply some preprocessing techniques on the input videos called fragmentation, our method does not require this pre-processing. In our study, we have avoided any transformation on the input videos because the preprocessing techniques such as fragmentation can alter semantic information and even distortion characteristics in the original videos.

D. PERFORMANCE ON CROSS-DATABASE

In addition to the individual performance study, we evaluated the performance of our model in a cross-database setting. We performed 10 simulations and reported the average results in terms of SROCC and PLCC in Table 2. Moreover, we compared the performance of our proposed method with the three other best methods from Table 1, i.e. CNN-TLVQM, MaxVQA and FAST-VQA [39].

TABLE 2. Performance results in terms of SROCC and PLCC for KoNViD-1k and LIVE-VQC in cross-database scenarios. Note that † are performances taken from original paper [38].

Training	LIVE-VQC		KoNViD-1k	
Testing	KoNViD-1k		LIVE-VQC	
	SROCC ↑	PLCC ↑	SROCC ↑	PLCC ↑
CNN-TLVQM [19]	0.643	0.632	0.711	0.745
FAST-VQA [39]	0.750	0.720	0.734	0.740
MaxVQA [38]†	**0.833**	**0.831**	**0.793**	**0.832**
Proposed	0.701	0.712	0.721	0.746

Globally, our method presents, for the two scenarios, the third-best performance after MaxVQA and FAST-VQA. The best performer is MaxVQA, exhibiting SROCC and PLCC values that are more than 0.05 higher than those of the second-best performer, FAST-VQA. Our proposed model demonstrates performance nearly equivalent to FAST-VQA when trained on KoNViD-1k and tested on LIVE-VQC. However, FAST-VQA significantly outperforms our model in terms of SROCC in the reverse scenario. We also note that the top three methods are based on Transformers. The strong performance of MaxVQA could be attributed to its use of pre-trained parameters adopted from the FAST-VQA method. Additionnally, MaxVQA is trained with more data than both FAST-VQA and our proposed method. Transformers are often described as "data-hungry" models, meaning that the volume of data on which they are trained has a significant impact on their performance.

E. ABLATION STUDY

We also performed an ablation study to verify the importance of each component of our model. Since our model extracts features from two different branches, we first studied their

individual impact. Secondly, we investigated the impact of some alternative distortion networks which are designed based on CNN and Swin Transformers models. Finally, we studied the advantages of our temporal memory block based on attention recurrent networks (GRU).

Our ablation tests are also performed on the two UGC databases selected in this work, the KoNViD-1k and LIVE-VQC databases. The same split strategy was also adopted, i.e. 80:20 where 80% of the data are used for the training and the 20% for testing. Similar to previous sections, we performed 10 simulations and reported the average performance in terms of SROCC and PLCC. Table 3 gives the performance of our compared scenarios which are described in the following paragraphs.

TABLE 3. Ablation Study: Performance in terms of SROCC and PLCC for different settings.

Methods	KoNViD-1k		LIVE-VQC	
	SROCC	PLCC	SROCC	PLCC
Encoder ViT	0.842	0.853	0.575	0.588
3D Swin-T	0.833	0.834	0.812	0.814
Proposed with CNN distortion	0.831	0.833	0.814	0.823
Proposed with Encoder Swin-T	0.870	0.860	0.810	0.840
Proposed without GRU	0.890	0.900	0.804	0.810
Proposed	**0.927**	**0.925**	**0.830**	**0.845**

1) STUDY OF THE EXTRACTION BRANCHES

In this scenario, we used only one category of features, i.e. either the spatial distortion or the spatio-temporal features extraction branch along with the temporal modeling block to predict the video quality. The performance of these methods referred to as *Encoder ViT* and *3D Swin-T* are reported in Table 3. We observe that the 3D Swin-T (spatio-temporal features) branch performs well on both KoNViD-1k and LIVE-VQC databases while the Encoder ViT branch yields a low performance on the LIVE-VQC database (SROCC = 0.575 and PLCC = 0.588). This could be explained by the distortions contained in this database. Actually, as mentioned in the literature review, the LIVE-VQC database contains a wide range of complex and authentic levels of distortions. Therefore our distortion encoder trained on the CSIQ image database [43] did not generalize well on all of these distortions.

2) STUDY OF THE SPATIAL DISTORTION NETWORK

To improve the performance of our Encoder ViT on the LIVE-VQC database, we implemented two other alternative distortion networks and compared their performance with our method. Firstly, we designed a distortion network based on CNN and, similar to our Encoder ViT, trained it on the same image distortion database (CSIQ image database). The performance of this method referred to as *Proposed with CNN distortion* is reported in Table 3. We observe that our method perform slightly better than the CNN-based method on the LIVE-VQC database (SROCC = 0.814 and PLCC = 0.823), while on KoNViD-1k database our method yields larger improvement of 0.096 and 0.092 in terms of SROCC and PLCC, respectively.

Lastly, we designed another distortion network based on Swin Transformers. Actually, the LIVE-VQC database mostly contains videos with high resolutions and degraded by capture or authentic distortions. To cover these two aspects of the LIVE-VQC database, we built a distortion network based on Swin Transformers and selected the CID2013 image database [53] which contains both high-resolution images (1600×1200) and authentic distortions (lightness, saturation, graininess, and sharpness) to train the network.

In Table 3, we report the performance of the method referred to as *Proposed with Encoder Swin-T* obtained by replacing our Encoder ViT with the distortion network based on Swin Transformers. We observe that this method does not perform well on the KoNViD-1k database. This could be explained by the fact that both the Encoder Swin-T (trained on authentic distortion database) and 3D Swin-T (trained on Kinetics database [49] which contains videos selected from YouTube) extract similar characteristics in the videos frames, and therefore this modification does not provide a major improvement for the final model.

3) STUDY OF TEMPORAL MEMORY BLOCK

In this scenario, we study the impact of the attention GRU network. Firstly, we removed the attention GRU network in our method. The performance of the resulting variant, called *Proposed without GRU*, is presented in Table 3. We note that this variant shows lower performance on both databases (KoNViD-1k and LIVE-VQC) compared to our proposed method. This confirms the importance of the attention GRU network for capturing the dependencies among the combined features. Through these ablation studies, we confirm the importance of each component of our method and its superiority compared to other considered alternatives.

F. COMPUTATIONAL EFFICIENCY

Finally, we evaluate the computational performance of our proposed method. Following the approach of the authors in [17], we selected twenty representative video sequences from the CVD2014 database, ten with low resolution (640×480) and ten with high resolution (1280×720). The videos have variable lengths ranging from 11 seconds to 28 seconds, with frame rates from 9 fps to 30 fps. We simulated our method and the three other top approaches, namely MaxVQA, FAST-VQA, and CNN-TLVQM, on a desktop computer equipped with an NVIDIA Quadro RTX 8000 with 4608 CUDA cores. In Table 4, we report the average computational complexity in terms of frames per second for these methods.

As shown in Table 4, our method achieves the second-best runtime. The best performer, FAST-VQA is approximately 3.4 times and 9 times faster than our method for low and high-resolution videos, respectively. Our method can process

TABLE 4. Average computational complexity (frames per second) on example videos.

Method	Low Resolution 640×480	High Resolution 1280×720
CNN-TLVQM [19]	10.20	4.23
FAST-VQA [39]	**45.71**	**39.05**
MaxVQA [38]	4.59	4.05
Proposed (frames per second)	13.56	4.31

33% more frames per second compared to CNN-TLVQM and nearly three times as many frames as the MaxVQA methods on low-resolution videos, while performing comparably to them on high-resolution videos. Our method and CNN-TLVQM consider all the frames in the videos, resulting in nearly the same complexity. In contrast, FAST-VQA and MaxVQA process only 128 frames per video in the considered databases (i.e. $clip\_len = 32$ and $num\_clips = 4$), which may affect the robustness of their solutions on other video material.

V. CONCLUSION

In this paper, we have proposed an objective NR-VQA method for UGC videos. The main contribution of our work is a new method based on two complementary categories of vision Transformers and attention recurrent networks to predict the video quality. Experiments on two UGC databases containing natural and complex distortions videos demonstrate the effectiveness of our proposed method.

Despite these promising results, the performance of our model could be improved. In this study, we did not take into consideration the motion-related distortion features in the videos, which could further boost the performance of our proposed method on databases such as LIVE-VQC. In future work, we plan to investigate vision Transformers models which could efficiently extract these video features.

REFERENCES

[1] T. Barnett, S. Jain, U. Andra, and T. Khurana, *CISCO Visual Networking Index (VNI) Complete Forecast Update, 2017–2022*. San Jose, CA, USA: Americas/EMEAR Cisco Knowledge Network (CKN), Presentation, 2018, pp. 1–30.

[2] F. Fatemi. How The Pandemic Has Changed Video Content and Consumption. Forbes.com. Accessed: Apr. 14, 2022. [Online]. Available: https://www.forbes.com/sites/falonfatemi/2021/02/01/how-the-pandemichas-changed-video-content-and-consumption/?sh=b910dec6ec00

[3] F. Yi, M. Chen, W. Sun, X. Min, Y. Tian, and G. Zhai, "Attention based network for no-reference UGC video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1414–1418, doi: 10.1109/ICIP42928.2021.9506420.

[4] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6816–6826, doi: 10.1109/ICCV48922.2021.00676.

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

[6] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 612–627, Feb. 2019.

[7] 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. 2013.

[8] 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.

[9] 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.

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

[11] 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, Jan. 2016.

[12] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. 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.

[13] K. Zhu, C. Li, V. Asari, and D. Saupe, "No-reference video quality assessment based on artifact measurement and statistical analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 4, pp. 533–546, Apr. 2015.

[14] X. Li, Q. Guo, and X. Lu, "Spatiotemporal statistics for video quality assessment," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3329–3342, Jul. 2016.

[15] 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.

[16] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman, and A. C. Bovik, "ChipQA: No-reference video quality prediction via space-time chips," *IEEE Trans. Image Process.*, vol. 30, pp. 8059–8074, 2021.

[17] 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.

[18] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021, doi: 10.1109/TIP.2021.3072221.

[19] 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.

[20] 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.

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

[22] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, and G. Shi, "Spatiotemporal representation learning for blind video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3500–3513, Jun. 2022.

[23] 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.

[24] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, "KonVid-150k: A dataset for no-reference video quality assessment of videos in-the-wild," *IEEE Access*, vol. 9, pp. 72139–72160, 2021.

[25] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RIRNet: Recurrent-in-recurrent network for video quality assessment," in *Proc. 28th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2020, pp. 834–842, doi: 10.1145/3394171.3413717.

[26] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2349–2353, doi: 10.1109/ICIP.2019.8803395.

[27] W. Wu, Q. Li, Z. Chen, and S. Liu, "Semantic information oriented no-reference video quality assessment," *IEEE Signal Process. Lett.*, vol. 28, pp. 204–208, 2021, doi: 10.1109/LSP.2020.3048607.

[28] D. Varga and T. Szirányi, "No-reference video quality assessment via pretrained CNN and LSTM networks," *Signal, Image Video Process.*, vol. 13, no. 8, pp. 1569–1576, Nov. 2019, doi: 10.1007/s11760-019-01510-8.

[29] R. Hou, Y. Zhao, Y. Hu, and H. Liu, "No-reference video quality evaluation by a deep transfer CNN architecture," *Signal Process., Image Commun.*, vol. 83, Apr. 2020, Art. no. 115782, doi: 10.1016/j.image.2020.115782.

[30] K. Kossi, S. Coulombe, C. Desrosiers, and G. Gagnon, "No-reference video quality assessment using distortion learning and temporal attention," *IEEE Access*, vol. 10, pp. 41010–41022, 2022, doi: 10.1109/ACCESS.2022.3167446.
[31] D. Varga, "No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion," *Sensors*, vol. 22, no. 6, p. 2209, Mar. 2022, doi: 10.3390/s22062209.
[32] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Jun. 2019, pp. 3285–3294, doi: 10.1109/ICCV.2019.00338.
[33] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N G. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[35] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "MANIQA: Multi-dimension attention network for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1190–1199.
[36] 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)*, Berlin, Germany, Jun. 2019, pp. 1–3, doi: 10.1109/QoMEX.2019.8743252.
[37] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. Eur. Workshop Vis. Inf. Process. (EUVIP)*, Paris, France, Jun. 2013, pp. 106–111.
[38] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Towards explainable in-the-wild video quality assessment: A database and a language-prompted approach," in *Proc. 31st ACM Int. Conf. Multimedia (MM)*, 2023, pp. 1045–1054, doi: 10.1145/3581783.3611737.
[39] H. Wu et al., "Efficient end-to-end video quality assessment with fragment sampling," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 538–554, doi: 10.1007/978-3-031-20068-7_31.
[40] 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.
[41] 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.
[42] S. Paul and P. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 2, Jun. 2022, pp. 2071–2081.
[43] 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.
[44] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201, doi: 10.1109/CVPR52688.2022.00320.
[45] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[46] *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*, VQEG, San Mateo, CA, USA, 2000.
[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2019, pp. 1–7.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[49] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[50] N.-W. Kwong, Y.-L. Chan, S.-H. Tsang, and D. P. Lun, "Quality feature learning via multi-channel CNN and GRU for no-reference video quality assessment," *IEEE Access*, vol. 11, pp. 28060–28075, 2023.
[51] P. Chen, L. Li, J. Wu, W. Dong, and G. Shi, "Contrastive self-supervised pre-training for video quality assessment," *IEEE Trans. Image Process.*, vol. 31, pp. 458–471, 2022.
[52] A. H. Bakhtiari and A. Mansouri, "Feature maps correlation-based video quality assessment," *Multimedia Tools Appl.*, vol. 83, no. 23, pp. 63309–63328, Jan. 2024.
[53] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A database for evaluating no-reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 390–402, Jan. 2015, doi: 10.1109/TIP.2014.2378061.

KOFFI KOSSI received the M.Eng. degree in software and IT engineering from the École de Technologie Supérieure (ÉTS), Université du Québec, Montréal, in 2013, where he is currently pursuing the Ph.D. degree. From 2007 to 2017, he was a Consultant, specialized in audio-video systems and performance, and also 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 the 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 was a Research Engineer with the Nokia Research Center, Dallas, TX, USA; and the Program Manager with the Audiovisual Systems Laboratory. He joined the École de Technologie Supérieure (ÉTS)–a constituent of the Université du Québec network, in 2004, where he is currently a Professor with the Department of Software and IT Engineering. From 2009 to 2018, he held the Vantrix Industrial Research Chair in Video Optimization. His research interests include video processing, compression and communication (transport) and systems with a recent focus on immersive video and machine learning for video applications.

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 the Department of Software and IT Engineering, ÉTS, University of Québec, as a Professor. 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.